http://dx.doi.org/10.35596/1729-7648-2024-22-4-76-83

Оригинальная статья
Original paper

УДК 004.415.533

МЕРА РАЗЛИЧИЯ ДЛЯ УПРАВЛЯЕМЫХ ВЕРОЯТНОСТНЫХ ТЕСТОВ

В. Н. ЯРМОЛИК[1], В. В. ПЕТРОВСКАЯ[1], Н. А. ШЕВЧЕНКО[2]

[1]Белорусский государственный университет информатики и радиоэлектроники (г. Минск, Республика Беларусь)
[2]Дармштадтский технический университет (г. Дармштадт, Германия)

Поступила в редакцию 05.03.2024

Аннотация. Исследована задача построения характеристик различия тестовых последовательностей. Обоснованы ее актуальность для генерирования управляемых вероятностных тестов и сложность нахождения мер отличия для символьных тестов. Показана ограниченность применения традиционных характеристик расстояния для получения меры различия тестовых наборов. Для двоичного случая определена новая мера различия $MH(T_i, T_k)$ двух символьных тестовых наборов T_i и T_k на основе классического расстояния Хэмминга. Данная мера представляет собой n компонент, каждая из которых определяется расстоянием Хэмминга между двоичным набором T_i и циклически сдвинутым на v бит набором T_k. Рассмотрены основные свойства предложенной меры различия и показана ее эффективность для классификации кандидатов в тесты при генерировании управляемых вероятностных тестов. Приведены экспериментальные результаты, подтверждающие эффективность меры различия.

Ключевые слова: мера различия, расстояние Хэмминга, расстояние Левенштейна, тест, тестовый набор.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Для цитирования. Ярмолик, В. Н. Мера различия для управляемых вероятностных тестов / В. Н. Ярмолик, В. В. Петровская, Н. А. Шевченко // Доклады БГУИР. 2024. Т. 22, № 4. С. 76–83. http://dx.doi.org/10.35596/1729-7648-2024-22-4-76-83.

DIFFERENCE MEASURE FOR CONTROLLED RANDOM TESTS

VYACHESLAV N. YARMOLIK[1], VITA V. PETROVSKAYA[1], MIKALAI A. SHAUCHENKA[2]

[1]Belarusian State University of Informatics and Radioelectronics (Minsk, Republic of Belarus)
[2]Darmstadt Technical University (Darmstadt, Germany)

Submitted 05.03.2024

Abstract. The task of constructing test sequences difference characteristics was studied. Its relevance for generating controlled random tests and complexity in finding difference measures for the case of symbolic tests were substantiated. The limitations of using traditional distance characteristics to obtain a measure of the difference between test sets are shown. For the binary case, a new measure of the difference $MH(T_i, T_k)$ of two character test sets T_i and T_k is defined based on the classical Hamming distance. This measure represents n components, each of which is determined by the Hamming distance between the binary set T_i and the pattern T_k cyclically shifted by v bits. The main properties of the proposed dissimilarity measure are reviewed and its effectiveness for classifying test candidates when generating controlled random tests is shown. Experimental results are presented that confirm the effectiveness of the proposed difference measure.

Keywords: measure of difference, Hamming distance, Levenshtein distance, test, test pattern.

Conflict of interests. The authors declare no conflict of interests.

For citation. Yarmolik V. N., Petrovskaya V. V., Shauchenka M. A. (2024) Difference Measure for Controlled Random Tests. *Doklady BGUIR*. 22 (4), 76–83. http://dx.doi.org/10.35596/1729-7648-2024-22-4-76-83 (in Russian).

Введение

Классическое определение расстояния Хэмминга предполагает сравнение двух последовательностей $T_i = t_{i,0}, t_{i,1}, ..., t_{i,n-1}$ и $T_k = t_{k,0}, t_{k,1}, ..., t_{k,n-1}$, включающих по n символов $t_{i,j}$ и $t_{k,j}$ из произвольного алфавита [1, 2]. Расстояние Хэмминга $H(T_i, T_k)$ между T_i и T_k как количество позиций, в которых $t_{i,j}$ и $t_{k,j}$ различаются, описывается соотношением [1]

$$H(T_i, T_k) = \sum_{j=0}^{n-1} \mathrm{I}(t_{i,j} \neq t_{k,j}). \quad (1)$$

Выражение $\mathrm{I}(t_{i,j} \neq t_{k,j})$ представляет собой индикаторную функцию, равную единице при $t_{i,j} \neq t_{k,j}$ и нулю в противном случае. Минимальное значение $\min H(T_i, T_k)$ равняется 0 при совпадении всех символов последовательностей T_i и T_k, а максимальное $\max H(T_i, T_k)$ равняется n при несовпадении всех n символов.

