ЛОГИЧЕСКОЕ ПРОЕКТИРОВАНИЕ
LOGICAL DESIGN

УДК 004.33.054
https://doi.org/10.37661/1816-0301-2022-19-4-7-26

Оригинальная статья
Original Paper

Мера различия для тестовых наборов при генерировании управляемых вероятностных тестов

В. Н. Ярмолик[1]✉, В. В. Петровская[1], И. Мрозек[2]

[1]Белорусский государственный университет информатики и радиоэлектроники, ул. П. Бровки, 6, Минск, 220013, Беларусь
✉E-mail: yarmolik10ru@yahoo.com

[2] Белостоцкий технический университет, ул. Вейска, 45А, 15-351, Белосток, Польша

Аннотация

Цели. Решается задача построения характеристик различия тестовых наборов, представляющих собой наборы символов, включая двоичные наборы. Обосновывается ее актуальность для генерирования управляемых вероятностных тестов и сложность нахождения мер различия для символьных тестов. Показывается ограниченность применения расстояния Хэмминга и Дамерау – Левенштейна для получения меры различия тестовых наборов.

Методы. На основе характеристики интервала, применяемого в теории строя цепи последовательных событий, определяется новая мера различия двух символьных тестовых наборов. В качестве меры различия рассчитывается расстояние $AD(T_i, T_k)$ между тестовыми наборами T_i и T_k, использующее характеристику интервала и основанное на определении независимых пар одинаковых (тождественных) символов, принадлежащих двум наборам, и вычислении интервалов между ними.

Результаты. Показывается комбинаторный характер вычисления предложенной меры различия для символьных тестовых наборов произвольного алфавита и размерности. Приводится пример вычисления данной меры для различных видов тестовых наборов, в том числе таких, как адресные тестовые наборы. Показываются возможные ее модификации и определяются некоторые свойства и ограничения. Рассматривается применение данной меры различия для случая многократного тестирования запоминающих устройств на основе адресных последовательностей pA с четным p повторением адресов. Для случая $p = 2$ приводятся математические соотношения вычисления интервалов и расстояния $AD(T_i, T_k)$ для последовательностей адресов $2A$, используемых для управляемого вероятностного тестирования запоминающих устройств. Основное внимание уделяется двоичным тестовым наборам, для которых задача вычисления данной метрики различия сводится к классической задаче о назначениях с использованием венгерского алгоритма. Вычислительная сложность венгерского алгоритма оценивается соотношением $O(n^4)$. Как альтернатива венгерскому алгоритму предлагается алгоритм вычисления рассматриваемой меры, сложность которого существенно меньше и имеет оценку $O(n^2)$. Проведенные экспериментальные исследования подтверждают эффективность рассмотренного алгоритма.

Заключение. Предложенная мера различия расширяет возможности генерирования тестовых последовательностей при генерировании управляемых вероятностных тестов. Показано, что тестовые наборы, неразличимые при использовании в качестве меры различия расстояния Хэмминга, имеют различные значения $AD(T_i, T_k)$, позволяющие более точно классифицировать формируемые случайным образом наборы, которые являются кандидатами в тестовые наборы.

Ключевые слова: управляемые вероятностные тесты, мера различия символьных наборов, адресные последовательности, задача о назначениях, венгерский алгоритм

Благодарности. Авторы выражают искреннюю благодарность старшему преподавателю кафедры программного обеспечения информационных технологий БГУИР Н. С. Петюкевич за участие в обсуждении результатов статьи, советы и рекомендации.

Для цитирования. Ярмолик, В. Н. Мера различия для тестовых наборов при генерировании управляемых вероятностных тестов / В. Н. Ярмолик, В. В. Петровская, И. Мрозек // Информатика. – 2022. − Т. 19, № 4. – С. 7–26. https://doi.org/10.37661/1816-0301-2022-19-4-7-26

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Поступила в редакцию | Received 28.07.2022
Подписана в печать | Accepted 02.09.2022
Опубликована | Published 29.12.2022

A measure of the difference between test sets for generating controlled random tests

Vyacheslav N. Yarmolik[1]✉, Vita V. Petrovskaya[1], Ireneusz Mrozek[2]

[1]Belarusian State University of Informatics and Radioelectronics,
st. P. Brovki, 6, Minsk, 220013, Belarus
✉*E-mail: yarmolik10ru@yahoo.com*

[2]Bialystok University of Technology,
Wiejska, 45A, 15-351, Białystok, Poland

Abstract

Objectives. The problem of constructing the characteristics of the difference between test sequences is solved. Its relevance for generating controlled random tests and the complexity of finding measures of difference for symbolic tests are substantiated. The limitations of using the Hamming and Damerau – Levenshtein distances to obtain a measure of the difference between test sets are shown.

Methods. Based on the characteristic of the interval used in the theory of the chain of successive events, a new measure of the difference between two symbolic test sets is determined. As a difference measure, the distance $AD(T_i, T_k)$ between the test sets T_i and T_k is calculated using the interval characteristic, which is based on determining independent pairs of same (identical) symbols belonging to two sets and calculating the intervals between them.

Results. The combinatorial nature of the calculation of the proposed difference measure for symbolic test sets of an arbitrary alphabet and dimension is shown. An example of calculating this measure for various types of test sets, including such as address test sets, is given. Possible modifications are shown and some properties and limitations are determined. The application of the measure of difference is considered for the case of repeated testing of storage devices based on address sequences pA with even p repetition of addresses. For the case $p = 2$, mathematical relations are given for calculating the intervals and distances $AD(T_i, T_k)$ for address sequences $2A$ used for controlled random testing of storage devices. The main attention is paid to binary test sets, when the task of calculating given difference metric is reduced to the classical assignment problem using the Hungarian algorithm. The computational complexity of the Hungarian algorithm is estimated by the relation $O(n^4)$. As an alternative to the Hungarian algorithm, an algorithm for calculating the considered difference measure is proposed, the complexity of which is much less and has an estimate equal to $O(n^2)$. The experimental studies confirm the effectiveness of the proposed algorithm.

Conclusion. The proposed difference measure extends the possibilities of generating test sequences when generating controlled random tests. It is shown that test sets, which are indistinguishable when Hamming distance is used as a measure of difference, have different values of $AD(T_i, T_k)$ that allows to make more accurate classification of randomly generated sets as candidates for test sets.

Keywords: controlled random tests, character set difference measure, address sequences, assignment problem, Hungarian algorithm

Acknowledgements. The authors express their sincere gratitude to N. S. Petyukevich, Senior Lecturer of the Department of Information Technology Software at BSUIR, for participating in the discussion of the results of the article, advice and recommendations.

For citation. Yarmolik V. N., Petrovskaya V. V., Mrozek I. *A measure of the difference between test sets for generating controlled random tests*. Informatika [*Informatics*], 2022, vol. 19, no. 4, pp. 7–26 (In Russ.). https://doi.org/10.37661/1816-0301-2022-19-4-7-26

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

Введение. Фундаментальный подход к тестированию современных вычислительных систем включает в себя построение тестовых наборов случайным образом из всех возможных входных данных объекта тестирования. Такой подход называется *вероятностным тестированием* (Random Testing) [1, 2]. Вероятностное тестирование зачастую бывает единственно возможным подходом к тестированию не только на стадии эксплуатационного тестирования, где оценивается надежность вычислительной системы, но и на стадии отладочного и других видов тестирования. Вероятностный тест задается количеством q тестовых наборов $T_i = t_{i,0}, t_{i,1}, ..., t_{i,n-1}$, $i \in \{0, 1, \ldots, q-1\}$, данных $t_{i,j}$, $j \in \{0, 1, \ldots, n-1\}$, их множеством, определяемым заданным алфавитом данных и числом n в наборе, а также законом распределения наборов.

Для повышения эффективности вероятностных тестов были предложены различные их модификации, которые получили общее название *управляемые (адаптивные) вероятностные тесты* (Adaptive Random Tests) [1–3]. Основная предпосылка появления подобных тестов заключается в том, что в целях достижения более высокого покрытия неисправностей, обнаруживаемых тестом, и минимизации его временной сложности необходимо целенаправленно выбирать очередной тестовый набор в зависимости от ранее сгенерированных наборов. Существует большое разнообразие как принципов генерирования подобных тестов, так и учета специфики объекта тестирования при формировании тестовых наборов для управляемых вероятностных тестов [3–7]. Под управляемыми вероятностными тестами понимают вероятностные тесты, в которых очередной тестовый набор формируется с учетом ранее сгенерированных наборов и которые формально определяются следующим образом [3, 4, 7]: управляемым вероятностным является тест, состоящий из q наборов $T_i = t_{i,0}, t_{i,1}, ..., t_{i,n-1}$ тестовых данных $t_{i,j}$, $j \in \{0, 1, \ldots, n-1\}$, сгенерированных случайным образом так, что очередной тестовый набор T_i удовлетворяет заданным критериям, полученным на основании ранее сгенерированных наборов $T_0, T_1, ..., T_{i-1}$.

Согласно общепринятому представлению об управляемых вероятностных тестах ключевой особенностью генерирования тестовых наборов T_i является информация, которая извлекается в виде некоторых характеристик (метрик) из ранее сгенерированных тестовых наборов и используется для формирования очередного набора [3, 7]. Основная идея управляемых вероятностных тестов заключается в том, что очередной тестовый набор T_i формируется максимально удаленным (высокой степени различия) от ранее сгенерированных наборов $T_0, T_1, ..., T_{i-1}$ в терминах заранее определенных мер различия. Таким образом, принимается гипотеза, что для двух тестовых наборов T_i и T_k, имеющих минимальное различие, количество обнаруживаемых неисправностей будет минимальным и, наоборот, для максимально различных тестовых наборов обнаруживающая способность будет максимальной [3, 7]. В качестве меры различия тестового набора T_i от предыдущих наборов $T_0, T_1, T_2, ..., T_{i-1}$ чаще всего используются *расстояние Хемминга* и *расстояние Евклида* [3–9]. Эти характеристики наиболее эффективны для двоичного случая тестовых данных и малопродуктивны для произвольных данных [10, 11].

Проблема сравнения тестовых наборов, которые в общем случае представляют собой символьные последовательности, актуальна для различных областей науки. В работе [10] показано, что в пространстве символьных последовательностей сложно ввести метрику различия. Формально подобная метрика в таком пространстве существует – это *расстояние Хэмминга* [3, 10, 11]. Отметим, что расстояние Хэмминга $HD(T_i, T_k)$ между двумя наборами T_i и T_k равняется числу их несовпадающих компонент (символов) $t_{i,j}$ и $t_{k,j}$. В ряде работ отмечалось, что эта метрика малоэффективна, так как позволяет различать лишь полностью совпадающие последовательности

при $HD(T_i, T_k) = n$ и все остальные несовпадающие [10–13]. В настоящее время чаще всего обсуждается метод сравнения символьных последовательностей, основанный на их выравнивании (метод *редакционного расстояния*), и его модификации [10]. Наиболее известными из них являются методы, использующие *расстояние Дамерау – Левенштейна* (Damerau – Levenshtein distance) [14]. Они основаны на сравнении одной последовательности символов T_i с другой T_k с помощью операций вставки, замены (либо удаления) и транспозиции так, чтобы эти две последовательности совпали. Недостатком методов выравнивания является необходимость назначения и обоснования системы штрафных (весовых) функций и выбора опорной последовательности, относительно которой проводится выравнивание. Кроме того, их эффективность снижается экспоненциально при увеличении длины строк символов [10, 14]. Оценку схожести (подобия) символьных последовательностей можно получить, используя *сходство Джаро – Винклера* (Jaro – Winkler distance), представляющее собой меру схожести строк для измерения расстояния между двумя последовательностями символов [15]. Чем меньше расстояние Джаро – Винклера для двух строк, тем больше сходства между этими строками. Данная мера расстояния чаще всего применяется для оценки схожести наборов символов, а не их различия и также характеризуется достаточно большой вычислительной сложностью.

