

УДК 621.391 : 519.72

© 2023 г. И.В. Воробьев

**ПОКРЫВАЮЩИЕ КОДЫ ДЛЯ МЕТРИКИ ЛЕВЕНШТЕЙНА
ФИКСИРОВАННОЙ ДЛИНЫ¹**

Покрывающим кодом или покрытием называется множество кодовых слов, такое что объединение шаров с центрами в этих кодовых словах покрывает все пространство. Как правило, задача состоит в минимизации мощности покрывающего кода. Для классической метрики Хэмминга размер минимального покрывающего кода фиксированного радиуса R известен с точностью до постоянного множителя. Аналогичный результат был недавно получен для кодов с R вставками и кодов с R удалениями. В данной статье изучаются покрытия пространства для метрики Левенштейна фиксированной длины, т.е. для R вставок и R удалений. Для $R = 1$ и 2 доказываются новые нижние и верхние оценки минимальной мощности покрывающего кода, которые отличаются лишь в константу раз.

Ключевые слова: покрывающие коды, граница сферической упаковки, метрика Левенштейна, вероятностный метод.

DOI: 10.31857/S055529232302002X, **EDN:** PPPXTQ

§ 1. Введение

Покрывающие коды являются важным объектом в теории кодирования и дискретной математике. Они нашли применения в таких областях, как сжатие данных [1], схемная сложность [2], поиск ближайшего соседа [3], в различных задачах на решетках [4] и даже в футбольных тотализаторах [5]. Подробное описание этих приложений и многих других можно найти в книге [1].

Большинство работ о покрывающих кодах посвящено метрике Хэмминга, т.е. заменам. В последнее время резко возрос интерес к биологическим приложениям теории кодирования, в частности к хранению данных с помощью молекул ДНК (см. [6–8]). Известно, что в ДНК-хранилищах наряду с заменами распространены также вставки и удаления, что является мотивацией для изучения кодов, исправляющих такие ошибки.

Основным отличием задачи построения покрывающих кодов для случая вставок и удалений от кодов для замен является отсутствие симметрии. В метрике Хэмминга объем шара радиуса R не зависит от центра. Это же верно и для случая вставок. Однако при появлении удалений это свойство нарушается, и объем шара получается разным для разных центров, что усложняет задачу.

В табл. 1 мы приводим известные результаты о минимальной мощности покрывающего кода для случаев R замен, R удалений и R вставок. Величины $V_H^q(n, R)$

¹ Исследование выполнено за счет гранта Российского научного фонда № 22-41-02028.

Границы на минимальную мощность покрывающих кодов

Тип покрывающего кода	Нижняя граница	Верхняя граница	Ссылка
Одна замена, q – степень простого числа	$\frac{q^n}{(q-1)n+1}$	$\frac{q^n(1+o(1))}{(q-1)n+1}$	[9]
R замен	$\frac{q^n}{V_H^q(n, R)}$	$\frac{cR \log R q^n (1+o(1))}{V_H^q(n, R)}$	[10]
R удалений	$\frac{q^n R! (1+o(1))}{n^R (q-1)^R}$	$\frac{cR \log R q^n R! (1+o(1))}{n^R (q-1)^R}$	[11]
R вставок	$\frac{q^{n+R}}{V_I^q(n, R)}$	$\frac{cR \log R q^{n+R} (1+o(1))}{V_I^q(n, R)}$	[11]

и $V_I^q(n, R)$ обозначают объемы шаров радиуса R для замен и вставок соответственно, т.е.

$$V_H^q(n, R) = \sum_{i=0}^R \binom{n}{i} (q-1)^i,$$

$$V_I^q(n, R) = \sum_{i=0}^R \binom{n+R}{i} (q-1)^i.$$

Отметим, что для каждого из этих трех случаев верхняя и нижняя границы отличаются в $O(R \log R)$ раз. Зависимость же мощности кода от длины одинакова в верхней и нижней границах. Для случая одной замены и размера алфавита q , равного степени простого числа, верхняя и нижняя границы асимптотически эквивалентны.

В данной статье мы рассматриваем случай, в котором возможны одновременно и вставки, и удаления. Мы концентрируемся на случае, когда количество вставок и удалений одинаково и равно R . Основным результатом статьи является получение нижних и верхних оценок минимального размера покрывающего кода, которые имеют одинаковую (а значит, оптимальную) зависимость от длины кода n для $R = 1, 2$.

Оставшаяся часть статьи имеет следующую структуру. В § 2 введены основные обозначения и определения, а § 3 посвящен доказательству нижней границы для радиуса 1. Затем в § 4 получена верхняя граница минимального размера покрывающего кода. Случай радиуса $R > 1$ рассматривается в § 5. В § 6 даны заключительные замечания.

§ 2. Определения и обозначения

Алфавит $\{0, 1, \dots, q-1\}$ из q символов будем обозначать через Σ_q . Множество последовательностей длины n с символами из алфавита Σ_q обозначим через Σ_q^n . Кодом $\mathcal{C}$ будем называть произвольное подмножество в Σ_q^n . Знак $\triangleq$ будет означать равенство по определению.

Определение 1. *Расстоянием в метрике Левенштейна фиксированной длины* (fixed length Levenstein metric) между двумя последовательностями $\mathbf{x}$ и $\mathbf{y}$ одинаковой длины назовем минимальное число R , такое что из $\mathbf{x}$ можно получить $\mathbf{y}$ с помощью R удалений и R вставок. Будем обозначать такое расстояние через $d_{\text{FLL}}(\mathbf{x}, \mathbf{y})$.

