Лекция №8

Списочное декодирование. Приложение теории кодирования в вычисл. биологии.

Линейный $[n,k,d]$-код исправляет $\lfloor \frac{d-1}{2} \rfloor$ ошибок

Мы можем увеличить радиус декодирования, позволяя алг-мам вернуть <u>список</u> векторов.

I Списочное декодирование кода Рида-Соломона

$$RS_{\mathbb{F}, S=\{\alpha_1 \dots \alpha_n\}}[n,k] = \{(p(\alpha_1) \dots p(\alpha_n)) \in \mathbb{F}^n \mid p \in \mathbb{F}[x], \deg p < k\}$$

$d = n-k+1$

<u>Задача поиска двух многочленов:</u> Положим $p_1(x), p_2(x) \in \mathbb{F}[x]$, $\deg p_1(x) = \deg p_2(x) = k-1$, $n \geq 4k$ - чётное

$\alpha_1 \dots \alpha_n$ - различны. $\exists\, T \subset \{1 \dots n\}$, $|T| = n/2$.

Далее пусть нам даны

$$y_i = \begin{cases} p_1(\alpha_i) & , i \in T \\ p_2(\alpha_i) & i \in \{1 \dots n\} \setminus T \end{cases}$$

Задача состоит в вычислении $p_1(x), p_2(x)$ по данным парам $\{(\alpha_i, y_i)\}_{i \leq n}$

<u>Связь с декодированием RS</u> Мы могли бы считать $(y_i)_{i \leq n}$ — полученным словом, полагая, например, м-н $p_1(x)$ - исходным сообщением. Однако, $p_1(x)$ не совпадает с $\vec{y}$ на $\frac{n}{2}$ значениях α_i : $\frac{n}{2} > \lfloor \frac{d-1}{2} \rfloor = \lfloor \frac{n-k}{2} \rfloor$.

Поэтому, обычные алг-мы кода RS не подходят.

Решение: списочное декодирование

Из задачи поиска двух многочленов: $(y_i - p_1(d_i))(y_i - p_2(d_i)) = 0 \ \forall i \le n$

Положим, $Q(x,y) = (y - p_1(x))(y - p_2(x)) = y^2 - \underbrace{(p_1(x) + p_2(x))}_{B(x)} y + \underbrace{p_1(x) \cdot p_2(x)}_{C(x)} =$

$= y^2 - B(x) \cdot y + C(x)$, где $\deg B(x) = k-1$
$\deg C(x) = 2(k-1)$

$$Q(d_i, y_i) = 0 \quad \forall i \le n.$$

Алгоритм

I. Составим систему $Q(d_i, y_i) = 0$ из $\{b_0 \dots b_{k-1}, c_0, \dots c_{2(k-1)}\}$ неизвестных (всего $3k-1$) и n уравнений. Т.к. $n \ge 4k$, то система будет иметь решение. Получим мн-ны $B(x), C(x)$ в явном виде. Сложность $O(n^\omega)$, $2 < \omega \le 3$

II. Факторизуем мн-н $Q(x,y) = (y - f_1(x))(y - f_2(x))$.
Вернуть $f_1(x), f_2(x)$ факторизация мн-ов от двух переменных степени d над $\mathbb{F}_q$, $O(d^5 \log q)$

Корректность на шаге I. мы всегда отыщем какие-либо $B(x), C(x)$, т.к. $\exists$ решение $B(x) = p_1(x) + p_2(x)$, $C(x) = p_1(x) \cdot p_2(x)$.
докажем, что на II м шаге мы получим корректные $p_1(x), p_2(x)$.

Лемма $\forall Q(x,y)$, полученного на шаге I, справедливо $(y - p_1(x)) \mid Q(x,y)$ и $(y - p_2(x)) \mid Q(x,y)$

◁ докм утверждение для $p_1(x)$.
Заметим, $Q(x,y)$ - унитарный (от y) мн-н. Для того, чтобы показать, что $(y - \beta)$ делит Q, достаточно показать, что β - корень Q $(Q(\beta) = 0) \Rightarrow$ чтобы показать, что $y - p_1(x) \mid Q(x,y)$ достаточно показать, что $Q(x, p_1(x)) = 0$.

$* \; R(x) := Q(x, p_1(x))$, $\deg R(x) \leq 2(k-1)$.

$$\overset{\|}{p_1(x)^2 - (p_1(x)+p_2(x))p_1(x) + p_1(x)p_2(x)}$$

Заметим, что $\exists \; \frac{n}{2} \geq 2k$ различных α_i, т.ч. $p_1(\alpha_i) = y_i$

Для таких α_i, $R(\alpha_i) = Q(\alpha_i, p_1(\alpha_i)) = Q(\alpha_i, y_i) = 0 \Rightarrow$

мы нашли $\geq 2k$ корней мн-на степени $\leq 2(k-1) \Rightarrow R(x) = 0$ ▶

II Задача: групповое тестирование (group testing)

<u>Пример</u>: даны N человек, $s \ll N$ из которых заражены. Задача: выявить заражённых за min. число тестов.

Тривиальное решение: выполнить N тестов. Проблема: анализ может быть дорогим.

