ПРИКЛАДНАЯ ТЕОРИЯ КОДИРОВАНИЯ

DOI 10.17223/20710410/2/26

УДК 519.72

АРИФМЕТИЧЕСКОЕ КОДИРОВАНИЕ СООБЩЕНИЙ С ИСПОЛЬЗОВАНИЕМ СЛУЧАЙНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ[1]

В.Н. Потапов

Институт математики им. С.Л. Соболева СО РАН, г. Новосибирск

E-mail: vpotapov@math.nsc.ru

Предлагается модификация метода арифметического кодирования сообщений, использующая некоторую случайную последовательность как секретный ключ. Доказано, что предлагаемый метод достигает теоретической границы сжатия текстов не только при кодировании обыкновенных сообщений, но и при кодировании частично определённых данных.

Ключевые слова: *арифметическое кодирование*, *сжатие данных*, *энтропия источника сообщений*, *частично определённые данные.*

Арифметическое кодирование является одним из наиболее известных методов, применяемых для сжатия текстов. В качестве математической модели текста обычно рассматривают множество конечных слов в некотором алфавите A с заданным на множестве слов распределением вероятностей. Кодированием называется инъективное отображение f множества слов в алфавите A в множество двоичных слов. Основными критериями эффективности метода кодирования являются степень сжатия текста и трудоёмкость кодирования и декодирования.

Идея арифметического кодирования, которая будет изложена ниже, была высказана П. Элайесом. Й. Риссанен [1, 2] предложил первый эффективный алгоритм вычисления арифметического кода с почти линейной относительно длины слова трудоёмкостью кодирования и декодирования. В дальнейшем этот метод был развит в многочисленных работах, в частности в [3, 4]. Подробное описание эффективного алгоритма арифметического кодирования имеется в [5].

В настоящей работе предлагается новое воплощение идеи арифметического кодирования с использованием в качестве параметра произвольной случайной последовательности. Новый метод обеспечивает такую же теоретическую степень сжатия сообщений, как и классическая реализация, при этом позволяя использовать произвольную случайную последовательность чисел из интервала (0,1) как секретный ключ. Кроме того, предлагаемый метод позволяет эффективно сжимать не только обычные, но и частично определённые данные, введённые в работах Л.А. Шоломова [6, 7].

Источником сообщений называется пара из конечного алфавита A и распределения вероятностей P на множестве A^* конечных слов в алфавите A. Распределение вероятностей должно удовлетворять естественным соотношениям $\sum P(wa)=P(w)$, где сумма берётся по всем продолжениям wa, $a\in A$, слова $w\in A^*$, и $\sum P(w) = 1$, где сумма берётся по всем словам одинаковой длины. Источник сообщений называется источником без памяти (источником Бернулли), если вероятность буквы в слове не зависит от контекста, т. е. $P(a_1, \ldots, a_n) = P(a_1)\ldots P(a_n)$. Величина $I(w)= -\log P(w)$ называется сложностью слова $w\in A^*$ относительно источника сообщений (A, P). Из теоремы Шеннона (см., например, [8]) следует, что величина $I(w)$ является длиной двоичного кода слова при оптимальном префиксном кодировании.

Арифметическое кодирование сообщений заключается в следующем. Все слова $w\in A^n$ упорядочим лексикографически и для каждого слова определим величины $L(w)= \sum P(x)$, где сумма берётся по всем словам $x\in A^n$, лексикографически меньшим, чем слово w, и $R(w) = L(w) + P(w)$. Разделим полуинтервал [0,1) на полуинтервалы $i(w) = [L(w), R(w))$. В каждом полуинтервале $i(w)$ длины $P(w)$ найдётся двоично-рациональное число $q(w)$ со знаменателем, не большим, чем $2^{\lceil -\log P(w)\rceil}$, поскольку разность между ближайшими числами с таким знаменателем не превосходит длины полуинтервала. В качестве кода $f(w)$ слова $w\in A^n$ рассмотрим

[1] Исследование выполнено при финансовой поддержке РФФИ (проект 08-01-00671).