Определение 2. *Шаром $B(\mathbf{x}, R)$* с центром в $\mathbf{x}$ и радиусом R в метрике Левенштейна фиксированной длины будем называть множество всех последовательно-

стей, которые находятся на расстоянии не более R от $\mathbf{x}$, т.е.

$$B(\mathbf{x}, R) \triangleq \{\mathbf{y} \in \Sigma_q^n \mid d_{\text{FLL}}(\mathbf{x}, \mathbf{y}) \leq R\}.$$

В случае $R = 1$ будем обозначать шар через $B(\mathbf{x})$.

Определение 3. Будем говорить, что код $\mathcal{C}$ является *покрывающим кодом* (или просто *покрытием*) радиуса R в метрике Левенштейна фиксированной длины, если объединение шаров радиуса R с центрами в кодовых словах покрывает все пространство, т.е. иными словами, если $\bigcup_{\mathbf{x} \in \mathcal{C}} B(\mathbf{x}, R) = \Sigma_q^n$.

В дальнейшем будем называть такой код покрывающим кодом радиуса R , а слова о метрике Левенштейна фиксированной длины будем опускать. Следующие понятия являются важными характеристиками, влияющими на размер шара.

Определение 4. *Серией* (run) в строке $\mathbf{x}$ будем называть максимальный по включению непрерывный подотрезок строки $\mathbf{x}$, состоящий из одинаковых символов. Количество серий в строке $\mathbf{x}$ будем обозначать через $r(\mathbf{x})$. Например, $r(01101) = 4$, $r(000000) = 1$.

Определение 5. *Отрезком чередования* (alternating segment) в строке $\mathbf{x}$ называется максимальный по включению непрерывный подотрезок строки $\mathbf{x}$, состоящий из двух чередующихся символов алфавита, т.е. на четных позициях стоит один символ, а на нечетных другой. Например, в строке 021211 есть три отрезка чередования: 02, 2121 и 1.

§ 3. Нижняя граница мощности покрытия радиуса 1

Начнем с доказательства простейшей нижней границы мощности покрывающего кода.

Предложение 1. *Мощность покрывающего кода $\mathcal{C}$ радиуса 1 не меньше чем*

$$\frac{q^n(1 + o(1))}{(q - 1)n^2}.$$

Доказательство. Если у нас есть слово длины n , то количество удалений не больше n , количество последующих вставок не больше $n(q - 1) + 1$. Таким образом, количество покрываемых слов, считая исходное, не больше $n^2(q - 1)(1 + o(1))$.

Поэтому необходимо не менее $\frac{q^n(1 + o(1))}{(q - 1)n^2}$ кодовых слов. ▲

Эту границу можно улучшить, воспользовавшись методом сферической упаковки для несимметричных шаров, предложенным в [12, 13].

Лемма 1 [13]. *Для любого слова $\mathbf{x}$ определим $S(\mathbf{x})$ как множество слов, которые покрываются шаром с центром в $\mathbf{x}$. Пусть задана весовая функция $w: \Sigma_q^n \rightarrow \mathbb{R}$, причем для всех $\mathbf{x}$ выполняются неравенства*

$$\sum_{\mathbf{y} \in S(\mathbf{x})} w(\mathbf{y}) \leq 1. \tag{1}$$

Тогда размер любого покрытия $\mathcal{C}$ удовлетворяет неравенству

$$|\mathcal{C}| \geq \sum_{\mathbf{x} \in \Sigma_q^n} w(\mathbf{x}).$$

Докажем следующую оценку.

Теорема 1. *Мощность покрывающего кода $\mathcal{C}$ радиуса 1 не меньше чем*

$$\frac{q^{n+1}(1+o(1))}{(q-1)^2n^2}.$$

Отметим, что эта оценка асимптотически в $\frac{q}{q-1}$ раз превосходит оценку из предложения 1.

Доказательство. Воспользуемся леммой 1 для доказательства оценки. В качестве множеств $S(\mathbf{x})$ возьмем шары $B(\mathbf{x})$. Определим весовую функцию

$$w(\mathbf{x}) = \frac{1}{\max_{\mathbf{y} \in B(\mathbf{x})} |B(\mathbf{y})|},$$

тогда условие (1) выполняется, так как суммируется $|B(\mathbf{x})|$ слагаемых, каждое из которых не больше $|B(\mathbf{x})|^{-1}$.

Для объема шара радиуса 1 известна следующая формула [14]:

$$|B(\mathbf{x})| = r(\mathbf{x})(n(q-1)-1) + 2 - \sum_{i=1}^{a(\mathbf{x})} \frac{(s_i-1)(s_i-2)}{2}, \quad (2)$$

где $a(\mathbf{x})$ – количество отрезков чередования в $\mathbf{x}$, а s_i – длины этих отрезков. Воспользуемся следующей верхней оценкой объема шара:

$$|B(\mathbf{x})| \leq r(\mathbf{x})(q-1) + 1.$$

Несложно видеть, что количество серий при одном удалении и одной вставке может измениться не более чем на 2. Таким образом,

$$w(\mathbf{x}) \geq \frac{1}{(r(\mathbf{x})+2)n(q-1)+1}.$$

Нам осталось оценить снизу сумму весов всех слов. Число слов с k сериями равно

$$q(q-1)^{k-1} \binom{n-1}{k-1},$$

так как есть q способов выбрать первый символ, а потом надо выбрать позиции начал всех серий, кроме первого, а также символы для этих серий. Поэтому из леммы 1 вытекает следующее неравенство:

$$|\mathcal{C}| \geq \sum_{k=1}^n \frac{q(q-1)^{k-1} \binom{n-1}{k-1}}{(k+2)n(q-1)+1}.$$

Нам понадобится следующее утверждение.