Основной проблемой управляемых вероятностных тестов является их большая вычислительная сложность, связанная с необходимостью перечисления возможных тестовых наборов и вычисления меры различия для каждого потенциального кандидата в тесты [10–13]. Чаще всего при формировании очередного тестового набора первоначально генерируется достаточное количество кандидатов в тесты (ориентировочно в диапазоне от 10 до 100), представляющих собой равномерно распределенные случайные наборы. Для каждого из них вычисляются метрики различия, с учетом которых и выбирается наилучший из кандидатов в тесты в качестве очередного тестового набора T_i.

Таким образом, главная задача управляемого вероятностного тестирования состоит в нахождении меры различия для тестовых наборов T_i и T_k, которая максимально адекватно показывает их различия и характеризуется невысокой вычислительной сложностью. Вычисление мер различия тестовых наборов, в общем случае представляющих собой символьные последовательности, в свою очередь, сводится к задаче их сравнения [10].

1. Мера различия для построения управляемых вероятностных тестов. Рассматриваемая мера различия основана на *теории строя*, которая предназначена для формального описания и анализа последовательностей данных (символов) любой природы [16]. В общем случае строй может представлять собой последовательность символов (*цепь событий*) любого алфавита, произвольной структуры и длины, а его анализ может осуществляться по различным правилам. Однако основной числовой характеристикой строя является *интервал*, определяемый как расстояние от выделенного в строе символа до другого ближайшего, отмеченного в направлении просмотра такого же символа. Содержание характеристики интервала весьма близко к *числу транспозиций*, используемому для вычисления расстояния Джаро – Винклера [15].

Характеристика интервала, которая служит основой для оценки различных свойств последовательностей символов, была использована для определения меры различия (степени несовпадения) двух тестовых наборов, показывая их удаленность (либо близость) друг от друга [11]. В общем случае для тестовых наборов T_i и T_k, каждый из которых состоит из n_i и n_k данных $t_{i,j}$, $j \in \{0, 1, \ldots, n_i - 1\}$, и $t_{k,r}$, $r \in \{0, 1, \ldots, n_k - 1\}$, интервалом для пары $t_{i,j} = t_{k,r}$ является значение $D(t_{i,j}, t_{k,r})$, которое в дальнейшем будем отождествлять с расстоянием между данными $t_{i,j}$ и $t_{k,r}$. Предполагая, что определяется расстояние (интервал) между $t_{i,j}$ и $t_{k,r}$, первоначально вычисляются значения $|j - r|$ и $\max(n_i, n_k) - |j - r|$. Минимальное $\min[|j - r|, \max(n_i, n_k) - |j - r|]$ из приведенных значений принимается в качестве расстояния $D(t_{i,j}, t_{k,r})$ между $t_{i,j}$ и $t_{k,r}$.

Как отмечалось ранее, подобная оценка расстояния необходима для синтеза управляемых вероятностных тестов, когда очередной тестовый набор формируется максимально удаленным от ранее сгенерированных наборов. Формально эта характеристика описана в работе [11] и соответствует следующему определению.

Определение 1. Мера различия $AD(T_i, T_k)$ тестовых наборов T_i и T_k, каждый из которых состоит из n_i и n_k данных $t_{i,j}, j \in \{0, 1, \ldots, n_i - 1\}$, и $t_{k,r}, r \in \{0, 1, \ldots, n_k - 1\}$, использующая характеристику интервала, основана на определении независимых пар одинаковых данных, принадлежащих двум наборам. Независимость пар означает участие каждого значения данных $t_{i,j}$ и $t_{k,r}$ тестовых наборов T_i и T_k только в одной паре. Процедура формирования подобных пар носит комбинаторный характер и заключается в нахождении такого их сочетания, для которого сумма расстояний $D(t_{i,j}, t_{k,r})$ также минимальна. При отсутствии пары для очередного значения данных $t_{i,j}$ в наборе T_k разность величин индексов, т. е. расстояние $D(t_{i,j}, –)$, принимается равным $\min(n_i, n_k)$. Такое же значение расстояний задается и для данных набора T_k, для которых отсутствует пара в T_i.

Минимальное и максимальное значения меры различия $AD(T_i, T_k)$ зависят от алфавита данных $t_{i,j}$ и $t_{k,r}$ и соотношения их значений в тестовых наборах T_i и T_k, а также размерности n_i и n_k сравниваемых наборов. Минимальное значение метрики $AD(T_i, T_k)$ определяется выражением $\min AD(T_i, T_k) = |n_i - n_k| \cdot \min(n_i, n_k)$, что свидетельствует о максимальном сходстве двух наборов T_i и T_k. При выполнении равенства $n_i = n_k$ значение $\min AD(T_i, T_k)$ равняется нулю, что свидетельствует о тождественности сравниваемых наборов. Максимальное значение $AD(T_i, T_k)$ определяется как $\max AD(T_i, T_k) = (n_i + n_k)\cdot\min(n_i, n_k)$, что свидетельствует о полном (максимальном) отличии наборов T_i и T_k. Чем ближе численное значение $AD(T_i, T_k)$ к его максимальной величине $\max AD(T_i, T_k)$, тем больше степень различия между тестовыми наборами T_i и T_k, и, наоборот, близость этой характеристики к $\min AD(T_i, T_k)$ говорит о их максимальном сходстве. В качестве примера рассмотрим тестовые данные, представляющие собой адресные тестовые наборы, удовлетворяющие следующему определению [7, 17].

Определение 2. Под тестовым набором адресов $T_i = t_{i,0}, t_{i,1}, \ldots, t_{i,n-1}$ (последовательностью адресов A) понимают последовательность из $n = 2^m$ m-битовых векторов $t_{i,j} \in \{0, 1, \ldots, 2^m - 1\}$, $j \in \{0, 1, \ldots, 2^m - 1\}$, каждый из которых принимает одно из 2^m возможных значений.

Приведенное выше определение меры различия позволяет сравнить два тестовых набора адресов T_i и T_k. В качестве примера набора T_i рассмотрим счетчиковую последовательность для $m = 3$, а в качестве T_k используем последовательность отраженного кода Грея (табл. 1) [18].

Таблица 1
Примеры адресных тестовых последовательностей

Table 1
Examples of address test sequences

T_i	j	0	1	2	3	4	5	6	7
	$t_{i,j}$	000	001	010	011	100	101	110	111
T_k	r	0	1	2	3	4	5	6	7
	$t_{k,r}$	000	001	011	010	110	111	101	100

Анализ тестовых наборов, приведенных в табл. 1, показывает их одинаковую размерность $n_i = n_k = n = 2^m = 8$ и структуру, характеризующуюся наличием в каждом наборе по одному трехбитовому вектору из восьми возможных. Таким образом, для каждого данного в тестовых последовательностях адресов T_i и T_k существует пара, и она единственная. Для $t_{i,0} = 000$ набора T_i тождественное значение 000 в наборе T_k имеет тот же индекс $r = j = 0$. Соответственно, расстояние $D(t_{i,0}, t_{k,0})$, равное разности их индексов, принимает значение 0, т. е. $D(t_{i,0}, t_{k,0}) = 0$. Аналогично и для 001 имеем $D(t_{i,1}, t_{k,1}) = 0$. Для $t_{i,2} = t_{k,3} = 010$ существуют уже два значения положительных разностей индексов $|2 - 3| = 1$ и $8 - |2 - 3| = 7$, минимальное из которых и определяет расстояние $D(t_{i,2}, t_{k,3}) = 1$. Продолжая пример, приведенный в табл. 1, получим $D(t_{i,3}, t_{k,2}) = \min(|3 - 2|, 8 - |3 - 2|) = 1$. Далее имеем $D(t_{i,4}, t_{k,7}) = \min(|4 - 7|, 8 - |4 - 7|) = 3$; $D(t_{i,5}, t_{k,6}) = \min(|5 - 6|, 8 - |5 - 6|) = 1$; $D(t_{i,6}, t_{k,4}) = \min(|6 - 4|, 8 - |6 - 4|) = 2$ и $D(t_{i,7}, t_{k,5}) = \min(|7 - 5|, 8 - |7 - 5|) = 2$. Окончательно для рассматриваемого примера мера различия $AD(T_i, T_k)$ вычисляется как $AD(T_i, T_k) = D(t_{i,0}, t_{k,0}) + D(t_{i,1}, t_{k,1}) + D(t_{i,2}, t_{k,3}) + D(t_{i,3}, t_{k,2}) + D(t_{i,4}, t_{k,7}) + D(t_{i,5}, t_{k,6}) + D(t_{i,6}, t_{k,4}) + D(t_{i,7}, t_{k,5}) = 0 + 0 + 1 + 1 + 3 + 1 + 2 + 2 = 10$.

Мера различия $AD(T_i, T_k)$ для двух произвольных T_i и T_k тестовых адресных последовательностей, удовлетворяющих определению 2, вычисляется согласно соотношению [11]

$$AD(T_i, T_k) = \sum_{j=0}^{2^m-1} \sum_{r=0}^{2^m-1} \mathrm{I}(t_{i,j} = t_{k,r}) \cdot \min\left[\left| j - r \right|, 2^m - \left| j - r \right| \right]. \quad (1)$$

Выражение $\mathrm{I}(t_{i,j} = t_{k,r})$ представляет собой индикаторную функцию, равную единице при $t_{i,j} = t_{k,r}$ и нулю в противном случае. Минимальное значение $\min AD(T_i, T_k) = 0$ при совпадении последовательностей T_i и T_k, а максимальное значение $\max AD(T_i, T_k) = 2^{2m-1}$ для случая максимальных величин $\min(|j - r|, 2^m - |j - r|) = 2^{m-1}$ для всех данных в сравниваемых последовательностях.

В ряде случаев формирование тестового набора адресов T_k, функционально зависимого от набора T_i, позволяет аналитически вычислять значение меры различия $AD(T_i, T_k)$. Например, в случае генерирования набора $T_k = T_i(s)$ – циклически сдвинутой вправо на $s \in \{0, 1, \ldots, 2^m-1\}$ позиций копии исходного набора T_i – существенно упрощается соотношение (1). Соответственно, получим равенство

$$AD(T_i, T_i(s)) = \begin{cases} s \cdot 2^m, s = 0,1, ...,2^m-1; \\ (2^m-1 - s \bmod 2^m-1) \cdot 2^m, s = 2^m-1 + 1, 2^m-1 + 2, ..., 2^m - 1. \end{cases} \quad (2)$$

Очевидно, что $\min AD(T_i, T_i(s)) = 0$ при $s = 0$ и $\max AD(T_i, T_i(s)) = 2^{2m-1}$ при $s = 2^{m-1}$.

Рассматриваемая мера различия основана на определении независимых пар тождественных данных, принадлежащих наборам T_i и T_k. Структура адресных тестовых наборов, соответствующих определению 2, упрощает задачу нахождения независимых пар, так как для каждого $t_{i,j}$ набора T_i всегда существует единственное значение $t_{k,r} = t_{i,j}$ в наборе T_k, где $j, r \in \{0, 1, \ldots, 2^m-1\}$, что и определяет независимость пар тождественных данных.

