Кодирование сообщений, порожденных неизвестным Марковским источником, при различных длительностях передаваемых сигналов.

В.К.Трофимов, Т.В. Храмова

ГОУ ВПО «Сибирский государственный университет телекоммуникаций и информатики»

e-mail: trofimov@sibsutis.ru, e-mail: tvkhramova@gmail.com

Аннотация

Предложен метод оптимального универсального кодирования сообщений, порожденных марковским источником с бинарным алфавитом для каналов с различной стоимостью букв кодового алфавита. Получены асимптотически совпадающие оценки избыточности универсального кодирования сообщений порожденных неизвестным марковским источником. Для получения верхней оценки избыточности кодирования предложен конструктивный метод такого кодирования.

Ключевые слова: энтропия, стоимость кодирования, пропускная способность, избыточность.

Передача и сжатие информации – задачи, актуальные в наше время в любой сфере деятельности человека. При решении конкретной проблемы рассматриваются соответствующие ей типы источников, порождающих информацию, различные модели каналов передачи информации и виды кодирования (сжатия) информации. В качестве базовой модели канала передачи информации рассмотрим канал, при передаче по которому информация не подвергается искажению, т.е. канал без шума. При этом, источник информации может быть как статистически изученным (т.е. известным), так и неизвестным. Кодирование источника заключается в отображении слов, порожденных входным алфавитом в слова кодового алфавита.

Ключевыми понятиями, определяющими эффективность того или иного вида кодирования для различных источников, являются пропускная способность канала, стоимость кодирования и энтропия источника сообщений.

Канал связи с выходным (кодовым) алфавитом $Y = \{y_1, y_2, \cdots, y_m\}$, как доказано в [1] имеет пропускную способность

$$C = \log \omega_0,$$

где ω_0 - наибольший положительный корень уравнения $\omega^{-t_1} + \omega^{-t_2} + \ldots + \omega^{-t_k} = 1$. В частности, когда $t_1 = t_2 = \ldots = t_k = 1$ имеет место равенство $C = \log m$. Здесь и далее $\log x = \log_2 x$, $0\log 0 = 0$; M^N- множество всех последовательностей длины N и M^*- множество всех последовательностей конечной длины, состоящих из элементов множества M.

При равномерном по входу кодировании, последовательность букв, порожденных источником Θ, разбивается на слова длины N, которые, с помощью отображения φ, переводятся в слова выходного алфавита $Y=\{y_1,y_2,\cdots,y_m\}$, где каждый символ y_i, $i=\overline{1,m}$ имеет свою стоимость передачи или, другими словами, $t(y_i)=t_i$, $i=\overline{1,m}$. В результате кодирования, каждому блоку $w_i\in X^N$ ставится в соответствие некоторое кодовое слово $\varphi(w_i)\in Y^*$. При этом, длительность $t(\varphi(w_i))$ кодового слова $\varphi(w_i)$ определяется как сумма длительностей входящих в него символов. Если $\varphi(w_i)=y_{i_1}y_{i_2}\cdots y_{i_l}$, то $t(\varphi(w_i))=\sum_{j=1}^{l}t(y_{i_j})$, где $\varphi(w_i)=y_{i_1}y_{i_2}\cdots y_{i_l}$. Среднюю длительность кодирования φ для источника Θ, обозначим

$$\tilde{L}(N,\Theta,\varphi)=\sum_{w\in X^N}p(w)t(\varphi(w)).$$

Стоимость описанного выше блочного кодирования - это средняя длительность букв выходного алфавита, приходящаяся на одну букву входного алфавита, очевидно равна

$$L(N,\Theta,\varphi)=\tilde{L}(N,\Theta,\varphi)/N.$$

В данной работе, мы рассматриваем произвольный марковский источник Θ с памятью s входной алфавит которого бинарный, т.е. $X=\{0;1\}$. Вероятность появления некоторого слова $w=\underbrace{x_1x_2\cdots x_s}_{w^0}x_{s+1}\cdots x_N$ для марковского источника определяется вектором вероятностей начальных блоков $p_w^0=p_\Theta(w^0)$ и матрицей $(p_{vi})_{v\in X^s;i=0,1}$ где $p_{vi}=p_\Theta(x_i|v)$ условная вероятность появления буквы x_i после блока v. Обозначим $n_0=n_0(w)$ и $n_1=n_1(w)$ - число вхождений букв 0 и 1 в блок $w=\underbrace{x_1x_2\cdots x_s}_{w^0}x_{s+1}\cdots x_N$, $n_{vi}=n_{vi}(w)$ - число вхождений буквы x_i после блока v и $n_v=n_v(w)=\sum_{x_i\in X}n_{vi}$. Тогда вероятность появления слова w может быть вычислена по любой из формул:

