ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

УДК 519.244.2, 519.244.8
DOI: 10.18101/2304-5728-2018-4-3-15

О КРИТЕРИИ ПРОВЕРКИ ВЛОЖЕНИЯ С ДОПУСКОМ ДЛЯ ДИСКРЕТНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ

© Меженная Наталья Михайловна
кандидат физико-математических наук, доцент,
Московский государственный технический университет им. Н. Э. Баумана
Россия, 105005, г . Москва, ул. 2-я Бауманская, 5
E-mail: Natalia.mezhennaya@gmail.com

Последовательность X является подпоследовательностью с допуском d последовательности Y, если X получается из Y удалением несмежных отрезков не более чем из d знаков. В этом случае говорят, что X может быть вложена в Y с допуском d. В работе предложен последовательный критерий проверки гипотезы о вложении с допуском d для дискретных случайных последовательностей над конечным алфавитом и изучены его свойства. Вероятность ошибки первого рода (вероятность отклонения верной гипотезы о вложении с допуском) построенного критерия равна нулю. Получено выражение для вероятности ошибки второго рода при альтернативной гипотезе о том, что рассматриваемые дискретные последовательности образованы независимыми в совокупности случайными величинами с равномерными распределениями на конечном алфавите. Найдено значение среднего числа знаков вкладываемой последовательности, используемых критерием до принятия решения при альтернативной гипотезе. Трудоемкость предложенной процедуры при верной гипотезе о плотном вложении пропорциональна длине вкладываемой последовательности или меньше при альтернативной гипотезе, что по порядку намного меньше трудоемкости тотального опробования. Приведены численные значения вероятности ошибки второго рода и среднего количества используемых знаков при различных значениях d и размерах алфавита.

Ключевые слова: плотное вложение; вложение с допуском; последовательный критерий; гипотеза о независимости; вероятности ошибок первого и второго рода; дискретная случайная последовательность.

Введение

Пусть $X_n = (x_1,...,x_n)$ и $Y_m = (y_1,\ldots,y_m)$ — последовательности элементов множества $A_N = \{0,...,N-1\}$, $N \geqslant 2$, длин n и m соответственно. Согласно [1], последовательность X_n может быть *плотно вложена* в последовательность Y_m, если существуют такие натуральные числа $1 \leqslant j_1 < j_2 < \ldots < j_n \leqslant m$, $j_{k+1} - j_k \in \{1,2\}$, $k = 1,\ldots,n-1$, что $x_k = y_{j_k}$,

$k = 1,\ldots,n.$ В этом случае X_n является *плотной подпоследовательностью* Y_m. Если $j_1 = 1,$ то будем говорить, что X_n может быть плотно вложена *в начало* Y_m.

Впервые задача о плотном вложении одной дискретной последовательности в другую рассмотрена в [1]. Найдена верхняя оценка для вероятности того, что заданная двоичная последовательность может быть плотно вложена в последовательность независимых двоичных случайных величин с равномерными распределениями.

В работе [2] результаты, полученные в [1], обобщены на последовательности со значениями в произвольном конечном алфавите. В [2] получены неулучшаемые нижняя и верхняя оценки для вероятности плотного вложения и указаны классы последовательностей, на которых они достигаются.

Будем говорить, что $X_n = (x_1,...,x_n)$ может быть *вложена* в последовательность $Y_m = (y_1,\ldots,y_m)$ *с допуском* $d \geqslant 1$, если существуют такие натуральные числа

$$1 \leqslant j_1 < j_2 < \ldots < j_n \leqslant m,\ j_{k+1} - j_k \in \{1,2,\ldots,d+1\},\ k = 1,\ldots,n-1, \quad (1)$$

что $x_k = y_{j_k},\quad k = 1,\ldots,n.$ В этом случае X_n является подпоследовательностью Y_m с допуском d. Если $j_1 = 1,$ то будем говорить, что X_n может быть вложена с допуском $d \geqslant 1$ *в начало* Y_m.

При $d = 1$ определения плотного вложения и вложения с допуском совпадают.

Длиной вложения называется $j_n - j_1 + 1$ — длина отрезка последовательности $Y_m,$ в который вкладывается X_n.

Понятие вложения с допуском введено в [3]. Задача о вложениях двоичных последовательностей и ее практическое применение рассмотрены в [4]. Отмечена связь этих задач со свойствами поточных шифров [5]. Поточные генераторы используют функциональное преобразование, которое на основе короткого ключа строит шифрующую последовательность, производя быстрое шифрование сообщения при помощи поразрядного исключающего или между элементами сообщения и элементами шифрующей последовательности. Несколько примеров таких генераторов и их свойства описаны в работах [6–8]. Известны различные усовершенствования генерирующего алгоритма, чтобы получить последовательности с более высокой степенью защиты [9]. Вероятностные свойства выходных последовательностей в разных математических моделях рассматривались в [10–12].

Задача о вложении последовательностей также связана с задачами о поиске цепочек с заданным свойством в дискретных случайных последовательностях [13]. Такие задачи возникают в самых различных областях: при анализе генома [14], машинном обучении, распознавании образов [15] и так далее.

