2013 Теоретические основы прикладной дискретной математики №1(19)

DOI 10.17223/20710410/19/3

УДК 621.391: 519.728

ДВОИЧНЫЕ ПРЕДСТАВЛЕНИЯ НЕДООПРЕДЕЛЁННЫХ ДАННЫХ И ДИЗЪЮНКТИВНЫЕ КОДЫ[1]

Л. А. Шоломов

Институт системного анализа РАН, г. Москва, Россия

E-mail: sholomov@isa.ru

Рассматриваются достаточно компактные представления недоопределённых данных, позволяющие полностью восстановить исходные данные (а не только их доопределения). Их построение основано на введённых и изученных в данной работе специальных матрицах, названных селективными. Они обобщают широко применяемые в информатике дизъюнктивные матрицы. Исследованы свойства селективных матриц и получены оценки длины представления данных в функции от некоторых параметров. Рассмотрены сложностные вопросы, связанные с построением представлений.

Ключевые слова: *недоопределённые данные, сжатие, двоичное представление, базис системы множеств, длина представления, дизъюнктивная матрица, дизъюнктивный код, свободное от покрытий семейство, полиномиальный алгоритм.*

Введение

Для полностью определённых данных термины «сжатие» и «архивация» обычно понимаются одинаково. Они означают экономное кодирование данных, обеспечивающее их полное восстановление. Для недоопределённых данных эти понятия будем различать.

В качестве недоопределённых данных мы рассматриваем последовательности недоопределённых символов. Каждому такому символу соответствует некоторое множество основных (полностью определённых) символов, одним из которых он может быть замещён (доопределён). Имеются две естественные постановки задачи кодирования недоопределённых данных.

В первой постановке кодирование должно обеспечить восстановление какого-либо доопределения исходных данных (но не их самих). Она соответствует случаю, когда допустимо любое доопределение данных. При второй постановке от кодирования требуется, чтобы оно позволяло полностью восстановить исходные недоопределённые данные. К такой постановке приводит задача их хранения. Термин «сжатие» будем связывать с первой постановкой, термин «архивация» — со второй.

До сих пор мы имели дело с первой постановкой (например, в [1]). На её основе введена энтропия недоопределённых данных как характеристика степени сжатия, определены другие информационные показатели. Данная работа инициирована второй постановкой; часть её результатов имеется в [2].

Рассмотрим побуквенное представление недоопределённых последовательностей. Символы основного алфавита кодируются наборами из 0 и 1, а недоопределённые сим-

[1]Работа выполнена при поддержке ОНИТ РАН по проекту «Теоретические основы эффективного использования недоопределённой информации» программы «Интеллектуальные информационные технологии, системный анализ и автоматизация».

волы представляются наборами из 0, 1 и $*$, где $*$ — неопределённый символ. Ставится условие, чтобы отношение доопределимости между основными и недоопределёнными символами переносилось на представляющие их наборы. Это позволяет решать задачи, связанные со сжатием и доопределением данных, работая непосредственно с представлениями.

Если Λ и $\tilde{\Lambda}$ — матрицы, столбцами которых являются соответственно представления основных и недоопределённых символов, то при мощности m основного алфавита матрица Λ состоит из m столбцов, а число столбцов матрицы $\tilde{\Lambda}$ может достигать 2^m-1. Это делает процедуры представления недоопределённых последовательностей, применяющие матрицу $\tilde{\Lambda}$, неэффективными. В работе предложен эффективный способ, использующий лишь матрицу Λ. В результате этого задача экономного представления недоопределённых последовательностей свелась к построению допустимых матриц кодирования Λ с возможно малым числом строк.

Работа посвящена характеризации и изучению свойств матриц кодирования Λ, их построению и оценке параметров. Показано, что эти матрицы обладают некоторым свойством, названным селективностью, и их частным случаем являются дизъюнктивные матрицы. Последние введены в работе [3] как средство описания дизъюнктивных (superimposed) кодов. Эти коды широко применяются в различных задачах прикладной математики и информатики (см., например, [3–7] и цитированные там источники). Учитывая направленность журнала ПДМ, отметим применение дизъюнктивных кодов в криптографии для задачи распределения ключей [5, 8–9]. Дизъюнктивные коды интенсивно изучаются в течение многих лет, и некоторые из полученных для них результатов используются в данной работе.

1. Двоичное представление

Задан алфавит $A_0 = \{a_0, a_1, \ldots, a_{m-1}\}$ *основных символов*. Пусть $M = \{0, 1, \ldots, m-1\}$ и каждому непустому $T \subseteq M$ поставлен в соответствие символ a_T. Символ a_T называется *недоопределённым* и его *доопределением* считается всякий основной символ a_i, $i \in T$. Символ a_M, доопределимый любым основным символом, называется *неопределённым* и обозначается $*$. Под *доопределением последовательности* $a_{T_1} \ldots a_{T_N}$ недоопределённых символов понимается любая последовательность $a_{i_1} \ldots a_{i_N}$ основных символов, полученная из исходной заменой всех символов какими-либо доопределениями. Пусть выделена система $\mathcal{T}$ некоторых непустых подмножеств T множества M и с ней связан *недоопределённый алфавит* $A = A_{\mathcal{T}} = \{a_T : T \in \mathcal{T}\}$.

Задавшись натуральным числом s, припишем каждому $a_i \in A_0$ некоторый набор $\lambda_i = (\lambda_i(1), \lambda_i(2), \ldots, \lambda_i(s)) \in \{0,1\}^s$ — *код символа* a_i, а каждому $a_T \in A$ — некоторый набор $\lambda_T = (\lambda_T(1), \lambda_T(2), \ldots, \lambda_T(s)) \in \{0, 1, *\}^s$. Обозначим через Λ матрицу, столбцами которой являются наборы (точнее, транспонированные наборы) λ_i, $i \in M$, а через $\tilde{\Lambda}$ — матрицу со столбцами λ_T, $T \in \mathcal{T}$. Скажем, что пара $(\Lambda, \tilde{\Lambda})$ задает *двоичное представление алфавита* A, если для любых $T \in \mathcal{T}$ и $i \in M$

$$\lambda_i \textit{ доопределяет } \lambda_T \Leftrightarrow i \in T.$$

(Фактически используется расширенное двоичное представление — к 0 и 1 добавляется символ $*$, но удобно говорить о двоичном представлении.)

Утверждение 1. Всякий алфавит A двоично представим, т. е. обладает представлением указанного вида.

Такое представление можно получить, назначив $s = m$ и взяв в качестве λ_i, $i \in M$, набор $(\lambda_i(v), v = 1, \ldots, m)$, в котором компонента $\lambda_i(v)$ равна 1 для $v - 1 = i$ и 0 для

$v-1 \neq i$, а в качестве λ_T, $T \in \mathcal{T}$, — набор $(\lambda_T(v),\ v = 1, \ldots, m)$, в котором компонента $\lambda_T(v)$ равна $*$ для $v-1 \in T$ и 0 для $v-1 \notin T$. ■

Двоичное представление алфавита $A_{\mathcal{T}}$ допускает интерпретацию в терминах вложения системы $\mathcal{T}$ подмножеств множества M в булев куб $\{0,1\}^s$ (подходящей размерности s). При вложении каждому $i \in M$ сопоставляется некоторая точка λ_i куба, каждому $T \in \mathcal{T}$ — некоторый подкуб λ_T, и этот подкуб содержит те и только те точки λ_i, которые соответствуют значениям $i \in T$. Параметр s будем называть *размерностью представления* (вложения). Наименьшее s, при котором для алфавита A имеется двоичное представление размерности s, обозначим $s(A)$ и назовем *размерностью алфавита* A.