Предложение 2. *Число серий в случайной строке из Σ_q^n лежит в интервале*

$$I = \left(\frac{(q-1)n}{q} - \sqrt{n} \ln n, \frac{(q-1)n}{q} + \sqrt{n} \ln n \right)$$

с вероятностью $1 - n^{-\ln n(1+o(1))}$.

Доказательство. Заметим, что число серий случайного слова дается выражением $r(\mathbf{x}) = 1 + \xi$, где ξ – биномиальная случайная величина $\text{Bin}\left(n-1, \frac{q-1}{q}\right)$.

Таким образом, вероятность того, что количество серий не попадает в интервал I , можно оценить по неравенству Хёфдинга [15]:

$$\mathbf{P} \left(\left| r(\mathbf{x}) - \frac{(q-1)n}{q} \right| \geq \sqrt{n} \ln n \right) \leq 2 \exp \left\{ \frac{-2n \ln^2 n}{n} (1 + o(1)) \right\} < n^{-\ln n(1+o(1))}. \quad \blacktriangle$$

Теперь оценим мощность кода C :

$$|C| \geq \sum_{k=1}^n \frac{q(q-1)^{k-1} \binom{n-1}{k-1}}{(k+2)n(q-1)+1} \geq \sum_{k \in I} \frac{q(q-1)^{k-1} \binom{n-1}{k-1}}{(k+2)n(q-1)+1}.$$

Заметим, что при $k \in I$ знаменатель равен $\frac{(q-1)^2 n^2}{q} (1+o(1))$. Сумма же выражений в числителе по $k \in I$ равна вероятности того, что количество серий в случайной строке принадлежит отрезку I , умноженной на q^n . Воспользовавшись только что доказанным предложением 2, получаем следующее:

$$\begin{aligned} |C| &\geq \sum_{k \in I} \frac{q(q-1)^{k-1} \binom{n-1}{k-1}}{(k+2)n(q-1)+1} = \frac{q^n (1 - n^{-\ln n(1+o(1))})}{(q-1)^2 n^2 q^{-1} (1+o(1))} \geq \\ &\geq \frac{q^{n+1}}{(q-1)^2 n^2} (1+o(1)), \end{aligned}$$

что завершает доказательство теоремы. $\blacktriangle$

§ 4. Верхняя оценка минимальной мощности покрытия радиуса 1

В этом параграфе мы докажем верхнюю оценку минимальной мощности покрытия, т.е. докажем существование покрытия определенной мощности.

Нам понадобятся некоторые факты о размерах шаров. В работах [16, 17] доказано, что размер случайного шара радиуса 1 и 2 в метрике Левенштейна фиксированной длины достаточно близок к своему математическому ожиданию с вероятностью, стремящейся к 1. Нам нужна чуть более сильная оценка, которая, тем не менее, доказывается похожим образом.

Введем дополнительное определение.

Определение 6. Назовем слово $\mathbf{x} \in \Sigma_q^n$ *плохим*, если выполнено хотя бы одно из следующих условий:

1. Число серий не принадлежит интервалу

$$I = \left(\frac{q-1}{q}n - \sqrt{n} \ln n, \frac{q-1}{q}n + \sqrt{n} \ln n \right).$$

2. Максимальный отрезок чередования имеет длину не менее $4 \log n$.

Множество плохих слов будем обозначать через $\mathcal{B}$. Все остальные слова назовем *хорошими* и будем обозначать через $\mathcal{G}$.

Лемма 2. *Справедливы следующие утверждения:*

1. Количество плохих слов не превосходит $o\left(\frac{q^n}{n^2}\right)$;
2. Количество хороших слов не менее $q^n(1 - o(n^{-2}))$;

3. Размер шара $B(\mathbf{x})$ для $\mathbf{x} \in \mathcal{G}$ имеет порядок

$$|B(\mathbf{x})| \sim \frac{n^2(q-1)^2}{q}.$$

Доказательство. Докажем утверждение 1. Как было показано выше в предложении 2, количество серий $r(\mathbf{x})$ не принадлежит интервалу $I = \left(\frac{q-1}{q}n - \sqrt{n} \ln n, \frac{q-1}{q}n + \sqrt{n} \ln n \right)$ с вероятностью $\leq n^{-\ln n(1+o(1))}$.

Теперь оценим вероятность появления отрезка чередования длины $k = 4 \log n$. Вероятность конкретного отрезка чередования в конкретном месте равна q^{-k} . Оценим вероятность существования отрезка чередования длины k как сумму вероятностей по всем позициям и по всем парам символов алфавита:

$$\mathbf{P}(\text{существование отрезка чередования длины } k) \leq q^2 q^{-k} n = o(n^{-2}).$$

Утверждение 2 тривиально следует из предыдущего.

Утверждение 3 следует из определения хороших слов и формулы (2) для размера шара. $\blacktriangle$

Теперь докажем случайную верхнюю границу.

Теорема 2. *Существует покрытие $\mathcal{C}$ радиуса 1 и мощности не более*

$$\frac{2q^{n+1} \ln n}{(q-1)^2 n^2}.$$

Доказательство. Опишем покрывающий код $\mathcal{C}$. Все плохие слова сразу добавим в код $\mathcal{C}$. Хорошие слова будем включать в код случайным образом, а именно каждое хорошее слово будем включать в код независимо с вероятностью p . Для каждого хорошего слова $\mathbf{x}$ определим вероятность p_0 того, что оно не покрыто. Тогда

$$p_0 \leq (1-p)^{|B(\mathbf{x})|} = (1-p)^{\frac{n^2(q-1)^2}{q}(1+o(1))}.$$

В конце добавим все непокрытые слова в код.

