ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 27 МАТЕМАТИКИ Выпуск 1
2020

В. О. Миронкин, **М. М. Михайлов** (Москва, Национальный исследовательский университет «Высшая школа экономики», Лаборатория ТВП). **Об энтропии последовательной процедуры опробования дискретной вероятностной схемы.**

УДК 519.722

Резюме: В работе применен теоретико-информационный подход к исследованию процедуры последовательного опробования элементов дискретной вероятностной схемы до наступления «успеха». Выписаны точные и оценочные выражения для математического ожидания и дисперсии энтропии Шеннона процедуры последовательного опробования.

Ключевые слова: дискретная вероятностная схема, энтропия Шеннона, количество информации, последовательное опробование.

Введение. Задачи, связанные с исследованием алгоритмов опробования элементов произвольных дискретных множеств, возникают в ряде практических приложений криптографической защиты информации, например, при синтезе и анализе парольных систем, алгоритмов хеширования [1] и т. д. Характеристики подобных алгоритмов существенно зависят от структуры опробуемых множеств, заданных на них вероятностных распределений, а также от используемых моделей опробования (выбора с возвращением или без него [2]). Так, в частности, в [3, 4] вычислены математическое ожидание и оценки объема работы алгоритма последовательного опробования ключевой информации до наступления «успеха», а также его модификации (алгоритма усеченного опробования) для различных распределений на заданном множестве.

В настоящей работе для процедуры последовательного опробования элементов произвольного дискретного множества, с заданным на нем равновероятным распределением, используется теоретико-информационный подход. Исследуемой характеристикой при этом является выраженное в битах значение энтропии, достаточное для наступления соответствующего «успеха».

1. Теоретико-вероятностная модель. Для произвольного $n \in \mathbb{N}$ рассмотрим вероятностное пространство $(\Omega, \mathcal{F}, \mathbf{P})$, где пространство элементарных исходов Ω — произвольное множество $\mathcal{S}_n = \{a_1, \ldots, a_n\}$ мощности $n \in \mathbb{N}$, алгебра событий $\mathcal{F}$ — множество всех подмножеств Ω, а вероятностная мера $\mathbf{P}$ задана следующим образом:

$$\mathbf{P}\{\omega\} = \frac{1}{n} \quad \forall\, \omega \in \Omega. \tag{1}$$

Пусть далее последовательно, в порядке возрастания индексов элементов множества $\mathcal{S}_n$, выполняется процедура, состоящая в проверке условия $f(\omega) = 1$ для некоторого заранее заданного индикаторного отображения $f\colon \mathcal{S}_n \to \{0,1\}$:

$$f(\omega) = \begin{cases} 1, & \omega = \omega_0, \\ 0, & \omega \neq \omega_0, \end{cases}$$

где $\omega_0 \in \mathcal{S}_n$, выполнение которого будем называть «успехом» (см., например, монографию А. Файнстейна [5]):

Алгоритм
Вход: $\mathcal{S}_n = \{a_1, \ldots, a_n\}$.

1. $i = 1$;
2. Если $f(a_i) = 1$, то полагаем $\omega_0 = a_i$, и алгоритм завершает работу; в противном случае полагаем $i = i + 1$ и переходим на шаг 2.

Выход: ω_0.

С учетом (1) исходными данными для проведения процедуры опробования является вероятностная схема [6]

$$\mathcal{A}_1 = \begin{pmatrix} a_1 & a_2 & \ldots & a_n \\ \dfrac{1}{n} & \dfrac{1}{n} & \ldots & \dfrac{1}{n} \end{pmatrix}.$$

Через H_k обозначим среднее количество информации, достаточное для определения искомого элемента ω_0 на k-м шаге указанной процедуры, $k \in \{1, \ldots, n\}$.

Согласно [5] имеем

$$H_1 = H\left(\frac{1}{n}, 1 - \frac{1}{n}\right). \tag{2}$$

где $H\left(\frac{1}{n}, 1 - \frac{1}{n}\right)$ — энтропия Шеннона случайной величины с распределением $\mathcal{B}i\left(1, \frac{1}{n}\right)$ (см. [7]).