Новым развитием адресных последовательностей являются последовательности pA для произвольного четного p, удовлетворяющие следующему определению [19, 20].

Определение 3. Под тестовым набором адресов $T_i = t_{i,0}, t_{i,1}, \ldots, t_{i,n-1}$ (последовательностью pA) понимают последовательность из $n = p2^m$ m-битовых векторов $t_{i,j} \in \{0, 1, \ldots, 2^m-1\}$, $j \in \{0, 1, \ldots, p2^m-1\}$, каждый из которых формируется ровно p раз.

Для случая адресных наборов T_i и T_k, соответствующих определению 3, вычисление метрики $AD(T_i, T_k)$ сопряжено с нахождением минимальной суммы минимальных расстояний для всех сочетаний p независимых пар одинаковых значений $t_{i,j} = t_{k,r}$ адресов $t_{i,j}, t_{k,r} \in \{0, 1, \ldots, 2^m-1\}$; $j, r \in \{0, 1, \ldots, p2^m-1\}$, входящих в T_i и T_k. Основная сложность определения данной метрики заключается в получении $p!$ сочетаний пар для каждого адреса $t_{i,j} = t_{k,r}$ в T_i и T_k. Для $p = 2$ количество подобных сочетаний равняется двум. Рассмотрим задачу вычисления $AD(T_i, T_k)$ для случая последовательностей $2A$, представленных в табл. 2.

Таблица 2
Примеры T_i и T_k тестовых последовательностей адресов $2A$

Table 2
Examples of 2A test sequences for addresses T_i and T_k

T_i	j	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	$t_{i,j}$	000	001	010	011	000	001	010	011	100	101	110	111	100	101	110	111
T_k	r	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
	$t_{k,r}$	110	111	101	100	101	000	001	011	010	000	010	011	100	110	001	111

В приведенных тестовых наборах адресов T_i и T_k каждый из восьми адресов повторяется дважды и каждый из них входит в два возможных сочетания пар адресов. Например, адрес 010

входит в последовательность T_i как два ее адреса $t_{i,2}$ и $t_{i,6}$, а в последовательность T_k, соответственно, как $t_{k,8}$ и $t_{k,10}$. Таким образом, первое и второе сочетания пар адресов имеют следующий вид: $\{(t_{i,2}, t_{k,8}), (t_{i,6}, t_{k,10})\}$ и $\{(t_{i,6}, t_{k,8}), (t_{i,2}, t_{k,10})\}$.

В общем случае для каждого значения адреса $A \in \{0, 1, 2, \ldots, 2^m-1\}$ в адресных последовательностях T_i и T_k при $p = 2$ используются четыре одинаковых значения $t_{i,j} = t_{i,l} = t_{k,r} = t_{k,g}$, образующие два сочетания пар адресов. Следовательно, можно предположить, что пара $\{(t_{i,j}, t_{k,r}), (t_{i,l}, t_{k,g})\}$ представляет собой первое сочетание, а пара $\{(t_{i,l}, t_{k,r}), (t_{i,j}, t_{k,g})\}$ – второе. Каждое сочетание пар адресов характеризуется суммой минимальных интервалов для каждой пары сочетания: в первом случае $D(t_{i,j}, t_{k,r}) + D(t_{i,l}, t_{k,g})$ и во втором $D(t_{i,l}, t_{k,r}) + D(t_{i,j}, t_{k,g})$. Далее для каждого адреса $A = t_{i,j} = t_{i,l} = t_{k,r} = t_{k,g}$ определяется минимальное значение суммы минимальных интервалов согласно выражению [11]

$$D(A) = \min \begin{Bmatrix} \min[\,|j-r|, 2^{m+1} - |j-r|\,] + \min[\,|l-g|, 2^{m+1} - |l-g|\,], \\ \min[\,|l-r|, 2^{m+1} - |l-r|\,] + \min[\,|j-g|, 2^{m+1} - |j-g|\,] \end{Bmatrix}. \quad (3)$$

Значение 2^{m+1} в выражении (3) представляет собой размерность n последовательностей T_i и T_k, которая для примеров, приведенных в табл. 2, принимает значение $2^{3+1} = 16$. Соотношение $\min[|j - r|, 2^{m+1}-|j - r]$ в выражении (3) является минимальным расстоянием $D(t_{i,j}, t_{k,r})$ между адресами $t_{i,j}$ и $t_{k,r}$.

Ранее показывалось (см. табл. 2), что адрес 010 образует в последовательностях T_i и T_k два сочетания одинаковых адресов $\{(t_{i,2}, t_{k,8}), (t_{i,6}, t_{k,10})\}$ и $\{(t_{i,6}, t_{k,8}), (t_{i,2}, t_{k,10})\}$. Соответствующее расстояние между адресами пары $(t_{i,2}, t_{k,8})$ первого сочетания принимает значение $D(t_{i,2}, t_{k,8}) =$ $= \min(|2-8|, 16-|2-8|) = \min(6, 10) = 6$, а второй пары – значение $D(t_{i,6}, t_{k,10}) = \min(|6-10|, 16-|6-10|) = \min(4, 12) = 4$. Для второго сочетания пар адресов 010 получим: $D(t_{i,6}, t_{k,8}) =$ $= \min(|6-8|, 16-|6-8|) = \min(2, 14) = 2$; $D(t_{i,2}, t_{k,10}) = \min(|2-10|, 16-|2-10|) = \min(8, 8) = 8$. Далее для каждого сочетания в соответствии с выражением (3) вычисляется сумма полученных расстояний, которая для первого сочетания адресов 010 равняется $D(t_{i,2}, t_{k,8}) + D(t_{i,6}, t_{k,10}) = 6 + 4 = 10$, а для второго – $D(t_{i,6}, t_{k,8}) + D(t_{i,2}, t_{k,10}) = 2 + 8 = 10$. Окончательно для адреса 010 согласно (3) получим значение $D(010) = 10$. Для адреса 000 существуют два сочетания $\{(t_{i,0}, t_{k,5}), (t_{i,4}, t_{k,9})\}$ и $\{(t_{i,4}, t_{k,5}), (t_{i,0}, t_{k,9})\}$ (см. табл. 2). Соответствующие интервалы для каждой из пар двух сочетаний принимают значения $D(t_{i,0}, t_{k,5}) = 5$, $D(t_{i,4}, t_{k,9}) = 5$, $D(t_{i,4}, t_{k,5}) = 1$, $D(t_{i,0}, t_{k,9}) = 7$, тогда $D(000) = 8$. Для рассмотренного выше примера $AD(T_i, T_k) = D(000) + D(001) + D(010) + D(011) + D(100) +$ $+ D(101) + D(110) + D(111) = 8 + 4 + 10 + 8 + 5 + + 10 + 5 + 6 = 56$.

Из приведенного примера видно, что в общем случае для наборов адресов T_i и T_k, соответствующих определению 3, метрика расстояния $AD(T_i, T_k)$, показывающая степень их различия, определяется как минимальная сумма сумм минимальных расстояний $D(A)$ (3) для всех значений адресов $\{0, 1, \ldots, 2^m-1\}$. Расстояние $AD(T_i, T_k)$ для произвольных последовательностей T_i и T_k, удовлетворяющих условиям определения 3, по аналогии со схожей метрикой, рассмотренной в работах [19, 20], принимает значения в диапазоне $0 \le AD(T_i, T_k) \le 2^{2m}$. Минимальное значение $\min AD(T_i, T_k) = 0$ достигается для $T_i = T_k$. Оценка максимального значения $AD(T_i, T_k)$ основана на анализе расстояния между адресами $t_{i,j} = t_{i,l} = t_{k,g} = t_{k,r}$. Очевидно, что $D(t_{i,j}, t_{i,l})$ и $D(t_{k,r}, t_{k,g})$ для произвольных тождественных адресов $t_{i,j}$ и $t_{i,l}$ в T_i , а также $t_{k,r}$ и $t_{k,g}$ в T_k не превышают 2^m. Задача максимизации согласно выражению (3), по сути, заключается во взаимном расположении $t_{i,j} = t_{i,l} = t_{k,g} = t_{k,r}$ в пространстве значений индексов j, l, g и r. Предположив, что $j < g < l < r$, максимизация значения (3) достигается для $l - j = r - g = 2^m$ и $g - j = r - l = 2^{m-1}$. Таким образом, для каждого адреса $\{0, 1, \ldots, 2^m-1\}$ максимальное значение расстояния (3) равняется 2^m. Учитывая, что для тестовых наборов адресов, удовлетворяющих определению 3 для $p = 2$, их количество равняется 2^m, окончательно получим $\max AD(T_i, T_k) = 2^{2m}$.

Для рассмотренного выше примера $AD(T_i, T_k) = 56$, а $\max AD(T_i, T_k) = 2^{2 \cdot 3} = 64$, что свидетельствует о большой степени различия адресных тестовых последовательностей T_i и T_k, представленных в табл. 2.

2. Мера различия $AD(T_i, T_k)$ для двоичного случая. Более широкая трактовка тестового набора по сравнению с тестовыми наборами адресов существенно усложняет процедуру вычисления меры различия, определенную для общего случая. Даже для двоичных тестовых данных комбинаторный характер определения этой меры значительно увеличивает временную сложность ее вычисления. В качестве примера, иллюстрирующего данное утверждение, рассмотрим случай двоичных тестовых наборов $T_i = t_{i,0}, t_{i,1}, ..., t_{i,n-1}$ и $T_k = t_{k,0}, t_{k,1}, ..., t_{k,n-1}$, $i, k \in \{0, 1, \ldots, q-1\}$, $t_{i,j}, t_{k,r} \in \{0, 1\}$, приведенных в табл. 3 для величины $n = 8$.

