Защита информации

Павел Юдаев

МГТУ им. Баумана, Кафедра ИУ-9

Москва, 2014

Раздел 22 - Скрытое получение информации

Передача данных (битов) по двоичному каналу с ошибкой. Ошибка аддитивна.

Вер-ть ошибки в каждом бите: $p < 1/2$. Ошибки независимы.

Для испр. ошибки потребуется передавать избыточную информацию (кодировать исх. данные.). Либо кодировать всч сообщ. целиком, либо поблочно. Размер блока - 1 бит и более.

Векторы - строки.

Опр.

Расстояние Хэмминга между бинарными векторами.

Опр. (Двоичный блочный код)

$B = \{0, 1\}$. Код: $B^k \to C \subset B^n$.

Опр.

k/n - скорость передачи кода.

Опр.

$d = \min\limits_{c_1, c_2 \in C} \rho(c_1, c_2)$ - кодовое расстояние.

Очевидно, $[n, k, d]$ код замечает $d - 1$ ошибку и исправляет $[(d-1)/2]$ ошибок.

Пример: $0 \to 000, 1 \to 111$.

Декодировние - поблочно. Считаем, что число ошибок в блоке не превосходит $[(d-1)/2]$.

$w(x)$ - вес вектора x, число единиц.
Пусть $y \in B^n$ - принятый вектор. Цель - найти к.с.
$x : y = x \oplus e : x = \arg\min\limits_{x \in C} w(e)$.

Мажоритарное декодирование: к ближайшему кодовому слову. (Тем или иным способом: по таблицам для B^n, или путем записи нескольких СЛАУ для битов кодового слова и выбора значений i-го бита из них: большинство систем дает ответ a - значит, этот бит кодового слова равен a.)

Можно декодировать эффективней.

Линейные коды.

Опр.

Линейным блоковым кодом длины n наз. лин. подпр-во C лин. пр-ва B^n.

Замеч.: $0 \in C$.

Способы задания п-пр-ва: через его базис или через базис ортогонального п-пр-ва.

Опр.

Пусть H - двоичн. матр. $(n-k) \times n$. Линейн. блок. кодом длины n с проверочной матрицей H наз. множ.
$C = \{x \in B^n | Hx^T = 0\}$

Параметры линейного блочного кода:

- n - длина кода
- k - размерность п-пр-ва C - размерность кода.
- кодовое расстояние d лин. кода C равно мин. весу Хэм. ненулевых кодовых слов.

Код с этими пар-ми наз. $[n, k, d]$ кодом.

Теорема 1

Код C с пров. матр. H имеет код. расст. $d \Leftrightarrow \forall d-1$ стб. м-цы H лин/незав., но сущ. d лин/зав. стб.

Док-во

Рассм. $\forall x \in C$. Переведем его в 0 линейным отображением $M : B^n \to B^n$, $\exists M^{-1}$.

$\forall x' \in C : w(x') \leq d-1 \; HM^T x^T \neq 0$, но
$\exists x \in C : \exists x' \neq x \in C : w(x') = d, HM^T x'^T = 0$.

Опр.

Порождающей матрицей линейного $[n, k, d]$ кода C называется матрица G $(k \times n)$, строками которой являются базисные векторы линейного пространства C. С помощью порождающей матрицы код можно представить в виде $C = \{x = aG | \forall a \in B^k\}$.

Т.е. кодирование - суть умножение справа на матрицу G.

Задача

Пусть G, H - порожд. и провер. м-цы кода C. Доказать, что $HG^T = GH^T = 0$.

Опр.

Пусть $[n, k, d]$ код C имеет проверочную матрицу H. Пусть y - любой вектор длины n. Синдромом вектора y называется вектор $S(y) = Hy^T$.

Если $y \in C$, то $S(y) = 0$.

Свойство: Пусть вектор $y = x \oplus e$, где $x \in C$, а e - вектор ошибок. Тогда $S(y) = He^T = \sum_{i:e_i=1} H_i$ - сумма некоторых столбцов матрицы H.

