УДК 519.677

ОПРЕДЕЛЕНИЕ ОГРАНИЧЕНИЙ ДЛЯ МЕТОДА БИНОМИАЛЬНОГО НУМЕРАЦИОННОГО СЖАТИЯ

***И.А. Кулик**, канд. техн. наук, доцент;*
***С.В. Костель**, ассистент;*
***Е.М. Скордина**, аспирант,*
Сумский государственный университет, г. Сумы

В статье рассматриваются граничные условия для метода биномиального нумерационного сжатия, при которых выполняется сжатие информации. Приводится доказательство существования ограничений, налагаемых на число единиц в зависимости от длины двоичного сообщения. Предлагаются выражения для нахождения граничного значения числа единиц.

***Ключевые слова:** биномиальное нумерационное кодирование, биномиальный коэффициент, метод наименьших квадратов, коэффициент корреляции.*

У статті розглядаються граничні умови для методу біноміального нумераційного стиснення, при яких виконується стиснення інформації. Наводиться доказ існування обмежень щодо кількості одиниць залежно від розміру двійкового повідомлення. Пропонуються співвідношення для знаходження граничного значення числа одиниць.

***Ключові слова:** біноміальне нумераційне кодування, біноміальний коефіцієнт, метод найменших квадратів, коефіцієнт кореляції.*

ПОСТАНОВКА ПРОБЛЕМЫ

В работах [1,2] был рассмотрен метод сжатия, основанный на использовании биномиальных чисел. В данном методе сжатия используется нумерационное кодирование с использованием биномиальной числовой функции [3,4]. Нумерационное кодирование преобразует двоичную равномерную последовательность A, состоящую из n битов, в двоичную последовательность B, состоящую из

$$l(k) = \left\lceil \log_2 (n+1) \right\rceil + \left\lceil \log_2 \left(C_n^k \right) \right\rceil \quad (1)$$

двоичных разрядов.

Сжатие исходной двоичной последовательности A с использованием биномиального нумерационного кодирования осуществляется только при выполнении следующего условия:

$$\left\lceil \log_2 (n+1) \right\rceil + \left\lceil \log_2 \left(C_n^k \right) \right\rceil < n . \quad (2)$$

Для случая, когда

$$\left\lceil \log_2 (n+1) \right\rceil + \left\lceil \log_2 \left(C_n^k \right) \right\rceil = n , \quad (3)$$

сжатие не происходит. Закодированная двоичная последовательность B будет иметь ту же длину, что и исходная A.

При выполнении условия

$$\left\lceil \log_2 (n+1) \right\rceil + \left\lceil \log_2 \left(C_n^k \right) \right\rceil > n \quad (4)$$

нумерационное кодирование при помощи биномиальных чисел будет вносить избыточную информацию, и количество битов в закодированном сообщении B будет превышать количество битов исходного сообщения A.

Поскольку число разрядов n в исходной последовательности A величина постоянная, то возникает задача определения числа единиц k в сообщении, при котором выполняется условие (2) и происходит сжатие информации.

ИСХОДНЫЕ ДАННЫЕ К ИССЛЕДОВАНИЮ

Покажем, что для всякого натурального $n \in N$ и целого $k \in \overline{0,n}$ справедливо неравенство

$$\log_2\left(C_n^k\right) < n. \tag{5}$$

Поскольку функция $\log_2(x)$ является положительной и непрерывно возрастающей, можно переписать неравенство (5) в виде

$$C_n^k < 2^n. \tag{6}$$

Воспользовавшись свойством биномиальных коэффициентов [4]

$$\sum_{k=0}^{n} C_n^k = 2^n, \tag{7}$$

в результате получим очевидное неравенство

$$C_n^k < \sum_{k=0}^{n} C_n^k, \tag{8}$$

которое подтверждает истинность неравенства (5).

