УДК 123.456.789

Оценка избыточности универсального кодирования произвольного множества источников без памяти

В. К. Трофимов

Получена нижняя оценка избыточности универсального кодирования произвольного множества источников без памяти, совпадающая по порядку убывания с верхней оценкой, полученной автором ранее.

Ключевые слова: кодирование, избыточность, энтропия, хранение и обработка информации, источник сообщений.

1. Введение

В настоящей работе рассматривается вопрос о кодировании блоков, порожденных источником без памяти, словами различной длины в том случае, когда источник не известен, но известно множество, которому он принадлежит.

Если источник полностью известен, то данный вопрос решен в [2, 3], если же источник полностью не известен, то оптимальное кодирование предложено в [4, 5]. При этом скорость убывания избыточности в первом случае равна c_1/n, во втором – $(c_2 \log n)/n$, где n – длина кодированного блока, c_1, c_2 – постоянные, не зависящие от n.

В [1] автором предложен метод универсального кодирования, который зависит от массивности множества источников, определяемой ε-энтропией этого множества. Как было показано в [1], в зависимости от массивности заданного множества источников скорость убывания избыточности изменяется в пределах от c_1/n до $(c_2 \log n)/n$.

Целью настоящей работы является получение нижней оценки скорости убывания избыточности, совпадающей с верхней оценкой из [1], и установление факта, является ли предложенное в [1] кодирование асимптотически оптимальным.

2. Основные определения

В данной работе мы будем использовать определения и обозначения, введенные в работе [1]. Напомним кратко основные из них. Как и прежде, буквы конечного входного алфавита $A=\{a_1, a_2, ..., a_k\}$, $2 \le k < \infty$ порождаются бернуллиевским источником, т.е. источником без памяти, с вероятностями $P_\theta(a_i)=\theta_i$, $i=\overline{1,k}$, $\theta_1+\theta_2+...\theta_k=1$.

Таким образом, источник θ однозначно определяется числами θ_i, $i=\overline{1,k}$, сумма которых равна единице. Верно и обратное утверждение. Как обычно, A^n – множество всех слов (блоков) длины n в алфавите A. В этом случае произвольная полубесконечная последовательность букв алфавита A, порождаемая источником θ, однозначно разбиваются на слова (блоки) длины n.

Каждое из слов u, $u \in A^n$ с помощью отображения φ отображается в слово $\varphi(u)$ конечного выходного алфавита B. Не уменьшая общности, можно считать, что $B = \{0,1\}$. Мы берем только такие изображения φ, чтобы $\varphi(A^n) = \{\varphi(u), u \in A^n\}$ являлось префиксным, т.е. рассматриваются только дешифруемые кодирования [6]. Стоимость кодирования φ обозначим как $c(n, \theta, \varphi)$ [6]. Здесь и в дальнейшем:

$$c(n,\theta,\varphi) = \frac{1}{n} \sum_{u \in A^n} P_\theta(u) \cdot |\varphi(u)|, \tag{1}$$

где $P_\theta(u)$ – вероятность порождения слова u длины n источником θ, $|\varphi(u)|$ – число букв выходного алфавита в слове $\varphi(u)$.

Энтропию источника θ обозначим $H(\theta)$. По определению

$$H(\theta) = -\sum_{i=1}^{k} \theta_i \log \theta_i,$$

здесь и в дальнейшем $\log x = \log_2 x$, $0 \log 0 = 0$. Разность между $c(n, \theta, \varphi)$ и $H(\theta)$ обозначим через $r(n, \theta, \varphi)$ и назовем избыточностью кодирования φ. Таким образом:

$$r(n,\theta,\varphi) = c(n,\theta,\varphi) - H(\theta) \tag{2}$$

Величина $R(n, \Omega)$, определяемая равенством

$$R(n,\Omega) = \inf_{\varphi} \sup_{\theta \in \Omega} r(n,\theta,\varphi), \tag{3}$$

называется избыточностью универсального кодирования для множества источников Ω, $\Omega \subset \Omega_0$ при заданной сложности n. В равенстве (3) нижняя грань берется по всем дешифруемым кодированиям φ, определенным на множестве входных слов A^n. Если Ω содержит единственный источник θ_0, то поведение $R(n, \theta_0)$ полностью изучено в [2, 3], где установлено, что $R(n, \theta_0)$, как правило, убывает со скоростью c/n. Если же множество источников Ω совпадает с Ω_0, то в [4] доказано асимптотическое равенство

$$R_n(\Omega_0) \sim \frac{k-1}{2} \cdot \frac{\log n}{n},$$

здесь и в дальнейшем $f(n) \sim g(n)$, если $\lim_{n \to \infty} \frac{f(n)}{g(n)} = 1$; $f(n) \le g(n)$, если $\lim_{n \to \infty} \frac{f(n)}{g(n)} \le 1$.