Преобразуем (2) с помощью IV аксиомы Хинчина [6]. Тогда с учетом равенства $H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) = \log_2 n$ (здесь и далее количество информации измеряется в битах) получаем

$$H_1 = H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right) - \left(1 - \frac{1}{n}\right) H\left(\frac{1}{n-1}, \ldots, \frac{1}{n-1}\right)$$
$$= \log_2 n - \left(1 - \frac{1}{n}\right) \log_2 (n-1).$$

Если $f(a_1) \neq 1$, а это событие происходит с вероятностью $1 - \frac{1}{n}$, то, как было указано выше, процесс опробования продолжается. При этом распределение на оставшемся множестве опробуемых элементов пересчитывается с использованием формулы Байеса [2]:

$$\mathbf{P}\left\{f(a_i) = 1 \,\middle|\, f(a_1) \neq 1\right\}$$
$$= \frac{\mathbf{P}\left\{f(a_1) \neq 1 \,\middle|\, f(a_i) = 1\right\} \mathbf{P}\{f(a_i) = 1\}}{\mathbf{P}\{f(a_1) \neq 1\}} = \frac{1}{n-1}, \qquad i = 2, \ldots, n,$$

формируя новую вероятностную схему

$$\mathcal{A}_2 = \begin{pmatrix} a_2 & a_3 & \ldots & a_n \\ \dfrac{1}{n-1} & \dfrac{1}{n-1} & \ldots & \dfrac{1}{n-1} \end{pmatrix},$$

которая определяет промежуточные данные для дальнейшего проведения процедуры опробования. При этом с учетом первого шага выполняется цепочка соотношений

$$H_2 = H\left(\frac{1}{n}, 1 - \frac{1}{n}\right) + \left(1 - \frac{1}{n}\right) H\left(\frac{1}{n-1}, 1 - \frac{1}{n-1}\right) = H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$
$$- \left(1 - \frac{1}{n}\right) H\left(\frac{1}{n-1}, \ldots, \frac{1}{n-1}\right) + \left(1 - \frac{1}{n}\right) H\left(\frac{1}{n-1}, \ldots, \frac{1}{n-1}\right)$$
$$- \left(1 - \frac{1}{n}\right)\left(1 - \frac{1}{n-1}\right) H\left(\frac{1}{n-2}, \ldots, \frac{1}{n-2}\right) = H\left(\frac{1}{n}, \ldots, \frac{1}{n}\right)$$
$$- \left(1 - \frac{2}{n}\right) H\left(\frac{1}{n-2}, \ldots, \frac{1}{n-2}\right) = \log_2 n - \left(1 - \frac{2}{n}\right) \log_2 (n-2).$$

Рассуждая далее аналогично, для произвольного $k \in \{1, \ldots, n-1\}$ получаем

$$H_k = \log_2 n - \left(1 - \frac{k}{n}\right) \log_2 (n-k), \quad H_n = \log_2 n.$$

При этом, $H_i < H_j$ для произвольных $1 \leqslant i < j \leqslant n$.

Через C_k обозначим событие, заключающееся в совпадении среднего количества информации, достаточного для определения искомого элемента ω_0 с помощью указанной процедуры, с H_k, где $k \in \{1, \ldots, n\}$.

Очевидным образом выполняются равенства

$$\mathbf{P}\{C_k\} = \left(1 - \frac{1}{n}\right) \cdot \ldots \cdot \left(1 - \frac{1}{n-k+2}\right) \frac{1}{n-k+1} = \frac{1}{n}, \quad \mathbf{P}\{C_n\} = \frac{1}{n}.$$

Тогда математическое ожидание случайной величины μ с распределением

$$\mu \sim \begin{pmatrix} H_1 & H_2 & \ldots & H_n \\ \frac{1}{n} & \frac{1}{n} & \ldots & \frac{1}{n} \end{pmatrix}$$

представляет собой среднее количество информации, достаточное для определения искомого элемента ω_0, принадлежащего множеству $\mathcal{S}_n$, с заданным на нем распределением (1), с использованием указанной выше процедуры.