Будем рассматривать побуквенное представление недоопределённых последовательностей, при котором последовательности $a_{T_1}a_{T_2}\ldots a_{T_N}$ соответствует представление $\lambda_{T_1}\lambda_{T_2}\ldots\lambda_{T_N}$. Его построение и восстановление по нему исходной недоопределённой последовательности использует матрицу $\tilde{\Lambda}$, число столбцов которой может достигать $2^m - 1$. Это делает процедуру неэффективной (неполиномиальной). Покажем, что задачи представления и восстановления недоопределённых последовательностей можно эффективно решать, пользуясь лишь матрицей Λ, имеющей m столбцов.

Дальше, говоря о множестве столбцов T матрицы Λ, будем иметь в виду столбцы λ_i с индексами $i \in T$. Недоопределённый набор $\dot{\lambda}_T \in \{0,1,*\}^s$ назовём *T-селектором* для матрицы Λ, если все столбцы множества T являются его доопределениями и никакой другой столбец матрицы его не доопределяет. Матрицу Λ будем называть *$\mathcal{T}$-селективной*, если при каждом $T \in \mathcal{T}$ для нее существует T-селектор. В случае, когда система $\mathcal{T}$ образована всеми t-элементными подмножествами множества M, $\mathcal{T}$-селективную матрицу будем называть *t-селективной*.

Утверждение 2. Двоичное представление алфавита $A_{\mathcal{T}}$, использующее матрицу кодирования Λ, существует тогда и только тогда, когда она $\mathcal{T}$-селективна.

Действительно, если Λ $\mathcal{T}$-селективна, то в качестве столбцов матрицы $\tilde{\Lambda}$ в двоичном представлении могут быть взяты T-селекторы $\dot{\lambda}_T$, и наоборот, если представление $(\Lambda, \tilde{\Lambda})$ существует, то роль T-селекторов $\dot{\lambda}_T$ могут играть столбцы матрицы $\tilde{\Lambda}$. ■

Число строк матрицы Λ будем обозначать $s(\Lambda)$. По матрице кодирования Λ с $s(\Lambda) = s$ построим наборы $\hat{\lambda}_T \in \{0,1,*\}^s$, $T \in \mathcal{T}$, положив компоненту $\hat{\lambda}_T(v)$, $v = 1, \ldots, s$, набора равной $\tau \in \{0,1\}$, если $\lambda_i(v) = \tau$ для всех $i \in T$, и равной $*$, если найдутся $i, j \in T$, для которых $\lambda_i(v) \neq \lambda_j(v)$.

Утверждение 3. Если матрица Λ $\mathcal{T}$-селективна, то при каждом $T \in \mathcal{T}$ набор $\hat{\lambda}_T$ является T-селектором.

Доказательство. Для всех $i \in T$ столбцы λ_i доопределяют $\hat{\lambda}_T$ по построению. Рассмотрим $j \notin T$. Возьмем произвольный T-селектор $\dot{\lambda}_T$. Он не доопределим до λ_j, т. е. при некоторых v и τ имеет место $\dot{\lambda}_T(v) = \tau$, $\lambda_j(v) = \bar{\tau}$. Все столбцы λ_i, $i \in T$, доопределяют $\dot{\lambda}_T$, а потому имеют $\dot{\lambda}_i(v) = \tau$. Отсюда $\hat{\lambda}_T(v) = \tau \neq \lambda_j(v)$ и, следовательно, λ_j не доопределяет $\hat{\lambda}_T$. ■

Отметим, что если столбцы λ_i матрицы Λ интерпретировать как точки куба $\{0,1\}^s$, то набор $\hat{\lambda}_T$ соответствует минимальному (по включению) подкубу, содержащему все точки λ_i, $i \in T$.

Следствие 1. Если алфавит A двоично представим с матрицей кодирования Λ, то одним из его двоичных представлений является $(\Lambda, \hat{\Lambda})$, где $\hat{\Lambda}$ — матрица, образованная столбцами $\hat{\lambda}_T$, $T \in \mathcal{T}$.

Представление $(\Lambda, \hat{\Lambda})$, будем называть *каноническим*. С использованием матрицы Λ нахождение канонического представления $\hat{\lambda}_T$ символа a_T и восстановление символа a_T по каноническому представлению $\hat{\lambda}_T$ выполнимы с линейной относительно площади ms матрицы сложностью. Дальше будем рассматривать канонические представления. В этом случае задача нахождения двоичного представления алфавита $A_{\mathcal{T}}$ сводится к задаче построения $\mathcal{T}$-селективной матрицы.

2. Сведение к задаче о конъюнктивном базисе

Скажем, что система $\mathcal{Z}$ подмножеств множества M образует *конъюнктивный базис* системы $\mathcal{T}$, если каждое множество $T \in \mathcal{T}$ может быть получено как пересечение некоторых множеств из $\mathcal{Z}$, и образует *обобщённый конъюнктивный базис*, если каждое $T \in \mathcal{T}$ может быть получено как пересечение каких-либо множеств из $\mathcal{Z}$ либо их дополнений (до M). При этом считается, что пересечение пустой совокупности множеств даёт M, так что множество M конъюнктивно порождаемо любой системой $\mathcal{Z}$.

Строки матрицы Λ будем обозначать через $\lambda(v)$, $v = 1, \ldots, s$. Строке $\lambda(v)$ сопоставим множество $Z_v \subseteq M$, $Z_v = \{i : \lambda_i(v) = 1\}$, характеристическим набором которого строка является. Положим $\mathcal{Z} = \{Z_1, \ldots, Z_s\}$.

Утверждение 4. Матрица кодирования $\mathcal{T}$-селективна тогда и только тогда, когда соответствующая ей система множеств $\mathcal{Z}$ образует обобщённый конъюнктивный базис системы $\mathcal{T}$.

Доказательство. Для $Z \subseteq M$ и $\tau \in \{0, 1\}$ под Z^τ будем понимать Z при $\tau = 1$ и $\bar{Z} = M \setminus Z$ при $\tau = 0$.

1. Пусть матрица кодирования Λ является $\mathcal{T}$-селективной, $\mathcal{Z}$ — соответствующая ей система множеств, $\hat{\Lambda}$ — матрица канонического представления, построенная по Λ. Рассмотрим произвольное $T \in \mathcal{T}$. Положим $V_T = \{v : \hat{\lambda}_T(v) \neq *\}$ и убедимся, что

$$T = \bigcap_{v \in V_T} Z^{\hat{\lambda}_T(v)}. \tag{1}$$

Если множество V_T пусто, то столбец $\hat{\lambda}_T$ образован элементами $*$ и его доопределяют все столбцы матрицы Λ. В этом случае $T = M$ и равенство (1) выполнено в силу соглашения относительно пересечения пустой системы множеств.

Дальше считаем $V_T \neq \varnothing$. Для любого $i \in T$ столбец λ_i доопределяет $\hat{\lambda}_T$ и, следовательно, при каждом $v \in V_T$ выполнено $\lambda_i(v) = \hat{\lambda}_T(v)$. Это означает, что i содержится в каждом множестве $Z_v^{\hat{\lambda}_T(v)}$, $v \in V_T$, а потому и в их пересечении. С учётом произвольности $i \in T$ заключаем, что T включено в пересечение из правой части (1). Обратное включение вытекает из того, что для всякого $j \notin T$ столбец λ_j не является доопределением столбца $\hat{\lambda}_T$, т. е. при некотором $v \in V_T$ имеет место $\lambda_j(v) \neq \hat{\lambda}_T(v)$, а потому $j \notin Z_v^{\hat{\lambda}_T(v)}$ и j не содержится в пересечении из (1). Равенство (1) показывает, что система $\mathcal{Z}$ образует обобщённый конъюнктивный базис для $\mathcal{T}$.