Таблица 3
Пример двоичных тестовых последовательностей

Table 3
Example of binary test sequences

T_i	j	0	1	2	3	4	5	6	7
	$t_{i,j}$	0	0	1	1	1	1	1	1
T_k	r	0	1	2	3	4	5	6	7
	$t_{k,r}$	1	0	1	1	1	0	1	1

В силу того что наборы T_i и T_k состоят из двоичных данных, в общем случае существует большое количество сочетаний пар тождественных данных. Для примера, приведенного в табл. 3, имеются два сочетания данных $\{(t_{i,0}, t_{k,1}), (t_{i,1}, t_{k,5})\}$ и $\{(t_{i,0}, t_{k,5}), (t_{i,1}, t_{k,1})\}$ для нулевого их значения $t_{i,j} = t_{k,r} = 0$. Для данных $t_{i,j} = t_{k,r} = 1$ число сочетаний пар единичных величин в наборах T_i и T_k равняется 6!, где 6 представляет собой число единиц в T_i и T_k. Далее из всевозможных сочетаний пар (как нулевых значений данных, так и единичных) необходимо выбрать по одному сочетанию. Для этих пар сумма их минимальных положительных разностей индексов (интервалов) также должна быть минимальной, как следует из определения 1. Используя ранее введенную метрику расстояния $D(t_{i,j}, t_{k,r}) = \min(|j - r|, n - |j - r|)$ между $t_{i,j}$ и $t_{k,r}$, для сочетания пар $\{(t_{i,0}, t_{k,1}), (t_{i,1}, t_{k,5})\}$ получим $D(t_{i,0}, t_{k,1}) = \min(|0 - 1|, 8 - |0 - 1|) = 1$ и $D(t_{i,1}, t_{k,5}) = \min(|1 - 5|, 8 - |1 - 5|) = 4$. Соответственно, $D(t_{i,0}, t_{k,1}) + D(t_{i,1}, t_{k,5}) = 1 + 4 = 5$. В то же время для другого сочетания пар $\{(t_{i,0}, t_{k,5}), (t_{i,1}, t_{k,1})\}$ нулевых значений данных в наборах T_i и T_k получим $D(t_{i,0}, t_{k,5}) + D(t_{i,1}, t_{k,1}) = \min(|0 - 5|, 8 - |0 - 5|) + \min(|1 - 1|, 8 - |1 - 1|) = 3$. Отсюда следует, что при определении $AD(T_i, T_k)$ для примера, приведенного в табл. 3, будет учтено второе сочетание пар $\{(t_{i,0}, t_{k,5}), (t_{i,1}, t_{k,1})\}$, для которого сумма минимальных расстояний $D(t_{i,j}, t_{k,r})$, принимающая значение 3, также минимальна. Минимальная сумма минимальных расстояний для единичных данных в рассматриваемом примере определяется путем анализа сумм расстояний для 6! сочетаний пар. Примерами подобных сочетаний могут быть следующие два сочетания: $\{(t_{i,2}, t_{k,2}), (t_{i,3}, t_{k,3}), (t_{i,4}, t_{k,4}), (t_{i,5}, t_{k,6}), (t_{i,6}, t_{k,7}), (t_{i,7}, t_{k,0})\}$ и $\{(t_{i,2}, t_{k,3}), (t_{i,3}, t_{k,4}), (t_{i,4}, t_{k,6}), (t_{i,5}, t_{k,7}), (t_{i,6}, t_{k,0}), (t_{i,7}, t_{k,2})\}$. Для первого сочетания сумма минимальных расстояний равняется 0+0+0+1+1+1=3, а для второго, соответственно, 1+1+2+2+2+3=11. Подобным образом анализируются все сочетания пар единичных данных. Оказывается, что минимальное значение минимальных сумм так же, как и для нулевых данных, равняется трем и, соответственно, $AD(T_i, T_k) = 6$.

Рассмотренный пример основан на двоичных тестовых наборах, состоящих из одинакового числа единичных и нулевых значений данных $t_{i,j}$ и $t_{k,r}$, что является маловероятным случаем. Чаще всего эти числа не будут совпадать, что приведет к появлению данных, например, в наборе T_i, для которых будут отсутствовать пары в наборе T_k, и наоборот. Примером, иллюстрирующим подобную ситуацию, могут быть наборы T_i = 10000000 и T_k = 01111111, для которых можно выделить по одной паре как нулевых, так и единичных данных в выбранном сочетании пар. Следуя определению 1, значение меры различия $AD(T_i, T_k)$ определяется суммой $D(t_{i,0}, t_{k,1}) + D(t_{i,1}, t_{k,0}) + D(t_{i,2}, -) + D(t_{i,3}, -) + D(t_{i,4}, -) + D(t_{i,5}, -) + D(t_{i,6}, -) + D(t_{i,7}, -) + D(t_{k,2}, -) + D(t_{k,3}, -) + D(t_{k,4}, -) + D(t_{k,5}, -) + D(t_{k,6}, -) + D(t_{k,7}, -)$, где значения расстояния $D(t_{i,j}, t_{k,r})$ для данных, не входящих в выбранные сочетания пар, принимаются равными $\min(n_i, n_k) = 8$ согласно определению 1 [11], что приводит к существенному увеличению меры $AD(T_i, T_k)$. При этом значимость расстояния между парами идентичных данных будет иметь меньшее влияние на конечное значение суммарной характеристики. Поэтому для двоичного случая примем следующие допуще-

ния. Не нарушая общности рассуждений, будем считать, что $n_i = n_k = n$ и его значение четно. В силу равенства размерностей наборов T_i и T_k количество данных $t_{i,j}$ и $t_{k,r}$, не имеющих пары в другом наборе, будет одинаковым. Поэтому в качестве слагаемых для пар данных, не вошедших в пары идентичных данных, т. е. с несовпадающими значениями, примем величину расстояния, равную $n/2$. Отметим, что $n/2$ представляет собой максимально возможное значение расстояния для пары идентичных данных.

Формально значения расстояний $D(t_{i,j}, t_{k,r}) = \min(|j - r|, n - |j - r|)$ между тождественными данными $t_{i,j} = t_{k,r}$ можно представить в виде двухмерной матрицы их величин (табл. 4).

Таблица 4
Значения расстояний $D(t_{i,j}, t_{k,r})$ между $t_{i,j} = t_{k,r}$ для четного n

Table 4
The values of the distances $D(t_{i,j}, t_{k,r})$ between $t_{i,j} = t_{k,r}$ for even n

Набор *Pattern*	T_i									
T_k	Данные *Data*	$t_{i,0}$	$t_{i,1}$	$t_{i,2}$	...	$t_{i,n/2-1}$	$t_{i,n/2}$	$t_{i,n/2+1}$	...	$t_{i,n-2}$	$t_{i,n-1}$
	$t_{k,0}$	0	1	2	…	$n/2-1$	$n/2$	$n/2-1$	…	2	1
	$t_{k,1}$	1	0	1	...	$n/2-2$	$n/2-1$	$n/2$		3	2
	$t_{k,2}$	2	1	0	…	$n/2-3$	$n/2-2$	$n/2-1$	…	4	3
	…	…	…	…	…	…	…	…	…	…	…
	$t_{k,n/2-1}$	$n/2-2$	$n/2-3$	$n/2-4$	…	0	1	2	…	$n/2$	$n/2-1$
	$t_{k,n/2}$	$n/2-1$	$n/2-2$	$n/2-3$	...	1	0	1	...	$n/2-1$	$n/2$
	$t_{k,n/2+1}$	$n/2$	$n/2-1$	$n/2-2$	…	2	1	0	…	$n/2-2$	$n/2-1$
	…	…	…	…	…	…	…	…	…	…	…
	$t_{k,n-2}$	2	3	4	…	$n/2-1$	$n/2-2$	$n/2-3$	…	0	1
	$t_{k,n-1}$	1	2	3	…	$n/2$	$n/2-1$	$n/2-2$	…	1	0

Структура данных приведенной матрицы расстояний размерностью $n \times n$ показывает их регулярность, а численное значение на пересечении столбца, помеченного $t_{i,j}$, и строки, соответствующей $t_{k,r}$, представляет собой значение $D(t_{i,j}, t_{k,r}) = \min(|j - r|, n - |j - r|)$ для $t_{i,j} = t_{k,r}$. Из табл. 4 видно, что, например, при $j = r$ $D(t_{i,j}, t_{k,r}) = 0$.

Очевидно, что минимальное значение меры различия $AD(T_i, T_k)$ достигается в случае тождественности наборов T_i и T_k и для принятых допущений равенства $n_i = n_k = n$ равняется нулю, так как $|n_i - n_k|$ будет равняться нулю и, соответственно, $\min AD(T_i, T_k) = |n_i - n_k| \cdot \min(n_i, n_k) = 0$ (см. определение 1). Для получения $\min AD(T_i, T_k) = 0$ выбирается совокупность пар совпадающих данных $t_{i,j} = t_{k,r}$ с $D(t_{i,j}, t_{k,r}) = 0$ (табл. 5) для случая идентичных наборов $T_i = T_k = 11101001$. Диагональные нулевые значения и определяют пары совпадающих данных, принимающих участие в совокупности пар для определения значения $AD(T_i, T_k)$.

Таблица 5
Значения расстояний $D(t_{i,j}, t_{k,r})$ между $t_{i,j} = t_{k,r}$ для $n = 8$

Table 5
The values of the distances $D(t_{i,j}, t_{k,r})$ between $t_{i,j} = t_{k,r}$ for $n = 8$

Набор *Pattern*	T_i								
T_k	Данные *Data*	**1**	**1**	**1**	**0**	**1**	**0**	**0**	**1**
	1	0	1	2		4			1
	1	1	0	1		3			2
	1	2	1	0		2			3
	0				0		2	3	
	1	4	3	2		0			3
	0				2		0	1	
	0				3		1	0	
	1	1	2	3		3			0

При заполнении матрицы расстояний указываются только расстояния $D(t_{i,j}, t_{k,r})$ для $t_{i,j} = t_{k,r}$, соответствующие значениям для общего случая (см. табл. 4). При их несовпадении, т. е. при $t_{i,j} \neq t_{k,r}$, соответствующий элемент матрицы остается пустым.

В терминах меры различия согласно определению 1 полностью различными, несовпадающими наборами T_i и T_k считаются те, в которых отсутствуют данные, используемые в другом наборе. Для рассматриваемого случая двоичных тестовых наборов T_i и T_k максимально различными являются наборы $T_i = 000\ldots0$ и $T_k = 111\ldots1$ либо, наоборот, $T_i = 111\ldots1$ и $T_k = 000\ldots0$, для которых $AD(T_i, T_k) = \max AD(T_i, T_k) = n(n/2) = n^2/2$. Последнее соотношение получено на основании ранее принятых допущений о максимальной удаленности на $n/2$ позиций данных в парах несовпадающих данных, которые не вошли в сочетание совпадающих данных, и их количество в этом случае равняется n. Сравнивая полученное значение $AD(T_i, T_k) = 6$ с $\min AD(T_i, T_k) = 0$ и $\max AD(T_i, T_k) = n^2/2 = 32$, можно сделать заключение о практической тождественности (равенстве) тестовых двоичных наборов T_i и T_k данных, приведенных в табл. 3.

Количественное совпадение единичных и нулевых значений в наборах T_i и T_k значительно уменьшает величину $AD(T_i, T_k)$, показывая их схожесть. Различие таких наборов зависит от расположения в них данных и соответствующих расстояний между ними. Очевидно, что большее различие между наборами T_i и T_k наблюдается при неравном количестве у них как единичных, так и нулевых данных. В этом случае появляются пары несовпадающих данных, для которых расстояние принимается равным $n/2$. Примером подобных данных могут служить $T_i = 01010101$ и $T_k = 11100011$, значения расстояний для которых даны в табл. 6.

Таблица 6
Значения расстояний $D(t_{i,i}, t_{k,r})$ между $t_{i,i} = t_{k,r}$ для $T_i = 01010101$ и $T_k = 11100011$

Table 6
The values of the distances $D(t_{i,j}, t_{k,r})$ between $t_{i,j} = t_{k,r}$ for $T_i = 01010101$ and $T_k = 11100011$

Набор *Pattern*	T_i								
	Данные *Data*	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
T_k	**1**		1		3		3		1
	1		0		2		4		2
	1		1		1		3		3
	0	3		1		1		3	
	0	4		2		0		2	
	0	3		3		1		1	
	1		3		3		1		1
	1		2		4		2		0

В приведенных наборах количество пар совпадающих нулевых значений равняется трем, так как для четвертого нуля набора T_i отсутствует четвертый нуль в наборе T_k, а количество пар совпадающих единичных данных равняется четырем. Следуя определению 1, для нулевых тождественных данных $t_{i,j} = t_{k,r} = 0$ находим сочетание пар, для которых сумма минимальных расстояний $D(t_{i,j}, t_{k,r})$ также минимальна. Это достигается путем выбора сочетания пар $\{(t_{i,2}, t_{k,3}), (t_{i,4}, t_{k,4}), (t_{i,6}, t_{k,5})\}$, для которых сумма $D(t_{i,2}, t_{k,3}) + D(t_{i,4}, t_{k,4}) + D(t_{i,6}, t_{k,5})$ минимальна и равняется двум. Отметим, что общее количество сочетаний пар нулевых данных в указанных наборах равняется 24. Аналогично выбирается сочетание пар единичных значений $\{(t_{i,1}, t_{k,1}), (t_{i,3}, t_{k,2}), (t_{i,5}, t_{k,6}), (t_{i,7}, t_{k,7})\}$, для которых сумма их расстояний принимает минимальное значение, равное двум. Для пары данных $t_{i,0}$ и $t_{k,0}$, которые не приняли участие в выбранных ранее сочетаниях пар, расстояние принимается равным $n/2 = 4$. Окончательно для $T_i = 01010101$ и $T_k = 11100011$ значение $AD(T_i, T_k)$ вычисляется как сумма $2 + 2 + 4 = 8$.