Расстояние Хэмминга было разработано для измерения различия двух последовательностей двоичных символов и в основном использовалось для целей помехоустойчивого кодирования [1]. Для случая, когда $t_{i,j}, t_{k,j} \in \{0,1\}$, индикаторная функция $\mathrm{I}(t_{i,j} \neq t_{k,j})$ принимает вид арифметического выражения $\left|t_{i,j} - t_{k,j}\right|$ или булевой функции $t_{i,j} \oplus t_{k,j}$. Простота и привлекательность этой меры различия предопределили ее применимость для более широкого диапазона практических задач. В первую очередь это касается метода сравнения символьных последовательностей, основанного на их выравнивании (метод редакционного расстояния), и его модификаций [3, 4]. Классическое расстояние Хэмминга обобщалось и модифицировалось с той же целью редактирования (преобразования) исходного графического изображения в целевое изображение [5]. Это расширение, называемое обобщенным расстоянием Хэмминга, рассматривалось для исследования свойств сопряженных (conjugates) последовательностей символов, когда $T_i = G, V$, $T_k = V, G$, где $G = t_{i,0}, t_{i,1}, ..., t_{i,l}$, $V = t_{i,l+1}, t_{i,l+2}, ..., t_{i,n-1}$ [6].

Расстояние Хэмминга как мера различия тестовых наборов широко применяется для формирования управляемых вероятностных тестов [7–9]. Характерной особенностью управляемого генерирования вероятностных тестовых наборов является информация, которая извлекается в виде некоторых характеристик (мер) из ранее сгенерированных тестовых наборов $T_0, T_1, ..., T_{i-1}$ и используется для формирования следующего набора T_i [7]. Очередной тестовый набор T_i управляемого вероятностного теста формируется максимально удаленным (различным) от ранее сгенерированных наборов $T_0, T_1, ..., T_{i-1}$ в терминах заранее выбранных мер различия. Основная операция при определении различия между наборами T_i и T_k – операция сравнения их символов с использованием расстояния Хэмминга $H(T_i, T_k)$ [7–9].

Как отмечалось в ряде работ, расстояние Хэмминга является малоэффективной мерой, поскольку позволяет различать лишь полностью совпадающие последовательности при $H(T_i, T_k) = 0$ и все остальные несовпадающие [3]. Примером несовпадающих двоичных наборов могут быть наборы $T_k \in \{00110011, 01010110, 10000100\}$, которые не совпадают с набором $T_i = 11110000$. Действительно, во всех трех случаях имеем одно и то же значение $H(T_i, T_k) = 4$, что свидетельствует об одинаковом различии трех наборов T_k от T_i, хотя последовательности T_k различны, и их структуры существенно отличаются. Приведенный пример показывает необходимость использования более эффективных мер сравнения двоичных последовательностей символов, позволяющих полнее оценивать их различие.

Мера различия двоичных тестовых наборов

Не нарушая общности дальнейшего изложения, предположим, что тестовые наборы T_i и T_k представляют собой двоичные последовательности, т. е. их символы $t_{i,j}, t_{k,j} \in \{0, 1\}$. Основываясь на классическом расстоянии Хэмминга [1] и его обобщениях [6, 10], введем меру различия, соответствующую следующему определению: мера различия $MH(T_i, T_k)$ двоичных тестовых набо-

ров $T_i = t_{i,0}, t_{i,1}, ..., t_{i,n-1}$, и $T_k = t_{k,0}, t_{k,1}, ..., t_{k,n-1}$, где $t_{i,j}, t_{k,j} \in \{0, 1\}$, для произвольного целого n состоит из n компонент $MH_0, MH_1, \ldots, MH_{n-1}$, формируемых согласно соотношению:

$$MH_v = MH_v(T_i, T_k) = \sum_{j=0}^{n-1} (t_{i,j} \oplus t_{k,(j+v) \bmod n}), v = \overline{0,1,...,(n-1)}. \quad (2)$$

Выражение (2) при $v = 0$ соответствует соотношению (1), используемому для двоичного случая. Для других значений v компоненты меры различия $MH(T_i, T_k)$ определяют расстояния Хэмминга между двоичным набором T_i и циклически сдвинутыми на v бит наборами T_k. В табл. 1 приведены результаты вычисления компонент предложенной меры различия для нескольких пар наборов T_i и T_k разрядностью $n = 8$.

Таблица 1. Значения компонент $MH_0, MH_1, \ldots, MH_7$ меры различия $MH(T_i, T_k)$ для наборов T_i и T_k
Table 1. Values of the difference measure $MH(T_i, T_k)$ components $MH_0, MH_1, \ldots, MH_7$ for patterns T_i and T_k

T_i, T_k	MH_0	MH_1	MH_2	MH_3	MH_4	MH_5	MH_6	MH_7
11110000, 00110011	4	4	4	4	4	4	4	4
11110000, 01010110	4	4	4	2	4	4	4	6
11110000, 10000100	4	6	4	4	4	2	4	4

Как видно из табл. 1, последовательности $T_k \in \{00110011, 01010110, 10000100\}$ имеют различные значения набора компонент меры $MH(T_i, T_k)$ по отношению к набору $T_i = 11110000$. Это позволяет более полно соотнести данные наборы с T_i, так как они стали различимыми между собой.