Соотношение

$$\gamma(k) = \left\lceil \log_2\left(C_n^k\right) \right\rceil \tag{9}$$

определяет количество битов, необходимое для хранения номера двоичной последовательности A длиной n разрядов с количеством k единиц при использовании метода биномиального нумерационного кодирования. Исходя из (5), можно записать, что

$$\left\lceil \log_2\left(C_n^k\right) \right\rceil < n. \tag{10}$$

Неравенство (10) будет выполняться для всех натуральных $n > 2$ и целых $k \in \overline{0,n}$. Таким образом, при помощи биномиального нумерационного кодирования можно сжимать равновесные комбинации заданной длины n с постоянным весом k.

Теперь покажем, что для всякого натурального $n \in N$, $n > 1$ будет справедливо неравенство

$$\log_2(n+1) + \log_2\left(C_n^{n/2}\right) > n. \quad (11)$$

Взяв антилогарифм от обеих частей неравенства (11), получим

$$(n+1)\cdot C_n^{n/2} > 2^n. \quad (12)$$

Левую часть неравенства (12) перепишем в виде

$$(n+1)\cdot C_n^{n/2} = \sum_{k=0}^{n} C_n^{n/2}, \quad (13)$$

а правую часть неравенства (12) перепишем в соответствии с (7). В результате получим неравенство

$$\sum_{k=0}^{n} C_n^{n/2} > \sum_{k=0}^{n} C_n^k. \quad (14)$$

Поскольку $C_n^{n/2} \geq C_n^k$, неравенство (14), а следовательно и неравенство (11), выполняется для всех натуральных $n \in N$, $n > 1$.

Соотношение

$$\left\lceil \log_2(n+1) \right\rceil + \left\lceil \log_2\left(C_n^{n/2}\right) \right\rceil > n \quad (15)$$

так же верно, поскольку операция округления вверх только увеличивает значения величин в левой части неравенства.

Если в формуле (1) использовать минимальное значение биномиального коэффициента C_n^k, которое соответствует значению параметра $k = 0$ или $k = n$, то для всех натуральных $n > 1$ будет выполняться следующее неравенство:

$$\left\lceil \log_2(n+1) \right\rceil + \left\lceil \log_2\left(C_n^0\right) \right\rceil < n, \quad (16)$$

или

$$\left\lceil \log_2(n+1) \right\rceil < n.$$

Проанализировав неравенства (5), (15) и (16), приходим к выводу, что функция (1) имеет точки пересечения с линией $f(k) = n$.

Учитывая, что функция для биномиального коэффициента симметрична относительно точки $k = n/2$ и согласно со свойством симметрии биномиальных коэффициентов ($C_n^k = C_n^{n-k}$), точек пересечения будет две. Обозначим эти точки как $k_{\min} = k_m$ и $k_{max} = n - k_m$.

Графики функций (1) и (9) для значения $n = 64$ представлены на рисунке 1. На интервалах $[0, k_m]$ и $[n - k_m, n]$ метод биномиального нумерационного кодирования позволяет сжимать двоичные последовательности, состоящие из n двоичных разрядов. Поэтому

нахождение значений k_m является важной задачей при построении метода сжатия на основе биномиальных чисел.

ОПРЕДЕЛЕНИЕ ГРАНИЧНЫХ УСЛОВИЙ

Определить значения k_m для различных n можно путем последовательного программного перебора, пока не выполнится условие (3). Некоторые значения k_m, полученные в результате подбора, для различных значений n представлены в таблице 1.

При значениях n от 1 до 11 целого значения k_m, которое соответствовало бы условию (3), нет. Так же важно отметить, что величина k_m для указанного интервала n имеет малое значение, что говорит о нецелесообразности использования биномиального нумерационного сжатия для малых n. При больших n значений k_m, которые соответствуют условию (3), может быть несколько. Это связанно с операцией округления вверх, производимой над логарифмом биномиального коэффициента в формуле (1).

