

Кодирование сообщений, порождённых произвольным марковским источником, при неизвестной статистике сообщений *

В. К. Трофимов

Сибирский гос. унив. телекоммуникаций и информатики (СибГУТИ)

Аннотация: Доказано существование универсального кодирования произвольного множества марковских источников, а также произведена оценка его эффективности в зависимости от ε -энтропии множества источников.

Ключевые слова: источник сообщений, универсальное кодирование, марковский источник, энтропия, стоимость кодирования, избыточность кодирования.

Для цитирования: Трофимов В. К. Кодирование сообщений, порождённых произвольным марковским источником, при неизвестной статистике сообщений // Вестник СибГУТИ. 2023. Т. 17, № 4. С. 109–115. <https://doi.org/10.55648/1998-6920-2023-17-4-109-115>.



Контент доступен под лицензией
Creative Commons Attribution 4.0
License

© Трофимов В. К., 2023

Статья поступила в редакцию 20.05.2023;
принята к публикации 25.07.2023.

1. Введение

Данная работа посвящена обобщению результатов автора [1] на произвольное множество марковских источников с конечной памятью и конечным алфавитом. В этой работе доказано существование универсального кодирования произвольного множества марковских источников в зависимости от его структуры. Дана оценка избыточности этого кодирования в зависимости от ε -энтропии множества источников. В качестве следствия получаются оценки для множества всех источников без памяти, полученных Р. Е. Кричевским [2], и для множества всех марковских источников, изученных Ю. М. Штарковым [3] и автором [4].

2. Основные определения

Пусть буквы конечного алфавита $A = \{a_1, a_2, \dots, a_k\}$, $2 \leq k < \infty$, называемого в дальнейшем входным, порождаются марковским источником с памятью $0 \leq s < \infty$. В этом случае источник θ однозначно определяется переходными вероятностями:

$$p(a_i a_\alpha) = \theta_{ai}, \text{ где } a_\alpha \in A^s, i = \overline{1, k}.$$

* Работа выполнена в рамках государственного задания № 071-03-2023-001 от 19.01.2023.

При $\alpha = (i_1, i_2, \dots, i_s)$, $a_\alpha = a_{i_1}a_{i_2} \dots a_{i_s}$, $a_\alpha \in A^s$, $t = \overline{1, s}$. Очевидно, всегда выполняется равенство $p(a_{\alpha 1}) + p(a_{\alpha 2}) + \dots + p(a_{\alpha k}) = 1$, при любом $a_\alpha \in A^s$. Таким образом, каждый марковский источник с памятью s однозначно определяется набором (вектором) переходных вероятностей. Размерность этого вектора равна $k^s(k-1)$. Верно и обратное утверждение. Отсюда видно, что множество марковских источников с памятью s является единичным кубом размерности $k^s(k-1)$.

Множество всех слов A^n , взятых в алфавите A , назовем кодовым. Тогда произвольная полубесконечная последовательность букв входного алфавита, порожаемая источником, однозначно разбивается на последовательность слов $u \in A^n$. Число букв, входящих в произвольное слово u , будем обозначать $|u|$ и называть его длиной.

Произвольная полубесконечная последовательность букв входного алфавита, порожаемая источником θ , однозначно разбивается на последовательность слов длины n . Каждому слову $u \in A^n$ с помощью отображения φ поставим в соответствие слово $\varphi(u)$ в выходном алфавите B . Не умаляя общности, можно считать B – двоичным, т.е. $B = \{0, 1\}$.

В работе рассматриваются только дешифруемые кодирования, т.е. множество слов в выходном алфавите $\varphi(A^n) = \{|\varphi(u)|, u \in A^n\}$ являются префиксными или, что то же самое, что для совокупности чисел $\{|\varphi(u)|, u \in A^n\}$ выполняется неравенство Крафта [5]. В этом случае существуют различные алгоритмы построения дешифруемых кодов. Мы не ставим задачи построить код, наша задача – оценить эффективность этих кодов.

Стоимостью кодирования φ при кодировании блоков длины n источником θ назовем величину $C(n, \theta, \varphi)$, которая определяется как среднее число букв выходного алфавита, приходящихся на одну входную [6], т.е.

