УДК 123.456.789

Универсальное кодирование произвольного множества источников без памяти

В. К. Трофимов

Предложен метод универсального кодирования произвольного множества источников без памяти. Получена оценка универсального кодирования в зависимости от ε-энтропии множества источников, описывающей массивность этого множества.

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

1. Введение

Уменьшение объёма передаваемой информации – одна из проблем теории информации. Уменьшение объёма, например, эквивалентно увеличению скорости передачи, уменьшению процессорного времени, необходимого для её обработки. В настоящее время существует два направления сокращения избыточности: сжатие данных и кодирование дискретных источников. В работе [1] заложены основы теории передачи информации, с помощью которой получены различные алгоритмы устранения избыточности как при известной, так и при неизвестной статистике сообщений. Достаточно подробную библиографию по этому вопросу можно найти в [2]. Кодирование дискретных источников используется при факсимильной передаче изображений [3], в задачах поиска и хранения данных [4], в теории управления [5], при выявлении скрытой информации [6]. Решение этой проблемы значимо при создании большемасштабных распределённых вычислительных систем [7].

В данной работе предлагается сжатие информации, порожденной источником без памяти символами различной длины, в том случае, когда источник неизвестен, но известно множество, которому он принадлежит. Например, перечислено конечное множество известных источников, которые могут порождать информацию или наложены определённые ограничения на вероятности источника и т.д.

Цель данной работы – построить кодирование, которое будет зависеть от массивности множества источников. Эта массивность определяется ε-энтропией этого множества. В зависимости от массивности множества источников скорость убывания избыточности изменяется от $\frac{C_1}{n}$ до $\frac{C\log n}{n}$, где n – длина кодируемого блока.

2. Основные определения. Постановка задачи

Пусть буквы конечного алфавита $A = \{a_1, a_2, \dots, a_k\}, 2 \le k \le \infty$, порождаются источником θ независимо с вероятностями $P_\theta(a_i) = \theta_i, i = \overline{1,k}, \theta_1 + \theta_2 + \dots + \theta_k = 1$. В этом случае считаем, что θ – бернуллиевский источник, однозначно определённый числами $\theta_i, i = \overline{1,k}, \theta_i \ge 0$, сумма которых равна единице. Верно и обратное утверждение: любой набор чисел $\theta_i, i = \overline{1,k}$, удовлетворяющий перечисленным выше условиям, однозначно определяет бернуллиевский источник. Множество слов, взятых в произвольном алфавите, называ-

ется префиксным, если никакое слово не является началом другого. Множество слов A^n, взятых в алфавите A, назовём кодовым. В этом случае произвольная полубесконечная последовательность символов букв входного алфавита, порождаемая источником, однозначно разбивается на последовательность слов длины n. Из неравенства Крафта–Макмиллана [8] следует: самое общее из всех возможных дешифруемых кодирований φ состоит в том, что полубесконечная последовательность букв, порожденная источником, разбивается в соответствии с кодовым множеством A^n на слова, переведённые с помощью отображения φ в слова выходного алфавита B, который, не уменьшая общности, можно считать двоичным. При этом множество слов в выходном алфавите $\varphi(A^n) = \{\varphi(u), u \in A^n\}$ является префиксным.

Среднее число букв выходного алфавита, приходящееся на одну букву входного при кодировании φ, назовём стоимостью кодирования φ и обозначим $C(n, \theta, \varphi)$. Величина $C(n, \theta, \varphi)$ определяется равенством [8]

$$C(n, \theta, \varphi) = \frac{1}{n} \sum_{u \in T} P_\theta(u) |\varphi(u)|, \tag{1}$$

где $P_\theta(u)$ – вероятность порождения слова u источником θ; n – длина кодового слова.

Через $H(\theta)$ обозначим энтропию источника θ. Для бернуллиевского источника θ величина $H(\theta)$ определяется выражением [1]

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

Здесь и в дальнейшем $\log x = \log_2 x$, $0 \log 0 = 0$.

