Нумерационное кодирование источников Маркова

В.А. Монарёв
ИВТ СО РАН
e-mail: viktor.monarev@gmail.com

Предложен новый метод нумерации последовательностей, порожденных источником Маркова. Нумеруются равновероятные последовательности фиксированной длины. Предполагается, что известна память источника, но не известны переходные вероятности. Метод также применим в идеальной стеганографической системе, описанной в работе Рябко Б.Я. и Рябко Д.Б.

1. Введение

Задача нумерации последовательностей, порожденных случайным источником известна давно. В работах [1],[2],[3] рассматривалась проблема нумерации последовательностей порожденных бернуллиевским источником. Лучший результат в этой области имеет практически линейную зависимость количества операций и объема памяти от длины последовательности. В работах [2], [3] был рассмотрен вариант марковского источника с памятью равной одному над алфавитом $\{0,1\}$. Общий случай марковского источника над конечным алфавитом был рассмотрен в [2], но сложность метода растет экспоненциально с ростом алфавита или памяти. В данной работе предлагается новый метод частичной нумерации последовательностей порожденных источником Маркова произвольного порядка над конечным алфавитом, скорость которого совпадает со скоростью нумерации источника без памяти. Несмотря на то, что новый подход не дает полного решения проблемы тем не менее он может быть полезен для решения некоторых задач криптографии и стеганографии.

2. Постановка задачи и результат

Введем необходимые обозначения. Пусть A – алфавит, тогда обозначим, через A^k множество всех слов длины k и через A^∞ множество всех слов произвольной длины. Если $x \in A^\infty$, то через $|x|$ обозначим длину слова x.

Рассмотрим источник источник Маркова M порядка t над алфавитом $A = \{a_1, ..., a_n\}$. Обозначим переходные вероятности:

$$P(x_i = a_i^{'} | x_{i-1} = a_{i-1}^{'}, \ldots, x_{i-t} = a_{i-t}^{'}) = p_{a_i^{'},\ldots,a_{i-t}^{'}}, \quad (1)$$

где $a_i^{'} \in A$, для всех i.

Пусть дана последовательность (слово) $X = x_0, x_1, x_2, \ldots, x_N$ порожденная источником M порядка t над алфавитом A. Обозначим, через $\bar{E} = \{E_1, ...E_m\}, E_i \in A^t$ множество состояний источника. Каждое состояние E_i соответствует последовательности из t букв и задает распределение следующего символа (или распределение вероятности следующего состояния). Ясно, что $m \leq n^t$. Поясним на примере.

Пример 1. Источник с памятью 2 над алфавитом $A = \{0, 1\}$ имеет состояния $\{E_1, E_2, E_3, E_4\}$, где E_1 соответствует слову 00, $E_2 = 01$ и т.д. Допустим этот источник породил последовательность 0011010101. Тогда будем говорить, что начальное состояние равно E_1 или 00, далее источник изменил состояние на $E_2 = 01$ и породил символ 1 и т.д. Если выписать последовательность изменения состояний источника то получим последовательность $E_1, E_2, E_4, E_3, E_2, E_3, E_2$....

Введем следующие обозначения: b – начальное состояние и c – конечное состояние. Обозначим через P_E – множество состояний из которых можно перейти в состояние E за один шаг. Элементы множества P_E будем называть входами в состояние E. В примере 1, $P_{E_2} = \{E_1, E_3\}$.

Замечание 2.1. *$E \in P_E \Leftrightarrow$ если состояние E это однобуквенное слово.*

Пусть X_E – последовательность букв, которые породил источник когда находился в состоянии E. Считаем, что X_E упорядочена в том порядке в каком появлялись буквы. Если рассмотреть Пример 1, то: $b = E_1 = 00$, $X_{E_1} = 1$, $X_{E_2} = 100$, $X_{E_3} = 111$, $X_{E_4} = 0$ и $c = E_2 = 01$.

Замечание 2.2. *а) Если даны все последовательности X_E и начальное состояние, то можно восстановить всю последовательность X.*
б) Если у двух последовательностей X и Y, порожденных марковским источником порядка t, совпадают начальные состояния и для любого состояния E последовательности X_E и Y_E отличаются только порядком букв (совпадают по длине и составу букв), то вероятности появления X и Y равны.

Пусть $\nu_x(a)$ – частота буквы a в слове x и $l(x)$ – последняя буква слова x. Обозначим через (*) условие, что состояние E равно b и не равно c и через (**) условие, что E равно c и не равно b.