Рассмотрим основные свойства меры различия $MH(T_i, T_k) = MH_0, MH_1, \ldots, MH_{n-1}$ для двоичных тестовых наборов $T_i = t_{i,0}, t_{i,1}, ..., t_{i,n-1}$ и $T_k = t_{k,0}, t_{k,1}, ..., t_{k,n-1}$ с весами $w_i = w_i(T_i)$ и $w_k = w_k(T_k)$. Весом двоичного вектора является количество в нем единичных символов, соответственно, $0 \le w_i \le n$ и $0 \le w_k \le n$.

Свойство 1. Первая компонента MH_0 меры различия $MH(T_i, T_k)$ принимает значение в диапазоне

$$|w_i - w_k| \le MH_0 \le \begin{cases} w_i + w_k, & \text{если } w_i + w_k \le n; \\ 2n - (w_i + w_k), & \text{если } w_i + w_k > n. \end{cases} \quad (3)$$

Данное свойство вытекает из определения расстояния Хэмминга для случая двоичных наборов.

Свойство 2. Компоненты $MH_0, MH_1, \ldots, MH_{n-1}$ меры различия $MH(T_i, T_k)$ принимают все нечетные либо все четные значения, а величина MH_j отличается от MH_l ($j \ne l$) на $2d$, где d – целое либо ноль.

Доказательство. Согласно определению расстояния Хэмминга, компонента MH_0 принимает произвольное значение от 0 до n, как четное, так и нечетное. Конкретная величина MH_0 зависит от сравниваемых, согласно (1) и (2), наборов символов T_i и T_k и их характеристик, включая w_i и w_k (свойство 1). Остальные компоненты $MH_1, MH_2, \ldots, MH_{n-1}$ формируются на основании T_i и циклически сдвинутых на v позиций копий $T_k(v)$ набора T_k. В общем случае обе последовательности бит состоят из перемежающихся наборов серий, состоящих из нулей и единиц. Изменяющееся соотношение этих серий в наборах в результате циклического сдвига приводит к изменению количества их несовпадающих символов.

Оценим влияние серии из единиц на две последующие компоненты MH_v и MH_{v+1} для всевозможных вариантов соотношения двоичных значений в наборах $T_k(v)$ и $T_k(v+1)$. Первоначально исследуем влияние на компоненты $MH(T_i, T_k)$ только одной серии, понимая, что таких серий может быть больше.

Рассмотрим серию единиц $S(v) = s_{c-2}s_{c-1}s_cs_{c+1}s_{c+2} \ldots s_{c+b}s_{c+b+1}s_{c+b+2} = 00111\ldots100$ как фрагмент двоичного набора $T_k(v)$ и ее сдвинутую версию $S(v+1) = 000111\ldots10$ для $T_k(v+1)$. Операция сдвига серии единиц $S(v)$ приводит к тому, что только два двоичных символа, связанных с рассматриваемой серией единиц, меняют свои значения на противоположные, а именно: s_c – с единицы на ноль, а s_{c+b+1} – наоборот, с нуля на единицу. Отметим, что двоичные символы a_c и a_{c+b+1} набора T_i в указанных позициях не меняют свои значения. Соответственно, изменение $\Delta MH_{v+1}(S) = MH_{v+1}(S) - MH_v(S)$ значения компоненты MH_{v+1} по отношению к MH_v за счет сдвига серии единиц $S(v)$ будет иметь следующий вид:

$$\Delta MH_{v+1}(S) = MH_{v+1}(S) - MH_v(S) =$$
$$= (|a_c - s_{c-1}| + |a_{c+b+1} - s_{c+b}|) - (|a_c - s_c| + |a_{c+b+1} - s_{c+b+1}|) =$$
$$= (|a_c - 0| + |a_{c+b+1} - 1|) - (|a_c - 1| + |a_{c+b+1} - 0|) =$$
$$= (a_c - |a_c - 1|) - (a_{c+b+1} - |a_{c+b+1} - 1|) = A_c - A_{c+b+1}. \quad (4)$$

В выражении (4) s_c, s_{c-1}, s_{c+b}, s_{c+b+1}, a_c и a_{c+b+1} являются двоичными булевыми переменными, а A_c и A_{c+b+1} – арифметическими, принимающими только два значения: –1 и +1. Таким образом, величина $\Delta MH_{v+1}(S)$ в зависимости от a_c и a_{c+b+1} принимает три возможных значения: –2, 0 и +2. Обобщая полученный результат для MH_v и MH_{v+1} на случай MH_j и MH_l, где $j \neq l \in \{0, 1, 2, \ldots, n-1\}$, можно заключить, что их численные значения будут отличаться на $2d$, где d – целое либо ноль. Тогда в зависимости от того, четное или нечетное значение MH_0, все остальные компоненты меры различия $MH(T_i, T_k)$ будут принимать соответственно четные или нечетные значения и отличаться друг от друга на величину $2d$.