1. Процедура проверки с апробацией всех вариантов вложения с допуском

Рассмотрим задачу о проверке гипотезы H_{0n} о том, что X_n является подпоследовательностью с допуском d последовательности Y_m независимых равномерно распределенных на множестве A_N случайных величин. Будем считать, что $m \geqslant 1+(d+1)(n-1)$.

Самый простой способ проверки гипотезы H_{0n} состоит в том, чтобы опробовать все $(d+1)^{n-1}$ наборов $(j_1, j_2, \ldots, j_n)$, удовлетворяющих (1) мест вложения с допуском d последовательности X_n в начало последовательности Y_m. При этом вероятность отклонить верную гипотезу H_{0n} равна нулю. К очевидным недостаткам такой процедуры естественно отнести ее большую вычислительную сложность. Поэтому желательно иметь критерий, не требующий перебора всех возможных мест вложения одной последовательности в другую.

В [2] показано, что при $d=1$ вероятность ошибочного принятия гипотезы H_{0n} убывает экспоненциально быстро. В работе [2, неравенство (3)] найдена верхняя граница для вероятности того, что последовательность X_n может быть плотно вложена в начало независимой от нее последовательности Y_m, которая при $n \leqslant m$ имеет вид

$$\frac{1}{2N^{2n}}\left(\left(N-\sqrt{N^2-N}\right)^n+\left(N+\sqrt{N^2-N}\right)^n\right), \qquad (2)$$

не зависит от последовательности X_n и достигается на последовательностях, в которых нет совпадений соседних знаков. Обобщить оценку (2) на вложения с допуском $d \geqslant 2$ в общем случае не удалось. Однако скорость убывания оценок, полученных тем же методом, также будет экспоненциальной.

В [3] получены нижние оценки для вероятности того, что последовательность X_n может быть вложена с допуском d в начало независимой от нее последовательности Y_m, которая при $m \geqslant 1+(d+1)(n-1)$ имеет вид

$$\frac{1}{N}\left(1-\left(\frac{N-1}{N}\right)^{d+1}\right)^{n-1},$$

достигается на последовательностях, состоящих из одинаковых знаков, а также убывает экспоненциально быстро.

Заметим, что все наборы $(j_1, j_2, \ldots, j_n)$, удовлетворяющие (1), могут быть построены по всем реализациям последовательности случайных величин $\{Z_k\}_{k=1}^{\infty}$ со значениями в множестве $\{1,\ldots,d+1\}$, $Z_1=1$ следующим образом: $j_1 = Z_1$, $j_2 = j_1 + Z_2, \ldots, j_k = j_{k-1} + Z_k$. Самым простым способом, естественно, будет рассмотреть последовательность $\{Z_k\}_{k=1}^{\infty}$, состоящую

из независимых и равномерно распределенных случайных величин. Такое распределение как раз соответствует процедуре вложения с допуском, так как не отдает предпочтения каким-либо вариантам вложения, кроме того, промежутки между соседними вкладываемыми знаками будут независимы между собой. Аналогичная задача о построении скрытой марковской цепи путем вычеркивания знаков из простой цепи Маркова рассмотрена в [12]. Найдена матрица переходных вероятностей новой цепи и исследованы ее свойства.

При описанной процедуре длина вложения с допуском представляется суммой $\sum_{j=1}^{n} Z_j$ независимых случайных величин. Ее распределение является симметричным относительно центра $d+1/2$ и хорошо аппроксимируется нормальным законом распределения.

Однако фактически, если мы выберем X_n как подпоследовательность с допуском в Y_m, а затем проведем процедуру вложения с тем же допуском X_n в Y_m, то полученная при этом длина вложения может быть и меньше, чем $\sum_{j=1}^{n} Z_j$. Поэтому в дальнейшем необходимо рассмотреть вопрос о ее законе распределения, особенно в связи с выбором критической границы критерия, рассматриваемого в следующем разделе.

2. Построение критерия и его свойства

Последовательный критерий согласия [16, с. 443] с гипотезой H_{0n} при $d=1$ был предложен в [17]. Обобщение полученных результатов на случай произвольного $d \geqslant 2$ представлено в [18]. В настоящей работе проведем доказательство приведенных в [18] результатов.

Пусть $j_1 = 1,\ j_k = \min\{t > j_{k-1} : x_k = y_t\},\ k = 2,\ldots,n,$

$$V_1 = 1, V_k = V_k(X_n) = j_k - j_{k-1},\ k = 2,...,n,\ T_k = V_2 + \ldots + V_k.$$

Построим критерий $\mathcal{T}$ по следующему правилу. Последовательно по $k = 2,\ldots,n$ вычисляем значение T_k. Если $x_1 = y_1$ и на k-м шаге неравенство

$$T_k \leqslant (d+1)(k-1) \tag{3}$$

не выполнено, то гипотеза H_{0n} отклоняется. В противном случае продолжаем проверку. Если $x_1 = y_1$ и при всех $k = 2,\ldots,n$ выполнено неравенство (3), то считаем, что гипотеза H_{0n} не противоречит результатам наблюдений.