Математическое ожидание размера кода оценивается следующим образом:

$$\mathbf{E} |\mathcal{C}| \leq q^n(p + p_0) + o(q^n/n^2).$$

Если взять $p = \frac{2q \ln n}{(q-1)^2 n^2}$, то получим

$$p_0 \leq e^{-2 \ln n(1+o(1))} = n^{-2+o(1)},$$

а математическое ожидание размера кода не превосходит

$$\frac{2q^{n+1} \ln n}{(q-1)^2 n^2} (1 + o(1)).$$

Очевидно, что существует код, размер которого не превосходит математического ожидания. $\blacktriangle$

Отметим, что случайная верхняя граница отличается от доказанной ранее нижней в $O(\ln n)$ раз. Эту границу можно улучшить, строя покрытия из покрытий в пространствах меньшей размерности, а также кодов, покрывающих не все пространство, но его существенную часть. Наше рассуждение во многом повторяет работу [11], где похожая лемма использовалась для построения покрывающих кодов для случая одной вставки. Также аналогичная стратегия использовалась в работе [18].

Лемма 3. Пусть

$$S, T \subset \Sigma_q^{n_1}, \quad T = \Sigma_q^{n_1} \setminus \bigcup_{x \in S} B(x).$$

Пусть код C_{n_2} длины n_2 является покрывающим кодом радиуса 1 в метрике Ле-венштейна фиксированной длины. Тогда код

$$(S \otimes \Sigma_q^{n_2}) \cup (T \otimes C_{n_2})$$

является покрывающим кодом радиуса 1, длины $n_1 + n_2$ и размера $|S|q^{n_2} + |T||C_{n_2}|$, где

$$A \otimes B = \{ab \mid a \in A, b \in B\},$$

$a \mathbf{ab}$ – конкатенация слов a и b .

Доказательство. Рассмотрим произвольное слово длины $n = n_1 + n_2$. Разобьем его на две части $\mathbf{x}\mathbf{y}$, где длина $\mathbf{x}$ равна n_1 , а длина $\mathbf{y}$ равна n_2 . Если $\mathbf{x}$ покрывается множеством S , то и все слово покрыто множеством $S \otimes \Sigma_q^{n_2}$. В противном случае $\mathbf{x} \in T$. Так как код C_{n_2} покрывает все слова $\mathbf{y}$ длины n_2 , то в этом случае $\mathbf{x}\mathbf{y}$ покрывается множеством $T \otimes C_{n_2}$.

Определим функцию $\mu(n, q)$ следующим образом. Пусть $\mu(1, q) = 1$. Для $n > 1$ определим $\mu(n, q)$ как минимальное число, такое что существуют натуральные числа n_1 и n_2 , $n = n_1 + n_2$, а также множества $S, T \subset \Sigma_q^{n_1}$, где S покрывает $\Sigma_q^{n_1} \setminus T$, удовлетворяющие неравенству

$$|S| + \frac{\mu(n_2, q)|T|}{n_2^2} \leq \frac{\mu(n, q)q^{n_1}}{n^2}. \quad (3)$$

Следующая лемма позволит нам доказать верхнюю границу мощности минимального покрытия с помощью индукции.

Лемма 4. Справедливы следующие утверждения:

1. Для любого $n > 1$ существуют натуральные числа n_1 и n_2 , $n = n_1 + n_2$, а также множества $S, T \subset \Sigma_q^{n_1}$, где S покрывает $\Sigma_q^{n_1} \setminus T$, удовлетворяющие неравенству (3);
2. Для любого q для верхнего предела $\bar{\mu}(q) \triangleq \limsup_{n \rightarrow \infty} \mu(n, q)$ справедливо неравенство $\bar{\mu}(q) \leq \mu(q)$, где $\mu(q)$ определяется как минимальное положительное значение

$$\mu(q) = \frac{c(1 - \alpha)^2}{\alpha^2 \left((1 - \alpha)^2 - e^{-\frac{c(q-1)^2}{q}} \right)} \quad (4)$$

по всем $c > 0$ и $\alpha \in (0, 1)$;

3. При $q \rightarrow \infty$ выполняется $\mu(q) = O(q^{-1})$. Точнее, подставляя $c = \frac{5,282}{q}$ и $\alpha = 0,851$, получаем $\mu(q) \leq \frac{9,459}{q}(1 + o(1))$.

Доказательство. Утверждение 1 леммы следует из определения функции $\mu(n, q)$.

Для доказательства утверждения 2 покажем, как можно строить множества S и T . На длине n_1 построим множество $S_0 \subset \mathcal{G} \subset \Sigma_q^{n_1}$ случайным образом, включая в него каждое хорошее слово с вероятностью p . Определим S как объединение множества S_0 с плохими словами $S = S_0 \cup \mathcal{B}$.