В работе автора [1] для произвольного множества $\Omega \subset \Omega_0$ предложено кодирование, которое позволяет получить асимптотическую оценку

$$R_n(n,\Omega) \le \frac{h_{1/\sqrt{n}}(\Omega)}{n} + \frac{c}{n}, \tag{4}$$

где $h_\varepsilon(\Omega)$ – ε-энтропия Ω [6].

В настоящей работе установлена нижняя оценка для $R_n(n, \Omega)$, асимптотически совпадающая с верхней оценкой (4), и тем самым доказано асимптотическое равенство, когда $h_{1/\sqrt{n}}(\Omega) \to \infty$, при $n \to \infty$:

$$R(n,\Omega) \sim \frac{h_{1/\sqrt{n}}(\Omega)}{n} + \frac{c}{n}. \tag{5}$$

В том случае, когда $|\Omega| = m,\ m < \infty$, то

$$\frac{\log m}{n} \le R(n,\Omega) \le \frac{\log m}{n} + \frac{c}{n}.$$

Наряду с минимаксным подходом, который характеризуется избыточностью универсального кодирования для множества источников Ω, определяемого равенством (3), можно использовать среднюю избыточность $\overline{R}(n,\Omega)$ универсального кодирования для множества источников Ω, определяемую равенством:

$$\overline{R}(n,\Omega) = \inf_{\varphi} \int_{\Omega} r(n,\theta,\varphi)dw(\theta), \tag{6}$$

где $w(\theta)$ – распределение вероятности на множестве источников Ω.

Хорошо известно, что для любого множества источников Ω и произвольной меры $w(\theta)$ всегда выполняется неравенство:

$$R(n,\Omega) \ge \overline{R}(n,\theta). \tag{7}$$

Именно этим неравенством мы будем пользоваться для получения нижней оценки для $R(n,\Omega)$.

3. Оценка избыточности универсального кодирования произвольного множества источников без памяти

Для доказательства основного результата настоящего параграфа докажем ряд вспомогательных утверждений.

Для $\varepsilon = 2/\sqrt{n}$ построим минимальную ε-сеть $\Omega\left(2/\sqrt{n}\right)$ для источников Ω. Тогда согласно [7] справедливы неравенства

$$\left\|\Omega\left(1/\sqrt{n}\right)\right\| \ge \left\|\Omega\left(2/\sqrt{n}\right)\right\| \ge \frac{1}{2}\left\|\Omega\left(1/\sqrt{n}\right)\right\|. \tag{8}$$

Каждый из источников $\Omega\left(2/\sqrt{n}\right)$ накроем кубом со стороной $1/\sqrt{n}$. Полученное множество источников обозначим $\overline{\Omega}\left(2/\sqrt{n}\right)$, а через $\mu\left(\overline{\Omega}\left(2/\sqrt{n}\right)\right)$ – меру множества $\overline{\Omega}\left(2/\sqrt{n}\right)$ с плотностью $f(\theta) = c/\sqrt{\Pi_{i=1}^{K}\theta_i}$.

Для множества источников Ω и $\overline{\Omega}\left(2/\sqrt{n}\right)$ выполняется соотношение $\Omega \in \overline{\Omega}\left(2/\sqrt{n}\right)$, поэтому

$$R\left(n,\overline{\Omega}\left(2/\sqrt{n}\right)\right) \ge R(n,\Omega). \tag{9}$$

Используя кодирование, предложенное в [1], для $R\left(n,\overline{\Omega}\left(2/\sqrt{n}\right)\right)$ и для $R(n,\Omega)$ в силу определения $\overline{\Omega}\left(2/\sqrt{n}\right)$ имеем

$$R\left(n,\overline{\Omega}\left(2/\sqrt{n}\right)\right) \le \frac{h_{1/\sqrt{n}}(\Omega)}{n} + \frac{c_1}{n}; \tag{10}$$

$$R(n,\Omega) \le \frac{h_{1/\sqrt{n}}(\Omega)}{n} + \frac{c_2}{n}. \tag{11}$$

Из [1] следует, что всякое универсальное кодирование для $R(n,\overline{\Omega}(2/\sqrt{n}))$, которое позволяет получить оценку (10), является универсальным кодированием и для множества источников Ω.

