Доклады Академии наук СССР
1970. Том 191, № 1

УДК 519.95

МАТЕМАТИКА

В. А. СЛЕПЯН

О ЧИСЛЕ ТУПИКОВЫХ ТЕСТОВ И О МЕРАХ ИНФОРМАТИВНОСТИ СТОЛБЦА ДЛЯ ПОЧТИ ВСЕХ БИНАРНЫХ ТАБЛИЦ

(Представлено академиком С. Л. Соболевым 2 VII 1969)

1. Пусть T_{ln} — бинарная таблица, имеющая l строк и n столбцов, а T_{lk} — таблица, составленная из k ($k \leqslant n$) столбцов таблицы T_{ln}.

Определение 1. Таблицу T_{lk} будем называть тестом таблицы T_{ln}, если она состоит из разных строк.

Определение 2. Тест будем называть тупиковым, если после удаления из него любого столбца, он перестает быть тестом. Понятие теста и тупикового теста дано в работе [1].

Таблице T_{ln} поставим в соответствие величины $N(T_{ln})$ и $N^i(T_{ln})$ — число всех тупиковых тестов и число тупиковых тестов, в которые входит i-й столбец, соответственно.

В тестовых алгоритмах распознавания образов [2] за меру информативности столбца принимается информационный вес столбца

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

Для так называемых алгоритмов «голосования» Ю. И. Журавлевым в качестве меры информативности предложена величина

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

где ρ_{ij} — расстояние Хэмминга между строками i и j в таблице T_{ln}, а $\tilde{\rho}_{ij}$ — то же расстояние после удаления одного из столбцов.

2. Будем теперь считать T_{ln} случайной таблицей, а именно таблицей, каждый элемент которой принимает значения 0 и 1 с вероятностью $1/2$. Тогда $N(T_{ln})$, $N^i(T_{ln})$, I и J являются случайными величинами.

В настоящей заметке для класса таблиц, удовлетворяющих соотношению

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

находится асимптотика среднего ($n, l \to \infty$) случайной величины $N(T_{ln})$, а для класса таблиц, удовлетворяющих условию

$$\log l / \alpha \leqslant n \leqslant l^\beta \tag{4}$$
$$(\beta < 2,\ \alpha = 1/2 - \lambda_0 / \log l,\ \lambda_0 \to \infty,\ \lambda_0 / \log l \to 0),$$

доказывается, что

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

Для таблиц T_{ln} любых размеров, если только выполнено одно из соотношений $k \leqslant \sqrt[3]{n}$, $l^2 2^{-k} \to \infty$, справедлива формула

$$Jn / k \xrightarrow{p} 1 \quad (l, n \to \infty).$$

Здесь символ $\xrightarrow{p}$ обозначает, как обычно, сходимость по вероятности.

3. Оценим вероятности тупикового теста

Рассмотрим таблицу T_{lk}. Дадим некоторые определения.

Определение 3. Две неравные строки таблицы T_{lk} образуют соединение на i-м столбце, если после удаления i-го столбца они становятся равными.

Будем говорить, что таблица T_{lk} имеет соединение на i-м столбце, если хотя бы одна пара строк T_{lk} образует соединение на i-м столбце; таблица T_{lk} имеет соединение на m столбцах, если она имеет соединение на каждом из m столбцов.

Введем обозначения: T_{lk}^*, T_{lk}' — таблица T_{lk} тест, тупиковый тест соответственно; R_l^0, R_l^1 — множество тестов, множество тестов, имеющих соединение на первом столбце, соответственно; r_l^i — множество тестов, имеющих i и только i соединений на первом столбце; $\mu(A)$ — мощность множества A.