2. Пусть задан обобщённый конъюнктивный базис $\mathcal{Z} = \{Z_1, \ldots, Z_s\}$ системы $\mathcal{T}$. Для $i \in M$ и $v = 1, \ldots, s$ положим $\lambda_i(v) = 1$, если $i \in Z_v$, и $\lambda_i(v) = 0$, если $i \notin Z_v$. Образуем матрицу $\Lambda = \|\lambda_i(v)\|$. Всякое множество $T \in \mathcal{T}$ представимо в виде пересечения некоторых множеств Z_v или их дополнений. Фиксируем одно из таких представлений и обозначим через V_T совокупность индексов v участвующих в нём множеств Z_v. Для $v \in V_T$ положим $\lambda_T(v)$ равным 1 или 0 в зависимости от того, в какой из форм — Z_v или $\bar{Z}_v$ — входит в пересечение это множество. Если множество Z_v системы $\mathcal{Z}$ не

участвует в рассматриваемом представлении, полагаем $\lambda_T(v) = *$. В принятых обозначениях представление записывается в виде

$$T = \bigcap_{v \in V_T} Z^{\lambda_T(v)}. \tag{2}$$

Образуем матрицу $\tilde{\Lambda} = \|\lambda_T(v)\|$. Пара $(\Lambda, \tilde{\Lambda})$ двоично представляет алфавит A, ибо столбец λ_i доопределяет λ_T тогда и только тогда, когда для любого $v \in V_T$ выполнено $\lambda_i(v) = \lambda_T(v)$, что эквивалентно соотношениям $i \in Z^{\lambda_T(v)}$, а это в силу (2) имеет место тогда и только тогда, когда $i \in T$. ■

Следствие 2. Двоичное представление алфавита $A = A_{\mathcal{T}}$, имеющее размерность s, существует тогда и только тогда, когда у системы $\mathcal{T}$ имеется обобщённый конъюнктивный базис мощности s.

Проверка, образует ли система $\mathcal{Z}$ обобщённый конъюнктивный базис для $\mathcal{T}$, выполнима с полиномиальной относительно мощностей систем $\mathcal{Z}$, $\mathcal{T}$ и множества M сложностью. Это следует из того, что множество $T \in \mathcal{T}$ представимо в виде пересечения некоторых множеств из $\mathcal{Z}$ или их дополнений тогда и только тогда, когда T совпадает с пересечением всех содержащих его множеств Z_v^τ, $Z_v \in \mathcal{Z}$, $\tau \in \{0, 1\}$.

Обозначим через $[\mathcal{T}]$ замыкание системы множеств $\mathcal{T}$ относительно операции пересечения. В $[\mathcal{T}]$, в частности, содержится множество M как пересечение пустой системы множеств. Из того, что обобщённый конъюнктивный базис системы $\mathcal{T}$ является таковым и для $[\mathcal{T}]$, вытекает следующий факт.

Следствие 3. $\mathcal{T}$-селективная матрица $[\mathcal{T}]$-селективна.

Это позволяет в задачах построения и упрощения $\mathcal{T}$-селективных матриц не принимать в расчет множества системы $\mathcal{T}$, являющиеся пересечениями других.

Задача о размерности алфавита состоит в том, чтобы по алфавиту A (эквивалентно, по системе $\mathcal{T}$) и числу s узнать, существует ли для A двоичное представление размерности s.

Утверждение 5. Задача о размерности алфавита NP-полна.

Доказательство. Согласно следствию 2, достаточно установить NP-полноту задачи об обобщённом конъюнктивном базисе. Переход от множеств системы $\mathcal{T}$ к их дополнениям превращает её в задачу об обобщённом дизъюнктивном базисе (определения дизъюнктивного и обобщённого дизъюнктивного базисов аналогичны соответствующим определениям конъюнктивного и обобщённого конъюнктивного базисов с тем отличием, что роль пересечения играет объединение). Укажем сведение к ней NP-полной задачи о дизъюнктивном базисе (задача MP7 из [11]).

Пусть требуется решить задачу о (дизъюнктивном) базисе системы $\mathcal{T}$ некоторых подмножеств T множества $M = \{0, 1, \ldots, m-1\}$. Можно ограничиться случаем $M \notin \mathcal{T}$, ибо общий случай сводится к этому очевидным образом. Образуем множество $M' = \{0, 1, \ldots, m-1, m\}$ и рассмотрим систему $\mathcal{T}'$, полученную из $\mathcal{T}$ добавлением одноэлементного множества $\{m\}$. Докажем, что система $\mathcal{T}$ обладает базисом мощности s (т. е. состоящим из s множеств) тогда и только тогда, когда у системы $\mathcal{T}'$ имеется обобщённый базис мощности $s+1$.

Если система $\mathcal{Z}$ мощности s является базисом для $\mathcal{T}$, то система $\{\mathcal{Z}, \{m\}\}$ имеет мощность $s+1$ и образует базис (и обобщённый базис) для $\mathcal{T}'$. Обратно, пусть у системы $\mathcal{T}'$ имеется обобщённый базис $\mathcal{Z}'$ мощности $s+1$. Чтобы покрыть множество $\{m\} \in \mathcal{T}'$, он должен содержать множество $\{m\}$ или его дополнение $M' \setminus \{m\} = M$.

Ни одно из них не может участвовать в покрытии множеств системы $\mathcal{T}$. Удалив их из $\mathcal{Z}'$, получим обобщённый базис $\mathcal{Z}''$ системы $\mathcal{T}$, состоящий из не более чем s множеств. Для всякого $Z'' \in \mathcal{Z}''$ лишь одно из множеств Z'' и $\bar{Z}'' = M' \setminus Z''$ не содержит элемента m и может участвовать в покрытии множеств системы $\mathcal{T}$. Совокупность таких множеств образует дизъюнктивный базис системы $\mathcal{T}$. Если он состоит из менее чем s множеств, произвольно дополним его до мощности s. ■

В связи с трудностью нахождения значения $s(A)$ представляют интерес оценки этой величины. Обозначим через $s(m,n)$ максимальное значение $s(A)$ по всем алфавитам A с $|A_0| = m$, $|A| = n$, где $|\cdot|$ означает мощность множества.

Утверждение 6. Имеет место равенство

$$s(m,n) = \min\{m,n\}.$$

Доказательство. Для алфавита $A = A_{\mathcal{T}}$ с $|A_0| = m$ и $|A| = n$ верхнюю оценку $s(A) \leqslant m$ получим, если в качестве конъюнктивного базиса системы $\mathcal{T}$ возьмем совокупность всех $(m-1)$-элементных подмножеств множества M, а верхнюю оценку $s(A) \leqslant n$ обеспечим, если в качестве конъюнктивного базиса возьмем саму систему $\mathcal{T}$.

В случае $m \leqslant n$ нижняя оценка $s(A) \geqslant m$ справедлива для алфавита $A = A_{\mathcal{T}}$, у которого система $\mathcal{T}$ состоит из всех $(m-1)$-элементных подмножеств множества M и $n-m$ любых других подмножеств, а при $n < m$ оценка $s(A) \geqslant n$ имеет место для алфавита $A = A_{\mathcal{T}}$, у которого в качестве $\mathcal{T}$ взята система, образованная множествами $M \setminus \{i\}$, $i = 0, 1, \ldots n-1$. ■

Мощность $|A|$ алфавита недоопределённых символов может достигать $2^{|A_0|} - 1$. В содержательных задачах обычно $|A_0| \leqslant |A|$, поэтому в качестве основного будем рассматривать случай

$$m \leqslant n. \tag{3}$$

При условии (3) утверждение 6 приобретает вид

$$s(m,n) = m. \tag{4}$$

3. Алфавит с ограниченным числом доопределений

Представляет интерес рассмотрение содержательных случаев, когда размерность алфавита существенно ниже гарантируемой утверждением 6. Недоопределённые данные, с которыми имеют дело в приложениях, обычно помимо неопределённого символа $*$ используют лишь символы, имеющие небольшое число доопределений. Так, например, если в качестве недоопределённого символа выступает нечётко написанная буква, то скорее всего она похожа лишь на небольшое число реальных букв.