Рассмотренный пример показал, что основной проблемой при вычислении $AD(T_i, T_k)$ в случае двоичных данных является необходимость анализа большого количества сочетаний дан-

ных. Эта задача носит комбинаторный характер, сложность ее реализации определяется количеством сочетаний независимых пар нулевых и единичных значений в двоичных наборах T_i и T_k. В общем случае для наборов T_i и T_k, имеющих одинаковую размерность n, общее количество Q сочетаний независимых пар данных можно представить соотношением

$$Q = \binom{u}{l} \cdot l! + \binom{n-l}{n-u} \cdot (n-u)! = \frac{u! + (n-l)!}{(u-l)!},\ l = \min[w(T_i), w(T_k)],\ u = \max[w(T_i), w(T_k)]. \quad (4)$$

Выражение $w(T_i)$ означает вес (количество единиц) вектора T_i, а $w(T_k)$ – вес вектора T_k. Первое слагаемое в выражении (4) определяет количество сочетаний независимых пар единичных данных $t_{i,j} = t_{k,r} = 1$, общее число которых равняется l, а второе слагаемое – количество сочетаний независимых $(n - u)$ пар нулевых данных. Отметим, что число данных в наборах T_i и T_k, для которых отсутствует пара с одинаковым значением, определяется разностью $u - l$. Для рассмотренного выше примера (см. табл. 6) $w(T_i) = 4$, $w(T_k) = 5$. Соответственно, $l = 4$, $u = 5$. Таким образом, в наборах T_i и T_k выделяются $l = 4$ пары единичных данных и $n - u = 8 - 5 = 3$ пары нулевых данных. Общее количество Q сочетаний независимых единичных и нулевых пар данных в соответствии с выражением (4) вычисляется как $(u! + (n - l)!)/(u-l)! = (5! + 4!)/(1!) = 144$. Одна пара данных в наборах T_i и T_k имеет несовпадающие значения, так как $u - l = 5 - 4 = 1$.

Приведенная оценка Q количества сочетаний независимых пар свидетельствует о большом их числе для случаев реальных тестовых наборов и невозможности их полного перебора, необходимого для вычисления $AD(T_i, T_k)$.

3. Сведение задачи вычисления меры различия $AD(T_i, T_k)$ к задаче о назначениях. Приведенная в разд. 2 формулировка задачи вычисления меры различия $AD(T_i, T_k)$ для двоичного случая вытекает из определения 1, сформулированного для общего случая в работе [11]. Важными уточнениями указанной задачи являются алфавит тестовых данных, декларирующий использование только двоичных значений, и одинаковая размерность n тестовых наборов T_i и T_k. Допущение, что каждый из наборов содержит по n двоичных данных $t_{i,j}$ и $t_{k,r}$, позволяет сформировать полное множество сочетаний из n пар. Все их множество будет состоять из пар совпадающих данных $t_{i,j} = t_{k,r}$, удовлетворяющих определению 1 и требующих решения сложной комбинаторной задачи, подробно описанной в разд. 2, и пар несовпадающих данных $t_{i,j} \neq t_{k,r}$. Сложность выбора оптимального сочетания пар, удовлетворяющих определению 1, следует из того факта, что в зависимости от выбора конкретной пары идентичных данных $t_{i,j} = t_{k,r}$ расстояние $D(t_{i,j}, t_{k,r})$ между ними для четных величин n принимает значение от 0 до $n/2$ в соответствии с табл. 4. Выбор оптимального сочетания пар идентичных данных заключается в нахождении такого их сочетания, для которого сумма расстояний будет минимальной. Для пар несовпадающих данных в силу того, что согласно определению 1 для каждого из этих данных задается фиктивное значение расстояния, следует возможность выбора любого их сочетания. Важным является только значение количества несовпадающих пар, для каждой из которых задается расстояние в виде константы $D(t_{i,j} \neq t_{k,r}) = n/2$. Величина приведенного значения принята равной максимальному расстоянию $D(t_{i,j}, t_{k,r})$ для совпадающих данных.

Для нечетных значений отличием является только максимальное значение расстояния $D(t_{i,j}, t_{k,r}) = \lfloor n/2 \rfloor$, а вид матрицы расстояний остается таким же, как и для четных значений n. Для произвольного целого n матрица $n \times n$ расстояний $D(t_{i,j}, t_{k,r})$ представляет собой циркулярную матрицу, пример которой для четных величин n приведен в табл. 4.

В качестве примера для $n = 9$ рассмотрим наборы двоичных данных $T_i = 111010111$ и $T_k = 111101011$, для которых значения расстояний для всевозможных пар данных приведены в табл. 7.

Отметим, что в отличие от примеров, описанных ранее в табл. 5 и 6, в приведенной квадратной матрице размерностью 9×9 определены все ее элементы. Так как $n = 9$ является нечетным числом, то $D(t_{i,j}, t_{k,r}) = \lfloor n/2 \rfloor = 4$. Очевидно, что и для произвольного случая матрицы размерностью $n \times n$ также будут определены все ее элементы. Для совпадающих данных и четного значения n они будут соответствовать величинам, приведенным в табл. 4, а для нечетных n строки

аналогичной матрицы являются циклическими сдвигами первой строки, состоящей из элементов 0, 1, 2, …, $\lfloor n/2 \rfloor - 1$, $\lfloor n/2 \rfloor$, $\lfloor n/2 \rfloor$, $\lfloor n/2 \rfloor - 1$, …, 3, 2, 1. Для всех пар несовпадающих данных соответствующий им элемент матрицы для произвольного n равняется $\lfloor n/2 \rfloor$.

Таблица 7
Значения расстояний $D(t_{i,j}, t_{k,r})$ для T_i = 111010111 и T_k = 111101011

Table 7
The values of the distances $D(t_{i,j}, t_{k,r})$ for T_i = 111010111 and T_k = 111101011

Набор *Pattern*	T_i									
	Данные *Data*	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**1**	**1**
T_k	**1**	0	1	2	4	4	4	3	2	1
	1	1	0	1	4	3	4	4	3	2
	1	2	1	0	4	2	4	4	4	3
	1	3	2	1	4	1	4	3	4	4
	0	4	4	4	1	4	1	4	4	4
	1	4	4	3	4	1	4	1	2	3
	0	4	4	4	3	4	1	4	4	4
	1	2	3	4	4	3	4	1	0	1
	1	1	2	3	4	4	4	2	1	0

Таким образом, в случае двоичных тестовых наборов T_i и T_k размерностью n бит задача вычисления меры различия $AD(T_i, T_k)$ может быть сформулирована в виде следующего определения.

Определение 4. Мера различия $AD(T_i, T_k)$, соответствующая определению 1, для двоичных тестовых наборов T_i и T_k, каждый из которых состоит из n бит $t_{i,j}, t_{k,r} \in \{0, 1\}$, $j, r \in \{0, 1, \ldots, n-1\}$, основана на построении циркулярной матрицы расстояний размерностью $n \times n$. Первая строка данной матрицы имеет вид 0, 1, 2, …, $n/2-1$, $n/2$, $n/2-1$, …, 3, 2, 1 для четных значений n и 0, 1, 2, …, $\lfloor n/2 \rfloor - 1$, $\lfloor n/2 \rfloor$, $\lfloor n/2 \rfloor$, $\lfloor n/2 \rfloor - 1$, …, 3, 2, 1 – для нечетных. С использованием циркулярной матрицы строится матрица расстояний таким образом, что в случае совпадающих данных $t_{i,j} = t_{k,r}$ соответствующий элемент этой матрицы равняется элементу циркулярной матрицы, а в случае их несовпадения принимает значение расстояния, равное $\lfloor n/2 \rfloor$. Численное значение $AD(T_i, T_k)$ определяется минимальной суммой n элементов матрицы расстояний, покрывающих все строки и все столбцы матрицы расстояний.

Согласно определению 4 строится исходная квадратная матрица расстояний, которая аналогична исходной квадратной матрице стоимости, используемой в задаче о назначениях [21]. В обоих случаях решение задачи вычисления меры различия и задачи о назначениях заключается в нахождении n элементов матрицы, которые покрывают все строки и все столбцы, а сумма их значений минимальна. Отметим, что каждый из n найденных элементов определяет пару, состоящую из строки и столбца с соответствующими индексами. При этом каждый столбец и каждая строка входят только в одну пару.

Задачу о назначениях можно описать, если сформулировать ее, используя двудольный граф. Следовательно, задача нахождения меры различия $AD(T_i, T_k)$ тестовых наборов T_i и T_k, каждый из которых состоит из n_i и n_k данных $t_{i,j}, j \in \{0, 1, \ldots, n_i - 1\}$, и $t_{k,r}$, $r \in \{0, 1, \ldots, n_k - 1\}$, в терминах задачи на графах формулируется следующим образом. Дан полный двудольный граф G с n вершинами, соответствующими данным $t_{i,j}$ набора T_i, и n вершинами, соответствующими данным $t_{k,r}$ набора T_k. Стоимость (вес) каждого ребра графа неотрицательна и равняется величине расстояния $D(t_{i,j}, t_{k,r})$ между данными $t_{i,j}$ и $t_{k,r}$. Требуется найти совершенное, или полное, паросочетание с наименьшей стоимостью, равной сумме расстояний $D(t_{i,j}, t_{k,r})$ и определяемой значением меры различия $AD(T_i, T_k)$.

Приведенное определение соответствует задаче нахождения оптимального сочетания пар данных $t_{i,j}$ и $t_{k,r}$ в общем случае для произвольных наборов T_i и T_k. Наиболее близким известным решением подобной задачи является решение задачи о назначениях с помощью венгерского алгоритма [21]. Главная идея данного алгоритма была разработана Х. Куном и нашла широкое применение на практике [22]. С помощью одного из онлайн-приложений (URL: https://math.semestr.ru/nazn/index.php) для решения задачи о назначениях в случае определения меры различия $AD(T_i, T_k)$, где T_i = 111010111 и T_k = 111101011, а матрица расстояний $D(t_{i,j}, t_{k,r})$ приведена в табл. 7, было получено следующее решение:

$$\begin{gathered} Cmin = 0 + 0 + 0 + 1 + 1 + 1 + 1 + 0 + 0 = 4. \\ \textit{Путь}: (1;1), (2;2), (3;3), (4;5), (5;4), (6;7), (7;6), (8;8), (9;9). \end{gathered} \quad (5)$$

В терминах задачи вычисления меры различия *Cmin* (5) представляет собой искомое значение $AD(T_i, T_k) = 4$ для T_i = 111010111 и T_k = 111101011, а множество *Путь* определяет оптимальное сочетание пар $\{(t_{i,1}, t_{k,1}), (t_{i,2}, t_{k,2}), (t_{i,3}, t_{k,3}), (t_{i,4}, t_{k,5}), (t_{i,5}, t_{k,4}), (t_{i,6}, t_{k,7})\}$ и $\{(t_{i,7}, t_{k,6}), (t_{i,8}, t_{k,8}), (t_{i,9}, t_{k,9})\}$. Временная сложность оригинального алгоритма о назначениях имеет полиномиальную оценку $O(n^4)$ [21]. Поэтому даже в случае примера, решение для которого приведено в (5) (n = 9), требуются существенные временные затраты. Уменьшение временной сложности алгоритма до $O(n^3)$ [22] также не позволяет применять его с целью вычисления меры различия для реальных случаев управляемых вероятностных тестов.