Пусть Q_0 и Q — некоторые множества, элементами которых являются таблицы T_{lk}, $T_{(l-1)k}$ соответственно, а α_0 и α_1 — целые числа. Под записью $\alpha_0 Q_0 \to \alpha_1 Q_1$ условимся понимать следующее: из α_0 множеств Q_0 путем удаления по одной строке из каждой таблицы можно построить α_1 множеств Q_1. Для множеств r_l^i, r_{l-1}^i, r_{l-1}^{i-1} нетрудно доказать справедливость соотношений

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

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

Введенные обозначения позволяют записать равенства

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

На основании (5) и (6) всегда можно найти такое целое число $A_l > 0$, что справедливо представление

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

где $\alpha_i, \beta_i \geqslant 0$; $\alpha_i + \beta_i = 1$; m_i, n_i — целые положительные числа.

Доказывается, кроме того, что α_i, β_i удовлетворяют уравнению

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

Отсюда и из формул (7) следует

Лемма 1. *Существуют такие целые положительные числа A_l, B_l, что*

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

Следствием леммы 1 является

Лемма 2. *Вероятность тупикового теста удовлетворяет неравенству*

$$p(T_{lk}') \geqslant p(T_{lk}^*) \prod_{i=1}^{k} p_i, \quad p_i = 1 - \frac{2^{l-i+1} C_{2^{k-1}}^{l-i+1}}{C_{2^k}^{l-i+1}}.$$

Пусть таблица T_{lk} образует тест. Зафиксируем m столбцов.

Определение 4. Назовем гирляндой группу из $i+1$ строк, образующих между собой соединение на i столбцах, принадлежащих m указанным столбцам.

Определение 5. Максимальной гирляндой на j-м столбце назовем гирлянду, образующую соединения на максимальном числе столбцов из m фиксированных, включая j-й.

Обозначим через X_l^m тест T_{lk}^*, имеющий соединение на m фиксированных столбцов, в число которых входит первый, а через $Y_{i\max}$ — тест T_{lk}^*, содержащий максимальную гирлянду из $i+1$ строк на первом столбце.

Тогда

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

Из утверждения леммы 1 и того факта, что вероятность соединения на m столбцах не увеличивается при уменьшении числа строк таблицы, оцениваются сверху вероятности, входящие в правую часть (8). В результате имеем

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

Далее методом математической индукции, доказывается

Лемма 3. *Вероятность тупикового теста удовлетворяет неравенству* $p(T_{lk}') \leqslant p(T_{lk}^*) p_1^k (1+\delta)^k$, *где*

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

Следствием лемм 2 и 3 является

Теорема 1. *Если* $k^2 l^{-1} \to 0$, *то* $p(T_{lk}') \sim p(T_{lk}^*)(1 - \exp(-l^2 2^{-k-1}))$.

Для тех k, при которых условие теоремы 1 не выполнено, приведем следующую верхнюю оценку вероятности тупикового теста:

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

4. Среднее число тупиковых тестов может быть представлено суммой

$$\bar{N} = \sum_{k=[\log l]}^{\min(n, l-1)} \bar{N}_k, \quad \bar{N}_k = C_n^k p(T_{lk}'). \tag{10}$$

где $\bar{N}_k$ — среднее число тупиковых тестов длины k (длина теста — число столбцов в нем). С использованием теоремы 1 и формул (9) и (10) доказывается

Теорема 2. 1) *Если* $n \leqslant l^2 (\log l)^\gamma$ $(\gamma < 3)$, *то* $\bar{N} \sim \bar{N}_{k_0} + \bar{N}_{k_1}$.

2) *Если* $n \leqslant 2^{l^\alpha}$ $(\alpha < 1/2)$, *то*

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

где $\bar{N}_k \sim \tilde{N}_k = C_n^k p(T_{lk}^*)(1-x)^k$; $x = \exp(-l^2 2^{-k-1})$; $\eta_k = \bar{N}_k / \bar{N}_{k+1}$; k_0 — *корень уравнения* $\eta_k = 1$; $k_1 = k_0 - 1$ *или* $k_0 + 1$; $\eta_{k_0} \geqslant 1$, $\eta_{k_0-1} \leqslant 1$; $0 < \alpha < \alpha_0 < 1/2$; $0 \leqslant a \leqslant 1/4$.