Обозначим через $s(m,n;t)$ максимальную из размерностей $s(A)$ алфавитов A, для которых $|A_0| = m$, $|A| = n$ и каждый символ $a_T \in A$ имеет не более t доопределений либо является неопределённым.

Утверждение 7. Справедливы оценки

$$s(m,n;2) \leqslant 4\ln(mn) + 1; \tag{5}$$

$$s(m,n;3) \leqslant 8\ln(mn) + 1; \tag{6}$$

$$s(m,n;t) \leqslant e(t+1)\ln(mn) + 1. \tag{7}$$

Доказательство. Рассмотрим произвольный алфавит A с параметрами m, n и t. Пусть $A = A_{\mathcal{T}}$. Поскольку любое двоичное представление реализует символ $* = a_M$ автоматически, можно считать, что $M \notin \mathcal{T}$ и для любого $T \in \mathcal{T}$ выполнено $|T| \leqslant t$. Согласно утверждению 4, достаточно оценить сверху мощность обобщённого конъюнктивного базиса для системы $\mathcal{T}$.

Пары (T, j), $j \in M \setminus T$, будем называть *фрагментами*. Скажем, что множество Z *реализует* фрагмент (T, j) *прямо*, если $T \subseteq Z$, $j \notin Z$, и реализует *инверсно*, если $T \cap Z = \varnothing$, $j \in Z$. Множество *реализует* фрагмент, если оно его реализует прямо или инверсно. Ясно, что система $\mathcal{Z}$ образует обобщённый конъюнктивный базис для $\mathcal{T}$ тогда и только тогда, когда множествами системы $\mathcal{Z}$ реализуются все фрагменты (T, j), $T \in \mathcal{T}$, $j \in M \setminus T$. Для построения и оценки мощности системы с указанным свойством используем вероятностный метод.

Рассмотрим систему s случайных подмножеств множества M, в каждое из которых элементы $i \in M$ включаются независимо с одинаковой вероятностью p, которая будет выбрана позже. Вероятность того, что случайное множество реализует заданный фрагмент (T, j), составляет $p^{|T|}(1-p) + (1-p)^{|T|}p$, а вероятность, что ни одно из s множеств системы его не реализует, равна $\bigl(1 - (p^{|T|}(1-p) + (1-p)^{|T|}p)\bigr)^s$. Эта величина не превосходит $\bigl(1 - (p^t(1-p) + (1-p)^t p)\bigr)^s$. Поскольку число фрагментов меньше $|\mathcal{T}| \cdot |M| = nm$, вероятность отсутствия реализации хотя бы у одного из них меньше

$$mn\bigl(1 - (p^t(1-p) + (1-p)^t p)\bigr)^s.$$

Отсюда и из соотношения $\ln(1-x) \leqslant -x$ следует, что при любом

$$s \geqslant \frac{\ln(mn)}{p^t(1-p) + (1-p)^t p}$$

эта вероятность меньше 1 и, следовательно, существует обобщённый конъюнктивный базис такой мощности. Это дает

$$s_t(m, n) \leqslant \frac{\ln(mn)}{\psi_t(p)} + 1, \tag{8}$$

где $\psi_t(p) = p^t(1-p) + (1-p)^t p$.

Выберем значение p. Для $t = 2$ и 3 функция $\psi_t(p)$ достигает максимума при $p = 1/2$. Подстановка этого значения в (8) приводит к оценкам (5) и (6). При произвольном t с учетом того, что $\psi_t(p) \geqslant p^t(1-p)$ и величина $p^t(1-p)$ при $p = t/(t+1)$ принимает значение $t^t/(t+1)^{t+1}$, имеем

$$\psi_t\Bigl(\frac{t}{t+1}\Bigr) \geqslant \frac{t^t}{(t+1)^{t+1}} \geqslant \frac{1}{e(t+1)}.$$

Подставляя эту оценку в (8), приходим к (7). ■

В основном случае (3) соотношение (7) приводит к оценке

$$s(m, n; t) \leqslant 2e(t+1)\ln n + 1. \tag{9}$$

Справедлива тривиальная нижняя оценка

$$s(m, n; t) \geqslant \log_3 n,$$

вытекающая из мощностного соотношения $3^{s(m,n;t)} \geqslant |\mathcal{T}| = n$. Она отличается от верхней оценки (9) по порядку в t раз. Дальше этот разрыв будет сокращен до логарифма t (утверждение 14).

4. Строго двоичные представления

В двоичных представлениях рассматриваемого вида недоопределённым последовательностям в алфавите A соответствуют недоопределённые двоичные последовательности, фактически использующие трёхбуквенный алфавит $\{0, 1, *\}$. Обычно, когда речь идет о сжатии данных, характеристики сжатия связываются с параметрами двоичного кодирования. Данный пункт посвящён представлениям в двухбуквенном алфавите.

Двоичное представление $(\Lambda, \tilde{\Lambda})$ алфавита A назовем *строго двоичным*, если $\tilde{\Lambda}$ — матрица в двухбуквенном алфавите. Этим алфавитом может быть $\{0, *\}$ либо $\{1, *\}$. Оба случая равноценны, но удобнее иметь дело с $\{0, *\}$, и дальше будем рассматривать этот случай. Построенное в утверждении 1 двоичное представление является строгим, поэтому всякий недоопределённый алфавит строго двоично представим. Если интерпретировать строгое двоичное представление алфавита $A_{\mathcal{T}}$ как вложение системы множеств $\mathcal{T}$ в булев куб (с. 18), то подкубы, содержащие множества $T \in \mathcal{T}$, должны обладать дополнительным свойством — проходить через вершину $(0,\ldots,0)$.

Опишем вид матриц кодирования Λ, возникающих при строго двоичном представлении. Под *дизъюнкцией* $\lambda \vee \lambda'$ *двоичных наборов* $\lambda = (\lambda(1), \ldots, \lambda(s))$ и $\lambda' = (\lambda'(1), \ldots, \lambda'(s))$ понимается набор с компонентами $\lambda(i) \vee \lambda'(i)$, $i = 1, \ldots, s$. Скажем, что набор λ' *покрывает* λ, если $\lambda \vee \lambda' = \lambda'$, и что некоторое множество наборов покрывает λ, если λ покрывается их дизъюнкцией. Матрицу Λ со столбцами λ_i, $i \in M$, назовем $\mathcal{T}$*-дизъюнктивной*, если для любого $T \in \mathcal{T}$ множество столбцов T не покрывает ни одного столбца, не входящего в это множество. В случае, когда система $\mathcal{T}$ состоит из всех t-элементных подмножеств множества M, $\mathcal{T}$-дизъюнктивную матрицу называют t*-дизъюнктивной*. Множество наборов, соответствующих столбцам t-дизъюнктивной матрицы, образует t*-дизъюнктивный код*. Дизъюнктивные (superimposed) коды, введённые в работе [3], находят широкое применение в информатике; $\mathcal{T}$-дизъюнктивные матрицы общего вида встречались в [12].

Утверждение 8. Строго двоичное представление алфавита $A_{\mathcal{T}}$, использующее матрицу кодирования Λ, существует тогда и только тогда, когда она $\mathcal{T}$-дизъюнктивна.