В качестве иллюстрации рассмотренного свойства в табл. 2 приведены значения $\Delta MH_{v+1}(S)$ изменений компоненты $MH_{v+1}(S)$ по отношению к $MH_v(S)$ для серии из двух единиц в наборе $T_k(v)$ и всевозможных комбинаций значений двоичных переменных a_c и a_{c+b+1}. Одноименные биты наборов T_i и $T_k(v)$, а также T_i и $T_k(v+1)$, участвующие в коррекции $MH_{v+1}(S)$ по отношению к $MH_v(S)$, выделены подчеркиванием.

Таблица 2. Значения изменений $\Delta MH_{v+1}(S)$ компоненты $MH_{v+1}(S)$ меры различия $MH(T_i, T_k)$
Table 2. Values of changes $\Delta MH_{v+1}(S)$ components $MH_{v+1}(S)$ measures of difference $MH(T_i, T_k)$

T_i	0<u>0</u>0<u>0</u>0	0<u>0</u>0<u>1</u>110	0<u>0</u>1<u>1</u>10	0<u>1</u>1<u>1</u>0	0<u>1</u>1<u>1</u>0<u>0</u>
$T_k(v)$	0<u>1</u>1<u>0</u>0	0<u>1</u>1<u>0</u>000	0<u>1</u>1<u>0</u>00	0<u>1</u>1<u>0</u>0	00<u>1</u>1<u>0</u>0
$MH_v(S)$	2	5	3	1	1
T_i	0<u>0</u>0<u>0</u>0	0<u>0</u>0<u>1</u>110	0<u>0</u>1<u>1</u>10	0<u>1</u>1<u>1</u>0	0<u>1</u>1<u>1</u>0<u>0</u>
$T_k(v+1)$	0<u>0</u>1<u>1</u>0	0<u>0</u>1<u>1</u>000	0<u>0</u>1<u>1</u>00	0<u>0</u>1<u>1</u>0	00<u>0</u>1<u>1</u>0
$MH_{v+1}(S)$	2	3	1	1	3
$\Delta MH_{v+1}(S)$	0	–2	–2	0	+2

Как видно из табл. 2, во всех случаях изменения компоненты $MH_{v+1}(S)$ по отношению к $MH_v(S)$ соответствуют свойству 2.

Свойство 3. Сумма компонент MH_0, MH_1, …, MH_{n-1} меры различия $MH(T_i, T_k)$ равняется $n(w_i + w_k) - 2w_iw_k$.

Доказательство. Предположим, что двоичный набор T_i содержит w_i единиц и $n - w_i$ нулей, а T_k, соответственно, w_k единиц и $n - w_k$ нулей. Каждая из w_i единиц набора T_i совместно с каждым нулевым значением набора T_k порождает единичное слагаемое для одной из сумм (2). Это объясняется циклическими сдвигами T_k, когда каждый символ одного набора только один раз участвует в качестве слагаемого с каждым символом второго набора при вычислении компонент меры различия $MH(T_i, T_k)$. Количество таких ненулевых слагаемых для единичных символов набора T_i равняется $w_i(n - w_k)$. В свою очередь, число подобных слагаемых за счет единиц набора T_k определяется как $w_k(n - w_i)$. Общее количество единичных значений сумм (2), участвующих в определении всех компонент MH_0, MH_1, …, MH_{n-1}, равняется $w_i(n - w_k) + w_k(n - w_i) = n(w_i + w_k) - 2w_iw_k$.

Например, для T_i = 11110000 и T_k = 10000100 имеем $w_i = 4$, а $w_k = 2$; соответственно, сумма $MH_0 + MH_1 + MH_2 + MH_3 + MH_4 + MH_5 + MH_6 + MH_7 = 8 \cdot (4 + 2) - 2 \cdot 4 \cdot 2 = 32$. Два рассмотренных двоичных набора и их значения компонент меры различия $MH(T_i, T_k)$ приведены в табл. 1.

Важным выводом свойства 3 является постоянство суммы компонент меры $MH(T_i, T_k)$ для двоичных наборов T_i и T_k с весами w_i и w_k независимо от структуры этих наборов. Различие сравниваемых наборов наблюдается на уровне величин компонент MH_0, MH_1, …, MH_{n-1}. Как, например, это видно для двух наборов T_k = 00110011 и T_k = 01010110, которые сравниваются с набором T_i = 11110000 (табл. 1).

Сумма компонент меры различия $MH(T_i, T_k)$, определяемая весами w_k и w_i сравниваемых двоичных наборов T_i и T_k, позволяет вычислить среднее значение $MH_{avr} = MH_{avr}(w_i, w_k, n)$ для различного сочетания аргументов w_i, w_k и n

$$MH_{avr} = MH_{avr}(w_i, w_k, n) = \left(\sum_{j=0}^{n-1} MH_j \right) / n = w_i + w_k - \frac{2w_i w_k}{n}. \quad (5)$$

Для T_i = 11110000 и T_k = 10000100 имеем MH_{avr} = 4, аналогичное значение средней величины компонент (5) меры $MH(T_i, T_k)$ будет и в случае двух других пар наборов, приведенных в табл. 1. Это видно из следствия, вытекающего из соотношения (5), заключающегося в том, что если $w_i = n/2$, то для любого четного n, независимо от величины w_k, $MH_{avr} = n/2$.

