ДИСКРЕТНЫЙ АНАЛИЗ И ИССЛЕДОВАНИЕ ОПЕРАЦИЙ
Апрель—июнь 2004. Серия 1. Том 11, № 2, 66–72

УДК 519.12

ТЕСТЫ ДЛЯ СЛОВ*)

В. К. Леонтьев

Приводится несколько постановок задач о различимости слов. Изучаются длины фрагментов, которые позволяют различать двоичные слова длины n. На парах таких слов определяется случайная величина и исследуется ее распределение.

Как различать слова? Этот вопрос не лишен прагматического смысла и в разных контекстах встречается в математической генетике, теории информации, математической лингвистике и т. д. Определенные понятия "похожести" моделируются с помощью понятия "расстояния" [4], однако выражение "внутреннего" сходства требует иной терминологии. Мы рассматриваем несколько постановок задач разной степени сложности и интереса и приводим некоторые результаты. Ранее подобные задачи изучались нами в [2]. В содержательном плане основная задача может быть сформулирована следующим образом. Заданы два бинарных слова одинаковой длины. Какова длина фрагмента, различающего эти слова? Таким образом, на парах слов из B^n определяется случайная величина и требуется найти или оценить ее распределение.

Пусть $B = \{0, 1\}$ — двоичный алфавит и $B^n = \{0, 1\}^n$ — множество всех бинарных слов длины n.

Определение. Если $a = \alpha_1\alpha_2 \dots \alpha_n \in B^n$, то любое слово $a' = \alpha_{i_1}, \alpha_{i_2} \dots \alpha_{i_k}$, где $1 \leqslant i_1 < i_2 \dots < i_k \leqslant n$, называется *фрагментом* длины k слова a.

Тот факт, что слово a' является фрагментом слова a, будем записывать в виде $a' \subseteq a$.

Обозначим через T_a множество всех фрагментов слова a и для произвольной пары $(a, b) \in B^n \times B^n$ определим два множества

$$T(a, b) = T_a \backslash (T_a \cap T_b),$$
$$T(b, a) = T_b \backslash (T_a \cap T_b).$$

*)Исследование выполнено при финансовой поддержке Российского фонда фундаментальных исследований (проект 02-01-00716).

Каждое из множеств $T(a,b)$ и $T(b,a)$ будем называть *тестовым* для пары слов (a,b). По определению в $T(a,b)$ входят все те слова, которые являются фрагментами a и не являются фрагментами b, а в $T(b,a)$ — наоборот. Следовательно,

$$T(a,b) \cup T(b,a) = T_a \Delta T_b.$$

Любое слово $x \in T_a \Delta T_b$ будем называть *тестовым* для пары (a,b) и говорить, что слово x различает слова a и b. Пусть, как обычно, $|x|$ — длина слова x.

Лемма 1. *Если $|x| = k$, то слово x различает ровно $\sum\limits_{i=k}^{n} \binom{n}{i} \left[2^n - \sum\limits_{i=k}^{n} \binom{n}{i}\right]$ пар слов из $B^n \times B^n$.*

Доказательство. Так как $|x| = k$, то согласно [1] имеется $\sum\limits_{i=k}^{n} \binom{n}{i}$ слов из B^n, содержащих x в качестве фрагмента. Остальные $2^n - \sum\limits_{i=k}^{n} \binom{n}{i}$ слов из B^n не содержат x как фрагмент. Это и завершает доказательство леммы 1.

Следствие. *Любое слово длины $k = \lfloor n/2 \rfloor$ различает ровно 2^{n-2} пар слов из $B^n \times B^n$. Это число является максимальным.*

Пусть $\bar{T}_n$ — среднее число различных фрагментов всех длин при равномерном распределении на множестве слов из B^n, т. е. $\bar{T}_n = \frac{1}{2^n} \sum\limits_{a \in B^n} |T_a|$.