двоичную запись числителя дроби $q(w)$ длиной $\lceil -\log P(w) \rceil$, в случае необходимости перед двоичной записью числителя следует приписать недостающие нули. Полученное отображение f является инъективным на множестве A^n по построению, причём для длины кода справедливо неравенство $|f(w)| \le I(w) + 1$, т. е. длина кода каждого слова является оптимальной с точностью до неизбежного округления к ближайшему целому. Нетрудно видеть, что отображение f будет инъективным и на A^*, если каждое продолжение произвольного слова w имеет вероятность меньше, чем $P(w)/2$. В противном случае отображение f можно преобразовать в инъективное и даже префиксное кодирование без существенного увеличения длины кода, приписывая $Cod(n)$ в качестве префикса к $f(w)$ при $w \in A^n$, где Cod – префиксное кодирование натуральных чисел, удовлетворяющее свойству $|Cod(n)| = \log n(1 + o(1))$. Примеры таких префиксных кодов натурального ряда имеются, например, в [8, 9].

Предлагается следующая модификация арифметического кодирования. Пусть $x_1, \ldots, x_n, \ldots$ – последовательность чисел из интервала (0,1). Рассмотрим слово $w \in A^n$. Пусть m – номер первого из чисел последовательности, попавших в полуинтервал $i(w)$, т.е. $x_m \in i(w)$. В качестве кода слова w рассмотрим $g_x(w) = Cod(m)$. Нетрудно видеть, что отображение g_x является инъективным отображением на A^n.

Теорема 1. Пусть $\xi_1, \ldots, \xi_m, \ldots$ – последовательность независимых равномерно распределённых на (0,1) случайных величин. Тогда $\mathbf{E}|g_\xi(w)| \le I(w)(1 + o(1))$ при $|w| \to \infty$.

Доказательство. Пусть $A \subseteq (0, 1)$. По условию $P(\xi_i \in A) = P(A)$ – мера Лебега множества A для всех $i \in 1, \ldots, n$ и $P(\xi_1, \ldots, \xi_{n-1} \notin A, \xi_n \in A) = (1 - P(A))^{n-1} P(A)$. Рассмотрим случайную величину m на измеримых подмножествах $A \subseteq (0, 1)$, равную номеру первой из случайных ξ_i, такой, что $\xi_i \in A$. Очевидно $\mathbf{E}(m) = \sum_{n=1}^{\infty} n(1 - P(A))^{n-1} P(A) = 1/P(A)$. Поскольку $\log t$ – выпуклая вверх функция, имеем $\mathbf{E}(\log m) \le \log(\mathbf{E}(m))$. Пусть $A = i(w)$. Тогда $\mathbf{E}|g_\xi(w)| = \mathbf{E}|Cod(m)| = (1 + o(1))\mathbf{E}(\log m) \le (1 + o(1)) \log(\mathbf{E}(m)) = I(w)(1 + o(1))$ при $|w| \to \infty$. Теорема доказана.

Так же как известное арифметическое кодирование f, предлагаемое кодирование g_ξ можно преобразовать в префиксное без асимптотического увеличения длины кода. Из теоремы Шеннона следует, что полученная в теореме 1 оценка не улучшаема ни для какого префиксного кодирования, так как средняя длина префиксного кода слова не может быть меньше величины $I(w)$.

Описанная выше модификация арифметического кодирования может быть использована для кодирования частично определённых данных. Задача кодирования частично определённых данных состоит в следующем. Пусть $B = 2^A$ – множество подмножеств алфавита A. Слово $w = a_1 \ldots a_n$ называется доопределением сообщения $W = b_1 \ldots b_n$, если $a_i \in b_i$ для всех $i = 1 \ldots n$. Кодированием частично определённых данных называется отображение F из множества B^* в множество двоичных слов, обладающее свойством: найдётся такая функция (декодирование) F^o, действующая из множества двоичных слов в множество слов в алфавите A, что $F^o(F(W))$ является доопределением слова $W \in B^*$. Таким образом, результатом декодирования будет не исходное частично определённое сообщение, а некоторое его доопределение. Источник частично определённых сообщений без памяти определяется как пара (B, P), где P – некоторое распределение вероятностей на множестве B.