Доказательство. Если строго двоичное представление с матрицей кодирования Λ существует, то у Λ при каждом $T \in \mathcal{T}$ имеется некоторый T-селектор $\dot{\lambda}_T$ в алфавите $\{0, *\}$. Столбцы множества T доопределяют его, и их единицы находятся в позициях, где $\dot{\lambda}_T$ принимает значение $*$. Любой другой столбец содержит 1 в некотором нулевом разряде T-селектора $\dot{\lambda}_T$ и потому не покрывается столбцами множества T. В силу произвольности $T \in \mathcal{T}$ это означает $\mathcal{T}$-дизъюнктивность матрицы.

Обратно, если матрица $\mathcal{T}$-дизъюнктивна, то T-селектор в алфавите $\{0, *\}$ для $T \in \mathcal{T}$ можно образовать, взяв дизъюнкцию столбцов из множества T и заменив в ней все элементы 1 на $*$. Очевидно, что столбцы λ_i, $i \in T$, доопределяют этот набор, а любой столбец λ_j, $j \notin T$, не доопределяет, ибо имеет 1 в некотором разряде, где дизъюнкция равна 0. Наличие для каждого $T \in \mathcal{T}$ T-селектора в алфавите $\{0, *\}$ обеспечивает строго двоичную представимость. ■

Из этого утверждения, в частности, следует, что $\mathcal{T}$-дизъюнктивные матрицы $\mathcal{T}$-селективны. Применяя в качестве столбцов λ_T матрицы $\tilde{\Lambda}$ при строго двоичном представлении T-селекторы из второй части доказательства, можно, пользуясь лишь матрицей Λ, эффективно строить строгое представление недоопределённой последовательности и эффективно восстанавливать по нему исходную последовательность. Заменив в этом представлении символы $*$ на 1, получим последовательность в алфавите $\{0, 1\}$,

которую можно рассматривать как код исходной последовательности. При этом кодом символа a_T будет дизъюнкция столбцов λ_i, $i \in T$. Заметим, что по той же матрице Λ можно построить каноническое двоичное представление, но оно, как правило, не будет строгим.

Никакой столбец t-дизъюнктивной матрицы не покрывается t другими, поэтому такие матрицы также называют *свободными от t-покрытий* [13]; $\mathcal{T}$-дизъюнктивные матрицы естественно называть *свободными от $\mathcal{T}$-покрытий*. Представим в аналогичных терминах понятие $\mathcal{T}$-селективной матрицы. Под *инверсией двоичного набора* будем понимать набор, полученный из него заменой всех компонент их отрицаниями. Будем говорить, что некоторое множество наборов *инверсно покрывает* набор λ, если множество инверсий этих наборов покрывает инверсию набора λ. Скажем, что множество наборов *дважды покрывает* набор λ, если оно покрывает и инверсно покрывает набор λ. Матрицу Λ назовем *свободной от двойных $\mathcal{T}$-покрытий*, если для любого $T \in \mathcal{T}$ множество столбцов λ_i, $i \in \mathcal{T}$, не покрывает дважды ни одного столбца, не входящего в это множество.

Утверждение 9. Матрица Λ $\mathcal{T}$-селективна тогда и только тогда, когда она свободна от двойных $\mathcal{T}$-покрытий.

Доказательство. Рассмотрим матрицу Λ и построенную по ней матрицу $\hat{\Lambda}$ канонического представления. Убедимся, что столбец λ_j доопределяет столбец $\hat{\lambda}_T$ тогда и только тогда, когда совокупность столбцов λ_i, $i \in T$, дважды покрывает λ_j. Пусть λ_j доопределяет $\hat{\lambda}_T$. Если для некоторой позиции v имеет место $\lambda_j(v) = \tau$, то $\hat{\lambda}_T(v) \in \{*, \tau\}$ и по построению $\hat{\Lambda}$ найдется $i \in T$, для которого $\lambda_i(v) = \tau$. Применив это рассуждение ко всем позициям v с $\tau = 1$, заключаем, что множество столбцов λ_i, $i \in T$, покрывает λ_j. Аналогичное рассмотрение позиций с $\tau = 0$ показывает, что множество инверсий этих столбцов покрывает инверсию столбца λ_j. Это означает, что покрытие двойное. Доказательство в другую сторону проводится обращением этих рассуждений.

Матрица Λ $\mathcal{T}$-селективна тогда и только тогда, когда алфавит $A_{\mathcal{T}}$ представим парой $(\Lambda, \hat{\Lambda})$ (утверждение 3). Это означает, что для $T \in \mathcal{T}$ ни один из столбцов λ_j, $j \notin T$, не доопределяет столбец $\hat{\lambda}_T$ и по доказанному выше не покрывается дважды множеством столбцов T, т. е. матрица Λ свободна от двойных $\mathcal{T}$-покрытий. ■

Используя это утверждение, продемонстрируем возможность улучшения параметров при переходе от дизъюнктивных матриц к селективным.

Утверждение 10. В результате дописывания к произвольной 2-дизъюнктивной матрице столбца, составленного из единиц, получается 2-селективная матрица.

Доказательство. Обозначим через Λ' матрицу, полученную из 2-дизъюнктивной матрицы Λ дописыванием единичного столбца $\tilde{1}$. Рассмотрим три разных столбца λ_i, λ_j, λ_k матрицы Λ'. Если среди столбцов λ_i, λ_j нет $\tilde{1}$, то они не покрывают λ_k в силу 2-дизъюнктивности Λ либо того, что $\lambda_k = \tilde{1}$. Если же, например, $\lambda_i = \tilde{1}$, то $\lambda_i \wedge \lambda_j = \lambda_j$ не покрывается столбцом λ_k, а потому столбцы λ_i, λ_j не покрывают инверсно λ_k. ■

Полученная матрица Λ' не дизъюнктивна, поскольку в дизъюнктивных матрицах отсутствуют поглощения столбцов.

Для $\mathcal{T}$-дизъюнктивных матриц справедлив аналог утверждения 4. Как и раньше, строкам $\lambda(v)$, $v = 1, \ldots, s$, матрицы Λ сопоставим множества $Z_v = \{i : \lambda_i(v) = 1\}$. Положим $\mathcal{Z}' = \{\bar{Z}_1, \ldots, \bar{Z}_s\}$, где $\bar{Z}_v = M \setminus Z_v$.

Утверждение 11. Матрица $\mathcal{T}$-дизъюнктивна тогда и только тогда, когда соответствующая ей система множеств $\mathcal{Z}'$ образует конъюнктивный базис системы $\mathcal{T}$.

Доказательство опускаем. Это свойство может быть использовано для построения и упрощения $\mathcal{T}$-дизъюнктивных матриц для конкретных систем $\mathcal{T}$.

Следующее утверждение указывает способ преобразования $\mathcal{T}$-селективной матрицы в $\mathcal{T}$-дизъюнктивную. Пусть $\bar{\Lambda}$ означает инверсию матрицы Λ — результат замены в ней всех элементов их отрицаниями.

Утверждение 12. Если матрица Λ $\mathcal{T}$-селективна, то матрица $\begin{bmatrix}\Lambda\\ \bar{\Lambda}\end{bmatrix}$, полученная дописыванием к строкам матрицы Λ строк её инверсии, $\mathcal{T}$-дизъюнктивна.