Лемма 2. *Справедливо равенство*

$$\bar{T}_n = 2\left(\frac{3}{2}\right)^n - 1. \tag{1}$$

Доказательство. Рассмотрим стандартную характеристическую функцию η_a^x свойства "быть фрагментом":

$$\eta_a^x = \begin{cases} 1, & \text{если } x \subseteq a, \\ 0, & \text{если } x \not\subseteq a. \end{cases}$$

Тогда с учетом леммы 1 получаем

$$\bar{T}_n = \frac{1}{2^n} \sum_{a \in B^n} \sum_{x} \eta_a^x = \frac{1}{2^n} \sum_{x} \sum_{a \in B^n} \eta_a^x = \frac{1}{2^n} \sum_{k=0}^{n} \sum_{|x|=k} \sum_{a \in B^n} \eta_a^x$$

$$= \frac{1}{2^n} \sum_{k=0}^{n} \sum_{i=k}^{n} \binom{n}{i}.$$

Далее имеем

$$\bar{T}_n = \frac{1}{2^n}\sum_{i=0}^{n}\binom{n}{i}\sum_{k=0}^{i}2^k = \frac{1}{2^n}\sum_{i=0}^{n}\binom{n}{i}\left(2^{i+1}-1\right) = 2\left(\frac{3}{2}\right)^n - 1.$$

Лемма 2 доказана.

Теперь найдем среднее число элементов в тестовом множестве $T_a \Delta T_b$. Пусть

$$\bar{H}_n = \frac{1}{2^{2n}}\sum_{a,b\in B^n}|T_a\Delta T_b|.$$

Лемма 3. *Справедливо соотношение*

$$\bar{H}_n = 4\left(\frac{3}{2}\right)^n - \frac{1}{2^{2n-1}}\sum_{i=0}^{n}2^i\binom{n}{i}^2 + \frac{1}{2^{2n}}\binom{2n}{n} - 2. \tag{2}$$

Доказательство. Используя предыдущие обозначения, получаем

$$\bar{H}_n = \frac{1}{2^{2n}}\sum_{a,\,b\in B^n}[|T_a| + |T_b| - |T_a\cap T_b|].$$

Далее из леммы 2 следует, что

$$\frac{1}{2^{2n}}\sum_{a,\,b\in B^n}[|T_a| + |T_b|] = 4\left(\frac{3}{2}\right)^n - 2 \tag{3}$$

и

$$\begin{aligned}\frac{1}{2^{2n}}\sum_{a,b\in B^n}[|T_a\cap T_b|] &= \frac{1}{2^{2n}}\sum_{a,\,b}\sum_{x}\eta_a^x\cdot\eta_b^x = \frac{1}{2^{2n}}\sum_{x}\sum_{a}\eta_a^x\sum_{b}\eta_b^x\\ &= \frac{1}{2^{2n}}\sum_{k=0}^{n}\sum_{|x|=k}\sum_{a}\eta_a^x\sum_{b}\eta_b^x = \frac{1}{2^{2n}}\sum_{k=0}^{n}2^k\sum_{i=k}^{n}\binom{n}{i}\\ &= \frac{1}{2^{2n}}\sum_{i=0}^{n}\binom{n}{i}^2\left(2^{i+1}-1\right) = \frac{2}{2^{2n}}\sum_{i=0}^{n}\binom{n}{i}^2 2^i - \frac{\binom{2n}{n}}{2^{2n}}.\end{aligned} \tag{4}$$

Из (3) и (4) следует (2).

Следствие. *Справедливы асимптотические неравенства*

$$4\left(\frac{3}{2}\right)^n - 2\left(\frac{3}{2}\right)^n\cdot\frac{\binom{n}{\lfloor n/2\rfloor}}{2^n} \lesssim \bar{H}_n \lesssim 4\cdot\left(\frac{3}{2}\right)^n. \tag{5}$$

Доказательство. Оценим сверху сумму в (2):

$$\sum_{i=0}^{n} 2^i \binom{n}{i}^2 \lesssim \binom{n}{\lfloor n/2 \rfloor} \sum_{i=0}^{n} \binom{n}{i} 2^i = \binom{n}{\lfloor n/2 \rfloor} 3^n. \tag{6}$$

Из (2) и (6) следует (5).