Декодирование с помощью синдрома.

Из канала принят вектор $y = x \oplus e$. Его синдром $S(y) = S(e) = He^T$.

Очевидно, $\forall e_1, e_2 : w(e_1), w(e_2) \leq t$ $S(e_1) \neq S(e_2) \Leftrightarrow$ код может исправлять t ошибок.

Пусть $\forall y$ $w(S(y)) < t = [(d-1)/2)]$.

Вычислим синдромы $\forall e \in B^n, w(e) \leq t$. Результаты занесем в таблицу. Теперь, если принят вектор y, то вычисляем его синдром и находим по таблице вектор e, которому соответствует этот синдром. Тогда $y \to x = e \oplus y$.

Т.е. $y \to e \to x$.

Теорема 2 (Граница Синглтона)

Пусть C - $[n, k, d]$-код. Тогда $n - k \geq d - 1$.

Док-во

Любую матрицу G $(n \times k)$ ранга k можно элементарными преобразованиями привести к виду $(I|M)$, I - единичная матрица $k \times k$.

Тогда любое кодовое слово состоит из исходного слова с приписанными к нему проверочными символами. Вес кодового слова, в котором только один бит исходного слова не равен нулю, не превосходит $n - k + 1$. $\Rightarrow d \leq n - k + 1$.

Опр.

Коды, которые лежат на границе Синглтона, называются кодами с максимальным расстоянием (сокращенно кодами МДР).

Раздел 22 - Скрытое получение информации

Private Information Retrieval

Блочные коды: избыточность *Rightarrow* исправление ошибок в блоке. Нужен фрагмент длинного сообщения - декодир. опред. блок. (Доступ к произвольным данным.)

Недост.: плохая устойчивость к концентрир. ошибкам.

Избавимся от недост.: кодир. всч сообщ. как один блок кода, исправляющего ошибки.

Получ. др. недост.: невозм. доступ к произвольным данным.

Локально декодируемые коды: эффективный доступ к произвольным данным и более высокая устойчивость к концентрированным ошибкам, чем у кодов с небольшими блоками. Декодир. одного бита по данным о небольшом кол-ве случайно выбранных бит кодового слова.

Цена: потеря эфф-ти. Меньшая скорость передачи данных, чем у классических блочных кодов, исправляющих ошибки.

Скрытое получение информации (PIR): вл-ц БД может мониторить запросы п-лей и узнать, чем интерес. отдельн. польз-ль.

Цель - не позволить это узнать.
Практических приложений пока нет.

PIR схема:

БД - строка X из n бит

$S_1, .., S_k$ - реплицир., некоммуницир. сервера

i - номер бита стр. X, знач. кот. хотим узнать.

Создаем несколько случайных чисел (“брос. монету”), запраш. сервера, выч. зн-е бита по ответам серв-в.

Каждый запрос случаен, не зав. от i. След., кажд. сервер не получ. никакой инф. об i.

Запросы - не обяз. запросы о зн-ях нек. битов. Это запросы о выч. опред. ф-й нек. битов. Напр., XOR.

Осн. пар-р эф-ти (историч.) - макс. сложность в смысле числа переданных по каналу бит. Макс. по всем вар-м знач. строки X и зн-ям генератора случ. чисел.

2 сервера: субэкспоненц., Chor (1998), с тех пор не улучш.

3 сервера: субполином., но растет быстрее,чем любая степень логарифма.

Вычислительные PIR.

- владелец БД должен решить сложную задачу, чтобы узнать, что запрашивал польз-ль.

- не требуют репликации БД

- если серверы обмен. инф-ей, это не угрожает без-ти польз-ля

- при работе - большой объем вычислений на сервере.

Пример выч. PIR, построенной на основе проблемы распознавания квадратичных вычетов.

Пусть $m = p_1 * p_2$. Проблема: явл. ли $a \in z_m^*$ QR или нет. Проблема выч. сложная, если не изв. факторизация m.