4. Алгоритм вычисления меры различия $AD(T_i, T_k)$. Приведенный выше анализ показал высокую вычислительную сложность определения меры различия $AD(T_i, T_k)$ для произвольных тестовых наборов T_i и T_k. Основная проблема заключается в необходимости рассмотрения большого количества Q сочетаний независимых пар тождественных данных. Только в этом случае возможно получение минимального значения $AD(T_i, T_k)$, которое показывает степень различия сравниваемых наборов. Получить близкое к оптимальному решение в смысле минимальности значения указанной меры можно с использованием более простых с точки зрения вычислительной сложности алгоритмов, в том числе и широко известных на практике, например жадного алгоритма (*Greedy algorithm*), суть которого состоит в нахождении локально оптимальных решений на каждом этапе, допуская, что конечное решение также окажется оптимальным [23].

Сущность предлагаемого алгоритма вычисления меры различия $AD(T_i, T_k)$ заключается в определении количества пар данных $t_{i,j}$ и $t_{k,r}$ тестовых наборов T_i и T_k, имеющих определенное значение расстояния $D(t_{i,j}, t_{k,r})$. Учитывая то, что для $n = \max(n_i, n_k)$ значение $D(t_{i,j}, t_{k,r})$ равняется минимальному значению из двух величин $|j - r|$ и $n - |j - r|$, необходимо рассмотреть только $\lfloor n/2 \rfloor + 1$ возможных значений $D(t_{i,j}, t_{k,r})$. В отличие от общей постановки задачи вычисления меры $AD(T_i, T_k)$ в предлагаемом алгоритме последовательно от 0 до $\lfloor n/2 \rfloor + 1$ определяются количества пар тождественных данных $t_{i,j} = t_{k,r}$, имеющих соответствующее значение расстояния $D(t_{i,j}, t_{k,r}) = 0 \div \lfloor n/2 \rfloor + 1$.

Исходными являются тестовые наборы T_i и T_k, каждый из которых состоит из n_i и n_k данных $t_{i,j}$, $j \in \{0, 1, \ldots, n_i - 1\}$, и $t_{k,r}$, $r \in \{0, 1, \ldots, n_k - 1\}$. Предположив, что $n_i \geq n_k$, расстояния между данными наборов T_i и T_k генерируются путем циклического сдвига T_i вправо и влево с последующим определением количества совпадающих данных, имеющих одинаковые индексы j и r. В этом случае индекс j для каждого данного $t_{i,j}$ набора T_i примет все возможные значения $j \in \{0, 1, \ldots, n_i - 1\}$ как результат операций циклического сдвига T_i. Таким образом, каждому данному $t_{k,r}$ набора T_k последовательно будут сопоставлены все данные $t_{i,j}$ и проанализированы на предмет совпадения. Совпадение данных $t_{i,j}$ и $t_{k,r}$ свидетельствует о наличии пары тождественных данных с расстоянием $D(t_{i,j}, t_{k,r})$, равным количеству циклических сдвигов набора T_i. Независимость пар тождественных данных обеспечивается исключением данных выявленной пары из дальнейшего рассмотрения. В результате подобных действий определяются пары тожде-

ственных данных $t_{i,j}$ и $t_{k,r}$ с фиксированными значениями расстояния $D(t_{i,j}, t_{k,r})$, на основании которых и вычисляется значение меры различия $AD(T_i, T_k)$.

Алгоритм определения $AD(T_i, T_k)$ состоит из следующих шагов:

1. В исходных тестовых наборах T_i и T_k сравниваются данные $t_{i,j}$ и $t_{k,r}$ для $j = r$ и определяется количество $q(0)$ пар совпадающих данных $t_{i,r} = t_{k,r}$. В результате идентифицируются данные наборов T_i и T_k, имеющие расстояние $D(t_{i,j}, t_{k,r}) = 0$, и их число $q(0)$. Выделенные данные, входящие в наборы T_i и T_k, исключаются из дальнейшего рассмотрения. При $q(0) = n_k$, т. е. при совпадении первых n_k данных набора T_i с данными второго набора T_k, задается $q(t) = n_i - n_k$. Переход к п. 5.

2. Последовательно для $v = 1, 2, \ldots, \lfloor n_i / 2 \rfloor - 1$ формируются циклические сдвиги тестового набора T_i относительно набора T_k на v позиций вправо и влево. Операция сдвига эквивалентна уменьшению и увеличению значения индекса j на величину v по модулю n_i. Далее определяется количество $q(v)$ тождественных пар данных с совпадающими индексами, а именно с модифицированным в результате сдвига $j = (j \pm v) \bmod n_i$ и индексом r. Величина $q(v)$ определяет количество пар данных $t_{i,j}$ и $t_{k,r}$, имеющих расстояние $D(t_{i,j}, t_{k,r}) = v$. При формировании очередного значения сдвига v набора T_i тождественные данные, выделенные при всех предыдущих сдвигах, меньших v, исключаются из рассмотрения с соблюдением последовательности их выделения. Выполнение данного шага прекращается в случае, если количество пар эквивалентных данных, выявленных в наборах T_i и T_k, достигло величины n_k. Переход к п. 5 после задания $q(t) = n_i - n_k$.

3. Если n_i, большее либо равное n_k, является нечетным числом, повторяются действия шага 2 для $v = \lfloor n_i / 2 \rfloor$, а если n_i четно – формируется только один сдвиг на $v = n_i/2$ вправо либо влево. В обоих случаях подсчитывается количество $q(\lfloor n_i / 2 \rfloor)$ совпадающих пар данных, имеющих расстояние $D(t_{i,j}, t_{k,r}) = \lfloor n_i / 2 \rfloor$.

4. Определяется суммарное число $q(t)$ данных $t_{i,j}$ и $t_{k,r}$ наборов T_i и T_k, которые не участвовали в парах тождественных данных.

5. Вычисляется значение меры различия $AD(T_i, T_k)$ согласно выражению

$$AD(T_i, T_k) = q(t) \cdot \min(n_i, n_k) + \sum_{v=1}^{\lfloor n_i / 2 \rfloor} v \cdot q(v). \qquad (6)$$

Полученное значение $AD(T_i, T_k)$ является искомой величиной меры различия наборов T_i и T_k.

Рассмотрим применение данного алгоритма для случая тестовых наборов адресов T_i и T_k, представленных в табл. 1, для которых двоичные значения адресов $t_{i,j}$ и $t_{k,r}$ заменим их десятичными эквивалентами:

1. Сравниваются данные $t_{i,j}$ и $t_{k,r}$ для $j = r$ и определяется количество пар $q(0)$ совпадающих данных $t_{i,r} = t_{k,r}$:

$j = r$	0	1	2	3	4	5	6	7
$t_{i,j}$	<u>0</u>	<u>1</u>	2	3	4	5	6	7
$t_{k,r}$	<u>0</u>	<u>1</u>	3	2	6	7	5	4

Как видно, данные $t_{i,0}$ и $t_{k,0}$ образуют первую пару тождественных значений $t_{i,0} = t_{k,0} = 0$ и $t_{i,1} = t_{k,1} = 1$, что свидетельствует о наличии второй пары совпадающих данных (обозначены подчеркиванием). Соответственно, $q(0) = 2$, а сами данные $t_{i,0}$, $t_{k,0}$, $t_{i,1}$ и $t_{k,1}$ исключаются из дальнейшего рассмотрения путем замены их значений символом X.

2. Так как $n_i = n_k = 8$, последовательно для $v = 1$, 2 и 3 формируются циклические сдвиги тестового набора T_i относительно набора T_k на v позиций вправо и влево. Определяется количество $q(v)$ тождественных пар данных для совпадающих индексов, а именно модифицированного $j = (j \pm v) \bmod n_i$ и индекса r. Последовательность выполнения данного шага алгоритма приведена в табл. 8.

Таблица 8
Последовательность получения значений $q(v)$ для $v = 1, 2$ и 3

Table 8
Sequence of getting $q(v)$ values for $v = 1, 2$ and 3

$j=(j+v)\bmod 8=r$		0	1	2	3	4	5	6	7	$j=(j-v)\bmod 8=r$		0	1	2	3	4	5	6	7	$q(v)$
$v=1$	$t_{i,j}$	7	X	X	**2**	3	4	**5**	6	$v=1$	$t_{i,j}$	X	2	**3**	4	5	6	7	X	$q(1)=3$
	$t_{k,r}$	X	X	3	**2**	6	7	**5**	4		$t_{k,r}$	X	X	**3**	2	6	7	5	4	
$v=2$	$t_{i,j}$	6	7	X	X	X	X	4	X	$v=2$	$t_{i,j}$	X	X	4	X	**6**	**7**	X	X	$q(2)=2$
	$t_{k,r}$	X	X	X	X	6	7	X	4		$t_{k,r}$	X	X	X	X	**6**	**7**	X	4	
$v=3$	$t_{i,j}$	X	X	X	X	X	X	X	**4**	$v=3$	$t_{i,j}$	X	4	X	X	X	X	X	X	$q(3)=1$
	$t_{k,r}$	X	X	X	X	X	X	X	**4**		$t_{k,r}$	X	X	X	X	X	X	X	4	

При формировании очередного значения сдвига v набора T_i тождественные данные, выделенные при всех предыдущих сдвигах, меньших v, помечаются символом X и исключаются из рассмотрения. После реализации сдвигов на $v = 3$ в наборах T_i и T_k все данные заменяются на X. Переход к п. 5 после принятия $q(t) = 0$.

5. В соответствии с выражением (6) вычисляется $AD(T_i, T_k) = 0 + 1\cdot 3 + 2\cdot 2 + 3\cdot 1 + 4\cdot 0 = 10$, что соответствует ранее полученному результату для тестовых наборов, представленных в табл. 1.

Для двоичного случая наборов данных T_i и T_k, согласно определению 4 и с учетом введенных допущений и ограничений, соотношение для вычисления $AD(T_i, T_k)$ принимает вид

$$AD(T_i, T_k) = q(t)\frac{n}{2} + \sum_{v=1}^{n/2} v \cdot q(v). \qquad (7)$$

В выражении (7) в отличие от (6) значение $q(t)$ определяет число пар несовпадающих данных в наборах, так как при равенстве числа данных $t_{i,r}$ и $t_{k,r}$ в наборах T_i и T_k для каждого из них существует пара либо тождественных данных, либо несовпадающих.