Замечание 1. Если H_{0n} верна, то существует набор чисел $j_1,\ldots,j_n$, удовлетворяющих (1), и $x_k = y_{j_k},\ k = 1,\ldots,n.$ Значит,

$$V_k = j_k - j_{k-1} \leqslant d+1, k = 2,\ldots,n, \text{ и } T_k = V_2 + \ldots + V_k \leqslant (d+1)(k-1).$$

Таким образом, вероятность ошибки первого рода (вероятность отклонить верную гипотезу H_{0n}, см. [16, с. 313]) критерия $\mathcal{T}$ равна нулю.

Нас интересует вероятность ошибки второго рода критерия $\mathcal{T}$, а также среднее число знаков, используемых критерием до принятия решения, при альтернативной гипотезе H_{1n} о том, что последовательность X_n не зависит от последовательности Y_m и тоже состоит из независимых равномерно распределенных на множестве A_N случайных величин.

Теорема 1. Вероятность ошибки второго рода критерия $\mathcal{T}$ при $n \geqslant 2$ равна

$$\mathbf{P}\{H_{0n} \mid H_{1n}\} = \frac{1}{N}\left(1 - \sum_{k=1}^{n-1} \sigma_k\right), \tag{4}$$

где последовательность чисел σ_k имеет производящую функцию

$$\sigma(s) = \sum_{k=0}^{\infty} \sigma_k s^k = 1 - \frac{1-s}{1-s/N} \exp\left\{\sum_{n=1}^{\infty} \frac{s^n}{nN^n} \sum_{m=1}^{dn} C_{n+m-1}^{n-1} \left(1 - \frac{1}{N}\right)^m\right\}. \tag{5}$$

Среднее число шагов до принятия решения при гипотезе H_{1n} равно

$$\frac{1}{N}\left(N - 1 + \sum_{k=1}^{n-2} (k+1)\sigma_k + n\left(1 - \sum_{k=1}^{n-2} \sigma_k\right)\right). \tag{6}$$

Доказательство. Вероятность ошибки второго рода критерия $\mathcal{T}$ равна

$$\mathbf{P}\{H_{0n} \mid H_{1n}\} = \mathbf{P}\{x_1 = y_1, T_i \leqslant (d+1)(i-1), i = 2, \ldots, n \mid H_{1n}\} = \mathbf{P}\{x_1 = y_1 \mid H_{1n}\} \times$$

$$\times \left(1 - \sum_{k=2}^{n} \mathbf{P}\{T_i \leqslant (d+1)(i-1), i = 2, \ldots, k-1, T_k > (d+1)(k-1) \mid H_{1n}\}\right) =$$

$$= \frac{1}{N}\left(1 - \sum_{k=2}^{n} \mathbf{P}\{T_i \leqslant (d+1)(i-1), i = 2, \ldots, k-1, T_k > (d+1)(k-1) \mid H_{1n}\}\right). \tag{7}$$

Для вычисления вероятностей в правой части (7) рассмотрим вспомогательную задачу.

Пусть $L_1, L_2, \ldots, L_n, \ldots$ — последовательность независимых случайных величин, каждая из которых имеет геометрический закон (см. [19, с. 238]):

$$\mathbf{P}\{L_i = k\} = pq^{k-1},\ k = 1, 2, \ldots,\ i = 1, 2, \ldots,\ q = 1 - p. \tag{8}$$

Пусть $S_n = L_1 + \ldots + L_n - (d+1)n$. Найдем

$$\tau_n = \mathbf{P}\{S_1 \leqslant 0, S_2 \leqslant 0, \ldots, S_{n-1} \leqslant 0, S_n > 0\} \text{ —} \tag{9}$$

вероятность того, что на n-м шаге впервые выполнено неравенство $\sum_{i=1}^{n} L_i > (d+1)n$ при $n \geqslant 1$. Так как случайная величина L_i равна номеру испытания Бернулли, в котором впервые произошел успех, то сумма

$\sum_{i=1}^{n} L_i$ — это номер опыта, в котором произошел n-й успех в испытаниях Бернулли. Поэтому

$$\mathbf{P}\left\{\sum_{i=1}^{n} L_i = n+k\right\} = C_{n+k-1}^{n-1} p^n q^k, \quad k = 0,1,2,\ldots. \tag{10}$$

Значит, при $m = -dn, -dn+1, \ldots$

$$\mathbf{P}\{S_n = m\} = \mathbf{P}\left\{\sum_{i=1}^{n} L_i = (d+1)n + m\right\} = C_{(d+1)n+m-1}^{n-1} p^n q^{dn+m}. \tag{11}$$

Известно (см. [20, с. 466]), что

$$\ln(1-\tau(s))^{-1} = \sum_{n=1}^{\infty} \frac{s^n}{n} \mathbf{P}\{S_n > 0\}. \tag{12}$$