Этот факт следует из утверждений 4 и 12, поскольку указанное преобразование матриц соответствует переходу от обобщённого конъюнктивного базиса к конъюнктивному базису за счет добавления дополнений множеств, входящих в обобщённый базис. ■

Укажем некоторые различия в свойствах $\mathcal{T}$-селективных и $\mathcal{T}$-дизъюнктивных матриц, легко проверяемые непосредственно. В $\mathcal{T}$-селективных матрицах допустима инверсия строк, а свойство $\mathcal{T}$-дизъюнктивности при инверсиях может нарушаться. Если $\vee_T$ означает дизъюнкцию столбцов множества T, то в $\mathcal{T}$-дизъюнктивных матрицах наборы $\vee_T$ для разных $T \in \mathcal{T}$ различны. Для $\mathcal{T}$-селективных матриц они могут совпадать, но различными являются пары $(\vee_T, \wedge_T)$, где $\wedge_T$ — конъюнкция столбцов множества T. Двойственным к понятию $\mathcal{T}$-дизъюнктивности является $\mathcal{T}$-конъюнктивность (для любых $T \in \mathcal{T}$ и $j \notin T$ конъюнкция столбцов множества T не покрывается столбцом λ_j). Класс $\mathcal{T}$-селективных матриц включает оба эти класса матриц и не исчерпывается ими.

Обозначим через $s_0(A)$ размерность алфавита A для строго двоичных представлений, определяемую аналогично $s(A)$. Очевидно, $s(A) \leqslant s_0(A)$. Следующий пример показывает, что неравенство может быть строгим.

Пример 1. Пусть основным алфавитом является $A_0 = \{a_0, a_1, a_2, a_3\}$, а алфавит A образован всеми символами, имеющими два доопределения (эквивалентно, не более двух — следствие 3). В силу утверждения 10 матрица

λ_0	λ_1	λ_2	λ_3
1	1	0	0
1	0	1	0
1	0	0	1

,

образованная из диагональной матрицы дописыванием столбца из единиц, 2-селективна. Она обеспечивает для алфавита A размерность 3. Из приведённой в [4] нижней оценки Л. А. Бассалыго следует, что при использовании 2-дизъюнктивных матриц наименьшей достижимой размерностью здесь является 4.

На основе $s_0(A)$ введём функции $s_0(m, n)$ и $s_0(m, n; t)$, аналогичные тем, которые были определены применительно к $s(A)$. Для них справедливы аналоги утверждений 5, 6 и формулы (4). Оценка функции $s_0(m, n; t)$ проводится, как в утверждении 7, но вместо $\psi_t(p)$ в формуле (8) возникает функция $\psi_t^0(p) = p^t(1-p)$, принимающая наибольшее значение $t^t/(t+1)^{t+1}$ при $p = t/(t+1)$. Аналог оценки (7) сохраняет тот же вид, поскольку в процессе установления (7) вместо $\psi_t(p)$ фактически использовалась функция $\psi_t^0(p)$, а константы в аналогах (5) и (6) возрастают. Подсчёт показывает, что имеет место следующий факт.

Утверждение 13. Справедливы оценки

$$\begin{aligned} s_0(m,n;2) &\leqslant 6{,}75\ln(mn)+1,\\ s_0(m,n;3) &\leqslant 9{,}48\ln(mn)+1,\\ s_0(m,n;t) &\leqslant e(t+1)\ln(mn)+1. \end{aligned}$$

Тривиальной нижней оценкой является $s_0(m,n;t) \geqslant \log n$ (она будет улучшена). Всюду под $\log n$ понимается $\log_2 n$.

5. Нижние оценки

В конце п. 3 и 4 приведены тривиальные (мощностные) нижние оценки функций $s(m,n;t)$ и $s_0(m,n;t)$. Далее получим нетривиальные нижние оценки на основе установленного в [14] соотношения

$$m \leqslant t + \binom{s}{\left\lceil \frac{2(s-t)}{t(t+1)} \right\rceil}, \tag{10}$$

связывающего параметры t-дизъюнктивной матрицы. Здесь m — число столбцов, s — число строк. Подробнее об оценках для t-дизъюнктивных матриц и их авторстве будет сказано далее.

Начнём со случая строго двоичных представлений.

Утверждение 14. При выполнении условия

$$t = \mathrm{o}\left(\frac{\log n}{\log\log n}\right) \tag{11}$$

справедлива оценка

$$s_0(m,n;t) \gtrsim \frac{(t+1)\log n}{2(2\log t + c)}, \quad c = \log\frac{3e}{4} < 1{,}027. \tag{12}$$

Доказательство. Поскольку n не превосходит числа подмножеств множества M, имеющих мощность не выше t, справедливо неравенство

$$n \leqslant \sum_{u\leqslant t}\binom{m}{u}. \tag{13}$$

Пусть m' — максимальное значение, удовлетворяющее условию $n \geqslant \sum\limits_{u\leqslant t}\binom{m'}{u}$. Очевидно, что $m' \leqslant m$. В силу максимальности m' выполнено

$$n \leqslant \sum_{u\leqslant t}\binom{m'+1}{u} \leqslant (m'+1)^t.$$

Отсюда с учётом (11) получаем

$$\log(m'+1) \geqslant \frac{\log n}{t} \gg \log\log n \tag{14}$$

(соотношение $u \gg v$ означает $v = \mathrm{o}(u)$), а потому

$$m' \gg \log n \gg t. \tag{15}$$

Положим $M' = \{0, 1, \ldots, m' - 1\}$ и возьмём в качестве $\mathcal{T}$ некоторую систему, которая состоит из n подмножеств множества M мощности не выше t и включает все t-элементные подмножества множества M'. Рассмотрим произвольную $\mathcal{T}$-дизъюнктивную матрицу Λ, первые m' столбцов которой соответствуют элементам множества M'. Эти столбцы образуют t-дизъюнктивную матрицу и для неё справедливо соотношение (10). В силу тривиальной нижней оценки $s \geqslant \log n$ и условия (11) выполнено $s \gg t$. Прологарифмировав (10), получаем с учётом неравенства $\log \binom{u}{v} \leqslant \log \frac{eu}{v}$ и соотношений (14), (15) и $s \gg t$

$$\frac{\log n}{t} \lesssim \log m' \lesssim \left(\frac{2s}{t(t+1)} + 1\right) \log \frac{et(t+1)}{2}.$$

Отсюда

$$\frac{2s}{t(t+1)} + 1 \gtrsim \frac{\log n}{t \log \dfrac{et(t+1)}{2}}. \tag{16}$$

Так как в силу (11) выполнено

$$\frac{\log n}{t \log \dfrac{et(t+1)}{2}} \gg \frac{\log n}{\left(\dfrac{\log n}{\log \log n}\right) \log \log n} = 1,$$

из (16) находим

$$s \gtrsim \frac{t(t+1)\log n}{2t \log \dfrac{et(t+1)}{2}} = \frac{(t+1)\log n}{2\left(2\log t + \log \dfrac{(t+1)e}{2t}\right)} \geqslant \frac{(t+1)\log n}{2\left(2\log t + \log \dfrac{3e}{4}\right)}.$$

Отсюда следует (12). ■

При растущем t оценка (12) отличается от верхней оценки функции $s_0(m, n; t)$ по порядку в $\log t$ раз.

Рассмотрим теперь функцию $s(m, n; t)$, относящуюся к двоичным представлениям общего вида. Из утверждения 12 вытекает, что $s(A) \geqslant s_0(A)/2$, а потому $s(m, n; t) \geqslant \geqslant s_0(m, n; t)/2$ и для $s(m, n; t)$ справедлива нижняя оценка из (12), уменьшенная в 2 раза. Следующее утверждение содержит оценку, которая лучше этой при всех $t \geqslant 2$.

Утверждение 15. При выполнении условия (11) справедлива оценка

$$s(m, n; t) \gtrsim \frac{(t-1)\log n}{2(2\log(t-1) + c)}, \tag{17}$$

где c — константа из (12).