Таким образом, если мы докажем, что для $R(n,\overline{\Omega}(2/\sqrt{n}))$ выполняется асимптотическое равенство

$$R(n,\overline{\Omega}(2/\sqrt{n})) \sim \frac{h_{1/\sqrt{n}}(\Omega)}{n}, \tag{12}$$

то для $R(n,\Omega)$ будет выполняться то же самое тождество. Таким образом,

$$R(n,\overline{\Omega}(2/\sqrt{n})) \sim R(n,\Omega). \tag{13}$$

Для доказательства асимптотического равенства (12) нам необходимо получить нижнюю оценку для $R(n,\overline{\Omega}(2/\sqrt{n}))$, асимптотически совпадающую с верхней оценкой из (10).

Для получения нижней оценки $R(n,\overline{\Omega}(2/\sqrt{n}))$ воспользуемся неравенством (7). Докажем утверждение, позволяющее получить нижнюю оценку для $\bar{R}(n,\overline{\Omega}(2/\sqrt{n}))$.

Лемма. *Для средней избыточности* $\bar{R}(n,\overline{\Omega}(2/\sqrt{n}))$ *множества источников* $\overline{\Omega}\left(2/\sqrt{n}\right)$ *выполняется неравенство:*

$$\bar{R}(n,\overline{\Omega}(2/\sqrt{n})) \ge \frac{h_{1/\sqrt{n}}(\Omega)}{n} + \frac{c}{n}.$$

Доказательство. Обозначим через $\mu(\overline{\Omega}(2/\sqrt{n}))$ меру множества источников $\overline{\Omega}(2/\sqrt{n})$ с плотностью $f(\theta) = c/\sqrt{\Pi_{i=1}^{K}\theta_i}$. Согласно определению средней избыточности получаем:

$$\bar{R}(n,\overline{\Omega}(2/\sqrt{n})) = \inf_{\varphi} \int_{\overline{\Omega}\left(2/\sqrt{n}\right)} \frac{f(\theta)}{\mu(\overline{\Omega}(2/\sqrt{n}))} r(n,\theta,\varphi)\, d(\theta)\ . \tag{14}$$

Согласно теореме Шеннона [2, 6] правая часть (13) достигает своего минимума при кодировании φ^0, которое каждому слову u, $u \in A^n$ ставит в соответствие слово $\varphi^0(u)$ длины

$$\left|\varphi^0(u)\right| = -\log \int_{\overline{\Omega}\left(2/\sqrt{n}\right)} \frac{f(\theta)}{\mu(\overline{\Omega}(2/\sqrt{n}))} P_0(u)\, d(\theta).$$

Так как для любой положительной функции $g(\theta)$ выполняется неравенство

$$\int_{\overline{\Omega}\left(2/\sqrt{n}\right)} g(\theta)\, d\theta \le \int_{\Omega_0} g(\theta)\, d\theta,$$

то

$$\left|\varphi^0(u)\right| \ge -\log \int_{\Omega_0} f(\theta) P_{\theta(u)} d\theta + \log \mu(\overline{\Omega}(2/\sqrt{n})).$$

Принимая во внимание равенство

$$\mu(\bar{R}(2/\sqrt{n})) = c\left\|\Omega(2/\sqrt{n})\right\| / N^{k-1/2}$$

и определение $h_{1/\sqrt{n}}(\Omega)$, получаем:

$$\left|\varphi^0(u)\right| \geq -\log f(\theta)P_0(u)d\theta + h_{1/\sqrt{n}}(\Omega) - \frac{k-1}{2}\log n + c.$$

Умножив обе части на $P_0(u)$ и просуммировав по u, получаем

$$\sum_{u\in A^n} P_\theta(u)\cdot\left|\varphi^0(u)\right| \geq -\sum_{u\in A^n} P_\theta(u)\log f(\theta)P_\theta(u)d\theta + h_{1/\sqrt{n}}(\Omega) - \frac{k-1}{2}\log n + c.$$