Здесь $\tau(s)$ — производящая функция последовательности (9) (см. [21, с. 43]). Найдем производящую функцию $\tau(s)$, вычислив правую часть (12). Из (11) получаем, что

$$\mathbf{P}\{S_n > 0\} = 1 - \sum_{m=-dn}^{0} \mathbf{P}\{S_n = m\} = 1 - \sum_{m=-dn}^{0} C_{dn+m-1}^{n-1} p^n q^{dn+m} = 1 - p^n \sum_{u=0}^{dn} C_{n+u-1}^{n-1} q^u.$$

Подставив полученное выражение в правую часть (12), получим

$$\ln(1-\tau(s))^{-1} = \sum_{n=1}^{\infty} \frac{s^n}{n}\left(1 - p^n \sum_{m=0}^{dn} C_{n+m-1}^{n-1} q^m\right) = \sum_{n=1}^{\infty} \frac{s^n}{n}\left(1 - p^n - p^n \sum_{m=1}^{dn} C_{n+m-1}^{n-1} q^m\right).$$

Теперь воспользуемся разложением логарифма в ряд Тейлора $\sum_{n=1}^{\infty} \frac{s^n}{n} = -\ln(1-s)$ при $|s|<1$. Имеем

$$\ln(1-\tau(s))^{-1} = -\ln(1-s) + \ln(1-sp) - \sum_{n=1}^{\infty} \frac{(sp)^n}{n} \sum_{m=1}^{dn} C_{n+m-1}^{n-1} q^m.$$

Значит,

$$\tau(s) = 1 - \frac{1-s}{1-ps} \exp\left\{\sum_{n=1}^{\infty} \frac{(sp)^n}{n} \sum_{m=1}^{dn} C_{n+m-1}^{n-1} q^m\right\}. \tag{13}$$

Вернемся к нашему критерию. Так как знаки последовательности Y_m независимы и распределены на множестве $\{0,\ldots,N-1\}$ равномерно, то распределения случайных величин $V_k(X_n)$ одинаковы при всех X_n. Случайные величины $V_2,\ldots,V_n$ независимы (в совокупности) и $\mathbf{P}\{V_k = l \mid H_{1n}\} = N^{-1}(1-N^{-1})^{l-1}$, $l \geqslant 1$, $k = 2,\ldots,n$ (см., например, [19, с. 327–328]). Закон распределения случайных величин $V_2,\ldots,V_n$ (при ги-

потезе H_{1n}) — это тот же геометрический закон распределения (8) с $p=1/N$ и $q=1-1/N$. Обозначим $\sigma_n=\tau_n\big|_{p=1/N,q=1-1/N}$ и $\sigma(s)=\sum_{n=1}^{\infty}\sigma_n s^n$. Очевидно, что $\sigma(s)=\tau(s)\big|_{p=1/N,q=1-1/N}$, где $\tau(s)$ определена формулой (13). В этих обозначениях равенство (7) можно записать в виде

$$\mathbf{P}\{H_{0n}\mid H_{1n}\}=\frac{1}{N}\left(1-\sum_{k=1}^{n-1}\sigma_k\right),\ n\ge 2,\ \text{и}\ \mathbf{P}\{H_{0n}\mid H_{1n}\}=\frac{1}{N},\ n=1.$$

Отсюда получаем (4).

Теперь перейдем к вычислению среднего для числа ϑ_n знаков последовательности X_n, используемых критерием. Так как

$$\mathbf{P}\{\vartheta_n=1\mid H_{1n}\}=1-N^{-1},\quad \mathbf{P}\{\vartheta_n=k+1\mid H_{1n}\}=\sigma_k N^{-1},\quad k=1,\ldots,n-2,$$

$$\mathbf{P}\{\vartheta_n=n\mid H_{1n}\}=\frac{1}{N}\sum_{k=n-1}^{\infty}\sigma_k=\frac{1}{N}\left(1-\sum_{k=1}^{n-2}\sigma_k\right),$$

то $\mathbf{E}\vartheta_n$ задается формулой (6). *Теорема 1 доказана.*

Замечание 2. Обозначим $F_{n,p}(x)$ функцию распределения отрицательного биномиального закона с вероятностями отдельных значений, задаваемыми формулой (10). Тогда (5) можно переписать в виде:

$$\sigma(s)=1-(1-s)\exp\left\{\sum_{n=1}^{\infty}\frac{s^n}{n}F_{n,1/N}(n(d+1)+1)\right\}.$$

Замечание 3. Согласно [20, теорема 2 §2 гл. XII, с. 448–449] (см. также [22]), случайная величина с законом распределения, соответствующим производящей функции (5), является собственной, если $\mathbf{E}V_2\geqslant d+1$ (закон распределения V_2 задается формулой (14)) и имеет конечное математическое ожидание, если $\mathbf{E}V_2>d+1$. Очевидно, $\mathbf{E}V_2=N$. Значит, $\sigma(1)=1$ при $N\geqslant d+1$ и $\sigma'(1)<\infty$ при $N\geqslant d+2$.