Доказательство.

1. Рассмотрим сначала случай, когда $(m - 1, n, t - 1)$ — допустимая тройка, т. е. существует алфавит с такими параметрами. Покажем, что в этом случае

$$s(m, n; t) \geqslant s_0(m - 1, n; t - 1). \tag{18}$$

Пусть A — произвольный алфавит с параметрами $(m-1, n, t-1)$ и ему соответствуют основной алфавит $A_0 = \{a_0, \ldots, a_{m-2}\}$ и система множеств $\mathcal{T}$. Обозначим через A'_0

основной алфавит, полученный из A_0 присоединением нового символа a_{m-1}. Путём добавления к каждому множеству T системы $\mathcal{T}$ элемента $m-1$ образуем множества T' и систему всех таких множеств обозначим через $\mathcal{T}'$. Алфавит $A' = A_{\mathcal{T}'}$ имеет параметры (m, n, t).

Рассмотрим $\mathcal{T}'$-селективную матрицу, на которой достигается $s(A')$. Путем инверсирования её строк добьемся того, чтобы столбец, соответствующий элементу $m-1$, общему для всех множеств T', стал нулевым; полученную матрицу обозначим Λ'. Все T'-селекторы матрицы Λ' являются наборами в алфавите $\{0, *\}$ и потому она T'-дизъюнктивна. Матрица Λ, образованная из Λ' удалением нулевого столбца, T-дизъюнктивна. В результате получаем

$$s_0(A) \leqslant s(\Lambda) = s(\Lambda') = s(A') \leqslant s(m, n; t).$$

Учитывая произвольность алфавита A с параметрами $(m-1, n, t-1)$, приходим к неравенству (18), которое совместно с (12) обеспечивает в рассматриваемом случае оценку

$$s(n, m; t) \gtrsim \frac{t \log n}{2(2 \log(t-1) + c)}. \tag{19}$$

2. Пусть тройка $(m-1, n, t-1)$ недопустима. Положим $n' = \sum\limits_{u \leqslant t-1} \binom{m-1}{u}$.

Тройка $(m-1, n', t-1)$ допустима и справедливо неравенство $n > n'$. Параметр n удовлетворяет условию (13). Правая часть (13) в силу соотношения $\binom{m}{u} < m\binom{m-1}{u-1}$ не превосходит tmn'. Поэтому $n \leqslant tmn'$. Логарифмируя это неравенство и учитывая (11), получаем

$$\log n \lesssim \log m + \log n'. \tag{20}$$

Из (13) вытекает, что $n \leqslant m^t$, а потому $\log n \leqslant t \log m$. Подстановка этого соотношения в (20) даёт $t \log n' \gtrsim (t-1) \log n$. С учётом этого из оценки (19), применённой к допустимой тройке $(m-1, n', t-1)$, вытекает цепочка соотношений

$$s(m, n; t) \geqslant s(m, n'; t) \gtrsim \frac{t \log n'}{2(2 \log(t-1) + c)} \gtrsim \frac{(t-1) \log n}{2(2 \log(t-1) + c)},$$

приводящая к требуемому результату. ■

При растущем t разрыв между верхней и нижней оценками (9) и (17) функции $s(m, n; t)$ имеет порядок $\log t$.

6. Применение результатов о дизъюнктивных кодах

Естественным способом задания недоопределённого символа a_T является указание характеристического набора (длины m) для множества T. В данной работе этому способу соответствует задание набором длины m в алфавите $\{0, *\}$ (см. утверждение 1). Из (4) видно, что значение m не понижается и при использовании представлений в алфавите $\{0, 1, *\}$. Если недоопределённый алфавит A имеет параметры (m, n, t), то для двоичного представления его символа достаточно $s(m, n; t)$ символов 0, 1 и $*$, а для строго двоичного представления достаточно $s_0(m, n; t)$ символов 0 и $*$. Оценки этих величин в п. 3, 4 и 5 выражены через параметры n и t. Чтобы их сравнивать с длиной m естественного представления, исключим в функциях $s(m, n; t)$ и $s_0(m, n; t)$ параметр n и вместо них будем рассматривать

$$s(m; t) = \max_n s(m, n; t), \quad s_0(m; t) = \max_n s_0(m, n; t),$$

где максимумы берутся по всем допустимым значениям n. Очевидно, что значения $s(m;t)$ и $s_0(m;t)$ достигаются на алфавите A, состоящем из всех символов, допускающих не более t доопределений, и символа $*$. В силу следствия 3 и его аналога для строго двоичных представлений можно в качестве A рассматривать алфавит, образованный всеми символами, имеющими ровно t доопределений. Это означает, что величины $s(m;t)$ и $s_0(m;t)$ совпадают с минимальным числом строк t-селективной и t-дизъюнктивной матриц. В качестве следствия утверждений 7 и 13 получаем следующий факт.

Утверждение 16. Справедливы оценки

$$\begin{aligned} &s(m;2) \leqslant 12\ln m, \quad s_0(m;2) \leqslant 20{,}25\ln m,\\ &s(m;3) \leqslant 32\ln m, \quad s_0(m;3) \leqslant 37{,}92\ln m,\\ &s(m;t) \leqslant s_0(m;t) \leqslant e(t+1)^2\Big(\ln\frac{m}{t}+2\Big). \end{aligned}$$

Доказательство. Будем рассматривать алфавит, образованный всеми символами, имеющими ровно t доопределений. Для него $n = \binom{m}{t}$. Подстановка в (5), (6) и (7) верхних оценок для n, равных соответственно $m^2/2$, $m^3/6$ и $(me/t)^t$, приводит к указанным верхним оценкам для $s(m;2)$, $s(m;3)$ и $s(m;t)$. Аналогично из утверждения 13 получаются верхние оценки для $s_0(m;2)$, $s_0(m;3)$ и $s_0(m;t)$. ■

Отметим, что оценки, асимптотически совпадающие с оценками утверждений 7, 13 и 16, могут быть получены также с использованием жадного алгоритма (в другой терминологии — градиентной процедуры). Применительно к оценке для $s(m;t)$ см. [2].

В случае t-дизъюнктивных матриц метод случайного кодирования для верхней оценки величины $s_0(m;t)$ применялся многими авторами (например, в [4, 5, 13, 15]), а жадный алгоритм — в работе [16]. Все полученные оценки имеют одинаковый порядок $\mathrm{O}(t^2\log m)$, но лучшими являются оценки работы [4]. Оценка для $s_0(m;t)$ из утверждения 16 асимптотически совпадает с оценкой этой величины из [4], но способ её получения путём оценки мощности конъюнктивного базиса отличен от использованного там.

Впервые немощностная (и к данному моменту наилучшая) нижняя оценка для $s_0(m;t)$ установлена в [4]. Она относится к случаю $t = \mathrm{const}$, $m \to \infty$ и может быть представлена в виде

$$s_0(m;t) \gtrsim \frac{t^2\log m}{2\log t + c}, \tag{21}$$

где c — константа из (12).

Эта оценка доказывается достаточно сложно индукцией по t. Позднее в работе [17] было найдено чисто комбинаторное доказательство в 4 раза более слабой оценки. Затем в заметке [14] было представлено очень простое комбинаторное доказательство оценки, которая хуже (21) в 2 раза. Она вытекает из установленного в [14] соотношения (10). Мы будем основываться на последней работе, поскольку её результат удаётся распространить на случай растущего t. Имеет место следующий факт.

Утверждение 17. При условии

$$t = \mathrm{o}(\log m) \tag{22}$$

справедливы оценки

$$s_0(m;t) \gtrsim \frac{t(t+1)\log m}{2(2\log t + c)}; \tag{23}$$

$$s(m;t) \gtrsim \frac{(t-1)t\log m}{2(2\log t + c)}, \tag{24}$$

где c — константа из (12).