Эффективность кодирования φ оценивается разностью между стоимостью кодирования $C(n, \theta, \varphi)$, определяемой равенством (1), и энтропией источника $H(\theta)$. Эта разность называется избыточностью кодирования и обозначается $r(n, \theta, \varphi)$. Таким образом, по определению

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

Избыточностью универсального кодирования для заданного множества источников $\Omega \subseteq \Omega_0$ и сложностью n назовём величину

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

где нижняя грань берётся по всем кодированиям φ, определённым на множестве входных слов A^n. При известной статистике сообщений $R(n, \Omega)$ изучена в [1, 9], где было доказано, что избыточность известного источника без памяти, как правило, убывает со скоростью $\frac{C}{n}$. В [10] доказано, что для избыточности $R(n, \Omega_0)$ универсального равномерного по входу кодирования источников без памяти Ω_0 имеет место асимптотическое равенство:

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

В настоящей работе получена оценка сверху для избыточности универсального кодирования $R(n, \Omega)$, где $\Omega \subseteq \Omega_0$. Доказано, что

$$R(n, \Omega) \leq \frac{H_{\frac{1}{\sqrt{n}}}(\Omega)}{n} + \frac{C}{n}, \tag{4}$$

где $H_\varepsilon(\Omega)$ – ε-энтропия множества Ω, определяемая обычным образом [11].

3. ε-энтропия множества источников без памяти

Пусть Ω – произвольное подмножество источников из Ω_0. Произвольный источник θ из Ω_0 однозначно определяется набором (вектором) вероятностей $(\theta_1, \theta_2, \dots, \theta_k)$, $\theta_1 + \theta_2 + \dots + \theta_k = 1$, где $\theta_i = P_\theta(a_i), i = \overline{1, k}$. Определим евклидово расстояние между векторами из Ω как обычно, ε-энтропия равна логарифму числа элементов минимальной ε-сети для множества источников Ω [11].

Приведём некоторые примеры:

а) если Ω_0 – множество всех k-буквенных источников без памяти, то

$$h_\varepsilon(\Omega_0) = (k-1)\log\frac{1}{\varepsilon} + o\left(\log\frac{1}{\varepsilon}\right);$$

б) если Ω – конечное множество источников и $\|\Omega\|$ – мощность множества Ω, то

$$h_\varepsilon(\Omega) = \log\|\Omega\|;$$

в) если хотя бы одна из вероятностей источников из Ω заполняет некоторый интервал, то

$$h_\varepsilon(\Omega) = O\left(\log\frac{1}{\varepsilon}\right).$$

4. Метод кодирования и оценка его эффективности

Пусть Ω – произвольное подмножество источников из Ω_0. Предположим, что для любого источника $\theta,\ \theta \in \Omega$, выполняется неравенство $\theta_i \geq \nu \geq 0,\ i = \overline{1,k}$. Зададим произвольное сколь угодно малое число $\varepsilon,\ \varepsilon > 0$. Построим ε-сеть для множества Ω и через $\Omega(\varepsilon)$ обозначим элементы минимальной ε-сети множества Ω. Тогда для каждого источника $\theta,\ \theta \in \Omega$, найдется по крайней мере один источник $\tilde{\theta},\ \tilde{\theta} \in \Omega(\varepsilon)$, такой, что выполняются соотношения

$$\tilde{\theta}_i - \theta_i = \Delta_i, \qquad |\Delta_i| \leq \varepsilon\,. \tag{5}$$

Просуммировав последнее равенство по i и учитывая, что $\sum_{i=1}^k \tilde{\theta}_i = \sum_{i=1}^k \theta_i = 1$, получаем равенство

$$\sum_{i=1}^k \Delta_i = 0. \tag{6}$$

Справедливо утверждение:

Лемма. *Для произвольного множества источников Ω, $\Omega \subseteq \Omega_0$, и произвольного $\varepsilon,\ \varepsilon < \min\limits_{\theta\in\Omega} \theta_i$, существует ε-сеть множества источников Ω, такая, что для любого источника θ, $\theta \in \Omega$, найдется такой источник $\tilde{\theta},\ \tilde{\theta} \in \Omega(\varepsilon)$, что*

$$0 \leq -\sum_{i=1}^k \theta_i\left(log\,\tilde{\theta}_{\alpha i} - log\,\theta_{\alpha i}\right) \leq C\varepsilon^2, \tag{7}$$

C не зависит от θ.

Доказательство. Из равенства $\sum_{i=1}^k \tilde{\theta}_i = 1$ и теоремы К. Шеннона [1, 8] для каналов без шума вытекает справедливость нижней оценки в (6).