В табл. 3 приведены примеры значений компонент меры различия $MH(T_i, T_k)$ для n = 16, которые соответствуют свойству 3 и его следствию. Действительно, средняя величина $MH_{avr} = n/2 = 8$, так как во всех случаях $w_i = n/2 = 8$.

Таблица 3. Численные значения компонент MH_0, MH_1, ..., MH_{15} меры различия $MH(T_i, T_k)$
Table 3. Numerical values of the difference measure $MH(T_i, T_k)$ components MH_0, MH_1, …, MH_{15}

T_i, T_k	MH_0, MH_1, ..., MH_{15}															
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
0000000011111111 1111111100000000	16	14	12	10	8	6	4	2	0	2	4	6	8	10	12	14
0000000011111111 1111111111111111	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8	8
0011001100110011 1100110011001100	16	8	0	8	16	8	0	8	16	8	0	8	16	8	0	8
0011001100110011 0111111111111110	8	10	8	6	8	10	8	6	8	10	8	6	8	10	8	6
0101010101010101 1010101010101010	16	0	16	0	16	0	16	0	16	0	16	0	16	0	16	0
0101010101010101 0000000000000010	9	7	9	7	9	7	9	7	9	7	9	7	9	7	9	7

Анализ численных величин MH_0, MH_1, ..., MH_{15} меры различия $MH(T_i, T_k)$ показывает наличие среди них нулевых значений, что является признаком сопряженных последовательностей символов. Действительно, для T_i = 0000000011111111, предположив, что G = 00000000 и V = 11111111, соответственно имеем $T_i = G$, V и $T_k = V$, G (табл. 3).

Приведенная мера различия $MH(T_i, T_k)$ позволяет идентифицировать подобие структур сравниваемых наборов по равенству нулю ее компонент. Наличие одного нулевого значения $MH_v = 0$ означает, что T_k является копией набора T_i, циклически сдвинутого на v позиций, т. е. $T_i = T_k(v)$. Для первых двух наборов, представленных в табл. 3, имеем $MH_8 = 0$, что свидетельствует об идентичности T_i и $T_k(8)$. Наличие большего количества нулевых значений MH_v свидетельствует о еще большем подобии сравниваемых наборов, заключающемся в идентичности их составных фрагментов.

Также важным свойством численных значений компонент MH_0, MH_1, ..., MH_{n-1} меры различия $MH(T_i, T_k)$ является наличие в них периодически повторяющихся значений. Это свидетельствует о периодичности одного из сравниваемых двоичных наборов. Так, в двух последних примерах, приведенных в табл. 3, один из сравниваемых наборов имеет период значений, равный 2, соответственно, и численные значения меры различия повторяются с таким же периодом.

Применение меры различия для формирования управляемых вероятностных тестов

Как отмечалось ранее, сущность управляемых вероятностных тестов заключается в том, что очередной тестовый набор T_i формируется максимально отличающимся от сгенерированных ранее наборов T_0, T_1, ..., T_{i-1} [7–9]. Для этого на каждом шаге формирования очередного тестового набора осуществляется его выбор из множества кандидатов в тесты. Процедура выбора основана на вычислении численного значения меры различия между наборами T_i и T_k, один из которых, например, первый, является тестовым набором, а другой – одним из кандидатов в тесты. В качестве очередного тестового набора выбирается тот кандидат, для которого величина меры различия принимает максимальное значение. Однако чаще всего этому критерию отвечают несколько кандидатов в тесты, которые являются неразличимыми в рамках используемой меры различия.

В этом случае классическая методика формирования управляемых вероятностных тестов предполагает использование любого из кандидатов в тесты, отвечающего выбранному критерию [7].

Рассмотренная мера различия $MH(T_i, T_k)$ позволяет оценить степень различия двух тестовых наборов T_i и T_k, которые могут быть неразличимыми при использовании других мер различия, например, расстояния Хэмминга. Для подтверждения факта неразличимости тестовых наборов был реализован следующий эксперимент. Для заданного тестового набора T_i, полученного случайным образом, формировалось множество из кандидатов в тесты T_k, которые также генерировались случайным образом по равномерному закону распределения. Затем рассчитывались расстояния Хэмминга $H(T_i, T_k)$ между T_i и всеми тестовыми наборами из списка кандидатов T_k и определялось подмножество кандидатов в тесты, которые имели максимальное значение $H(T_i, T_k) = MH_0$. Отметим, что MH_0 является компонентой предложенной меры различия $MH(T_i, T_k)$. Эксперименты проводились для разных величин разрядности n наборов T_i, T_k и различного количества кандидатов в тесты. В качестве примера в табл. 4 приведены результаты вычисления компонент меры различия $MH(T_i, T_k)$ при $n = 8$. Для набора $T_i = 10000010$ было сформировано 100 кандидатов в тесты T_k, среди которых оказалось восемь наборов T_k с максимальным значением $H(T_i, T_k) = 6$.