Можно ли смешать образцы крови так, чтобы было возможно определить за меньшее число тестов?

<u>Идея</u>

- ассоциируем с каждым человеком кодовое слово $c \in RS_{\mathbb{F},S}[n,k]$, $|S| = n$
- имеем $n \cdot |\mathbb{F}|$ колб / тестов, сформированных в матрицу $|\mathbb{F}| \times n$, $\mathbb{F}$ где столбцы пронумерованы $\{c_i\}_i$
- $d = n - k + 1$ - min. расстояние $RS \Rightarrow \forall$ два c_i, c_j отличаются на как min. d позициях
- образцы крови человека c помещаются в колбы $(c[i], c_i)$

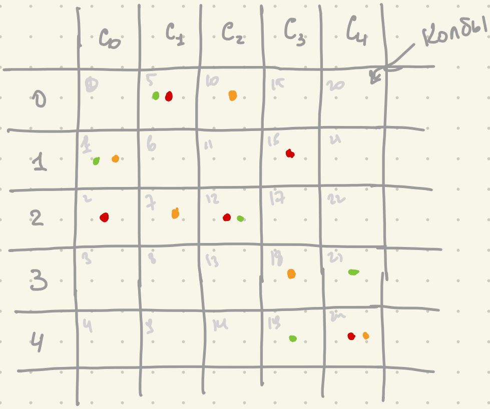
<u>Пример</u> $RS_{\mathbb{F}_5}$, $n = 5$, $k = 3$, $d = 3$ $\quad s \leq 2$

$|RS_{\mathbb{F}_5}| = 5^3 = 125 \Rightarrow$ макс. 125 человек

$c^* = (c_0^*, c_1^*, c_2^*, c_3^*, c_4^*)$ - заражённый №1
$\quad\;\; (2, 0, 2, 1, 4)$

$c^{**} = (1, 2, 0, 3, 3)$ - заражённый №2

$c = (1, 0, 2, 4, 3)$ - здоровый



Если $s=1 \Rightarrow \exists$ 3 негативных теста $\forall$ здорового

Если $s=2 \Rightarrow \exists$ 1 негативный тест

Для здорового человека $\exists$ хотя бы 1 негативный тест.

Для больного все тесты будут положительные.

Формально

N человек
s заражённых $(s \ll N)$
T — число тестов

A_i — подмн-во $\{1, .., T\}$ тестов, в которых человек i принимает участие

$\mathfrak{A}$ — мн-во таких подм-в $\mathfrak{A} = \{A_1, A_2 .., A_N\}$

$\mathfrak{A}$ называется s-дизъюнктивным, если ни одно мн-во из $\mathfrak{A}$ не содержится в объединении s других мн-в

(в примере, $\mathfrak{A}$ — 2-дизъюнктивное)

Теорема Для $1 \le s \ll N$, $\exists$ s-дизъюнктивное мн-во $\mathfrak{A}$ для T-тестов, удовлетворяющее

$$T = \Theta\left(s^2 \cdot \left(\frac{\lg N}{\lg s}\right)^2\right)$$

$\triangleleft$ $\exists$ Код RS с пар-ми $[q, k, q-k+1]_q$, $|RS| = q^k$

Каждому кодовому слову сопоставим бинарный вектор $a \in \mathbb{F}_2^{q^2}$, где i-ый блок длины q в a — индикатор значения $c[i]$

(в примере, $c[0]=1 \Rightarrow (\underbrace{01000}_{0\text{-блок}})$)

В итоге, получим вектор a с q единицами.

Построим мн-во $\mathfrak{A}$ для $T=q^2$, $N=q^k$, где каждому слову ставится в соответствие $\underbrace{\text{мн-во ненулевых координат}}_{\text{колбы}}$ вектора a

(в примере $C \to \{1, 5, 12, 13, 23\}$

Тогда 1) $\min\limits_{1 \le j \le N} |A_j| = q$ (число ненулевых координат)

2) $\max\limits_{1 \le j_1 < j_2 \le N} |A_{j_1} \cap A_{j_2}| = q - \min\limits_{c_i \ne c_j} \Delta(c_i, c_j) = q - (q - k + 1) = k - 1$

$\Rightarrow A$ – s-дизъюнктивно для $s = \left\lfloor \frac{q-1}{k-1} \right\rfloor$, т.к. объединение

$\left\lfloor \frac{q-1}{k-1} \right\rfloor$ содержит в себе $\left\lfloor \frac{q-1}{k-1} \right\rfloor \cdot (k-1) < q < \min |A_i|$ эл-тов

$\Rightarrow$ никакое мн-во не содержится в объединении $s = \left\lfloor \frac{q-1}{k-1} \right\rfloor$ мн-в.

Выберем в качестве $q = \text{const} \cdot s \cdot \frac{\lg N}{\lg s}$ (простое q для const $\in [2, 4]$ найдётся)

$\Rightarrow k - 1 \approx s(q-1) \Rightarrow k \approx s \cdot (q-1) \Rightarrow q \approx \frac{\lg N}{\lg s} \cdot s \Rightarrow T = q^2 = O\left(s^2 \cdot \frac{\lg^2 N}{\lg^2 s}\right)$