Для произвольного неоднобуквенного состояния E, выполнены условия:
1) Если не выполнено (*) и (**) или выполнены оба условия, то:

$$\sum_{E' \in P_E} \nu_{E'}(l(E)) = |X_E| \tag{2}$$

2) Если выполнено (*) и не выполнено (**), то:

$$\sum_{E' \in P_E} \nu_{E'}(l(E)) + 1 = |X_E| \tag{3}$$

3) Если выполнено (**) и не выполнено (*), то:

$$\sum_{E' \in P_E} \nu_{E'}(l(E)) - 1 = |X_E| \tag{4}$$

Для произвольного однобуквенного состояния E (обозначим эту букву через e), выполнены условия:
4) Если не выполнено (*) и (**) или выполнены оба условия, то:

$$\sum_{E' \in P_E, E' \neq E} \nu_{E'}(l(E)) = |X_E| - \nu_{X_E}(e) \tag{5}$$

5) Если выполнено (∗) и не выполнено (∗∗), то:

$$\sum_{E' \in P_E, E' \neq E} \nu_{E'}(l(E)) + 1 = |X_E| - \nu_{X_E}(e) \tag{6}$$

6) Если выполнено (∗∗) и не выполнено (∗), то:

$$\sum_{E' \in P_E} \nu_{E'}(l(E)) - 1 = |X_E| - \nu_{X_E}(e) \tag{7}$$

7) Если E не равно c, то $l(X_E) \neq e$

Утверждение 2.1. *Если заданы последовательности X_E и состояния b и c, то выполнение условий 1)-7) для любого состояния E необходимо для того, чтобы существовала X, но не достаточно.*

Доказательство. Если состояние не является конечным и не является начальным, то количество входов в это состояние должно быть равно количеству выходов из него. Это обеспечивает равенства (1) и (4) (в случае однобуквенного состояния). Если же выполнено условие (*), то количество попаданий из других состояний меньше на единицу количества выходов из состояния (см. равенства (2) и (5)). И наконец, если состояние является и конечным и начальным, то количество входов в состояние равно количеству выходов из других состояний плюс один вход из начального состояния. С другой стороны, общее число входов в состояние должно быть на единицу больше чем количество выходов (так как состояние конечное). Это обеспечивают условия 1) и 3). Кроме того, если состояние E однобуквенное и не равно c, то необходимым условием выхода из состояния является то, что соответствующая последовательность X_E не должна оканчиваться на ту букву из которой состоит слово E(см. условие 5)). Так как если оно будет оканчиваться на e, то выполнение 1)-7) влечет то, что последние символы X_E (которые равны e) не будут участвовать в восстановлении последовательности X.

Приведем пример, когда условия выполнены, но последовательность невозможно восстановить.

Пример 2. Память источника равна 2. Источник над алфавитом $A = \{0, 1\}$. Заданы последовательности, начальное состояние $b = 11$ и конечное $c = 10$

$$X_{E_1} = \{1\}; X_{E_2} = \{0, 1, 1\}; X_{E_3} = \{0, 1, 1\}; X_{E_4} = \{0, 0, 0, 1\}.$$

Восстанавливая последовательность X получим 110010110110. Последний символ последовательности X_{E_4} не был использован, то есть для данных последовательностей не существует исходной последовательности X. Следовательно выполнение условий 1)-7) недостаточно. Доказательство окончено.

Возьмем произвольное множество последовательностей $\{X_{E_{i_1}'}, ..., X_{E_{i_m}'}\}$, где $0 < m \leq n^t$ для которых выполнены условия 1)-7) и пусть задано начальное состояние b. Найдем условия необходимые и достаточные, для того чтобы последовательность X могла бы быть восстановлена по $\{X_{E_{i_1}'}, ..., X_{E_{i_m}'}\}$. Условия 1)-7) и задание состояния b влечет за собой однозначное определение c. Если будем восстанавливать последовательность X по этому множеству, то возможны два варианта. Первый, последовательность будет успешно восстановлена, то есть все символы последовательностей $\{X_{E_{i_1}'}, ..., X_{E_{i_m}'}\}$ были использованы, второй, восстановление последовательности было прервано на некотором состоянии (обозначим его через E') и часть некоторых символов последовательностей не была использована в восстановлении (аналогично с примером 2).

Замечание 2.3. *Если для некоторого конечного множества последовательностей* $\{X_{E'_{i_1}}, ..., X_{E'_{i_m}}\}$, *где* $0 < m \leq n^t$ *выполнены условия 1)-7) и задано начальное состояние* b, *то после восстановления последовательности* X *(возможно неуспешного, когда были использованы не все символы) состояние с равно* E' *и последовательность* X_c *будет использована полностью.*

Доказательство. Докажем от противного. Действительно, если восстановление последовательности закончилось в некотором состоянии E'(для однобуквенного доказывается аналогично) значит в последовательности $X_{E'}$ кончились символы, но этого не может быть поскольку количество переходов в состояние E'(если оно не равно c) равно количеству букв в последовательности X'_E(или количеству выходов из состояния). Таким образом, получим противоречие одному из условий 1)-7). Доказательство окончено.