В частности: 1) при $2\log l \leqslant n \leqslant l^\beta$ $(\beta < 2)$ $k_0 = [2\log l - \log(-\ln(n^{1/2\log l} - 1)) - 2]$; 2) при $l^\varphi \leqslant n \leqslant 2^{l^\alpha}$ $(\varphi > 2)$ $a = 1/4$, $k_0 = [1/2 \log(nl^2) - 1/2 \log\log(nl^2) - 1/2]$.

Рассмотрим множество таблиц $Q_n = \{T_{ln}\}$, где $2\log l \leqslant n \leqslant l^\beta$, $l \in (l_1, (1+a)l_1)$, $a > 0$.

Теорема 3. *Почти во всех таблицах множества* Q_n

$$\bar{N} \sim \bar{N}_{k_0} \quad (l \to \infty).$$

5. В таблице T_{ln} пронумеруем столбцы $(1, 2, \ldots, n)$. Рассмотрим две таблицы, составленные из k столбцов таблицы T_{ln}. Пусть в каждую из них входит m и только m столбцов с одинаковыми номерами. Обозначим через $p(m)$, $p'(m)$ вероятность того, что обе таблицы одновременно являются тестами, тупиковыми тестами соответственно. Относительно их справедлива

Лемма 4. 1) $p(m)$ *является неубывающей функцией* m.
2) $p'(m) \leqslant p(m) \leqslant \exp(-l^2 2^{-k} + \beta_m)$, *где*

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

3) $p'(m) \leqslant p^* \sim p(T_{lk}^*)(1-x)^k(1-x^{2^m})^{k-m}$ $(k^2 l^{-1} \to 0,\ l \to \infty)$.

Обозначим через DN_k дисперсию числа тупиковых тестов длины k. Тогда ([3]) справедлива формула

$$V_k = \frac{DN_k}{\overline{N}_k^2} = \sum_{m=\max(0,\,2k-n)}^{k} \frac{C_k^m C_{n-k}^{k-m}}{C_n^k} \frac{p'(m)}{p^2(T'_{lk})} - 1. \tag{11}$$

Используя лемму 4, можно показать, что справедлива

Лемма 5. *Для класса таблиц* (4) $V_{k_0} \to 0$ $(l, n \to \infty)$.

Далее будем рассматривать только те таблицы класса (4), для которых $\overline{N} \sim \overline{N}_{k_0}$ $(l \to \infty)$. Из неравенств Чебышева ([4]) и леммы 5 следует

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

Для случайной величины $N^i(T_{ln})$ можно доказать теорему, аналогичную теореме 4, и показать, что $\overline{N}^i \sim 2\log l\overline{N}/n$ $(n, l \to \infty)$. Отсюда и из теоремы 4 следует

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

6. Рассмотрим таблицу T_{ln}, размеры которой произвольны. Относительно случайной величны J справедлива

Теорема 6. *Если* $k \leqslant \sqrt[3]{n}$ *или* $l^2 2^{-k} \to \infty$, *то* $Jn/k \xrightarrow{p} 1$ $(l, n \to \infty)$.

Институт математики
Сибирского отделения Академии наук СССР
Новосибирск

Поступило
2 VII 1969

ЦИТИРОВАННАЯ ЛИТЕРАТУРА

[1] И. А. Чегис, С. В. Яблонский, Тр. Матем. инст. им. В. А. Стеклова АН СССР, 51, 270 (1958). [2] А. Н. Дмитриев, Ю. И. Журавлев, Ф. П. Кренделев, Дискретный анализ, в. 7, 3 (1966). [3] В. А. Слепян, Дискретный анализ, в. 12, 50 (1968). [4] Б. В. Гнеденко, Курс теории вероятностей, М., 1961.