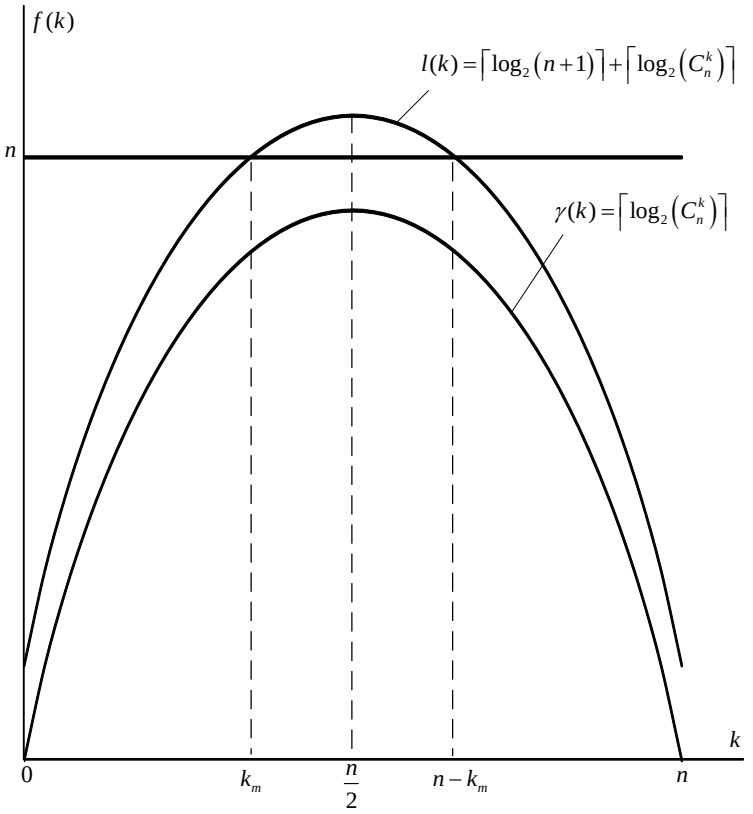

Рисунок 1 – Графики зависимостей $\gamma(k)$ и $l(k)$

По значениям, полученным в результате подбора, построим график зависимости k_m от n, который представлен на рисунке 2. Как видно из графика на рисунке 2, зависимость k_m от n имеет близкий к линейному вид. Найдем уравнение зависимости $k_m(n)$ при помощи математического метода наименьших квадратов [5]. Для анализа будем использовать значения параметра n в диапазоне от 8 до 1024.

Рассмотрим для начала линейное уравнение вида

$$y = a \cdot x + b\,. \tag{17}$$

В качестве x здесь выступает параметр n, а y соответствует значению k_m. Коэффициенты a и b для уравнения (17) находятся по формулам [5]:

$$a = \frac{\sum_{i=1}^{N} x_i \cdot y_i - \frac{1}{N}\sum_{i=1}^{N} x_i \cdot \sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\left(\sum_{i=1}^{N} x_i\right)^2} \tag{18}$$

и

$$b = \frac{1}{N}\left(\sum_{i=1}^{N} y_i - a \cdot \sum_{i=1}^{N} x_i\right), \tag{19}$$

где N - количество значений для анализа.