Пользуясь леммой 2, оценим математическое ожидание величины

$$W = |S| + \frac{\mu(n_2, q)|T|}{n_2^2},$$

равной выражению в левой части (3), при $n \rightarrow \infty$ следующим образом:

$$\mathbf{E} W \leq o\left(\frac{q^{n_1}}{n_1^2}\right) + q^{n_1}p + \frac{\mu(n_2, q)q^{n_1}(1-p)^{\frac{n_1^2(q-1)^2}{q}(1+o(1))}}{n_2^2}.$$

Пусть $p = \frac{c}{n_1^2}$, $n_1 = \alpha n$, $n_2 = (1 - \alpha)n$. Формально следовало бы написать $\lfloor \alpha n \rfloor$, но мы будем игнорировать целые части, так как мы все равно делаем оценку с точностью до умножения на $1 + o(1)$. Тогда неравенство можно переписать следующим образом:

$$\begin{aligned} \mathbf{E} W &\leq \frac{cq^{n_1}}{\alpha^2 n^2}(1 + o(1)) + \frac{q^{n_1}\mu(n_2, q)e^{-\frac{c(q-1)^2}{q}(1+o(1))}}{(1 - \alpha)^2 n^2} = \\ &= \frac{q^{n_1}}{n^2} \left(\frac{c}{\alpha^2} + \frac{\mu(n_2, q)e^{-\frac{c(q-1)^2}{q}}}{(1 - \alpha)^2} \right) (1 + o(1)), \end{aligned}$$

откуда следует, что выполняется неравенство

$$\mu(n, q) \leq \left(\frac{c}{\alpha^2} + \frac{\mu(n_2, q)e^{-\frac{c(q-1)^2}{q}}}{(1 - \alpha)^2} \right) (1 + o(1)). \quad (5)$$

Зафиксируем $\alpha \in (0, 1)$ и $c > 0$ и возьмем верхний предел левой и правой части при $n \rightarrow \infty$. Получим

$$\bar{\mu}(q) \leq \frac{c}{\alpha^2} + \frac{\bar{\mu}(q)e^{-\frac{c(q-1)^2}{q}}}{(1 - \alpha)^2}.$$

Пусть c и α выбраны так, что выполняется неравенство

$$\frac{e^{-\frac{c(q-1)^2}{q}}}{(1 - \alpha)^2} < 1.$$

Тогда $\bar{\mu}(q) < +\infty$, так как из неравенства (5) следует, что если для некоторого достаточно большого M выполняется $\mu(n_2, q) > M$, то $\mu(n, q) < \mu(n_2, q)$. Элементарные преобразования приводят к формуле (4). Выполнение неравенства $\frac{e^{-\frac{c(q-1)^2}{q}}}{(1 - \alpha)^2} < 1$ эквивалентно тому условию, что $\mu(q)$ должно быть положительно. Утверждение 2 доказано.

Для доказательства утверждения 3 убедимся, что если положить $c = \frac{A}{q}$, то

$$\mu(q) \leq \frac{A(1 - \alpha)^2}{q\alpha^2((1 - \alpha)^2 - e^{-A})}(1 + o(1)),$$

откуда следует $\mu(q) = O(q^{-1})$. Значения параметров A и α найдены с помощью компьютерного перебора. ▲

Для $q = 2$ константа $\mu(2) \approx 18,53$ при $c \approx 11,42$ и $\alpha \approx 0,85$. Для $q = 3$ константа $\mu(3) \approx 6,95$ при $c \approx 4,28$ и $\alpha \approx 0,85$.

Теперь мы готовы к доказательству верхней границы.

Теорема 3. *Существует покрытие $\mathcal{C}$ радиуса 1 мощности*

$$M \leq \frac{\mu(n, q)q^n}{n^2}.$$

Доказательство. Будем доказывать теорему по индукции. Для $n = 1$ функция $\mu(1, q) = 1$, и утверждение теоремы очевидно выполнено.

Допустим, что теорема доказана для всех длин не выше $n - 1$. Применяя лемму 4, получаем, что существуют множества $S, T \subset \Sigma_q^{n-1}$, где S покрывает $\Sigma_q^{n-1} \setminus T$, удовлетворяющие неравенству (3). Применяя к этим множествам лемму 3, получаем, что существует покрывающий код длины n и мощности $M \leq \frac{\mu(n, q)q^n}{n^2}$, что и требовалось. $\blacktriangle$

Сравнивая полученные в нашей статье нижнюю и верхнюю границу, получаем, что при $n \rightarrow \infty$ они отличаются в $\frac{\mu(q)(q-1)^2}{q}$ раз. Как следует из последнего утверждения леммы 4, при $q \rightarrow \infty$ это отношение стремится к константе, не большей 9,459.

§ 5. Случай радиуса, большего 1

В данном параграфе обсуждается обобщение полученных результатов на случай радиуса $R > 1$.

5.1. Нижняя граница. Снова начнем с наивной нижней границы мощности покрытия.

Предложение 3. *Мощность покрывающего кода $\mathcal{C}$ радиуса R не меньше чем*

$$\frac{q^n (R!)^2 (1 + o(1))}{(q-1)^R n^{2R}}.$$

Доказательство. Если у нас есть слово длины n , то количество удалений не больше $\binom{n}{R}$, а количество последующих вставок не больше

$$\sum_{i=0}^R \binom{n}{i} (q-1)^i = \binom{n}{R} (q-1)^R (1 + o(1)).$$

Таким образом, количество покрываемых слов не больше

$$\binom{n}{R}^2 (q-1)^R (1 + o(1)).$$

Поэтому необходимо не менее чем $\frac{q^n (R!)^2 (1 + o(1))}{(q-1)^R n^{2R}}$ кодовых слов.

Докажем более сильную оценку.

Теорема 4. *Мощность покрывающего кода $\mathcal{C}$ радиуса R не меньше чем*

$$\frac{q^{n+R} (R!)^2}{(q-1)^{2R} n^{2R}} (1 + o(1)).$$

Отметим, что эта оценка асимптотически в $\left(\frac{q}{q-1}\right)^R$ раз превосходит наивную.