Замечание 4. Асимптотическая формула для вероятности того, что случайное блуждание, образованное независимыми одинаково распределенными центрированными случайными величинами, впервые пересечет заданный уровень снизу вверх в момент n, найдена в работе [23]. Для нашего критерия эти результаты возможно применять при больших длинах последовательностей n и маленьких размерах алфавита N.

Замечание. Число знаков, используемых критерием $\mathcal{T}$, равно n при гипотезе H_{0n} и не превосходит n при гипотезе H_{1n}. Таким образом, трудоемкость предложенного критерия по порядку меньше трудоемкости тотального апробирования.

3. Численная иллюстрация

Ниже приведем численную иллюстрацию полученных результатов, так как их качественный анализ в общем случае провести затруднительно. Рассмотрим случай $d = 2$. Начнем с того, что при $N = 2,3$ значение $\sigma'(1) = \infty$ (в этом случае среднее число знаков, используемых критерием, неограниченно возрастает с ростом длины последовательности n). Поэтому вероятность ошибки второго рода β убывает медленно. Разница в поведении β при $N = 2$ и $N = 3$ объясняется тем, что в первом случае ряд, стоящий в экспоненте в формуле (5), расходится. При $N \geqslant 4$ значение $\sigma'(1) < \infty$ и вероятность ошибки второго рода быстро убывает. В табл. 1 приведены численные значения вероятности ошибки второго рода β при различных N в зависимости от длины вкладываемой последовательности n при $d = 2$, вычисленные по формуле (4). Аналогичные результаты имеют место и при $d = 3$ (табл. 2). Также ниже приведены значения $\mathbf{E}\vartheta_n$ (среднего числа знаков, используемых критерием) при различных значениях n и N для $d = 2$ (табл. 3) и $d = 3$ (табл. 4).

Таблица 1. Вероятность ошибки второго рода β при $d = 2$

N \ n	2	3	4	5	10	15
2	0.4375	0.4141	0.4023	0.3956	0.3847	0.3827
3	0.2346	0.1907	0.1647	0.1470	0.1035	0.0843
4	0.1445	0.1000	0.0750	0.0589	0.0241	0.0124
5	0.0976	0.0583	0.0382	0.0263	0.0060	0.0018
6	0.0702	0.0367	0.0212	0.0130	0.0017	$2.96 \cdot 10^{-4}$
7	0.0529	0.0246	0.0127	0.0070	$5.24 \cdot 10^{-4}$	$5.64 \cdot 10^{-5}$
8	0.0413	0.0172	0.0080	0.0040	$1.89 \cdot 10^{-4}$	$1.24 \cdot 10^{-5}$
9	0.0331	0.0125	0.0053	0.0024	$7.24 \cdot 10^{-5}$	$3.065 \cdot 10^{-6}$
10	0.0271	0.0094	0.0036	0.0015	$3.02 \cdot 10^{-6}$	$8.49 \cdot 10^{-7}$

Таблица 2. Вероятность ошибки второго рода β при $d = 3$

N \ n	2	3	4	5	10	15
2	0.4688	0.4609	0.4583	0.4572	0.4563	0.4563
3	0.2675	0.2415	0.2273	0.2185	0.2002	0.1945
4	0.1709	0.1375	0.1182	0.1052	0.0736	0.0599
5	0.1181	0.0845	0.0656	0.0533	0.0256	0.0152
6	0.0863	0.0553	0.0388	0.0287	0.0091	0.0037
7	0.0658	0.0380	0.0243	0.0164	0.0034	$9.57 \cdot 10^{-4}$
8	0.0517	0.0272	0.0159	0.0099	0.0014	$2.62 \cdot 10^{-4}$
9	0.0417	0.0201	0.0108	0.0062	$5.91 \cdot 10^{-4}$	$7.78 \cdot 10^{-5}$
10	0.0344	0.0153	0.0076	0.0040	$2.68 \cdot 10^{-4}$	$2.48 \cdot 10^{-5}$

Таблица 3. Среднее число проверенных знаков при $d = 2$

N \ n	2	3	4	5	6	7	8	9	10
2	1.50	1.94	2.35	2.75	3.15	3.54	3.93	4.32	4.70
3	1.33	1.57	1.76	1.92	2.07	2.20	2.33	2.44	2.55
4	1.25	1.39	1.49	1.57	1.63	1.68	1.72	1.75	1.78
5	1.20	1.30	1.36	1.39	1.42	1.44	1.45	1.46	1.47
6	1.17	1.24	1.27	1.29	1.31	1.32	1.32	1.33	1.33
7	1.14	1.20	1.22	1.23	1.24	1.24	1.25	1.25	1.25
8	1.13	1.17	1.18	1.19	1.20	1.20	1.20	1.20	1.20
9	1.11	1.14	1.16	1.16	1.16	1.17	1.17	1.17	1.17
10	1.10	1.13	1.14	1.14	1.14	1.14	1.14	1.14	1.14