В качестве еще одного примера применения предложенного алгоритма рассмотрим случай двоичных тестовых наборов $T_i = 01010101$ и $T_k = 11100011$, матрица весов для которых представлена в табл. 6:

1. Сравниваются данные $t_{i,j}$ и $t_{k,r}$ для $j = r$ и определяется количество пар $q(0)$ совпадающих данных $t_{i,r} = t_{k,r}$:

$j = r$	0	1	2	3	4	5	6	7
$t_{i,j}$	0	**1**	0	1	**0**	1	0	**1**
$t_{k,r}$	1	**1**	1	0	**0**	0	1	**1**

Данные $t_{i,1}$ и $t_{k,1}$ образуют первую пару тождественных значений, $t_{i,4} = t_{k,4} = 0$ – вторую, $t_{i,7} = t_{k,7} = 1$ – третью. Соответственно, $q(0) = 3$, а сами данные исключаются из дальнейшего рассмотрения путем замены их значений символом X.

2. Последовательно, начиная с $v = 1$, формируются циклические сдвиги тестового набора T_i относительно набора T_k на v позиций вправо и влево. Для каждого сдвига v определяется количество $q(v)$ тождественных пар данных.

Таблица 9
Последовательность получения значения $q(1)$

Table 9
Sequence of getting $q(1)$ value

$j=(j-v)\bmod 8=r$		0	1	2	3	4	5	6	7	$j=(j+v)\bmod 8=r$		0	1	2	3	4	5	6	7	$q(v)$
$v=1$	$t_{i,j}$	X	0	**1**	X	1	**0**	X	0	$v=1$	$t_{i,j}$	X	0	X	**0**	1	X	**1**	0	$q(1)=4$
	$t_{k,r}$	1	X	**1**	0	X	**0**	1	X		$t_{k,r}$	1	X	1	**0**	X	0	**1**	X	

После выполнения одной итерации второго шага алгоритма (см. табл. 9) находятся $q(1) = 4$ пары эквивалентных данных, которые выделены подчеркиванием и затем помечены символом X. В результате набор T_i принимает вид 0XXXXXXX, а набор T_k – вид 1XXXXXXX. На последующих шагах алгоритма сформируются значения $q(2)$ и $q(3)$, равные нулю. Далее в соответствии с шагом 3 алгоритма определяется $q(4) = 0$. Суммарное число $q(t)$ пар данных $t_{i,j}$ и $t_{k,r}$ наборов T_i и T_k, которые не участвовали в парах тождественных данных, равняется единице. Полученные значения подставляются в соотношение (7), и окончательно формируется значение $AD(T_i, T_k) = 8$, что соответствует ранее полученному результату вычисления $AD(T_i, T_k)$ согласно определению 1 и его уточнениям для двоичного случая. Такое же значение получается и в результате применения онлайн-приложения, реализующего венгерский алгоритм о назначениях (URL: https://math.semestr.ru/nazn/index.php). Как видно из выражения (8), значение *Cmin* тоже равняется восьми:

$$\begin{gathered} Cmin = 4 + 0 + 1 + 1 + 0 + 1 + 1 + 0 = 8. \\ Путь: (1;1), (2;2), (3;4), (4;3), (5;5), (6;7), (7;6), (8;8). \end{gathered} \tag{8}$$

Таким образом, предложенный алгоритм позволил получить аналогичный результат, который соответствует результату для меры различия, соответствующей определению 1. Однако применив данный алгоритм для случая $T_i = 111010111$ и $T_k = 111101011$, рассмотренного ранее, получим $AD(T_i, T_k) = 6$, что не соответствует оптимальному решению этой задачи (см. (6)) с использованием венгерского алгоритма.

Вычислительную сложность рассмотренного алгоритма в терминах количества операций сравнения данных в наборах T_i и T_k можно оценить величиной $O(n^2)$, где n – размерность анализируемых наборов. Это следует из того, что наборы, состоящие из n данных, анализируются на n итерациях алгоритма, в каждой из которых реализуется один из n сдвигов данных в наборе T_i относительно данных набора T_k.

5. Экспериментальные оценки меры различия. Рассмотренная мера позволяет оценить степень различия двух тестовых наборов T_i и T_k, которые могут быть неразличимыми при использовании других мер различия. В качестве иллюстрации данного утверждения рассмотрим пример двоичных наборов, второй из которых T_k является инверсией первого T_i. Для подобных наборов T_i и $T_k = \overline{T_i}$ расстояние Хэмминга $HD(T_i, \overline{T_i})$ всегда неизменно и равняется n, в то время как $AD(T_i, \overline{T_i})$ принимает различные значения. Например, $AD(00000000, 11111111) = 32$, $AD(00001111, 11110000) = 16$ и $AD(00110011, 11001100) = 8$, при том что расстояние Хэмминга во всех трех случаях равняется восьми.

В предыдущих разделах было показано, что точное вычисление $AD(T_i, T_k)$ для произвольных двоичных наборов T_i и T_k возможно только при использовании венгерского алгоритма. Предложенный авторами алгоритм имеет меньшую вычислительную сложность, однако не всегда повторяет значение $AD(T_i, T_k)$, соответствующее определениям 1 и 4. В то же время, как это видно для случая двоичных тестовых наборов T_i и T_k, представляющих собой случайные двоичные последовательности для n =32, отличий значений может и не быть либо они будут незначительными.

В качестве последовательностей T_i и T_k были использованы двоичные представления случайных целых чисел, сгенерированных в диапазоне [0 ÷ 4294967295]. Эксперимент проводился для 10 000 пар последовательностей T_i и T_k, для которых была получена мера различия $AD(T_i, T_k)$ согласно предложенному авторами алгоритму и результат *Cmin*, вычисленный в соответствии с венгерским алгоритмом.

На рисунке изображена диаграмма для $AD(T_i, T_k)$ и *Cmin* первых 50 пар T_i и T_k. Во многих случаях $AD(T_i, T_k)$ равняется *Cmin*, а при несовпадении значений их разность минимальна, как это видно на примере 15 пар T_i и T_k двоичных тестовых наборов, приведенных в табл. 10.

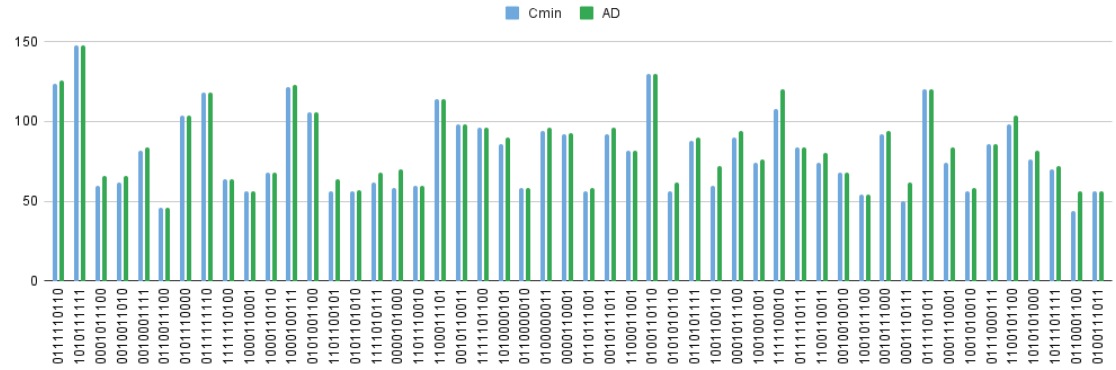

Диаграмма значений $AD(T_i, T_k)$ и *Cmin*

Value chart $AD(T_i, T_k)$ *and Cmin*

Таблица 10

Результаты вычислений для первых 15 пар тестовых наборов T_i и T_k

Table 10

Calculation results for the first 15 pairs of test cases T_i *and* T_k

T_i	T_k	*Cmin*	$AD(T_i, T_k)$
011111011001101011011111010111001	001110101101011101001000100000010	124	126
101011111101010000010011011000111	010100000000000011011111001010001	148	148
000101110001111010010000101101101	110010010110110111110000101101110	60	66
001001101001111001111001001000001	010100010110010110011011111000011	62	66
001000111100101111000010011011110	000001100100000011001010101001101	82	84
011001110000011001011110101011110	100110101000011100011101111111110	46	46
010111000011110110010001010111001	000000011001010000001111110001101	104	104
011111111000001101100010111111111	001011001001010010010110101010011	118	118
111111010000010101010101001000010	111000000110011010101010110000111	64	64
100011000101100111010011110101000	010101100110011011111111001100100	56	56
100011011001000001111001110001110	011111100101001000010011111001110	68	68
100010011111110000100011111111101	011010101110000100010001000011101	122	123
010100110000111111111110001010101	010111101011111110011110101111110	106	106
110110010100111101001101111110110	110001010100000111001111011111111	56	64
010110101001101001101111001011101	011011010011111100011110001111101	56	57

Для сравнительного анализа вычислительной сложности двух алгоритмов была написана программа на языке Python. Программа находит оптимальное решение задачи о назначениях по венгерскому алгоритму (*Cmin*) и вычисляет меру различия $AD(T_i, T_k)$ для двух последовательностей T_i и T_k по алгоритму, предложенному авторами. Для нахождения *Cmin* в программе используется модуль munkres (URL: https://software.clapper. org/munkres/). В табл. 11 представлены данные о времени выполнения венгерского алгоритма и алгоритма авторов. Для каждой пары последовательностей T_i и T_k было проведено 1000 тестов, в качестве времени выполнения принималось среднее арифметическое измеренных значений.

Результаты тестирования показали, что алгоритм вычисления меры различия $AD(T_i, T_k)$ характеризуется значительно меньшей вычислительной сложностью по сравнению с венгерским алгоритмом и практически не уступает в точности ее определения.

Таблица 11
Время вычисления меры различия по венгерскому алгоритму *Cmin* и авторскому $AD(T_i, T_k)$, мс

Table 11
The time for calculating the measure of difference according to the Hungarian algorithm Cmin and the author's $AD(T_i, T_k)$, *ms*

n	T_j	T_k	Венгерский алгоритм *Hungarian algorithm*	Авторский алгоритм *Author's algorithm*
8	11110000	00001111	0,33235	0,01918
	10101010	01010101	0,11701	0,01711
9	101010101	010101010	0,43508	0,01820
10	1111100000	0000011111	0,70302	0,02541
	1010101010	0101010101	0,15418	0,02301
11	10101010101	01010101010	0,55974	0,02512
12	111111000000	000000111111	1,06907	0,03751
	101010101010	010101010101	0,20669	0,03029
13	1010101010101	0101010101010	0,79060	0,03230
14	11111110000000	00000001111111	1,76948	0,04371
	10101010101010	01010101010101	0,25969	0,03858
15	101010101010101	010101010101010	1,07059	0,04072
16	1111111100000000	0000000011111111	2,39983	0,05518
	1010101010101010	0101010101010101	0,32111	0,04761

Заключение. Исследована мера различия тестовых наборов для управляемого вероятностного тестирования, основанная на вычислении суммы расстояний между совпадающими данными двух тестовых наборов. Показано, что расстояния между тождественными данными можно представить в виде двухмерной матрицы их величин, которая аналогична исходной квадратной матрице стоимости, используемой в классической задаче о назначениях. Отмечено, что наиболее близким известным решением задачи о назначениях является венгерский алгоритм, характеризующийся значительной вычислительной сложностью.

Как альтернатива венгерскому алгоритму предложен авторский алгоритм вычисления меры различия и приведена оценка его вычислительной сложности. Данный алгоритм, имеющий меньшую вычислительную сложность, позволяет получить значения меры различия, аналогичные венгерскому алгоритму, либо отличие значений, полученное с использованием обоих алгоритмов, будет незначительным. Проведенные экспериментальные исследования показали высокую временную эффективность и точность вычислений авторского алгоритма в сравнении с известными решениями.