Докажем справедливость верхней оценки. Как было отмечено выше, для каждого источника $\theta,\ \theta \in \Omega$, найдется источник $\tilde{\theta},\ \tilde{\theta} \in \Omega(\varepsilon)$, для которого справедливы соотношения (5). Левую часть (7) перепишем в виде:

$$-\sum_{i=1}^k \theta_i \log\tilde{\theta}_i + \sum_{i=1}^k \theta_i \log\theta_i = -\sum_{i=1}^k \theta_i \log\frac{\theta_i + \Delta_i}{\theta_i} = -\sum_{i=1}^k \theta_i \log\left(1 + \frac{\Delta_i}{\theta_i}\right).$$

Воспользовавшись разложением в ряд функции $\log(1+x)$ из последнего равенства и из (6) получаем:

$$-\sum_{i=1}^k \theta_i \log\tilde{\theta}_i + \sum_{i=1}^k \theta_i \log\theta_i \leq \left[\sum_{i=1}^k \Delta_i + \frac{1}{2}\sum_{i=1}^k \frac{\Delta_i^2}{\theta_i}\right]\log e = \left(\frac{1}{2}\sum_{i=1}^k \frac{\Delta_i^2}{\theta_i}\right)\log e.$$

Отсюда вытекает справедливость верхней оценки в (7). Лемма доказана.

Опишем метод кодирования множества источников без памяти и оценим его эффективность.

Теорема. *Для произвольного источника без памяти Ω и произвольного $\varepsilon,\ \varepsilon < \min\limits_{\substack{\theta\in\Omega \\ i=\overline{1,k}}} \theta_i$, для $R(n,\Omega)$ избыточности универсального кодирования множества источников Ω выполняется неравенство*:

$$R(n,\Omega) \le \frac{\log\|\Omega(\varepsilon)\|}{n} + C_1\varepsilon^2 + \frac{C_2}{n},$$

C_1 и C_2 не зависят от ε.

Доказательство. Возьмем произвольное ε, $\varepsilon > 0$, удовлетворяющее условию леммы, и построим минимальную ε-сеть $\Omega(\varepsilon)$. Элементы из множества $\Omega(\varepsilon)$ занумеруем произвольным образом: $\tilde{\theta}_1, \tilde{\theta}_2, \dots, \tilde{\theta}_{\|\Omega(\varepsilon)\|}$. Рассмотрим кодирование φ_ε, которое каждому слову u, $u \in A^n$, ставит в соответствие слово $\varphi_\varepsilon(u)$ длины

$$|\varphi_\varepsilon(u)| \le \left\lceil -\log \frac{\sum_{j=1}^{\|\Omega(\varepsilon)\|} P_{\tilde{\theta}_j}(u)}{\|\Omega(\varepsilon)\|} \right\rceil, \tag{8}$$

Здесь $\lceil x \rceil$ – наименьшее целое число, большее или равное x.

Существование дешифруемого кодирования φ_ε с длинами кодовых слов, удовлетворяющих равенствам (8), вытекает из выполнения неравенства Крафта [8] для чисел $|\varphi_\varepsilon(u)|$, $u \in A^n$. Оценим эффективность кодирования φ_ε. Из определения $r(n,\theta,\varphi_\varepsilon)$ избыточности кодирования φ_ε при заданном источнике θ и из (8) имеем:

$$r(n,\theta,\varphi_\varepsilon) = \frac{1}{n}\sum_{u\in A^n} P_\theta(u)|\varphi_\varepsilon(u)| - H(\theta) \le -\frac{1}{n}\sum_{u\in A^n} P_\theta(u)\log\frac{\sum_{j=1}^{\|\Omega(\varepsilon)\|} P_{\tilde{\theta}_j}(u)}{\|\Omega(\varepsilon)\|} - H(\theta) + \frac{1}{n}.$$

Пусть источник θ находится в ε-окрестности источника $\tilde{\theta}_{j_0}$, $\tilde{\theta}_{j_0} \in \Omega(\varepsilon)$. Тогда предыдущее неравенство можно переписать в виде:

$$r(n,\theta,\varphi_\varepsilon) \le \frac{\log\|\Omega(\varepsilon)\|}{n} - H(\theta) - \frac{1}{n}\sum_{u\in A^n} P_\theta(u)\log P_{\tilde{\theta}_{j_0}}(u) -$$

$$- \frac{1}{n}\sum_{u\in A^n} P_\theta(u)\log\left(1 + \sum_{\substack{j=1\\ j\ne j_0}}^{\|\Omega(\varepsilon)\|} \frac{P_{\tilde{\theta}_j}(u)}{P_{\tilde{\theta}_{j_0}}(u)}\right) + \frac{1}{n}. \tag{9}$$