В работе [7] доказан аналог теоремы Шеннона для источников (без памяти) частично определённых сообщений. А именно, доказано, что величина $H = \min\left(-\sum_{b \in B} P(b) \log\left(\sum_{a \in b} Q(a)\right)\right)$, где минимум берётся по всевозможным распределениям вероятностей Q на алфавите A, является нижней гранью стоимости кодирования, т. е. средней длины кода в расчёте на букву исходного недоопределённого сообщения $W \in B^*$. Естественно определить сложность $b \in B$ относительно (B, P) как $I(b) = -\log \sum_{a \in b} Q(a)$, а сложность $I(W)$ сообщения $W = b_1 \ldots b_n$ – как сумму сложностей $I(b_i)$.

Пусть Q – распределение вероятностей, на котором достигается минимум H. Для каждого слова $w \in A^n$ определим полуинтервал $i(w)$ в соответствии с распределением Q. Каждому сообщению $W \in B^*$ поставим в соответствие множество $i(W) = \cup\, i(w)$, где объединение берётся по всем словам $w \in A^*$, доопределяющим сообщение $W \in B^*$. Пусть $x_1, \ldots, x_n, \ldots$ – последовательность чисел из интервала (0,1). Пусть m – номер первого из чисел последовательности, попавших в множество $i(W)$, т.е. $x_m \in i(W)$. В качестве кода сообщения W рассмотрим $G_x(W) = Cod(m)$. Нетрудно видеть, что отображение G_x является инъективным отображением на B^n.

Теорема 2. Пусть $\xi_1, \ldots, \xi_m, \ldots$ – последовательность независимых равномерно распределённых на (0,1) случайных величин. Тогда $\mathbf{E}|G_\xi(W)| \le I(W)(1 + o(1))$ при $|W| \to \infty$.

Доказательство теоремы 2 аналогично доказательству теоремы 1. Из [7] следует, что предложенное кодирование обеспечивает асимптотически минимальную длину кодовых слов.

ЛИТЕРАТУРА

1. *Rissanen J.* Generalized Kraft inequality and arithmetic coding // IBM J. Res. Develop. 1976. V. 20. No. 3. P. 198 – 203.
2. *Rissanen J.*, *Langdon G.* Universal modelling and coding // IEEE Trans. Inform. Theory. 1981. V. IT-27. No. 1. P. 12 – 23.
3. *Рябко Б.Я.*, *Фионов А.Н.* Эффективный метод арифметического кодирования для источников с большими алфавитами // Проблемы передачи информации. 1999. Т. 35. Вып. 4. С. 95 – 108.
4. *Witten I.H.*, *Neal R.M.*, *Cleary J.G.* Arithmetic coding for data compression // Commun. ACM. 1987. V. 30. No. 6. P. 520 – 540.
5. *Потапов В.Н.* Арифметическое кодирование вероятностных источников // Дискретная математика и её приложения: Сб. лекций молодёжных научных школ по дискретной математике и её приложениям II. М.: Изд-во центра прикладных исследований при ММФ МГУ, 2001. С. 59 – 70.
6. *Шоломов Л.А.* О мере информации нечётких и частично определённых данных // Докл. РАН. 2006. Т. 140. № 1. С. 321 – 325.
7. *Шоломов Л.А.* Сжатие частично определённой информации // Нелинейная динамика и управление. Вып. 4. М.: Физматлит, 2004. С. 385 – 399.
8. *Потапов В.Н.* Теория информации. Кодирование дискретных вероятностных источников. Новосибирск: Изд. центр НГУ, 1999.
9. *Левенштейн В.И.* Об избыточности и замедлении разделимого кодирования натуральных чисел // Проблемы кибернетики. М.: Наука, 1968. Вып. 20. С. 173 – 179.