Замечание 2.4. *Если* $E \in P_c$ *и последовательность* X_E *оканчивается на* $l(c)$, *то последовательность* X_E *использовалась полностью при восстановлении.*

Доказательство. Очевидно, что если X_c использовалась полностью (это следует из замечания 2.3), то не может быть пропущено ни одного символа $l(c) \in X_E$, где $E \in P_E$, поскольку только с помощью них можно перейти в состояние c. Следовательно, если на конце последовательности X_E, $E \in P_c$ стоит символ $l(c)$, то он использовался в восстановлении, следовательно последовательность X_E использовалась в восстановлении полностью. Доказательство окончено.

Пусть $I = \{i'_1, ..., i'_m\}$ множество индексов всех возможных состояний. Обозначим через $I_0 = \{i'_j\}$, где $E_{i'_j} = c$ и через $I_1 \subset I = \{i'_1, ..., i'_m\}$ подмножество такое, что $i \in I_1 \Longleftrightarrow l(X_{E_i}) = l(c)$, и $E_i \in P_c$.

Аналогично по рекурсии определим множества индексов $I_2, I_3...$ следующим образом. Для I_n выполняются следующие условия:

1) $I_n \subset I_0 / \bigcup_{k=0}^{n-1} I_k$

2) $i \in I_n \Longleftrightarrow \exists k \in I_{n-1}$ такой что $E_i \in P_{E_k}$ и $l(X_{E_i}) = l(E_k)$.

Обозначим через m' индекс последнего непустого множества $I_{m'}$.

Утверждение 2.2. *Последовательность* X *можно восстановить по последовательностям* $\{X_{E'_{i_1}}, ..., X_{E'_{i_m}}\}$, *где* $0 < m \leq n$ *и заданному начальному состоянию* b *тогда и только тогда если выполнены условия 1)-7) и* $\bigcup_{k=0}^{m'} I_k = I$.

Доказательство. Из замечания 2.4 следует, что все последовательности X_{E_i}, где $i \in I_1$ используются полностью в восстановлении последовательности. По индукции докажем, что если $i \in \bigcup_{k=0}^{n+1} I_k$, то последовательность X_{E_i} использовалась при восстановлении полностью. Предположим, что верно для n докажем для $n+1$. Действительно, если $i \in I_n$, то выполнено условие 2), следовательно существует состояние E_k у которого последовательность X_{E_k} использовалась в восстановлении полностью, следовательно последний символ последовательности X_{E_i} участвовал в восстановлении (так как при его обработке мы переходим в состояние E_k). Можно сделать вывод, что последовательность X_{E_i} использовалась полностью. Таким образом, получаем что условий 1)-7) и $\bigcup_{k=0}^{m'} I_k = I_0$ достаточно, чтобы все последовательности использовались полностью.

Так как необходимость условий 1)-7) была доказана (см. утверждение 2.1), остается доказать необходимость условия $\bigcup_{k=0}^{m'} I_k = I$. Предположим, что последовательно можно восстановить полностью и существует такое непустое множество U отличное от I, что $\bigcup_{k=0}^{m'} I_k = I/U$. Из определения множества U следует, что для состояния E_i, $i \in I/U$ и E_i, где $i \in U$ несвязанные, а это противоречие с условием. Доказательство окончено.

Положим, что множество последовательностей $\{X_{E_{i_1'}}, ..., X_{E_{i_m'}}\}$, соответствуют некоторой последовательности X. Поясним, как занумеровать X. Заметим, что каждая последовательность $X_{E_{i_j'}}$ является последовательностью, порожденной бернуллиевским источником. Занумеруем последовательность $X_{E_{i_j'}}$ без последнего члена с помощью нумерации описанной в [1]. Таким образом, мы получим m номеров $\{n_1, ..., n_m\}$. Из утверждения 2 следует, что если фиксировать последний член каждой последовательности $X_{E_{i_j'}}$ и остальные члены последовательностей как угодно переставить, то полученные последовательности можно восстановить в некоторую последовательность, которая будет равновероятной с исходной (так как при перестановки сохраняются условия 1)-7)). Таким образом, каждому набору из m номеров (номера из соответсвующего диапозона значений) соответсвует некоторая последовательность порожденная марковским источником. m номеров очевидным образом можно занумеровать одним номером. Требуемая нумерация получена.

Список литературы

[1] Ryabko B.Ya. The fast enumeration of combinatorial objects // Discrete Math.and Applications. 1998 v.10, n2.

[2] T. M. Cover, Enumerative Source Encoding // IEEE Trans. Info. Theory IT-19. 1973 pp. 73-77.

[3] Cleary, J. G. and Witten, I. H. A comparison of enumerative and adaptive codes // IEEE Trans. Inform. Theory. IT-30. 1984 pp. 306–315.

[4] B. Ryabko, D.Ryabko. Asymptotically Optimal Perfect Steganographic Systems // Problems of Information Transmission, 2009, Vol. 45, No. 2, pp. 184-190.