$$p_\Theta^s(w)=p_w^0\prod_{i=s+1}^{N}p_\Theta(x_i|x_{i-s}x_{i-s+1}\cdots x_{i-1}),$$

$$p_\Theta^s(w)=p_w^0\prod_{v\in X^s}\left[p_{v0}^{n_{v0}}\cdot p_{v1}^{n_{v1}}\right],\qquad p_\Theta^s(w)=p_w^0\prod_{v\in X^s}\left[p_{v0}^{n_{v0}}\cdot(1-p_{v0})^{n_{v1}}\right].$$

Как известно [2,3], энтропия двоичного марковского источника с памятью s вычисляется по формуле

$$H_s(\Theta)=-\sum_{v\in X^s}p_v^0\left(p_{v0}\log p_{v0}+p_{v1}\log p_{v1}\right).$$

Величину $R^t(N,\Theta,\varphi)$, связывающую энтропию, стоимость и пропускную способность, назовем *избыточностью* кодирования φ источника Θ, и, по определению,

$$R^t(N,\Theta,\varphi) = L(N,\Theta,\varphi) - \frac{H(\Theta)}{C} \ .$$

Чем меньше избыточность, тем эффективнее кодирование. Величину

$$R^t_\Omega(N) = \inf_{\varphi} \sup_{\Theta\in\Omega} R^t(N,\Theta,\varphi)$$

назовем избыточностью универсального кодирования множества источников Ω. В работах [1,2,3,7,8] изучалось поведение избыточности кодирования для известных источников. Оценки избыточности универсального кодирования при одинаковых длительностях букв выходного алфавита получены в [4,5,6], а для различных в [9,10]. Цель настоящей работы – изучить поведение величины $R^t_{\Omega_s}(N)$ (Ω_s – двоичный марковский источник с памятью s) при различных длительностях букв выходного алфавита, для неизвестного источника. Основной результат работы можно сформулировать следующим образом.

Теорема. Для избыточности равномерного по входу кодирования φ некоторого двоичного марковского источника Ω_s в слова выходного алфавита с различными длительностями кодовых букв и, определяемой этим алфавитом, пропускной способностью канала передачи информации C имеют место асимптотические неравенства:

$$\frac{\lambda + 2^{s-1}\log(N)}{CN} \le R^t_{\Omega_s}(N) < \frac{T(s) + 2^{s-1}\log(N-s)}{CN} + \frac{t^{**}}{N},$$

где λ *— константа, не зависящая от* Θ и N, $T(s)$ *— константа, определяемая алфавитом источника,* t^{**} *— константа, зависящая только от канала.*

ЛИТЕРАТУРА

[1]. Шеннон К. Математическая теория связи. Работы по теории информации и кибернетике. – 1969. – Ил., М. – С.243 - 332.

[2]. Фано Р. Передача информации. Статистическая теория связи.// «Мир», М.: 1965.

[3]. Галлагер Р.Г. Теория информации и надежная связь. М.: Советское радио, 1974.

[4]. Фитингоф Б. М. Оптимальное кодирование при неизвестной и меняющейся статистике сообщений.// Проблемы передачи информации. 1966 т.2, №2. с. 3–11.

[5]. Кричевский Р.Е. Связь между избыточностью кодирования и достоверностью сведений об источнике // Проблемы передачи информации. 1968.- Т.4. № 3. – С.48-57.

[6]. Krichevsky R. E., Trofimov V. K. The performance of universal encoding // IEEE Transactions on Information Theory. 1981. - V. 27, № 2. - P. 199-207.

[7]. Кричевский Р.Е. Длина блока, необходимая для получения заданной избыточности. - ДАН СССР, 1965, т.171, № I, с. 37-40.

[8]. Чисар И. О каналах без шума. // Пробл. передачи информ. 1970.- Т.6. № 4. – С.3-15.

[9]. Храмова Т.В. Кодирование сообщений, порожденных неизвестным источником, при различных длительностях передаваемых сигналов.// Всероссийская научная конференция «Научное и техническое обеспечение исследований и освоения шельфа северного ледовитого океана», сборник трудов, Новосибирск. - 2010., с.60-62.

[10]. Трофимов В.К., Храмова Т.В. «Универсальное кодирование сообщений, порожденных неизвестным бернуллиевским источником буквами алфавита с различными длительностями» // Материалы российской научно-технической конференции «Информатика и проблемы телекоммуникаций» т.1. Новосибирск.- 2011., с.165-167.