Доказательство. Оценка (23) легко получается из соотношения (10) с использованием неравенства $\binom{u}{v} \leqslant (ue/v)^v$ и с учётом (22). Оценка (24) вытекает из неё и соотношения $s(m;t) \geqslant s_0(m-1;t-1)$, доказываемого подобно (18). ■

Метод случайного кодирования не дает конструкции t-селективной (либо t-дизъюнктивной) матрицы, а трудоёмкость жадного алгоритма экспоненциальна. Возникает задача нахождения эффективных и конструктивных алгоритмов, обеспечивающих «достаточно хорошие» верхние оценки величины $s_0(m;t)$. Детерминированный алгоритм считается эффективным, если его трудоёмкость ограничена полиномом от размера исходных данных, и считается конструктивным, если из него извлекается явное описание объекта.

В работе [3] представлены некоторые конструкции дизъюнктивных кодов. Один из предложенных там подходов основан на использовании q-ичных кодов с большим кодовым расстоянием. Реализация этого подхода в [5] с применением кодов Рида — Соломона позволила для дизъюнктивных матриц получить достаточно хорошую эффективную оценку величины m при заданных t и s (в [3] имеется указание на использование кодов Рида — Соломона, но оценки не приведены). Применительно к задаче двоичного представления недоопределённых данных из этой оценки вытекает следующий результат.

Утверждение 18. Существует эффективный и конструктивный метод, обеспечивающий оценку

$$s(m;t) \leqslant s_0(m;t) \lesssim \left(\frac{2(t+1)\log m}{\log(t\log m)}\right)^2. \tag{25}$$

Отношение оценки (25) к оценке, полученной методом случайного кодирования, имеет порядок $\dfrac{\log m}{(\log t + \log\log m)^2}$. При больших значениях t оценка (25) может оказаться даже лучше её. Так, например, при $t = m^c$, $c = \mathrm{const} < 1/2$, оценка (25) приобретает вид $s(m,t) \lesssim 4t^2/c^2$. Для сравнения укажем нижнюю оценку $s(m,t) \gtrsim t^2/4$, вытекающую из нижней оценки Бассалыго [4] для $s_0(m;t)$ и соотношения $s(m;t) \geqslant$ $\geqslant s_0(m;t)/2$.

В приложениях больший интерес представляют малые значения t. Малыми будем считать величины t, удовлетворяющие условию (22). При его выполнении оценка (25) хуже полученной методом случайного кодирования по порядку в $\dfrac{\log m}{(\log\log m)^2}$ раз. В работе [5] приведена ещё одна эффективная конструкция, позволившая при малых t улучшить оценку. Она использует некоторое семейство алгебро-геометрических кодов Гоппы. Из полученной в [5] оценки m в функции от t и s вытекает следующий результат для двоичных представлений.

Утверждение 19. Существует эффективный и конструктивный метод, обеспечивающий оценку

$$s(m;t) \leqslant s_0(m;t) = \mathrm{O}\left(\frac{t^3 \log m}{\log t}\right).$$

Эта оценка отличается от полученной методом случайного кодирования по порядку в $\dfrac{t}{\log t}$ раз и при условии (22) лучше оценки (25). Разрыв между детерминированными эффективными оценками и полученными методом случайного кодирования устранён в работе [6]. Из неё вытекает следующий факт.

Утверждение 20. Существует эффективный метод, обеспечивающий оценку

$$s(m;t) \leqslant s_0(m;t) = \mathrm{O}\left(t^2 \log m\right).$$

Однако этот результат нельзя считать вполне удовлетворительным, поскольку метод работы [6] эффективен (полиномиален), но не конструктивен — использует дерандомизацию метода случайного кодирования.

В [18] представлен эффективный и конструктивный метод построения t-дизъюнктивных кодов, который при некоторых значениях параметров приводит к лучшим результатам, чем метод случайного кодирования.

Введённые и исследованные в данной работе селективные матрицы являются обобщениями широко применяемых в информатике дизъюнктивных матриц. Основное назначение последних — выделение из больших множеств малых подмножеств. Эту задачу можно эффективно решать с помощью селективных матриц: в дизъюнктивных матрицах выделяются столбцы, покрываемые заданным набором, а в селективных — столбцы, доопределяющие заданный недоопределённый набор. Поскольку класс селективных матриц шире, от их применения можно ожидать некоторого выигрыша.

ЛИТЕРАТУРА

1. *Шоломов Л. А.* Элементы теории недоопределённой информации // Прикладная дискретная математика. Приложение. 2009. № 2. С. 18–42.
2. *Шоломов Л. А.* Двоичное представление недоопределённых данных // Доклады Академии наук. 2013. Т. 448. № 3. С. 275–278.
3. *Kautz W. H. and Singleton R. C.* Nonrandom binary superimposed codes // IEEE Trans. Inform. Theory. 1964. V. 10. No. 4. P. 363–377.
4. *Дьячков А. Г., Рыков В. В.* Границы длины дизъюнктивных кодов // Проблемы передачи информации. 1982. Т. 18. Вып. 3. С. 7–13.
5. *Kumar R., Rajagopalan S., and Sahai A.* Coding construction for blacklisting problems without computational assumptions // LNCS. 1999. V. 1666. P. 609–623.
6. *Porat E. and Rotshchild A.* Explicit non-adaptive combinatorial group testing schemes // Automata, Languages and Programming. LNCS. 2008. V. 5125. P. 748–759.
7. *Chuzhoy J. and Khanna S.* An $\mathrm{O}(k^3 \log n)$-approximation algorithm for vertex-connectivity survivable network design //Proc. 50th Annual IEEE Symp. Foundation Computer Science (FOCS). Washington: IEEE Computer Society, 2009. P. 437–441.
8. *Dyer M., Fenner T., Frieze A., and Thomason A.* On key storage in secure networks // J. Cryptology. 1995. V. 8. P. 189–200.
9. *Mitchell C. J. and Piper F. C.* Key storage in secure networks // Discr. Appl. Math. 1988. V. 21. P. 215–228.
10. *Quinn K. A. S.* Bounds for key distribution patterns // J. Cryptology. 1999. V. 12. P. 227–239.

11. *Гэри М., Джонсон Д.* Вычислительные машины и труднорешаемые задачи. М.: Мир, 1982. 416 с.
12. *Коспанов Э. Ш.* О кодировании (0,1)-матриц конъюнкциями // Дискретный анализ. Сборник трудов. Вып. 27. Новосибирск: Институт математики СО АН, 1975. С. 17–22.
13. *Erdos P., Frankl P., and Furedi Z.* Families of finite sets in which no set is covered by the union of r others // Israel J. Math. 1985. V. 51, No. 1–2. P. 79–89.
14. *Furedi Z.* On r-cover-free families // J. Combinator. Theory. Ser. A. 1996. V. 73. P. 172–173.
15. *Cheng V., Du D.-Z., and Lin G.* On the upper bounds of the minimum number of rows of disjunct matrices // Optimizat. Lett. 2009. V. 3. Iss. 2. P. 297–302.
16. *Hwang F. K. and Sos V. T.* Non-adaptive hypergeometric group testing // Studia Scientiarum Mathecarum Hungarica. 1987. V. 22. P. 257–263.
17. *Ruszinko M.* On the upper bound of the size of the r-cover-free families // J. Combinator. Theory. Ser. A. 1994. V. 66. P. 302–310.
18. *D'yachkov A. G., Macula A. J., and Rykov V. V.* New constructions of superimposed codes // IEEE Trans. Inform. Theory. 2000. V. 46. No. 1. P. 284–290.