Доказательство. Снова воспользуемся леммой 1 для доказательства оценки, только теперь $S(\mathbf{x}) = B(\mathbf{x}, R)$. Аналогично случаю $R = 1$ определим весовую функцию

$$w(\mathbf{x}) = \frac{1}{\max_{\mathbf{y} \in B(\mathbf{x}, R)} |B(\mathbf{y}, R)|},$$

что гарантирует выполнение условия (1).

Для объема шара радиуса $R > 1$ точной формулы не известно, но мы можем использовать следующую верхнюю оценку:

$$|B(\mathbf{x}, R)| \leq \binom{r(\mathbf{x}) + R - 1}{R} \sum_{i=0}^R \binom{n}{i} (q-1)^i,$$

где первый множитель – это максимальное количество слов, которое может получиться из $\mathbf{x}$ после R удалений [19], а второй – количество различных слов, которые могут получиться из одного слова длины $n - R$ с помощью R вставок [20].

Отметим, что количество серий при R удалениях и R вставках может измениться не более чем на $2R$. Таким образом,

$$w(\mathbf{x}) \geq \frac{1}{\binom{r(\mathbf{x}) + 3R - 1}{R} \sum_{i=0}^R \binom{n}{i} (q-1)^i}.$$

Вспомяная, что число слов с k сериями равно $q(q-1)^{k-1} \binom{n-1}{k-1}$, из леммы 1 получаем оценку

$$|C| \geq \sum_{k=1}^n \frac{q(q-1)^{k-1} \binom{n-1}{k-1}}{\binom{k+3R-1}{R} \sum_{i=0}^R \binom{n}{i} (q-1)^i} \geq \sum_{k \in I} \frac{q(q-1)^{k-1} \binom{n-1}{k-1}}{\binom{k+3R-1}{R} \sum_{i=0}^R \binom{n}{i} (q-1)^i}.$$

Заметим, что для $k \in I$

$$\binom{k+3R-1}{R} \sum_{i=0}^R \binom{n}{i} (q-1)^i = \frac{n^R (q-1)^R}{R! q^R} \frac{n^R (q-1)^R}{R!} (1 + o(1)). \quad (6)$$

В то же время из предложения 2 следует, что

$$\sum_{k \in I} q(q-1)^{k-1} \binom{n-1}{k-1} = q^n (1 - n^{-\ln n(1+o(1))}). \quad (7)$$

Используя формулы (6), (7), получаем

$$\begin{aligned} |C| &\geq \sum_{k \in I} \frac{q(q-1)^{k-1} \binom{n-1}{k-1}}{\binom{k+3R-1}{R} \sum_{i=0}^R \binom{n}{i} (q-1)^i} = \\ &= \frac{(R!)^2 q^R (1 + o(1))}{n^{2R} (q-1)^{2R}} \sum_{k \in I} q(q-1)^{k-1} \binom{n-1}{k-1} = \frac{q^{n+R} (R!)^2}{(q-1)^{2R} n^{2R}} (1 + o(1)), \end{aligned}$$

что завершает доказательство теоремы. $\blacktriangle$

5.2. Верхняя граница. Для доказательства верхней границы мощности минимального покрытия нам необходима нижняя граница мощности шара. В работе [17] была сформулирована следующая гипотеза об объеме шара в метрике Левенштейна фиксированной длины.

Гипотеза 1. Для размера шара $B(\mathbf{x}, R)$ со случайно выбранным согласно равномерному распределению центром выполняется следующее:

$$\begin{aligned} \mathbf{E} |B(\mathbf{x}, R)| &= \frac{(q-1)^{2R}}{(R!)^2 q^R} n^{2R} + o(n^{2R}), \\ \mathbf{P} \left(\left| |B(\mathbf{x}, R)| - \frac{(q-1)^{2R}}{(R!)^2 q^R} n^{2R} \right| = o(n^{2R}) \right) &= 1 - o(1). \end{aligned}$$

На текущий момент гипотеза доказана только для $R = 1, 2$ (см. [16, 17]). Но для нашего метода доказательства нам нужно более точное утверждение о концентрации мощностей шаров, а именно следующее.

Гипотеза 2. Для размера шара $B(\mathbf{x}, R)$ со случайно выбранным согласно равномерному распределению центром выполняется следующее:

$$\mathbf{P} \left(\left| |B(\mathbf{x}, R)| - \frac{(q-1)^{2R}}{(R!)^2 q^R} n^{2R} \right| > n^{2R-0.5} \ln n \right) = o(n^{-2R}).$$

Для случая $R = 1$ доказательство гипотезы 2 фактически приведено в § 4. Доказательство гипотезы 1 для случая $R = 2$, приведенное в работе [17], работает и для гипотезы 2 с минимальными изменениями. Для $R > 2$ задача доказательства гипотезы 2 остается открытой.

Мы же докажем, что из выполнения гипотезы 2 следует верхняя оценка мощности покрытия.

Теорема 5. *В случае выполнения гипотезы 2 существует покрытие $\mathcal{C}$ радиуса R и мощности $M \leq \frac{\lambda q^n}{n^{2R}}$ для некоторой константы λ .*

Доказательство. Доказательство проведем по индукции по длине кода n . Для любого фиксированного n_0 мы можем добиться выполнения утверждения теоремы выбором достаточно большой константы λ . Допустим, что теорема доказана для всех длин не выше $n - 1$, $n > n_0$, и докажем ее для длины n . Нам понадобится следующее утверждение, аналогичное лемме 3.