Таблица 4. Среднее число проверенных знаков при $d = 3$

N \ n	2	3	4	5	6	7	8	9	10
2	1.50	2.00	2.50	3.00	3.50	4.00	4.50	5.00	5.50
3	1.33	1.66	1.99	2.31	2.63	2.95	3.27	3.59	3.90
4	1.25	1.49	1.73	1.96	2.19	2.41	2.64	2.85	3.07
5	1.20	1.39	1.58	1.77	1.95	2.13	2.30	2.48	2.65
6	1.17	1.33	1.49	1.64	1.80	1.95	2.10	2.25	2.40
7	1.14	1.28	1.42	1.56	1.69	1.82	1.96	2.09	2.22
8	1.13	1.25	1.37	1.49	1.61	1.73	1.85	1.97	2.09
9	1.11	1.22	1.33	1.44	1.55	1.65	1.76	1.87	1.98
10	1.10	1.20	1.30	1.40	1.49	1.59	1.69	1.79	1.89

Заключение

В работе рассмотрен последовательный критерий проверки гипотезы о вложении с допуском одной дискретной последовательности в другую. Показано, что вероятность ошибки первого рода этого критерия равна нулю, найдено выражение для вероятности ошибки второго рода при альтернативной гипотезе о независимости и равномерной распределенности знаков рассматриваемых последовательностей. Проведено численное исследование вероятности ошибки второго рода.

Литература

1. Golic J. Dj. Constrained embedding probability for two binary strings // SIAM J. Discrete Math. 1996. V. 9, № 3. P. 360–364. DOI: 10.1137/S0895479894246917

2. Михайлов В. Г., Меженная Н. М. Оценки для вероятности плотного вложения одной дискретной последовательности в другую // Дискретная математика. 2005. Т. 17, № 3. С. 19–27. DOI: 10.4213/dm113

3. Меженная Н. М., Михайлов В. Г. Нижние оценки для вероятности вложения с произвольным допуском // Вестник Московского государственного технического университета им. Н. Э. Баумана. Сер. Естественные науки. 2012. № 2. С. 3–11.

4. Donovan D. M., Lefevre J., Simpson L. A. Discussion of constrained binary embeddings with applications to cryptanalysis of irregularly clocked stream ciphers // Balakrishnan R., Veni Madhavan C. (Eds.) Discrete mathematics. Proceedings of the international conference on discrete mathematics. Bangalore: Indian Institute of Science, 2006. P. 73–86.

5. Golic J. Dj. Embedding probabilities for the alternating step generator // IEEE Trans. Inform. Theory. 2005. V. 51(7). P. 2543–2553. DOI: 10.1109/TIT.2005.850114

6. Armknecht F., Mikhalev V. On lightweight stream ciphers with shorter internal states // Leander G. (Ed.) Fast Software Encryption. Proceedings of 22nd International Workshop, FSE 2015. Istanbul: Springer-Verlag, 2015. P. 451–470. DOI: 10.1007/978-3-662-48116-5_22

7. Asimi Y., Amghar A., Asimi A., Sadqi Y. New random generator of a safe cryptographic salt per session // Int. J. Netw. Secur. 2016. V. 18, № 3. P. 445–453.

8. El-Razouk H., Reyhani-Masoleh A., Gong G. New implementations of the WG stream cipher // IEEE Transactions on Very Large Scale Integration (VLSI) Systems. 2014. V. 22, № 9. P. 1865–1878. DOI: 10.1109/TVLSI.2013.2280092

9. Ma W., Feng D. Clock-controlled key-stream generator and its cryptographic properties // Front. Electr. Electron. Eng. China. 2008. Vol. 3, № 3. P. 327–332. DOI: 10.1007/s11460-008-0014-6

10. Jiang S., Tang Z., Wang M. On the CLD attack to a statistical model of a key stream generator // Int. J. Netw. Secur. 2016. Vol. 18, № 5. P. 987–992.

11. Deepthi P. P., Deepa S. J., Sathidevi P. S. Design and analysis of a highly secure stream cipher based on linear feedback shift register // Computer & Electrical Engineering. 2009. V. 35, № 2. P. 235–243. DOI: 10.1016/j.compeleceng.2008.06.005

12. Шуваев Д. В. О подпоследовательностях марковских последовательностей // Математические вопросы криптографии. 2016. Т. 7, № 4. С. 133–142. DOI: 10.4213/mvk208

13. Combinatorial Pattern Matching. Proceedings of CPM 2013, LNCS 7922 / J. Fischer, P. Sanders (Eds.). Bad Herrenalb: Springer-Verlag, 2013. 259 p. DOI: 10.1007/978-3-642-38905-4

14. Reinert G., Waterman M. S. On the length of the longest exact position match in a random sequence // IEEE/ACM transactions on computational biology and bioinformatics. 2007. V. 4, № 1. P. 153–156. DOI: 10.1109/TCBB.2007.1023

15. Сапаров А. Ю., Бельтюков А. П. Применение регулярных выражений в распознавании математических текстов // Вестник Удмуртского университета. Математика. Механика. Компьютерные науки. 2012. Вып. 2. С. 63–73. DOI: 10.20537/vm120206

16. Ивченко Г. И., Медведев Ю. И. Введение в математическую статистику. М.: Изд-во ЛКИ, 2010. 600 с.