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

где $P_\theta(u)$ – вероятность порождения слова u источником θ .

Энтропию источника θ обозначим $H(\theta)$. В нашем случае:

$$H(\theta) = - \sum_{a_\alpha \in A^s} \theta_{o\alpha} \sum_{i=1}^k \theta_{\alpha i} \log \theta_{\alpha i}, \quad (2)$$

где $\theta_{o\alpha}$ – начальные стационарные вероятности для блоков a_α , $\theta_{\alpha i}$ – вероятность получения буквы a_i после блока a_α , $a_\alpha \in A^s$, $i = \overline{1, k}$.

Здесь и далее: $\log x = \log_2 x$.

Как показано в [6]:

$$H(\theta) = \lim_{n \rightarrow \infty} -\frac{1}{n} \sum_{|u|=n} P_\theta(u) \log P_\theta(u).$$

Избыточность кодирования $r(n, \theta, \varphi)$ определяется разностью между стоимостью кодирования $C(n, \theta, \varphi)$ и энтропией источника $H(\theta)$, определенных равенствами (1) и (2) соответственно. Таким образом, по определению

$$r(n, \theta, \varphi) = C(n, \theta, \varphi) - H(\theta). \quad (3)$$

Избыточностью универсального кодирования для множества источников Ω , $\Omega \subseteq \Omega_s$ при кодировании слов длины n назовём, как обычно [7], величину

$$R(n, \Omega) = \inf_{\varphi} \sup_{\theta \in \Omega} r(n, \theta, \varphi), \quad (4)$$

где Ω_s – множество всех марковских источников с памятью s . История поведения $R(n, \theta, \varphi)$ подробно рассмотрена в [8].

В настоящей работе получена оценка сверху для $R(n, \Omega)$. Доказано, что

$$R(n, \Omega) \leq \frac{h_{1/\sqrt{n}}(\Omega)}{n} + \frac{\lambda}{n}. \quad (5)$$

Здесь и в дальнейшем $h_{\varepsilon}(\Omega)$ – ε -энтропия Ω , определяемая обычным образом [9], если принять во внимание, что источник θ определяется точкой в $k^s (k-1)$ единичном кубе. Расстояние между точками – обычное евклидово. Для конечного множества источников Ω при $\varepsilon \rightarrow 0$, очевидно, $H_{\varepsilon}(\Omega) = \log \Omega$, где Ω – число элементов в Ω .

3. Оценка избыточности универсального кодирования

Пусть Ω – произвольное подмножество источников из Ω_s , и пусть для источников из Ω выполняются неравенства $\theta_{\alpha i} \geq \nu > 0$, $\alpha \in A^s$ $i = \overline{1, k}$. Пусть ε , $\varepsilon > 0$, сколько угодно малое. Построим ε -сеть для Ω . Элементы минимальной ε -сети обозначим $\Omega(\varepsilon)$. Тогда для каждого источника θ из Ω найдется по крайней мере один источник из $\Omega(\varepsilon)$, такой что

$$\tilde{\theta}_{\alpha i} - \theta_{\alpha i} = \Delta_{\alpha i}, |\Delta_{\alpha i}| < \varepsilon, a_{\alpha} \in A^s, i = \overline{1, k}. \quad (6)$$

Докажем справедливость следующего утверждения.

Лемма. Для произвольного множества источников Ω и произвольного ε , $\varepsilon < \min \theta_{\alpha i}$, $\theta \in \Omega$, существует ε -сеть $\Omega(\varepsilon)$, такая что для любого источника θ , $\theta \in \Omega$, найдется источник $\tilde{\theta}$, $\tilde{\theta} \in \Omega(\varepsilon)$, для которого справедливы неравенства:

$$0 \leq - \sum_{a_{\alpha} \in A^s} \theta_{\alpha} \sum_{i=1}^k \theta_{\alpha i} (\log \tilde{\theta}_{\alpha i} - \log \theta_{\alpha i}) < \lambda \varepsilon^2, \quad (7)$$

где λ не зависит от θ .