Лемма 5. *Пусть*

$$S, T \subset \Sigma_q^{n_1}, \quad T = \Sigma_q^{n_1} \setminus \bigcup_{\mathbf{x} \in S} B(\mathbf{x}, R).$$

Пусть код C_{n_2} длины n_2 является покрывающим кодом радиуса R в метрике Левенштейна фиксированной длины. Тогда код

$$(S \otimes \Sigma_q^{n_2}) \cup (T \otimes C_{n_2})$$

является покрывающим кодом радиуса R , длины $n_1 + n_2$ и размера $|S|q^{n_2} + |T||C_{n_2}|$, где

$$A \otimes B = \{\mathbf{ab} \mid \mathbf{a} \in A, \mathbf{b} \in B\},$$

$\mathbf{a} \mathbf{b}$ – конкатенация слов $\mathbf{a}$ и $\mathbf{b}$.

Доказательство леммы полностью аналогично доказательству леммы 3.

Теперь построим множества S и T . Сначала добавим в множество S все слова $\mathbf{x}$, для которых выполнено неравенство

$$\left| |B(\mathbf{x}, R)| - \frac{(q-1)^{2R}}{(R!)^2 q^R} n_1^{2R} \right| > n_1^{2R-0,5} \ln n_1.$$

Остальные слова будем включать в множество независимо с вероятностью p .

Оценим математическое ожидание величины $W = |S| + \frac{\lambda|T|}{n_2^{2R}}$. Оценка математического ожидания

$$\mathbf{E} |S| < o\left(\frac{q^{n_1}}{n_1^{2R}}\right) + q^{n_1} p$$

следует из гипотезы 2. Для оценки математического ожидания мощности множества T отметим, что для попадания слова $\mathbf{x} \notin S$ в это множество все слова из шара $B(\mathbf{x}, R)$ должны быть не включены в множество S , т.е.

$$\mathbf{E} |T| \leq q^{n_1} (1-p)^{|B(\mathbf{x}, R)|} \leq q^{n_1} (1-p)^{\frac{(q-1)^{2R}}{(R!)^2 q^R} n_1^{2R} (1+o(1))},$$

где последнее неравенство следует из того, что $\mathbf{x}$ не лежит в множестве S , а значит, для него выполнено

$$\left| |B(\mathbf{x}, R)| - \frac{(q-1)^{2R}}{(R!)^2 q^R} n_1^{2R} \right| \leq n_1^{2R-0,5} \ln n_1.$$

Итак, математическое ожидание $\mathbf{E} W$ оценивается следующим образом:

$$\mathbf{E} W \leq o\left(\frac{q^{n_1}}{n_1^{2R}}\right) + q^{n_1} p + \frac{q^{n_1} \lambda (1-p)^{\frac{(q-1)^{2R}}{(R!)^2 q^R} n_1^{2R} (1+o(1))}}{n_2^{2R}}.$$

Пусть $p = \frac{c}{n_1^{2R}}$, $n_1 = \alpha n$, $n_2 = (1-\alpha)n$. Тогда неравенство можно переписать в виде

$$\begin{aligned} \mathbf{E} W &\leq \frac{cq^{n_1}}{\alpha^{2R} n^{2R}} (1+o(1)) + \frac{q^{n_1} \lambda e^{-\frac{c(q-1)^{2R}}{(R!)^2 q^R} (1+o(1))}}{(1-\alpha)^{2R} n^{2R}} = \\ &= \frac{q^{n_1}}{n^{2R}} \left(\frac{c}{\alpha^{2R}} + \frac{\lambda e^{-\frac{c(q-1)^{2R}}{(R!)^2 q^R}}}{(1-\alpha)^{2R}} \right) (1+o(1)). \end{aligned}$$

Подберем α , c и λ так, чтобы выполнялось неравенство

$$\frac{c}{\alpha^{2R}} + \frac{\lambda e^{-\frac{c(q-1)^{2R}}{(R!)^2 q^R}}}{(1-\alpha)^{2R}} < \lambda.$$

Тогда при достаточно большом $n > n_0$ будет выполнено

$$\mathbf{E} W \leq \frac{\lambda q^{n_1}}{n^{2R}}.$$

Возьмем множества S и T , для которых выполнено неравенство $W \leq \frac{\lambda q^{n_1}}{n^{2R}}$. Применяя лемму 5 к ним, а также к минимальному покрытию C_{n_2} , верхняя оценка на мощность которого известна из предположения индукции, получаем новое покрытие

радиуса R , мощность которого не превосходит

$$|S|q^{n_2} + |T||C_{n_2}| \leq \frac{\lambda q^n}{n^{2R}},$$

что завершает доказательство теоремы. ▲

Таким образом, для $R = 2$ верна верхняя оценка мощности минимального покрытия $\frac{\lambda q^n}{n^{2R}}$, которая отличается от нижней лишь в константу раз.

§ 6. Заключение

В статье доказаны новые нижние и верхние оценки минимальной мощности покрывающего кода для метрики Левенштейна фиксированной длины при $R = 1$ и 2 . Как и в случае покрывающих кодов для замен, вставок и удалений, нижние и верхние оценки отличаются лишь в константу раз. Доказательство аналогичных результатов для $R > 2$ сведено к гипотезе 2 о распределении размера шара в метрике Левенштейна фиксированной длины со случайно выбранным центром. Естественным следующим шагом исследования является доказательство гипотезы 2, которая интересна не только своей связью с покрывающими кодами, но и как самостоятельный факт.