Теорема. *Пусть на множестве* $\mathcal{S}_n$ *задано распределение* (1). *Тогда*

$$\mathbf{E}\mu = \log_2 \left(n \prod_{k=2}^{n-1} k^{-\frac{k}{n^2}} \right),$$

$$\log_4 n + \frac{1 + \frac{8\ln 2 - 4}{n^2}}{4 \ln 2} < \mathbf{E}\mu < \log_2 n - \left(1 - \frac{1}{n}\right)^2 \log_4 (n-1) + \frac{1 - \frac{2}{n}}{4 \ln 2}.$$

Если при этом $n \to \infty$

$$\mathbf{E}\mu \sim \log_4 n + \frac{1}{4 \ln 2}.$$

Следствие. *Пусть на множестве* $\mathcal{S}_n$ *задано распределение* (1). *Тогда*

$$\mathbf{D}\mu = \frac{1}{n} \sum_{k=0}^{n-1} \log_2^2 \left(n k^{-\frac{k}{n}} \right) - \log_2^2 \left(n \prod_{k=1}^{n-1} k^{-\frac{k}{n^2}} \right),$$

$$\mathbf{D}\mu > \frac{\log_2^2 n}{3} + \frac{5 \log_2 n}{18 \ln 2} + \frac{2}{27 \ln^2 2} + \frac{2 \log_2 n}{n^2} \left(2 - \frac{1}{\ln 2}\right) - \frac{8}{3n^3} \left(1 - \frac{2}{3 \ln 2} + \frac{2}{9 \ln^2 2}\right) - \left(\log_2 n - \frac{(n-1)^2 \log_2 (n-1)}{2n^2} + \frac{n-2}{4n \ln 2}\right)^2,$$

$$\mathbf{D}\mu < \log_2^2 n - \left(1 - \frac{1}{n}\right)^2 \log_2 n \log_2 (n-1) + \left(1 - \frac{1}{n}\right)^3 \frac{\log_2^2 (n-1)}{3} + \left(1 - \frac{2}{n}\right) \frac{\log_2 n}{2 \ln 2} - \left(1 - \frac{1}{n}\right)^3 \frac{2 \log_2 (n-1)}{9 \ln 2} + \frac{2(n-1)^3 - 2}{27 n^3 \ln^2 2} - \left(\frac{\log_2 n}{2} + \frac{n^2 + 8 \ln 2 - 4}{4n^2 \ln 2}\right)^2.$$

В таблице приведены приближенные значения $\mathbf{E}\mu$ и $\mathbf{D}\mu$ в зависимости от значения параметра n.

Таблица.

n	2^{16}	2^{32}	2^{64}	2^{128}	2^{256}	2^{512}	2^{1024}
$\mathbf{E}\mu$	$8,36$	$16,36$	$32,36$	$64,36$	128,36	256,36	512,36
$\mathbf{D}\mu$	22	$86,64$	$343,92$	$1370,49$	$5471,62$	$21865,9$	$87422,4$

СПИСОК ЛИТЕРАТУРЫ

1. *Лось А. Б.*, *Нестеренко А. Ю.*, *Рожков М. И.* Криптографические методы защиты информации: учебник для академического бакалавриата М.: Издательство Юрайт 2018, 473 с.
2. *Феллер В.* Введение в теорию вероятностей./ Пер. с англ. Ю. В. Прохорова М.: Мир, 1984, 528 с.
3. *Арбеков И. М.* Критерии секретности ключа. — Математические вопросы криптографии, 2016, т. 7, № 1, с. 39–56.
4. *Arbekov I. M.* Lower bounds for the practical secrecy of a key. — Математические вопросы криптографии, 2017, т. 8, № 2, с. 29–38.
5. *Файнстейн А.* Основы теории информации М.: Мир, 1960, 138 с.
6. *Духин А. А.* Теория информации: Учебное пособие М.: Гелиос АРВ, 2007, 248 с.
7. *Чечета С. И.* Введение в дискретную теорию информации и кодирования: учебное издание. М.: МЦНМО, 2011, 224 с.

UDC 519.722

Mironkin V. O., ***Mikhailov M. M.*** (Moscow, National Research University Higher School of Economics, TVP Laboratory). **On the entropy of a sequential procedure for testing elements of discrete probabilistic scheme**

Abstract: Information theoretic approach to solving the problem of assessing the complexity of the procedure for sequential testing of elements of a discrete probability scheme before the occurrence of “success” is applied. Exact and evaluative expressions for the mathematical expectation and variance of Shannon entropy of the sequential testing procedure are obtained. .

Keywords:discrete probability scheme, Shannon entropy, information content, sequential testing.