Доказательство. Для доказательства (7) достаточно установить справедливость соотношений

$$0 \leq - \sum_{i=1}^k \theta_{\alpha i} \log \tilde{\theta}_{\alpha i} + \sum_{i=1}^k \theta_{\alpha i} \log \theta_{\alpha i} \leq \lambda \varepsilon^2, a_{\alpha} \in A^s. \quad (8)$$

Из равенств: $\theta_{\alpha 1} + \theta_{\alpha 2} + \dots + \theta_{\alpha k} = 1$, $a_{\alpha} \in A^s$ и теоремы К. Шеннона для канала без шума [6] следует нижняя оценка в (8). Докажем справедливость верхней оценки. Как было отмечено выше, для любого источника из Ω найдется источник θ из $\Omega(\varepsilon)$, для которого выполнены условия (6). Соотношение (8) можно переписать в виде:

$$- \sum_{i=1}^k \theta_{\alpha i} \log \tilde{\theta}_{\alpha i} + \sum_{i=1}^k \theta_{\alpha i} \log \theta_{\alpha i} = - \sum_{i=1}^k \theta_{\alpha i} \log \left(1 + \frac{\Delta_{\alpha i}}{\theta_{\alpha i}} \right).$$

Воспользовавшись разложением в ряд функции $\log(1+x)$ из последнего неравенства, получаем:

$$-\sum_{i=1}^k \theta_{\alpha i} \log \tilde{\theta}_{\alpha i} + \sum_{i=1}^k \theta_{\alpha i} \log \theta_{\alpha i} \leq \left[\sum_{i=1}^k \Delta_{\alpha i} + \frac{1}{2} \sum_{i=1}^k \frac{\Delta_{\alpha i}^2}{\theta_{\alpha i}} \right] \log e. \quad (9)$$

Просуммировав общие части (7) по $i = \overline{1, k}$ и учитывая, что при любом α выполняются равенства $\sum_{i=1}^k \theta_{\alpha i} = \sum_{i=1}^k \tilde{\theta}_{\alpha i} = 1$, окончательно получаем, что при любом α выполняется равенство $\sum_{i=1}^k \Delta_{\alpha i} = 0$. Отсюда и из (9) вытекает справедливость верхней оценки в (8). Таким образом, утверждение леммы полностью доказано. $\square$

Перейдем к оценке универсального кодирования произвольного множества марковских источников.

Теорема. Для произвольного множества марковских источников Ω , у которых $\delta < \min \theta_{\alpha i}$, $\theta \in \Omega$, $\varepsilon < \delta$, для $R(n, \Omega, (\delta))$ избыточности универсального кодирования множества источников Ω выполняется неравенство:

$$R(n, \Omega(\delta)) \leq \frac{\log \Omega(\varepsilon)}{n} + C_1 \varepsilon^2 + \frac{C_2}{n}.$$

Здесь C_1 и C_2 не зависят от ε , ε – сколь угодно малое.

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

$$|\varphi_\varepsilon(u)| = -\log \left(\sum_{i=1}^{\Omega(\varepsilon)} P_{\tilde{\theta}_i}(u) \right) / \Omega(\varepsilon). \quad (10)$$

Существование дешифруемого кодирования φ_ε с длинами кодовых слов, удовлетворяющих равенствам (10), вытекает из выполнения неравенства Крафта [5] 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 (10) имеем:

$$\begin{aligned} r(n, \theta, \varphi_\varepsilon) &= 1/(n - \hat{s} + 1) \sum_{u=n} P_\theta(u) |\varphi_\varepsilon(u)| - H(\theta) \leq \\ &\leq -1/(n - \hat{s} + 1) \sum_{u=n} P_\theta(u) \log \left(\left[\sum_{i=1}^{\Omega(\varepsilon)} P_{\tilde{\theta}_i}(u) \right] / \Omega(\varepsilon) \right) - H(\theta) + 1/(n - \hat{s} + 1). \end{aligned}$$

Предположим, что источник θ находится в ε -окрестности источника θ_{i_0} , $\theta_{i_0} \in \Omega(\varepsilon)$, тогда предыдущее неравенство можно переписать в виде:

$$\begin{aligned} r(n, \theta, \varphi_\varepsilon) &\leq \log \Omega(\varepsilon) / (n - \hat{s} + 1) - H(\theta) - 1/(n - \hat{s} + 1) \sum_{u=n} P_\theta(u) \log P_{\theta_{i_0}}(u) + 1/(n - \hat{s} + 1) - \\ &- 1/(n - \hat{s} + 1) \sum_{u=n} P_\theta(u) \log \left(1 + \sum_{i=1, i \neq i_0}^{\Omega(\varepsilon)} P_{\tilde{\theta}_i}(u) / P_{\theta_{i_0}}(u) \right). \end{aligned} \quad (11)$$