СПИСОК ЛИТЕРАТУРЫ

1. *Cohen G., Honkala I., Listyn S., Lobstein A.* Covering Codes. Amsterdam: Elsevier, 1997.
2. *Smolensky R.* On Representations by Low-Degree Polynomials // Proc. 34th Annu. Symp. on Foundations of Computer Science. Palo Alto, CA, USA. Nov. 3–5, 1993. P. 130–138. <https://doi.org/10.1109/SFCS.1993.366874>
3. *Pagh R.* Locality-sensitive Hashing without False Negatives // Proc. 27th Annu. ACM-SIAM Symp. on Discrete Algorithms (SODA'2016). Arlington, VA, USA. Jan. 10–12, 2016. P. 1–9. <https://doi.org/10.1137/1.9781611974331.ch1>
4. *Micciancio D.* Almost Perfect Lattices, the Covering Radius Problem, and Applications to Ajtai's Connection Factor // SIAM J. Comput. 2004. V. 34. № 1. P. 118–169. <https://doi.org/10.1137/S0097539703433511>
5. *Hämäläinen H., Honkala I., Litsyn S., Östergård P.* Football Pools—A Game for Mathematicians // Amer. Math. Monthly. 1995. V. 102. № 7. P. 579–588. <https://doi.org/10.2307/2974552>
6. *Ceze L., Nivala J., Strauss K.* Molecular Digital Data Storage Using DNA // Nat. Rev. Genet. 2019. V. 20. № 8. P. 456–466. <https://doi.org/10.1038/s41576-019-0125-3>
7. *Bornholt J., Lopez R., Carmean D.M., Ceze L., Seelig G., Strauss K.* A DNA-Based Archival Storage System // Proc. 21st Int. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS'16). Atlanta, GA, USA. Apr. 2–6, 2016. P. 637–649. <https://doi.org/10.1145/2872362.2872397>
8. *Church G.M., Gao Y., Kosuri S.* Next-Generation Digital Information Storage in DNA // Science. 2012. V. 337. № 6102. P. 1628. <https://doi.org/10.1126/science.1226355>
9. *Кабатянский Г.А., Панченко В.И.* Упаковки и покрытия хэммингова пространства единичными шарами // Докл. АН СССР. 1988. Т. 303. № 3. С. 550–552. <https://www.mathnet.ru/rus/dan7368>
10. *Krivelevich M., Sudakov B., Vu V.H.* Covering Codes with Improved Density // IEEE Trans. Inform. Theory. 2003. V. 49. № 7. P. 1812–1815. <https://doi.org/10.1109/TIT.2003.813490>
11. *Lenz A., Rashtchian C., Siegel P.H., Yaakobi E.* Covering Codes Using Insertions or Deletions // IEEE Trans. Inform. Theory. 2020. V. 67. № 6. P. 3376–3388. <https://doi.org/10.1109/TIT.2020.2985691>
12. *Fazeli A., Vardy A., Yaakobi E.* Generalized Sphere Packing Bound // IEEE Trans. Inform. Theory. 2015. V. 61. № 5. P. 2313–2334. <https://doi.org/10.1109/TIT.2015.2413418>

13. *Applegate D., Rains E.M., Sloane N.J.A.* On Asymmetric Coverings and Covering Numbers // J. Combin. Des. 2003. V. 11. № 3. P. 218–228. <https://doi.org/10.1002/jcd.10022>
14. *Sala F., Dolecek L.* Counting Sequences Obtained from the Synchronization Channel // Proc. 2013 IEEE Int. Symp. on Information Theory (ISIT'2013). Istanbul, Turkey. July 7–12, 2013. P. 2925–2929. <https://doi.org/10.1109/ISIT.2013.6620761>
15. *Hoeffding W.* Probability Inequalities for Sums of Bounded Random Variables // J. Amer. Statist. Assoc. 1963. V. 58. № 301. P. 13–30. <https://doi.org/10.2307/2282952>. Reprinted in: The Collected Works of Wassily Hoeffding. New York: Springer, 1994. P. 409–426.
16. *Wang G., Wang Q.* On the Size Distribution of Levenshtein Balls with Radius One, *arXiv: 2204.02201 [cs.IT]*, 2022.
17. *He L., Ye M.* The Size of Levenshtein Ball with Radius 2: Expectation and Concentration Bound // Proc. 2023 IEEE Int. Symp. on Information Theory (ISIT'2023). Taipei, Taiwan. June 25–30, 2023. P. 850–855. <https://doi.org/10.1109/ISIT54713.2023.10206888>
18. *Cooper J.N., Ellis R.B., Kahng A.B.* Asymmetric Binary Covering Codes // J. Combin. Theory Ser. A. 2002. V. 100. № 2. P. 232–249. <https://doi.org/10.1006/jcta.2002.3290>
19. *Левенштейн В.И.* Двоичные коды с исправлением выпадений, вставок и замещений символов // Докл. АН СССР. 1965. Т. 163. № 4. С. 845–848. <https://www.mathnet.ru/rus/dan31411>
20. *Levenshtein V.I.* Efficient Reconstruction of Sequences from Their Subsequences or Supersequences // J. Combin. Theory Ser. A. 2001. V. 93. № 2. P. 310–332. <https://doi.org/10.1006/jcta.2000.3081>

Воробьев Илья Викторович
 Технический университет Мюнхена, Германия
vorobyev.i.v@yandex.ru

Поступила в редакцию
 14.10.2023
 После доработки
 28.11.2023
 Принята к публикации
 28.11.2023