Таким образом, в среднем слова из B^n имеют $2\,(3/2)^n$ различных фрагментов всех длин, а тестовое множество типичного множества содержит асимптотически $4\,(3/2)^n$ элементов.

Определение 1. Любое слово минимальной длины из множества $T_a \Delta T_b$ называется *минимальным тестом* для пары (a, b). Длина такого теста обозначается через $t(a, b)$.

Лемма 4. *Для любых двух различных слов $a, b \in B^n$ справедливо неравенство*

$$t(a, b) \leqslant \lfloor n/2 \rfloor + 1. \tag{7}$$

Доказательство. Известно (см., например, [3], глава 6), что для любых различных слов $a, b \in B^n$ можно построить такое слово x длины, не превосходящей $\left(\lfloor \frac{n}{2} \rfloor + 1\right)$, что $x \subseteq a$ и $x \not\subseteq b$. Отсюда непосредственно следует неравенство (7).

Замечание. Рассмотрим два "противоположных" слова

$$a = (1010 \ldots 10) \in B^n,$$
$$\bar{a} = (0101 \ldots 01) \in B^n.$$

Ясно, что любое слово длины не более $\lfloor n/2 \rfloor$ является как фрагментом слова a, так и фрагментом слова $\bar{a}$. Отсюда следует, что $t(a, \bar{a}) \geqslant \lfloor n/2 \rfloor + 1$. Значит, оценка (7) является достижимой.

Ниже рассматривается ряд функций, определенных на $B^n \times B^n$ и в той или иной мере связанных с тестами для слов.

Пусть Н.О.П.(a, b) — длина наибольшей общей подпоследовательности для слов $a, b \in B^n$.

Лемма 5. *Справедливо неравенство*

$$t(a, b) \leqslant \text{Н.О.П.}(a, b) + 1. \tag{8}$$

Определение 2. *Типом* $\delta(a)$ слова $a \in B^n$ называется такое максимальное натуральное число, что в слове a в качестве фрагментов содержатся все слова длины не более $\delta(a)$ и нет хотя бы одного слова длины $\delta(a) + 1$.

Примеры.

1. Слово $a = (1101)$ имеет тип $\delta(a) = 1$, так как в a нет фрагмента (00).

2. Слово $b = (011001110)$ имеет тип $\delta(b) = 4$, так как в b есть все фрагменты длины не более 4 и нет фрагмента (00000).

Лемма 6. *Справедливы неравенства:*

$$\begin{array}{l} t(a,b) \geqslant \min\{\delta(a),\delta(b)\} + 1; \\ \text{если } \delta(a) \neq \delta(b), \text{ то } t(a,b) \leqslant \max\{\delta(a),\delta(b)\}. \end{array} \tag{9}$$

Доказательство. Пусть $\delta(a) = x$, $\delta(a) = y$ и $x \geqslant y$. Тогда слова a и b неразличимы по словам длины y, т. е. $t(a,b) \geqslant y + 1$. С другой стороны, если $x > y$, то в слове a есть фрагменты длины x, которых нет в слове b, т. е.

$$t(a,b) \leqslant x.$$

Лемма 6 доказана.

Лемма 4 утверждает, что длина минимального теста для любой пары слов из B^n не превосходит $\lfloor n/2 \rfloor + 1$ и эта оценка достижима. Однако в "классических" ситуациях длина минимального теста в типичном случае отличается от максимальной на порядок.

Покажем, что средняя длина минимального теста линейно зависит от длины слова.

Пусть

$$\bar{t}_n = \frac{1}{2^{2n}} \sum_{a,b \in B^n} t(a,b) \tag{10}$$

— средняя длина минимального теста при равномерном распределении на $B^n \times B^n$.