Разделим обе части последнего равенства на n и вычтем из обеих частей неравенства энтропию $H(\theta)$. Учитывая определения, данные выше, получаем:

$$r(n,\theta,\varphi^0) \geq r(n,\theta,\overline{\varphi}^0) + \frac{h_{1/\sqrt{n}}(\theta)}{n} - \frac{k-1}{2}\frac{\log n}{n} + \frac{c}{n}, \qquad (15)$$

где кодирование $\overline{\varphi}^0$ каждому слову u ставит в соответствие слово $\overline{\varphi}^0(u)$, для которого

$$\left|\overline{\varphi^0}(u)\right| = -\log f(\theta)P_\theta(u)d\theta.$$

В [5] доказано, что для любого источника $\theta\in\Omega_0$ выполняется неравенство:

$$r(n,\theta,\overline{\varphi^0}) \geq \frac{k-1}{2}\frac{\log n}{n} + \frac{c}{n}. \qquad (16)$$

Из (15) с учетом (16) имеем:

$$r(n,\theta,\varphi^0) \geq \frac{h_{1/\sqrt{n}}(\theta)}{n} + \frac{c}{n}. \qquad (17)$$

Из (14), используя (17), получаем

$$\overline{R}(n,\overline{\Omega}(2/\sqrt{n})) \geq \frac{h_{1/\sqrt{n}}(\Omega)}{n} + \frac{c}{n}.$$

Лемма доказана. □

4. Избыточность универсального кодирования произвольного множества источников без памяти

Основное утверждение работы можно сформулировать в виде утверждения.

Теорема. *Для произвольного подмножества источников* Ω, $\Omega\subset\Omega_0$ *для избыточности* $R(n,\Omega)$ *универсального равномерного по входу кодирования выполняется асимптотическое равенство*

$$R(n,\Omega) \sim \frac{h_{1/\sqrt{n}}}{n} + \frac{c}{n}.$$

Доказательство. Верхняя оценка $R(n,\Omega)$ следует из асимптотического неравенства (4), доказанного в [1]. Нижняя оценка вытекает из леммы очевидного неравенства

$$R(n,\overline{\Omega}(2/\sqrt{n}) \geq \overline{R}(n,\overline{\Omega}(2/\sqrt{n})$$

и соотношения (13). Теорема доказана. □

5. Заключение

Получена нижняя оценка избыточности универсального кодирования через ε-энтропию множества источников, совпадающая с верхней оценкой, доказанной в [1]. Тем самым получено асимптотически точное значение избыточности для произвольного множества источников без памяти.

Литература

1. *Трофимов В. К.* Универсальное кодирование произвольного множества источников без памяти // Вестник СибГУТИ. 2018. № 4. С. 30–34.
2. *Шеннон К.* Математическая теория связи. Работы по теории информации и кибернетике. 1963. С. 243–332.
3. *Кричевский Р. Е.* Длина блока, необходимая для получения заданной избыточности // ДАН СССР. 1965. Т. 171, № 1. С. 37–40.
4. *Кричевский Р. Е.* Связь между избыточностью кодирования и достоверностью сведений об источнике // Проблемы передачи информации. 1968. Т. 4, № 3. С. 48−57.
5. *Krichevsky R. E., Trofimov V. K.* The performance of universal encoding // IEEE Transactions on Information Theory. 1981. V. 27, № 2. P. 199–207.
6. *Галлагер Р. Г.* Теория информации и надежная связь. М.: Советское радио, 1974.
7. *Витушкин А. Г.* Оценка сложности задачи табулирования. М.: Гос. изд. физико-математической лит., 1959. 228 с.

Статья поступила в редакцию 09.09.2019.

Трофимов Виктор Куприянович
д.т.н., профессор, зав. кафедрой высшей математики СибГУТИ (630102, Новосибирск, ул. Кирова, 86), e-mail: trofimov@sibsutis.ru.

Evaluation of the universal coding redundancy of an arbitrary set of sources without memory

V. K. Trofimov

This article considers a lower estimation of the universal coding redundancy of an arbitrary set of sources without memory coinciding in decreasing order with the upper estimation obtained by the author in the previous work.

Keywords: coding, redundancy, entropy, storage and processing of information, message source.