17. Меженная Н. М. О проверке гипотезы о плотном вложении для дискретных случайных последовательностей // Вестник БГУ. Математика, информатика. 2017. № 4. С. 9–20. DOI: 10.18101/2304-5728-2017-4-9-20

18. Меженная Н. М. Проверка гипотезы о вложении с допуском для дискретных случайных последовательностей // Прикладная дискретная математика. Приложение. 2018. Т. 11. С. 12–14. DOI: 10.17223/2226308X/11/3

19. Феллер В. Введение в теорию вероятностей и ее приложения: в 2 т. 2-е изд. М.: Либроком, 2010. Т. 1. 527 с.

20. Феллер В. Введение в теорию вероятностей и ее приложения: в 2 т. 2-е изд. М.: Либроком, 2010. Т. 2. 751 с.

21. Прохоров Ю. И., Пономаренко Л. С. Лекции по теории вероятностей и математической статистике. 2-е изд. М.: Изд-во Моск. ун-та, 2012. 256 с.

22. Рогозин Б. А. Замечание к одной теореме В. Феллера // Теория вероятностей и ее применения. 1969. Т. 14, № 3. С. 554–555. DOI: 10.1137/1114070

23. Могульский А. А., Рогозин Б. А. Локальная теорема для момента достижения фиксированного уровня случайным блужданием // Математические труды. 2005. Т. 8, № 1. С. 43–70.

ABOUT A TEST OF EMBEDDING WITH MARGIN
FOR DISCRETE SEQUENCES

Natalia M. Mezhennaya
Cand. Sci. (Phys. and Math.), A/Prof.,
Bauman Moscow State Technical University
5 2nd Baumanskaya St., Moscow 105005, Russia
E-mail: natalia.mezhennaya@gmail.com

Sequence X is a subsequence with margin d of sequence Y if X is constructed from Y by deleting non-adjacent segments consisted of at most d letters. In this case we say that X can be embedded into Y with margin d. The article presents a sequential test for the hypothesis of embedding with margin d for discrete random sequences over a finite alphabet and study its properties. The probability of type I error (the probability of rejection of true hypothesis of embedding with margin) of the constructed test is equal to zero. We derive an expression for the probability of type II error under the alternative hypothesis that the discrete sequences under consideration consist of mutually independent random variables with uniform distributions on finite alphabet. We find out the average number of letters of the embedded sequence used by test before the decision is made under the alternative hypothesis. The complexity of the proposed procedure is proportional to the length of the embedded sequence under true hypothesis of embedding with margin and is smaller under the alternative hypothesis which is less than complexity of total testing by order of magnitude. We have presented numerical values of the probability of type II error and the average number of used letters for different values of d and the alphabet size.

Keywords: dense embedding; embedding with margin; sequential test; hypothesis of independence; probabilities of type I and type II errors; discrete random sequence.

References

1. Golic J. Dj. Constrained Embedding Probability for Two Binary Strings. *SIAM J. Discrete Math.* 1996. V. 9. No. 3. Pp. 360–364. DOI: 10.1137/S0895479894246917

2. Mikhailov V. G., Mezhennaya N. M. Otsenki dlya veroyatnosti plotnogo vlozheniya odnoi diskretnoi posledovatelnosti v druguyu [Bounds for the Probability of a Dense Embedding of One Discrete Sequence into Another]. *Diskretnaya matematika — Discrete Mathematics.* 2005. V. 15. No. 4. Pp. 377–386. DOI: 10.1515/156939205774464864

3. Mezhennaya N. M., Mikhailov V. G. Nizhnie otsenki dlya veroyatnosti vlozheniya s proizvolnym dopuskom [Lower Bounds for Probabilities of Embedding with Arbitrary Margin]. *Vestnik Moskovskogo gosudarstvennogo tekhnicheskogo universiteta im. N. E. Baumana. Ser. Estestvennye nauki — Bulletin of Bauman Moscow State Technical University. Ser. Natural Sciences.* 2012. No. 2. Pp. 3–11.

4. Donovan D. M., Lefevre J., Simpson L. A. Discussion of Constrained Binary Embeddings with Applications to Cryptanalysis of Irregularly Clocked Stream Ciphers. Balakrishnan R., Veni Madhavan C. (Eds.) *Discrete Mathematics.* Proceedings of the International Conference on Discrete Mathematics. Bangalore: Indian Institute of Science, 2006. Pp. 73–86.

5. Golic J. Dj. Embedding Probabilities for the Alternating Step Generator. *IEEE Trans. Inform. Theory.* 2005. V. 51(7). Pp. 2543–2553. DOI: 10.1109/TIT.2005.850114

6. Armknecht F., Mikhalev V. On Lightweight Stream Ciphers with Shorter Internal States. Leander G. (Ed.) *Fast Software Encryption.* Proceedings of 22nd International Workshop, FSE 2015. Istanbul: Springer-Verlag, 2015. Pp. 451–470. DOI: 10.1007/978-3-662-48116-5_22

7. Asimi Y., Amghar A., Asimi A., Sadqi Y. New Random Generator of a Safe Cryptographic Salt per Session. *Int. J. Netw. Secur.* 2016. V. 18. No. 3. Pp. 445–453.