Протокол.

БД - строка X длины $n = s^2$ хр-ся в виде кв. м-цы (x_{ij}).

Польз-ль хочет получить зн-е нек. x_{ij}. Он выбир. произв. больш. $m = p_1 * p_2$. Созд. $s - 1$ QRs $\{a_t \in Z_m^* | 1 < t < s, t \neq j\}$, QNR $b_j \in Z_m^*$. Передает m и вектор $\{a_1, .., b_j, .., a_s\}$ на сервер. Тот принимает этот вектор как набор $u_1, .., u_s$.

Сервер возвр. набор $\pi_1, .., \pi_s$:

$$\pi_i = \prod_{k=1..s} u_t^{x_{ik}} \ mod\ m,\ i = 1..s$$

Заметим, что если $x_{ij} = 0$, то в произведении π_i нет QNR, иначе есть один QNR.

Вычисление x_{ij}: если сервер вернул π_i - QR, то $x_{ij} = 0$, иначе 1.

Сложность по объему перед. данных: $O(\sqrt{n})$.

Инф. - теоретические PIR протоколы.

В случ. одного сервера: с инф.-теоретичсекой точки зрения это невозм., единств сп. - получ. всей БД.

1998, Chor et al.: эффективн. протокол для реплицир. серверов. Каждый отд. сервер не получ. инф. о том, что интересно польз-лю, при усл, что сервера не обмен-ся инф-ей.

Все соврем. констр-и PIR: постр. локальн декодируемый код, конверт. его в PIR протокол.

Опр.

Локально декодируемый код с пар-рами (r, δ, ε):

кодирует k-bit сообщ. x в n-bit код. слово $C(x)$:

$\forall 1 < i < k$ значение x_i м.б. верно декодир. с вер-ю $(1-\varepsilon)$ рандомизированной процедурой декодир., кот. исп. r бит кодового слова. При этом принятое слово $y = C(x) + e$ имеет до δn ошибок (т.е. δ - макс. доля ошибок).

Рандомизир. - только относит. сл. чисел, генерир. на стороне декодера. (Т.е. верно для всякого искажения кодового слова)

Пример: код Адамара (Hadamard)

$(2, \delta, 2\delta)$ - (Hadamard) LDC that encodes k-bit messages to 2^k-bit codewords.

Обозн: [n] - это $\{1, .., n\}$

Кажд. бит кодового слова соотв. $XOR_{i \in S}(x_i)$ над одним из п-мн-в S мн-ва $[k]$.

Пусть y - искаж. код слова x. На вход декодера: y, i, $[k]$. Выб. случ. образом (равн. распр.) $S \subseteq [k]$ и получ. зн-я y_q, y_t в позициях, соотв. S и $(S \oplus \{i\})$.

δ - вер-ть искаж. одного бита к.с. Значит, вер-ть того, что оба запроса попали в неискаж. биты $= (1 - 2\delta)$.

Быстр. декодир. (2 бита), но огромн. длина к.с.

Семейства LDC

1) Низкая длина (сложность) запроса. $r = const$ или $r < log(k)$.
Прим-ся в криптогр., в PIR. Примеры:

- код Адамара
- код Рида-Маллера (Reed Muller Code)
- и др.

(*)
2) Высокая сложность запроса. $r = k^\epsilon$ для нек. $\epsilon > 0$.

Длина к.с. пропорц. длине сообщ. (коды с положительной скоростью передачи информации, positive rate codes)
После 2010 года исп-ся в нек. прилож. для передачи и хран-я данных. Ранние примеры:

- код Рида-Маллера с числом перем-х $k_{RM} = 1/\epsilon$, δ, $r = \Theta(\delta)$. Получ. LDC: $r = k^\epsilon$, скорость передачи инф. $\epsilon^{\Theta(1/\epsilon)}$, эта конст. всегда $< 1/2$.