Таблица 4. Значения компонент $MH_0, MH_1, \ldots, MH_7$ меры различия $MH(T_i, T_k)$
Table 4. Values of the difference measure $MH(T_i, T_k)$ components $MH_0, MH_1, \ldots, MH_7$

T_k		$T_i = 10000010$			
		$MH_0, MH_1, \ldots, MH_7$	$\Delta MH_v = MH_v - MH_{avr}$	$\Delta MH_1 + \Delta MH_7$	$\Delta MH_2 + \Delta MH_6$
$T_{k,1}$	01101111	6, 4, 6, 6, 4, 6, 4, 4	1, –1, 1, 1, –1, 1, –1, –1		
$T_{k,2}$	10111101	6, 6, 4, 6, 4, 4, 6, 4	1, 1, –1, 1, –1, –1, 1, –1		
$T_{k,3}$	11110101	6, 4, 4, 4, 6, 4, 8, 4	1, –1, –1, –1, 1, –1, 3, –1		
$T_{k,4}$	01010101	6, 2, 6, 2, 6, 2, 6, 2	2, –2, 2, –2, 2, –2, 2, –2	–4	
$T_{k,5}$	11111001	6, 4, 4, 4, 4, 6, 6, 6	1, –1, –1, –1, –1, 1, 1, 1		
$T_{k,6}$	00110101	6, 4, 4, 4, 4, 2, 6, 2	2, 0, 0, 0, 0, –2, 2, –2	–2	2
$T_{k,7}$	11101101	6, 4, 4, 6, 4, 6, 6, 4	1, –1, –1, 1, –1, 1, 1, –1		
$T_{k,8}$	01101001	6, 2, 4, 4, 2, 6, 4, 4	2, –2, 0, 0, –2, 2, 0, 0	–2	0

Соответственно, все восемь кандидатов в тесты оказались неразличимыми по критерию классического расстояния Хэмминга. В то же время значения компонент предложенной меры различия, как это видно из табл. 4, для них существенно разнятся. Анализ кандидатов в тесты, представленных в табл. 4, показывает их различие и по структуре. Наборы $T_{k,1}$, $T_{k,2}$, $T_{k,3}$, $T_{k,5}$ и $T_{k,7}$ имеют значение $w_k = 6$, а $T_{k,4}$, $T_{k,6}$ и $T_{k,8}$ – $w_k = 4$. Соответственно, в первом случае $MH_{avr} = 5$, а во втором – $MH_{avr} = 4$ (5). Для определения наиболее отличающегося по отношению к $T_i = 10000010$ кандидата в тесты приведем для каждого из них различие $\Delta MH_v = MH_v - MH_{avr}$ компонент меры $MH(T_i, T_k)$ по отношению к их средней величине MH_{avr}. Как видно из табл. 4, максимальное отличие для ΔMH_0, равное 2, достигается для $T_{k,4}$, $T_{k,6}$ и $T_{k,8}$. Для дальнейшего уменьшения количества кандидатов в тесты анализируются значения соседних по отношению к ΔMH_0 компонент, а именно: ΔMH_1 и ΔMH_7 (в общем случае ΔMH_{n-1}). Вычисляется их сумма, и по максимальному ее значению определяется тестовый набор. Как видно из табл. 4, максимальное значение $\Delta MH_1 + \Delta MH_7$ достигается для двух наборов $T_{k,6}$ и $T_{k,8}$. Подобная процедура повторяется для ΔMH_2 и ΔMH_6 (в общем случае ΔMH_{n-2}) и т. д. Как видно из рассмотренного в табл. 4 примера, величина $\Delta MH_2 + \Delta MH_6$ позволяет выбрать в качестве тестового набора $T_{k,6} = 00110101$.

Использование компонент MH_1 и MH_{n-1} при неразличимости наборов T_i и T_k по величине MH_0 объясняется их максимальной информативностью о структуре T_k, так как его циклически сдвинутая копия $T_k(1)$ в минимальной мере отличается от оригинала. Очевидно, что для выбора одного из кандидатов в тесты возможно применение и других критериев, основанных на предложенной мере различия.