Так как

$$-\log\left(1 + \sum_{\substack{j=1\\ j\ne j_0}}^{\|\Omega(\varepsilon)\|} \frac{P_{\tilde{\theta}_j}(u)}{P_{\tilde{\theta}_{j_0}}(u)}\right) \le 0,$$

то из (9) и определения $H(\theta)$ получим, что

$$r(n,\theta,\varphi_\varepsilon) \le \frac{\log\|\Omega(\varepsilon)\|}{n} - \sum_{i=1}^{k}\theta_i\left(\log\tilde{\theta}_i^{(j_0)} - \log\theta_i\right) + \frac{1}{n}, \tag{10}$$

где $\tilde{\theta}_i^{(j_0)}$ – параметры источника $\tilde{\theta}_{j_0}$.

Из леммы и (10) получаем:

$$r(n,\theta,\varphi_\varepsilon) \le \frac{\log\|\Omega(\varepsilon)\|}{n} + C_1\varepsilon^2 + \frac{C_2}{n}.$$

Отсюда и из определения $R(n,\Omega)$ вытекает справедливость утверждения теоремы.

Следствие. *Для избыточности $R(n,\Omega)$ универсального кодирования множества источников без памяти Ω выполняется неравенство:*

$$R(n,\Omega) \le \frac{h_{\frac{1}{\sqrt{n}}}(\Omega)}{n} + \frac{C}{n}.$$

Доказательство. Справедливость утверждения непосредственно вытекает из формулировки теоремы при $\varepsilon = \frac{1}{\sqrt{n}}$.

5. Заключение

Предложенный метод кодирования позволяет оценить избыточность универсального кодирования через ε-энтропию множества источников. В частности, если источник известен либо их конечное число, то избыточность универсального кодирования такого множества убывает со скоростью $\frac{C}{n}$, если же кодируется всё множество источников, то избыточность убывает со скоростью $\frac{C\log n}{n}$.

Литература

1. *Шеннон К.* Математическая теория связи. Работы по теории информации и кибернетике. 1963. С. 243–332.
2. *Krichevsky R. E., Trofimov V. K.* The performance of universal encoding // IEEE Transactions on Information Theory. 1981. V. 27, № 2. P. 199–207.
3. *Usubuchi T., Omachi T., Iinuma K.* Adaptive predictive coding for newspaper facsimile // Proc. IEEE. 1980. V. 68, № 7. P. 807–813.
4. *Бабкин В. Ф., Куделова К., Луценко В. Н. и др.* Опыт применения бортовой информационно-вычислительной системы для обработки данных и управления экспериментом «Интершок» // Космические исследования. 1986. Вып. 24, № 2. С. 210–216.
5. *Петров Б. Н., Добрушин Р. Л., Пинскер М. С. и др.* О некоторых взаимосвязях теории информации и теории управления // Проблемы управления и теории информации. 1976. Т. 5, № 1. С. 31−38.
6. *Жилкин М. Ю., Меленцова Н. А., Рябко Б. Я.* Методы выявления скрытой информации, базирующейся на сжатии данных // Вычислительные технологии. 2007. Т. 12. С. 26−31.
7. *Хорошевский В. Г.* Архитектура вычислительных систем. М.: МГТУ им. Н. Э. Баумана, 2005.
8. *Галлагер Р. Г.* Теория информации и надежная связь. М.: Советское радио, 1974.
9. *Кричевский Р. Е.* Длина блока, необходимая для получения заданной избыточности // ДАН СССР. 1965. Т. 171, № 1. С. 37–40.
10. *Кричевский Р. Е.* Связь между избыточностью кодирования и достоверностью сведений об источнике // Проблемы передачи информации. 1968. Т. 4, № 3. С. 48−57.
11. *Витушкин А. Г.* Оценка сложности задачи табулирования. М.: Гос. изд. физико-математической лит., 1959. 228 с.

Статья поступила в редакцию 15.09.2018.

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

Universal coding of an arbitrary set of sources without memory

V. K. Trofimov

The method of universal coding of an arbitrary set of sources without memory is proposed. An estimation of the universal coding is obtained depending on the ε-entropy of the set of sources describing the massive of this set.

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