Так как $\sum_{i=1, i \neq i_0}^{\Omega(\varepsilon)} P_{\tilde{\theta}_i}(u) / P_{\theta_{i_0}}(u) \geq 0$, то имеем неравенство

$$-\log \left(1 + \sum_{i=1, i \neq i_0}^{\Omega(\varepsilon)} P_{\tilde{\theta}_i}(u) / P_{\theta_{i_0}}(u) \right) \leq 0.$$

Поэтому из (11) и определения $H(\theta)$ получаем:

$$r(n, \theta, \varphi_\varepsilon) \leq \frac{\log \Omega(\varepsilon)}{n - \hat{s} + 1} + \frac{1}{n - \hat{s} + 1} - \sum_{\alpha} \theta_{o\alpha} \sum_{i=1}^k \theta_{\alpha i} (\log \tilde{\theta}_{\alpha i}^j - \log \theta_{\alpha i}).$$

Воспользовавшись леммой из (11), имеем:

$$r(n, \theta, \varphi_\varepsilon) \leq \frac{\log \Omega(\varepsilon)}{n - \hat{s} + 1} + C\varepsilon^2 + \frac{1}{n - \hat{s} + 1}.$$

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

Из доказанной теоремы вытекает:

Следствие 1. Для избыточности $r(n, \theta)$ универсального кодирования марковских источников справедливо неравенство:

$$r(n, \theta, \varphi_\varepsilon) \leq \frac{h_{1/\sqrt{n}}(\Omega)}{n} + \frac{C(\theta)}{n}.$$

Доказательство. Для фиксированного марковского источника существует δ , такое что $\theta_{\alpha i} > \delta > 0$, $a \in A^s$, $i = \overline{1, k}$. Отсюда и из теоремы вытекает справедливость следствия при $\varepsilon = 1/\sqrt{n}$. $\square$

Следствие 2. Для избыточности $R(n, \Omega(\delta))$ универсального кодирования произвольного множества марковских источников $\Omega(\delta)$, удовлетворяющих условию $\theta_{\alpha i} > \delta > 0$, $a_\alpha \in A^s$, $i = \overline{1, k}$, выполняется неравенство, где λ не зависит от источника:

$$R(n, \Omega(\delta)) \leq \frac{h_{1/\sqrt{n}}(\Omega(\delta))}{n} + \frac{\lambda}{n}.$$

Доказательство. Справедливость утверждения вытекает из теоремы и следствия 1 при $\varepsilon = 1/\sqrt{n}$. $\square$

4. Заключение

В настоящей работе было доказано существование универсального кодирования, зависящего от массивности множества марковских источников с конечной памятью. Получены оценки избыточности кодирования в зависимости от мощности множества. В частности, если источников конечное число, то избыточность убывает со скоростью $\frac{C}{n}$, если же кодируется всё множество источников, то избыточность убывает со скоростью $(k^s(k-1)\log n)/n$, а при $s = 0$ получаем известный результат Кричевского.

Литература

1. Трофимов В. К. универсальное кодирование произвольного множества источников без памяти // Вестник СибГУТИ. 2018. № 4. С. 30–34.
2. Кричевский Р. Е. Длина блока, необходимая для получения заданной избыточности // Доклады АН СССР. 1965. Т. 171, № 1. С. 37–40.
3. Штарьков Ю. М. Кодирование сообщений конечной длины на выходе источника с неизвестной статистикой // Труды V конференции по теории кодирования и передачи информации, 1972. Кн. 1. С. 147–152.
4. Трофимов В. К. Избыточность универсального кодирования произвольных марковских источников // Проблемы передачи информации. 1974. Т. X, № 4. С. 16–24.