Эксперимент с формированием тестового набора T_i и списка из 10-ти кандидатов в тесты T_k для $n = 8$ проводился 100 раз. Тестовый набор T_i и кандидаты в тесты T_k генерировались случайным образом по равномерному закону распределения. В каждом эксперименте рассчитывалось значение $H(T_i, T_k) = MH_0$ между T_i и всеми кандидатами. Далее рассматривались только те кандидаты в тесты, для которых MH_0 принимало максимальное значение. Если в списке с максимальным MH_0 оказывалось более одного кандидата, для каждого из них по мере необходимости рас-

считывалось MH_{avr}, вычислялись MH_1, MH_2, …, MH_7, а также ΔMH_v. Для кандидатов, у которых ΔMH_0 принимает наибольшее значение, рассчитывается сумма $\Delta MH_v + \Delta MH_{n-v}$, где $v = 1, \ldots n/2 - 1$. Итерации в рамках одного эксперимента продолжались до тех пор, пока в списке не оставался один кандидат в тесты или v не достигало значения $n/2 - 1$. Результаты экспериментов приведены в табл. 5.

Таблица 5. Результаты экспериментов с применением меры различия $MH(T_i, T_k)$ для $n = 8$
Table 5. Results of experiments using the difference measure $MH(T_i, T_k)$ for $n = 8$

$H(T_i, T_k)$	$MH(T_i, T_k)$			
	ΔMH_0	$\Delta MH_1 + \Delta MH_7$	$\Delta MH_2 + \Delta MH_6$	$\Delta MH_3 + \Delta MH_5$
61	20	15	3	1

Как видно из табл. 5, в 61 случае из проведенных 100 экспериментов кандидат в тесты был определен по значению расстояния Хэмминга. В остальных случаях использовалась предложенная мера различия $MH(T_i, T_k)$, которая позволила определить набор T_k, наиболее отличающийся от T_i по величине ΔMH_0 в 20 случаях, по значению суммы $\Delta MH_1 + \Delta MH_7$ в 15 экспериментах, а также в 3 и 1 случаях, соответственно, по суммам $\Delta MH_2 + \Delta MH_6$ и $MH_3 + \Delta MH_5$. Результаты аналогичных экспериментов для $n = 16$ и $n = 32$ бит, списков из 100 и 1000 кандидатов соответственно для 1000 итераций приведены в табл. 6, 7.

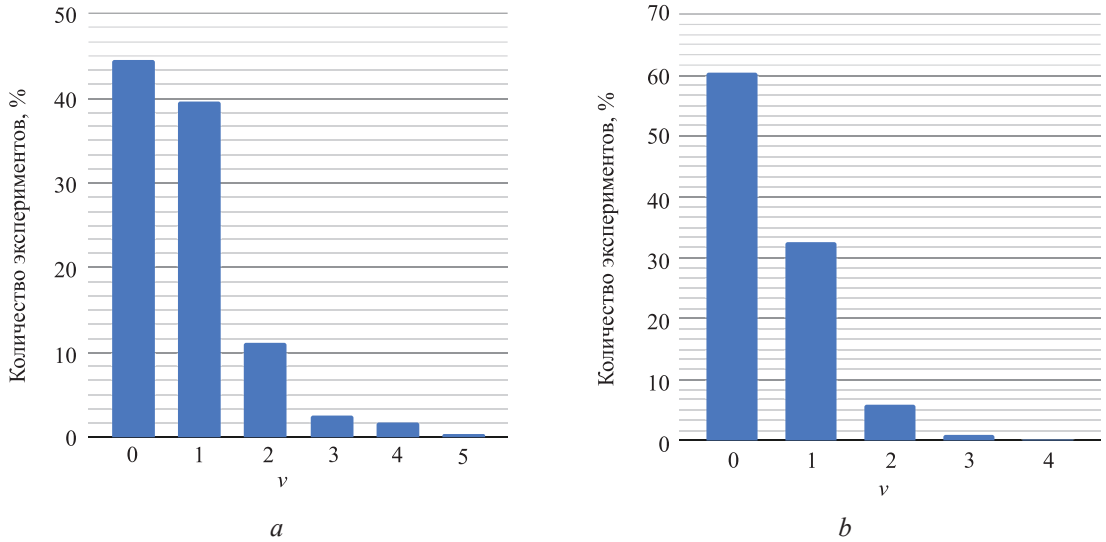
Таблица 6. Результаты экспериментов с применением меры различия $MH(T_i, T_k)$ для $n = 16$
Table 6. Results of experiments using the difference measure $MH(T_i, T_k)$ for $n = 16$

$H(T_i, T_k)$	$MH(T_i, T_k)$					
	ΔMH_0	$\Delta MH_1+\Delta MH_{15}$	$\Delta MH_2+\Delta MH_{14}$	$\Delta MH_3+\Delta MH_{13}$	$\Delta MH_4+\Delta MH_{12}$	$\Delta MH_6+\Delta MH_{11}$
533	208	185	52	12	8	2

Таблица 7. Результаты экспериментов с применением меры различия $MH(T_i, T_k)$ для $n = 32$
Table 7. Results of experiments using the difference measure $MH(T_i, T_k)$ for $n = 32$

$H(T_i, T_k)$	$MH(T_i, T_k)$				
	ΔMH_0	$\Delta MH_1+\Delta MH_{31}$	$\Delta MH_2+\Delta MH_{30}$	$\Delta MH_3+\Delta MH_{29}$	$\Delta MH_4+\Delta MH_{28}$
542	277	149	27	4	1