8. El-Razouk H., Reyhani-Masoleh A., Gong G. New Implementations of the WG Stream Cipher. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems.* 2014. V. 22. No. 9. Pp. 1865–1878. DOI: 10.1109/TVLSI.2013.2280092

9. Ma W., Feng D. Clock-Controlled Key-Stream Generator and Its Cryptographic Properties. *Front. Electr. Electron. Eng. China.* 2008. V. 3. No. 3. Pp. 327–332. DOI: 10.1007/s11460-008-0014-6

10. Jiang S., Tang Z., Wang M. On the CLD Attack to a Statistical Model of a Key Stream Generator. *Int. J. Netw. Secur.* 2016. V. 18. No. 5. Pp. 987–992.

11. Deepthi P. P., Deepa S. J., Sathidevi P. S. Design and Analysis of a Highly Secure Stream Cipher Based on Linear Feedback Shift Register. *Computer & Electrical Engineering.* 2009. V. 35. No. 2. Pp. 235–243. DOI: 10.1016/j.compeleceng.2008.06.005

12. Shuvaev D. V. O podposledovatelnostyakh markovskikh posledovatelnostei [On Subsequences of Markovian Sequences]. *Matematicheskie voprosy kriptografii — Mathematical Aspects of Cryptography.* 2016. V. 7. No. 4. Pp. 133–142. DOI: 10.4213/mvk208

13. *Combinatorial Pattern Matching.* Proceedings of CPM 2013, LNCS 7922. J. Fischer, P. Sanders (Eds.). Bad Herrenalb: Springer-Verlag, 2013. 259 p. DOI: 10.1007/978-3-642-38905-4

14. Reinert G., Waterman M. S. On the Length of the Longest Exact Position Match in a Random Sequence. *IEEE/ACM Transactions on Computational Biology and Bioinformatics.* 2007. V. 4. No. 1. Pp. 153–156. DOI: 10.1109/TCBB.2007.1023

15. Saparov A. Yu., Beltyukov A. P. Primenenie regulyarnykh vyrazhenii v raspoznavanii matematicheskikh tekstov [Application of Regular Expressions in the Recognition of Mathematical Text]. *Vestnik Udmurtskogo universiteta. Matematika. Mekhanika. Kompyuternye nauki — Bulletin of Udmurt University. Mathematics. Mechanics. Computer Science.* 2012. No. 2. Pp. 63–73. DOI: 10.20537/vm120206

16. Ivchenko G. I., Medvedev Yu. I. *Vvedenie v matematicheskuyu statistiku* [Introduction to Mathematical Statistics]. Moscow: LKI Publ., 2010. 600 p.

17. Mezhennaya N. M. O proverke gipotezy o plotnom vlozhenii dlya diskretnykh sluchainykh posledovatelnostei [About Testing the Dense Embedding Hypotesis for Discrete Random Sequences]. *Vestnik Buryatskogo gosudarstvennogo universiteta. Matematika, Informatika — Bulletin of Buryat State University. Mathematics, Informatics.* 2017. No. 4. Pp. 9–20. DOI: 10.18101/2304-5728-2017-4-9-20

18. Proverka gipotezy o vlozhenii s dopuskom dlya diskretnykh sluchainykh posledovatelnostei [Testing of Embedding with Margin for Discrete Random Sequences].

Prikladnaya diskretnaya matematika. Prilozhenie — Applied Discrete Mathematics. Supplement. 2018. V. 11. Pp. 12–14. DOI: 10.17223/2226308X/11/3

19. Feller W. *An Introduction to Probability Theory and Its Applications.* 2nd Ed. New-York: John Wiley and Sons Publ., 1968. V. 1. 528 p.

20. Feller W. *An Introduction to Probability Theory and Its Applications.* 2nd Ed. New-York: John Wiley and Sons, 1971. V. 2. 704 p.

21. Prokhorov Yu. V., Ponomarenko L. S. *Lektsii po teorii veroyatnostei i matematicheskoi statistike* [Lectures on Probability Theory and Mathematical Statistics]. 2nd Ed. Moscow: Moscow State University Publ., 2012. 256 p.

22. Rogozin B. A. A Remark to a Theorem due to W. Feller. *Theory Probab. Appl.* 1969. V. 14. No. 3. Pp. 554–555. DOI: 10.1137/1114070

23. Mogulskii A. A., Rogozin B. A. Lokalnaya teorema dlya momenta dostizheniya fiksirovannogo urovnya sluchainym bluzhdaniem [A Local Theorem for the First Hitting Time of a Fixed Level by a Random Walk]. *Matematicheskie trudy — Siberian Advances in Mathematics.* 2005. V. 8. No. 1. Pp. 43–70.