Теорема. *Справедливы неравенства*

$$\frac{n}{4} \lesssim \bar{t}_n \lesssim \frac{n}{2}. \tag{11}$$

Доказательство. Верхняя оценка средней длины минимального теста следует из леммы 4. Нижняя оценка есть следствие того обстоятельства, что среднее число серий в словах из B^n близко к $n/2$. Поэтому из (9) следует, что для любых слов из B^n с числом серий $\lfloor n/2 \rfloor$ длина минимального теста не меньше $n/4$. Ниже эти эвристические соображения принимают конкретную форму.

Итак, пусть

$$a = \gamma^{t_1}\bar{\gamma}^{t_2}\dots\gamma^{t_k}, \quad t_i \geqslant 1, \quad 1 \leqslant i \leqslant k,$$
$$b = \gamma^{s_1}\bar{\gamma}^{s_2}\dots\gamma^{t_s}, \quad t_s \geqslant 1, \quad 1 \leqslant i \leqslant s,$$

— представления слов a и b из B^n в "серийном" виде.

Тогда

$$\delta(a) \geqslant \lfloor k/2 \rfloor, \quad \delta(b) \geqslant \lfloor s/2 \rfloor. \tag{12}$$

Действительно, если $\lfloor k/2 \rfloor = m$, то произвольное слово длины m может быть получено из слова a удалением "лишних" букв. Теперь неравенства (12) прямо следуют из определения величины $\delta(x)$.

Пусть теперь $\eta(a)$ — случайная величина, равномерно распределенная на B^n и равная числу серий в слове $a \in B^n$. Элементарные выкладки показывают, что

$$M_\eta = (n+1)/2, \quad D_\eta = (n-1)/4. \tag{13}$$

Из (13) следует, что число слов длины n, для которых выполнено неравенство

$$\left|\eta(a) - \frac{n}{2}\right| \gtrsim \sqrt{n}\ \lg_2 n, \tag{14}$$

есть $o(2^n)$. Действительно, если в неравенстве Чебышева

$$P\left\{|\eta - M_\eta| \geqslant t\right\} < \frac{D_\eta}{t^2}$$

положить $t = \sqrt{n}\ \lg_2 n$, то с использованием (13) получаем (14).

Далее имеем

$$\bar{t}_n \gtrsim \frac{1}{2^{2n}} \sum_{\substack{|\eta(a) - M_\eta| \lesssim \sqrt{n}\lg_2 n \\ |\eta(b) - M_\eta| \lesssim \sqrt{n}\lg_2 n}} t(a,b). \tag{15}$$

Так как $M_\eta \sim n/2$, то из (12) вытекает, что общий член суммы (15) может быть оценен снизу следующим образом:

$$t(a,b) \gtrsim \min\left\{\eta(a), \eta(b)\right\} \gtrsim \frac{n}{4}. \tag{16}$$

Отсюда и из того, что число слагаемых в сумме (15) асимптотически равно 2^{2n}, получаем

$$\bar{t}_n \gtrsim \frac{n}{4}.$$

Теорема доказана.

Если для различения пар слов из $B^n \times B^n$ использовать информацию о кратностях вхождения фрагментов, то ситуация качественно изменится. Нетрудно показать, для различения почти всех пар слов из $B^n \times B^n$ достаточно фрагмента длины 1.

ЛИТЕРАТУРА

1. **Левенштейн В. И.** Элементы теории кодирования // Дискретная математика и математические вопросы кибернетики. М.: Наука, 1974. С. 207–305.
2. **Леонтьев В. К.** Задачи восстановления слов по фрагментам и их приложения // Дискрет. анализ и исслед. операций. 1995. Т. 2, № 2. С. 26–48.
3. Математические методы для анализа последовательностей ДНК. М.: Мир, 1999.
4. **Simon I.** Piecewise testable events // Automata theory and formal languages. Berlin etc.: Springer-Verl., 1975. P. 214–222 (Lecture Notes in Comput. Sci.; V. 33).

Адрес автора:

Вычислительный центр РАН
ул. Вавилова, 40,
117967 Москва,
Россия.
E-mail: vkleontiev@mtu-net.ru

Статья поступила
9 января 2004 г.,
переработанный вариант —
25 марта 2004 г.