- Multiplicity codes. (Мультипликативные коды.) Осн. на вычислении значений полиномов выс. степени от одной перем. и их производных. Пар-ры лучше, чем у кодов Рида-Маллера (в смысле избыточности кодир. и возм. соотн. пар-ров).

Построение IT-PIR по LDC.

Очев., набор r индексов, исп. при декодир., не д.б. постоянно из нек. окрестности i-го (или люб. др) бита.

Опр.

Коды, у кот распр. запросов явл. равномерным, наз. абсолютно гладкими.

Постр. r-server PIR по (r, δ, ε) абс. гладкому коду.

Пусть C - абс. гладк LDC: слово дл. k -> к.с. дл. n.

Препроцессинг: серв. $S_1, .., S_r$ кодир. x (дл. k бит) в n-битов. к.с.

Польз-ль случ. выбир. набор r запросов $q_1, .., q_r$: он может выч. x_i по $C(x)_{q_1}, .., C(x)_{q_r}$.

Запросы: q_j к S_j, j = 1..r.
Ответы: $C(x)_{q_j}$

Кажд. q_j - реализ. равн. распр. с.в. (на мн-ве координат к.с.), поэтому серверы не получают инф-ю о запросе.

Код Рида - Маллера
q-арный код.

Код опр-ся 3 пар-ми:
- р-р алфавита $q = p^n$ - степень простого числа. Поле F_q.
- степень полинома f от многих переменных над $GF(q)$ - число $d < q - 1$
- число переменных полинома n.

Пусть $m = (x_1, .., x_k)$ - сообщение.
Пусть $W = (w_1, .., w_k)$ - некоторые фикс. векторы из F_q^n.
Фиксируем полином: $f_m \in F_q[z_1, .., z_n]$ степени не более d:
$\forall i \in \{1, .., k\}$ $f(w_i) = x_i$. При опред. выборе W такой полином
$\exists$ $\forall m \in F_q^k$.

Коэффициенты полинома - решение СЛАУ
$f_m(w_i) = x_i, \ i = 1, .., k$.

Кодирование: $(x_1, ..., x_k) \to (f_m(a) \ \forall a \in F_q^n)$. Длина к.с.: q^n.

Исходное сообщение x имеет длину $k = C_{n+d}^d$ символов из F_q.
C_{n+d}^d - это количество разных одночленов степени от 0 до d от
n перем.

Декодир.: по номеру бита i и искаженному не более чем в $\delta \cdot q^n$ позициях кодовому слову - зн-ям полинома f_m в точках викт. пр-ва, надо найти зн-е f_m в точке w_i.

Проведем случайную прямую L через w_i.
Возьмем $v \in F_q^n$ - случайное.
L - одномерное п-пр-во, $L = \{w + \lambda v | \lambda \in F_q\}$.

Т.о., на L полином f_m становится полиномом одной перем. λ, степень по прежнему не более d. Значит, коэфф-ты полинома $f_m(\lambda)$ можно восстановить по зн-ям в $d+1$ точке.

Возьмем (искаженные) зн-я полин. f в $d+1$ точке $L \backslash \{w_i\}$.

Рассмотрим код Рида-Маллера как локально декодируемый код. Каждый запрос декодера идет в случайную, независ. точку. Какова вер-ть того, что все $d+1$ точек будут без искажений? Это $1-(d+1)\cdot\delta$.

$\Rightarrow$ это $(d+1, \delta, (d+1)\cdot\delta)$ LDC.

Литература к лекции:

1. Yekhanin, *Locally decodable codes: a brief survey* - обзор LDC

2. Reed-Muller code as LDC:
`http://people.mpi-inf.mpg.de/~csaha/lectures/lec3.pdf`

3*. Chor et al., *Private information retrieval*, 1998 - основополаг. статья по PIR
4*. Yekhanin, *Locally Decodable Codes* (книга, 2010 или 2011) - подробный рассказ о соврем. сост. LDC
5*. Yekhanin, *Private Information Retrieval* - применение LDC в криптогр., список лит-ры по мере развития LDC