Графически в процентном отношении результаты экспериментов для $n = 16$ и $n = 32$ представлены на рис. 1. Как видно из полученных данных, наибольшее количество результативного применения меры различия $MH(T_i, T_k)$ достигалось при анализе ΔMH_0 и $\Delta MH_1 + \Delta MH_{n-1}$, т. е. для $v = 0$ и $v = 1$.



a *b*

Рис. 1. Результаты экспериментов с применением меры различия $MH(T_i, T_k)$: *a* – $n = 16$; *b* – $n = 32$
Fig. 1. Results of experiments using the difference measure $MH(T_i, T_k)$: *a* – $n = 16$; *b* – $n = 32$

В исключительных случаях возникает необходимость в классификации кандидатов в тесты на основании предложенной меры различия $MH(T_i, T_k)$ для $v > 3$ (рис. 1), что свидетельствует о ее эффективности и невысокой вычислительной сложности.

Заключение

1. Предложена мера различия для построения управляемых вероятностных тестов, основанная на применении классического расстояния Хэмминга. Показана эффективность ее применения для построения управляемых вероятностных тестов.

2. Дальнейшие исследования целесообразно расширить в части свойств новой меры различия и ее применимости для различных прикладных задач. Наиболее интересным является применение данной меры различия в современных поисковых приложениях с оценкой вычислительной сложности ее получения.

Список литературы / References

1. Hamming R. W. (1950) Error Detecting and Error Correcting Codes. *Bell System Tech. J*. 29, 147–160.
2. Yarmolik V. N., Shevchenko N. A., Petrovskaya V. V. (2022) A Measure of Dissimilarity to Generate Controlled Random Tests. *Doklady BGUIR*. 20 (6), 52–60 (in Russian).
3. Sadovsky M. G. (2002) Comparison of Symbol Sequences: No Editing, No Alignment. *Open Systems & Information Dynamics.* 9 (1), 19–36.
4. Tannga M. J., Rahman S., Hasniati (2017) Comparative Analysis of Levenshnein Distance Algorithm and Jaro Winkler for Text Document Plagiarism Detection Application. *J. of Technology Research in Information System and Engineering.* 4 (2), 44–54.
5. Bookstein A., Klein S. T., Raita T. (2001) Fuzzy Hamming Distance: A New Dissimilarity Measure. *Proceedings of 12th Annual Symposium on Combinatorial Pattern Matching*, *CPM2001*. 1–4.
6. Shallit J. (2009) Hamming Distance for Conjugates. *Discrete Mathematics*. 309 (12), 4197–4199.
7. Yarmolik V. N. (2019) *Monitoring and Diagnostics of Computer Systems.* Minsk, Bestprint Publ. (in Russian).
8. Yarmolik V. N., Mrozek I., Yarmolik S. V. (2015) Controlled Method of Random Test Synthesis. *Automatic Control and Computer Sciences*. 49 (6), 395–403.
9. Levantsevich V. A., Yarmolik V. N. (2019) Multiple Controlled Random Testing. *Doklady BGUIR*. 121 (3), 65–69 (in Russian).
10. Volchikhin V. I., Ivanov A. I., Karpov A. P., Yunin A. P. (2019) Conditions for the Correct Calculation of the Entropy of Meaningful Long Passwords in the Hamming Convolution Space with Reference Texts in Russian and English. *Instruments and Methods of Measurement*. 29 (3), 33–38 (in Russian).

Вклад авторов / Authors' contribution

Авторы внесли равный вклад в написание статьи / The authors contributed equally to the writing of the article.

Сведения об авторах

Ярмолик В. Н., д-р техн. наук, проф., проф. каф. программного обеспечения информационных технологий, Белорусский государственный университет информатики и радиоэлектроники (БГУИР)

Петровская В. В., магистр техн. наук каф. программного обеспечения информационных технологий, БГУИР

Шевченко Н. А., студ., Дармштадтский технический университет

Адрес для корреспонденции

220013, Республика Беларусь,
г. Минск, ул. П. Бровки, 6
Белорусский государственный университет
информатики и радиоэлектроники
Тел.: +375 29 769-96-77
E-mail: yarmolik10ru@yahoo.com
Ярмолик Вячеслав Николаевич

Information about the authors

Yarmolik V.N., Dr. of Sci. (Tech.), Professor, Professor at the Department of Information Technology Software, Belarusian State University of Informatics and Radioelectronics (BSUIR)

Petrovskaya V. V., M. of Sci. at the Department of Information Technology Software, BSUIR

Shauchenka M. A., Student, Darmstadt Technical University

Address for correspondence

220013, Republic of Belarus,
Minsk, P. Brovki St., 6
Belarusian State University
of Informatics and Radioelectronics
Tel.: +375 29 769-96-77
E-mail: yarmolik10ru@yahoo.com
Yarmolik Vyacheslav Nikolaevich