5. Галлагер Р. Г. Теория информации и надежная связь. М.: Советское радио, 1974. 719 с.
6. Шеннон К. Математическая теория связи. Работы по теории информации и кибернетике. 1963. С. 243–332.
7. Кричевский Р. Е. Связь между избыточностью кодирования и достоверностью сведений об источнике // Проблемы передачи информации. 1968. Т. 4, № 3. С. 48–57.
8. Krichevsky R. E., Trofimov V. K. The performance of universal encoding // IEEE Transactions on Information Theory. 1981. V. 27, № 2. P. 199–207.
9. Витрушкин А. Г. оценка сложности задачи табулирования. М.: Гос. изд. физ.-мат. лит. С. 195–228.

Трофимов Виктор Куприянович

доктор технических наук, профессор, профессор кафедры высшей математики СибГУТИ (630102, Новосибирск, ул.Кирова 86), e-mail: trofimov@sibguti.ru.

*Автор прочитал и одобрил окончательный вариант рукописи.
Автор заявляет об отсутствии конфликта интересов.*

Encoding of Messages Generated by an Arbitrary Markov Source with Unknown Message Statistics

Viktor K. Trofimov

Siberian State University of Telecommunications and Information Science (SibSUTIS)

Abstract: The method of universal coding of an arbitrary set of Markov sources with finite 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, Markov sources, entropy, storage and processing of information, the source of messages.

For citation: Trofimov V. K. Encoding of messages generated by an arbitrary Markov source with unknown message statistics (in Russian). *Vestnik SibGUTI*, 2023, vol. 17, no. 4, pp. 109–115. <https://doi.org/10.55648/1998-6920-2023-17-4-109-115>.



Content is available under the license
Creative Commons Attribution 4.0
License

© Trofimov V. K., 2023

The article was submitted: 20.05.2023;
accepted for publication 25.07.2023.

References

1. Trofimov V. K. Universal'noe kodirovanie proizvol'nogo mnozhestva istochnikov bez pamyati [universal memoryless encoding of an arbitrary set of sources]. *Vestnik SibGUTI*, 2018, no.4, pp.30-34.
2. Krichevskii R. E. Dlina bloka, neobkhodimaya dlya polucheniya zadannoi izbytochnosti [Universal memoryless encoding of an arbitrary set of sources]. *Doklady akademii nauk SSSR*, 1965, vol.171, no. 1, pp.37-40.

3. Shtar'kov Yu. M. Kodirovanie soobshchenii konechnoi dliny na vykhode istochnika s neizvestnoi statistikoi [Encoding messages of finite length at the output of a source with unknown statistics]. *V konferentsiya po teorii kodirovaniya i peredachi informatsii*, 1972, Moscow, book 1, pp. 147-152.
4. Trofimov V. K. Izbytochnost' universal'nogo kodirovaniya proizvol'nykh markovskikh istochnikov [Redundancy of universal coding of arbitrary Markov sources]. *Problemy peredachi informatsii*, 1974, vol. X, no. 4, pp.16-24.
5. Gallager R. G. *Teoriya informatsii i nadezhnaya svyaz'* [Information Theory and Reliable Communication]. Moscow, Sovetskoe radio, 1974. p.719.
6. Shennon K. Matematicheskaya teoriya svyazi [Mathematical communication theory]. *Raboty po teorii informatsii i kibernetike*, 1963, Moscow, pp.243-332.
7. Krichevskii R. E. Svyaz' mezhdru izbytochnost'yu kodirovaniya i dostovernost'yu svedenii ob istochnike [The Relationship Between Coding Redundancy and Source Credibility]. *Problemy peredachi informatsii*, 1968, vol. 4, no. 3, pp.48-57.
8. Krichevsky R. E., Trofimov V. K. The performance of universal encoding. *IEEE Transactions On Information Theory*, 1981, v. 27, no. 2, pp.199-207.
9. Vitrushkin A. G. *Otsenka slozhnosti zadachi tabulirovaniya* [Otsenka slozhnosti zadachi tabulirovaniya]. Moscow, Fizmatlit, pp.195-228.

Trofimov Viktor Kupriyanovich

Dr. of Sci. (Engineering), Professor, Professor of the Department of Higher Mathematics, Siberian State University of Telecommunications and Information Science (SibSUTIS, Russia, 630102, Novosibirsk, Kirov St. 86), e-mail: trofimov@sibguti.ru.