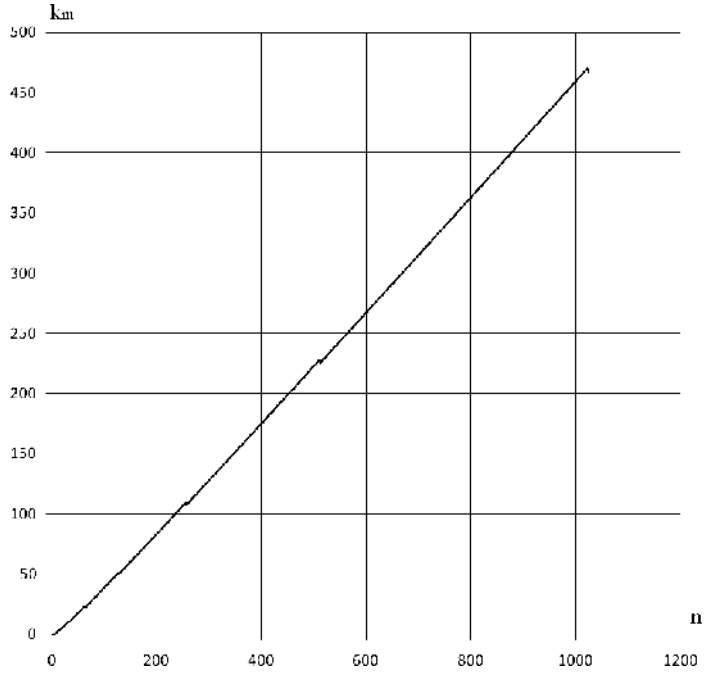

Рисунок 2 – График зависимости k_m от n

Подставив в соотношения (18) и (19) значения n и k_m, полученные опытным путем, получим значения $a = \mathbf{0,46284}$, $b = -10,25$. Тогда зависимость $k_m(n)$ будет иметь вид

$$k_{m\,l}(n) = \lceil \mathbf{0.46284} \cdot n - \mathbf{10.25} \rceil\,. \tag{20}$$

Поскольку значение k_m может иметь только целочисленные значения, в формулу (20) введена операция округления.

Оценку точности найденного по методу наименьших квадратов уравнения проводят путем вычисления коэффициента корреляции:

$$R = \frac{N \cdot \sum_{i=1}^{N}(a_i \cdot b_i) - \sum_{i=1}^{N}(a_i) \cdot \sum_{i=1}^{N}(b_i)}{\sqrt{\left(N \cdot \sum_{i=1}^{N}(a_i^2) - \left(\sum_{i=1}^{N}(a_i)\right)^2\right) \cdot \left(N \cdot \sum_{i=1}^{N}(b_i^2) - \left(\sum_{i=1}^{N}(b_i)\right)^2\right)}}, \qquad (21)$$

где a_i - значения, полученные опытным путем;

b_i - значения построенной функциональной зависимости.

Для найденного линейного уравнения (20) коэффициент корреляции составляет

$$R_{лин} = 0,99986 . \qquad (22)$$

Более точную зависимость $k_m(n)$, с использованием метода наименьших квадратов, можно получить, если использовать полиномиальное уравнение

$$y = a \cdot x^2 + b \cdot x + c . \qquad (23)$$

Так же, как и в линейной зависимости (17), в полиномиальной зависимости (23) x соответствует значению n, а y - значению k_m. В результате расчетов коэффициентов полиномиальная зависимость $k_m(n)$ будет иметь вид

$$k_{m\ p}(n) = \left\lceil 2,75 \cdot 10^{-5} \cdot n^2 + 0,4345 \cdot n - 5,3 \right\rceil . \qquad (24)$$

Для уравнения (24) коэффициент корреляции будет

$$R_{поли} = 0,999978 . \qquad (25)$$

Помимо соотношений (20) и (24), полученных методом наименьших квадратов, экспериментальным путем было получено соотношение, которое более точное описывает зависимость $k_m(n)$, и имеет следующий вид:

$$k_{m\ ex}(n) = \left\lceil \frac{n}{2}\left(1 - \sqrt{\frac{\ln(10 \cdot n)}{n}}\right) \right\rceil . \qquad (26)$$

Коэффициент корреляции уравнения (26) составляет

$$R_{ex} = 0,999983 . \qquad (27)$$

В таблице 1 для сравнения представлены значения k_m, найденные опытным путем, значения $k_{m\,l}$, полученные из уравнения (20), значения $k_{m\,p}$, полученные из уравнения (24) и значения $k_{m\,ex}$, полученные из уравнения (26).