Дальнейшие исследования целесообразно расширить в части свойств новой меры различия и ее применимости для различных прикладных задач.

Вклад авторов. *Ярмолик В. Н.* предложил алгоритм вычисления меры различия для тестовых наборов. *Петровская В. В.* провела экспериментальные исследования. *Мрозек И.* принял участие в обобщении и анализе полученных результатов.

Список использованных источников

1. An orchestrated survey on automated software test case generation / S. Anand [et al.] // J. of Systems and Software. – 2014. – Vol. C-39, no. 4. – P. 582–586.

2. Malaiya, Y. K. The coverage problem for random testing / Y. K. Malaiya, S. Yang // Proc. of ITC, Philadelphia, PA, USA, Oct. 1984. – Philadelphia, 1984. – P. 237–242.

3. A survey on adaptive random testing / R. Huang [et al.] // IEEE Transactions on Software Engineering. – 2021. – Vol. 47, no. 10. – P. 2052–2083.

4. Using the information: Incorporating positive feedback information into the testing process / K. P. Chan [et al.] // Proc. of the 7th Annual Intern. Workshop on Software Technology and Engineering Practice (STEP'03), Amsterdam, Netherlands, 19–21 Sept. 2003. – Amsterdam, 2003. – P. 71–76.

5. A preliminary study of adaptive random testing techniques / M. S. Roslina [et al.] // Intern. J. of Information Technology & Computer Science. – 2015. – Vol. 19, no. 1. – P. 116–127.

6. An empirical comparison of combinatorial testing, random testing and adaptive random testing / H. Wu [et al.] // IEEE Transactions on Software Engineering. – 2020. – Vol. 46, no. 3. – P. 302–320.

7. Ярмолик, В. Н. Контроль и диагностика вычислительных систем / В. Н. Ярмолик. – Минск : Бестпринт, 2019. – 387 с.

8. Ярмолик, В. Н. Многократные управляемые вероятностные тесты / В. Н. Ярмолик, В. А. Леванцевич, И. Мрозек // Информатика. – 2015. – № 2(46). – С. 63–76.

9. Mrozek, I. Multiple controlled random testing / I. Mrozek, V. N. Yarmolik // Fundamenta Informaticae. – 2016. – Vol. 144, no. 1. – P. 23–43.

10. Садовский, М. Г. О сравнении символьных последовательностей / М. Г. Садовский // Вычислительные технологии. – 2005. – № 3(10). – С. 106–116.

11. Ярмолик, В. Н. Мера различия для управляемых вероятностных тестов / В. Н. Ярмолик, Н. А. Шевченко, В. В. Петровская // Доклады БГУИР. – 2022. – № 6(20). – С. 52–60.

12. A modified similarity metric for unit testing of object-oriented software based on adaptive random testing / J. Chen [et al.] // Intern. J. of Software Engineering and Knowledge Engineering. – 2019. – Vol. 29, no. 4. – P. 577–606.

13. Chen, T, Y. Adaptive random testing based on distribution metrics / T. Y. Chen, F-C. Kuo, H. Liu // J. of Systems and Software. – 2009. – Vol. 82, no. 9. – P. 1419–1433.

14. Bard, G. V. Spelling-error tolerant, order-independent pass-phrases via the Damerau – Levenshtein string-edit distance metric / G. V. Bard // Proc. of the Fifth Australasian Symp. on ACSW Frontiers, Ballarat, Australia, 30 Jan. – 2 Feb. 2007. – Ballarat, 2007. – P. 117–124.

15. Tannga, M. J. Comparative analysis of Levenshnein distance algorithm and Jaro Winkler for text plagiarism detection application / M. J. Tannga, S. Rahman, Hasniati // J. of Technology Research in Information System and Engineering. – 2017. – Vol. 4, no. 2. – P. 44–54.

16. О мерах сходства расположения компонентов в массивах естественно упорядоченных данных / А. С. Гуменюк [и др.] // Тр. СПИИРАН. – 2019. – Vol. 18, no. 2. – P. 471–503.

17. Ярмолик, В. Н. Адресные последовательности для многократных маршевых тестов / В. Н. Ярмолик, С. В. Ярмолик // Автоматика и вычислительная техника. – 2006. – № 5. – С. 59–68.

18. Savage, C. A survey of combinatorial Gray code / C. Savage // SIAM Review. – 1997. – Vol. 3, no. 4. – P. 605–629.

19. Mrozek, I. Transparent memory tests based on the double address sequences / I. Mrozek, V. N. Yarmolik // Entropy. – 2021. – No. 23. – P. 894.

20. Неразрушающие тесты с четным повторением адресов для запоминающих устройств / В. Н. Ярмолик [и др.] // Информатика. – 2021 . – Т. 18, № 3. – С. 18–35.

21. Kuhn, H. W. Variants of the Hungarian method for assignment problems / H. W. Kuhn // Naval Research Logistics Quarterly. – 1956. – Vol. 3, no. 4. – P. 253–258.

22. Munkres, J. Algorithms for the assignment and transportation problems / J. Munkres // J. of the Society for Industrial and Applied Mathematics. – 1957. – Vol. 5, no. 1. – P. 32–38.

23. Blum, C. Metaheuristics in combinatorial optimization: Overview and conceptual comparison / C. Blum, A. Roli // ACM Computing Surveys. – 2003. – Vol. 35, no. 3. – P. 268–308.

References

1. Anand S., Burke E. K., Chen T. Y., Clark J., Cohen M. B., …, Zhu H. An orchestrated survey on automated software test case generation. *Journal of Systems and Software*, 2014, vol. C-39, no 4, pp. 582–586.

2. Malaiya Y. K., Yang S. The coverage problem for random testing. *Proceedings of the International Test Conference, Philadelphia, PA, USA, October 1984*. Philadelphia, 1984, pp. 237–242.

3. Huang R., Sun W., Xu Y., Chen H., Towey D., Xia X. A survey on adaptive random testing. *IEEE Transactions on Software Engineering*, 2021, vol. 47, no 10, pp. 2052–2083.

4. Chan K. P., Towey D., Chen T. Y., Kuo F.-C., Merkel R. G. Using the information: Incorporating positive feedback information into the testing process. *Proceedings of the 7th Annual International Workshop on Software Technology and Engineering Practice (STEP'03), Amsterdam, Netherlands, 19–21 September 2003*. Amsterdam, 2003, pp. 71–76.

5. Roslina M. S., Ghani A. A. A., Baharom S., Zulzazil H. A preliminary study of adaptive random testing techniques. *International Journal of Information Technology & Computer Science*, 2015, vol. 19, no. 1, pp. 116–127.

6. Wu H., Nie C., Petke J., Jia Y., Harman M. An empirical comparison of combinatorial testing, random testing and adaptive random testing. *IEEE Transactions on software engineering*, 2020, vol. 46, no. 3, pp. 302–320.

7. Yarmolik V. N. Control' i diagnostika vuchislitel'nuch system. *Computer Systems Testing and Diagnoses.* Minsk, Bestprint, 2019, 387 p. (In Russ.).

8. Yarmolik V. N., Levantsevich B. A., Mrozek I. *Multiple controlled random tests*. Informatika [*Informatics*], 2015, no. 2, pp. 63–76 (In Russ.).

9. Mrozek I., Yarmolik V. N. Multiple controlled random testing. *Fundamenta Informaticae*, 2016, vol. 144, no. 1, pp. 23–43.

10. Sadovskii M. G. *About symbolical sequences comparigion*. Vuchislitel'nue technologii [*Computational Technologise*], 2005, no. 3(10), pp. 106–116 (In Russ.).

11. Yarmolik V. N., Shauchenka M. A., Petrovskaya V. V. Distance measure for controlled random tests. Doklady BGUIR [*BSUIR Proceedings*], 2022, no. 6(20), pp. 52–60 (In Russ.).

12. Chen J., Kudjo P. K., Zhang Z., Su C., Guo Y., …, Song H. A modified similarity metric for unit testing of object-oriented software based on adaptive random testing. *International Journal of Software Engineering and Knowledge Engineering*, 2019, vol. 29, no. 4, pp. 577–606.

13. Chen T. Y., Kuo F.-C., Liu H. Adaptive random testing based on distribution metrics. *Journal of Systems and Software*, 2009, vol. 82, no. 9, pp. 1419–1433.

14. Bard G. V. Spelling-error tolerant, order-independent pass-phrases via the Damerau – Levenshtein string-edit distance metric. *Proceedings of the Fifth Australasian Symposium on ACSW Frontiers, Ballarat, Australia, 30 January – 2 February 2007*. Ballarat, 2007, pp. 117–124.

15. Tannga M. J., Rahman S., Hasniati. Comparative analysis of Levenshnein distance algorithm and Jaro Winkler for text plagiarism detection application. *Journal of Technology Research in Information System and Engineering*, 2017, vol. 4, no. 2, pp. 44–54.

16. Gumenyuk A. S., Skiba A. A., Pozdnichenko N. N., Shpynov S. N. *About similarity measures of components of naturaly ordered data arrays*. Trudy SPIIRAN [*SPIIRAS Proceedings*], 2019, vol. 18, no. 2, pp. 471–503 (In Russ.).

17. Yarmolik V. N., Yarmolik S. V. *Address sequences for multiple march tests*. Avtomatika i vuchislitel'naya technika [*Automation and Computer Science*], 2006, no 5, pp. 59–68 (In Russ.).

18. Savage C. A survey of combinatorial Gray code. *SIAM Review*, 1997, vol. 3, no. 4, pp. 605–629.

19. Mrozek I., Yarmolik V. N. Transparent memory tests based on the double address sequences. *Entropy*, 2021, no. 23, pp. 894.

20. Yarmolik V. N., Mrozek I., Levantsevich V. A., Demenkovets D. V. *Transparent memory tests with even repeating addresses for storage devices*. Informatika [*Informatics*], 2021, vol. 18, no. 3, pp. 18–35 (In Russ.).

21. Kuhn H. W. Variants of the Hungarian method for assignment problems. *Naval Research Logistics Quarterly*, 1956, vol. 3, no. 4, pp. 253–258.

22. Munkres J. Algorithms for the assignment and transportation problems. *Journal of the Society for Industrial and Applied Mathematics*, 1957, vol. 5, no. 1, pp. 32–38.

23. Blum C., Roli A. Metaheuristics in combinatorial optimization: Overview and conceptual comparison. *ACM Computing Surveys*, 2003, vol. 35, no. 3, pp. 268–308.

Информация об авторах

Ярмолик Вячеслав Николаевич, доктор технических наук, профессор, Белорусский государственный университет информатики и радиоэлектроники.
E-mail: yarmolik10ru@yahoo.com

Петровская Вита Владленовна, магистр технических наук, Белорусский государственный университет информатики и радиоэлектроники.
E-mail: vita.petrovskaya@gmail.com

Мрозек Иренеуш, доктор, адъюнкт, Белостокский технический университет.
E-mail: i.mrozek@pb.edu.pl

Information about the authors

Vyacheslav N. Yarmolik, D. Sc. (Eng.), Professor, Belarusian State University of Informatics and Radioelectronics.
E-mail: yarmolik10ru@yahoo.com

Vita V. Petrovskaya, M. Sc. (Eng.), Belarusian State University of Informatics and Radioelectronics.
E-mail: vita.petrovskaya@gmail.com

Ireneusz Mrozek, D. Sc., Lecture, Bialystok University of Technology.
E-mail: i.mrozek@pb.edu.pl