Таблица 1 – Значения k_m, $k_{m\,l}$, $k_{m\,p}$ и $k_{m\,ex}$ в зависимости от n

n	k_m	$k_{m\,l}$	$k_{m\,p}$	$k_{m\,ex}$
8	2	-6	-1	2
12	3	-4	0	3
16	4	-2	2	4
24	7	1	6	7
32	10	5	9	10
48	16	12	16	16
64	22	20	23	22
96	36	35	37	36
128	49	49	51	49
192	78	79	80	77
256	106	109	108	106
384	166	168	166	164
512	223	227	225	223
768	345	346	245	343
1024	464	464	469	464
1536	710	701	727	708
2048	953	938	1000	953

ВЫВОДЫ

В данной работе была доказана возможность применения биномиального нумерационного кодирования для сжатия равновесных комбинаций, а так же сжатия произвольных двоичных сообщений при некоторых ограничениях, налагаемых на число двоичных единиц в сообщении. Были получены соотношения (20), (24) и (26), которые позволяют вычислять граничные числа единиц в зависимости от длины двоичного сообщения. В результате сравнения коэффициентов корреляции (22), (25) и (27) можно сделать вывод, что $R_{ex} > R_{поли} > R_{лин}$. Это говорит о том, что уравнения (24) и (26) дают более точные решения, чем уравнение (20). Особенно это важно при малых значениях n, когда точность найденного значения k_m существенно влияет на возможность сжатия двоичных последовательностей. Соотношение (20) целесообразно использовать для значений n больше 128, поскольку в этом случае зависимость $k_m(n)$ отличается от линейной формы в незначительной степени. Соотношение (26) более точно описывает зависимость $k_m(n)$, особенно при n кратных степени двойки. Однако вычисление значения k_m по формуле (26) является более сложным, чем с использованием формулы (24). Поэтому соотношение (24) целесообразно применять для быстрой оценки значения k_m с использованием достаточно простых математических операций.

Из отношения k_m/n видно, что с ростом длины n сжимаемой двоичной последовательности уменьшается интервал от k_{min} до k_{max}, на

котором сжатие не происходит. В результате можно сделать вывод, что метод сжатия информации, построенный на основе биномиальных чисел, целесообразно использовать для относительно больших значений длины n сжимаемых двоичных последовательностей.

SUMMARY

DEFINE THE CONSTRAINT FOR THE METHOD OF BINOMIAL ENUMERATIVE COMPRESSION

I.A. Kulik, S.V. Kostel, E.M. Skordina,
Sumy State University, Sumy

The paper expounds the constraint for the method of binomial enumerative compression, under which the information is compressed. The proof of existence the constraint for the number of 1's depending on the length of binary information is contains in the paper. Expressions for finding the constraint of the number of 1's are considered.

Key words: *binomial enumerative coding, binomial coefficient, least-squares method, correlation coefficient.*

СПИСОК ЛИТЕРАТУРЫ

1. Кулик И.А. Использование биномиальных чисел для сжатия бинарных изображений / И.А. Кулик, С.В. Костель, Е.М. Скордина // Вісник Сумського державного університету. Серія Технічні науки. – 2009. – № 2. – С. 29-36.
2. Чередниченко В.Б. Метод сжатия двоичных кодов на основе биномиальных чисел / В.Б. Чередниченко // Вісник Сумського державного університету. Серія Технічні науки. – 2006. - №4(88). – С. 61-68.
3. Борисенко А.А. Биномиальное кодирование: монография / А.А. Борисенко, И.А. Кулик. – Сумы: СумГУ, 2010. – 206 с.
4. Борисенко А.А. Биномиальный счет. Теория и практика: монография/ А.А. Борисенко. – Сумы: ИТД "Университетская книга", 2004. – 170 с.
5. Гутер Р.С. Элементы численного анализа и математической обработки результатов опыта / Р.С. Гутер, Б.В. Овчинский. – М.: Наука, 1970. – 432 с.

Поступила в редакцию 20 апреля 2011 г.