http://dx.doi.org/10.35596/1729-7648-2023-21-2-86-94

Оригинальная статья
Original paper

УДК 004.415.533

СВОЙСТВА МЕРЫ РАЗЛИЧИЯ ДВОИЧНЫХ ТЕСТОВЫХ НАБОРОВ УПРАВЛЯЕМЫХ ВЕРОЯТНОСТНЫХ ТЕСТОВ

В. Н. ЯРМОЛИК, В. В. ПЕТРОВСКАЯ, А. А. ИВАНЮК

Белорусский государственный университет информатики и радиоэлектроники
(г. Минск, Республика Беларусь)

Поступила в редакцию 23.11.2022

Аннотация. Исследуется задача применения характеристик различия для двоичных тестовых последовательностей. Обосновывается их актуальность при генерировании управляемых вероятностных тестов. Рассматривается мера различия $AD(T_i, T_k)$ между тестовыми наборами T_i и T_k, использующая характеристику расстояния $D(t_{i,j}, t_{k,r})$ между $t_{i,j}$ и $t_{k,r}$, которая основана на определении независимых пар тождественных данных $t_{i,j} = t_{k,r}$, принадлежащих двум наборам T_i и T_k. Данная мера различия $AD(T_i, T_k)$ позволяет оценить степень различия двух тестовых наборов T_i и T_k, которые могут быть неразличимыми при использовании других мер различия, в том числе и расстояния Хэмминга. Получены верхние и нижние оценки меры различия для случая инверсных тестовых наборов и произвольных тестовых наборов T_i и T_k с различным сочетанием их весов w_i и w_k. Приводятся примеры вычисления граничных значений указанной меры различия и соотношения их значений. Экспериментальные результаты подтверждают корректность полученных граничных значений указанной меры различия $AD(T_i, T_k)$ и показывают возможность их применения для ее оценки.

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

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Для цитирования. Ярмолик, В. Н. Свойства меры различия двоичных тестовых наборов управляемых вероятностных тестов / В. Н. Ярмолик, В. В. Петровская, А. А. Иванюк // Доклады БГУИР. 2023. Т. 21, № 2. С. 86–94. http://dx.doi.org/10.35596/1729-7648-2023-21-2-86-94.

DISSIMILARITY MEASURE PROPERTIES OF BINARY TEST PATTERNS OF CONTROLLED RANDOM TESTS

VYACHESLAV N. YARMOLIK, VITA V. PETROVSKAYA, ALEXANDER A. IVANIUK

Belarusian State University of Informatics and Radioelectronics (Minsk, Republic of Belarus)

Submitted 23.11.2022

Abstract. The problem of application of the dissimilarity measures for binary test sequences is investigated. Their relevance in generating controlled random tests is substantiated. Dissimilarity measure $AD(T_i, T_k)$ between test sets T_i and T_k is considered, using the characteristic of the distance $D(t_{i,j}, t_{k,r})$ between $t_{i,j}$ and $t_{k,r}$, which is based on the determination of independent pairs of identical data $t_{i,j} = t_{k,r}$ belonging to two patterns T_i and T_k. This measure $AD(T_i, T_k)$ allows us to estimate the degree of difference between two test sets T_i and T_k, which may be indistinguishable when using other difference measures, including the Hamming distance. Upper and lower estimates for the measurement of dissimilarity are obtained for the case of inverse test patterns and arbitrary test

patterns T_i and T_k with different combinations of their weights w_i and w_k. Examples of calculating the boundary values of the specified dissimilarity measure and the ratio of their values are given. Experimental results confirm the correctness of the obtained boundary values of the indicated dissimilarity measure $AD(T_i, T_k)$ and show the possibility of their application for its evaluation.

Keywords: dissimilarity measure, Hamming distance, Levenshtein distance, test, test pattern, controlled random tests.

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

For citation. Yarmolik V. N., Petrovskaya V. V., Ivaniuk A. A. (2023) Dissimilarity Measure Properties of Binary Test Patterns of Controlled Random Tests. *Doklady BGUIR*. 21 (2), 86–94. http://dx.doi.org/10.35596/1729-7648-2023-21-2-86-94 (in Russian).

Введение

Основная задача управляемого вероятностного тестирования состоит в нахождении меры различия тестовых наборов, которая максимально полно показывает их отличие и характеризуется невысокой вычислительной сложностью [1–3]. Определение меры различия тестовых наборов, в общем случае представляющих собой символьные последовательности, в свою очередь, сводится к задаче их сравнения [4].

В [5] рассматривается мера различия (dissimilarity) конечных последовательностей $T_i = t_{i,0}, t_{i,1}, ..., t_{i,n-1}$ как объектов, представляющих упорядоченные тестовые наборы T_i из n данных (элементов) некоторого множества произвольной природы. Основой для построения меры различия $AD(T_i, T_k)$ является характеристика интервала, используемая в теории строя, которая применяется для формального описания и анализа последовательностей данных (символов) для любого алфавита [6]. Данная характеристика интервала была использована для определения меры различия или степени несовпадения двух тестовых наборов, показывая их удаленность либо близость друг от друга [7]. 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 совпадающих данных $t_{i,j} = t_{k,r}$ является значение расстояния $D(t_{i,j}, t_{k,r})$ между $t_{i,j}$ и $t_{k,r}$. Для вычисления величины расстояния $D(t_{i,j}, t_{k,r})$ первоначально определяются значения $|j - r|$ и $\max(n_i, n_k) - |j - r|$. Минимальное значение из приведенных величин принимается в качестве расстояния, т. е. $D(t_{i,j}, t_{k,r}) = \min[|j - r|, \max(n_i, n_k) - |j - r|]$. Для случая тестовых данных, когда $n = n_i = n_k$, расстояние $D(t_{i,j}, t_{k,r})$ определяется соотношением $\min[|j - r|, n - |j - r|]$. 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, описанная в [7], для случая двоичных тестовых наборов одинаковой размерности n 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 данных $t_{i,j}, t_{k,r} \in \{0, 1\}$, где $j, r \in \{0, 1, \ldots, n - 1\}$, основана на определении независимых пар одинаковых (тождественных) данных $t_{i,j} = t_{k,r}$, 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 их расстояний $D(t_{i,j}, t_{k,r})$ минимальна. При отсутствии пары для очередного значения данных $t_{i,j}$ в наборе T_k разность величин индексов, т. е. расстояние $D(t_{i,j}, -)$ принимает значение $\lfloor n / 2 \rfloor$. Показано, что приведенная мера различия $AD(T_i, T_k)$ тестовых наборов T_i и T_k удовлетворяет требованиям: тождественности ($AD(T_i, T_k) = 0$, если $T_k = T_i$), неотрицательности ($AD(T_i, T_k) \geq 0$) и симметричности ($AD(T_i, T_k) = AD(T_k, T_i)$) [5, 8].

Согласно уточненному определению меры различия для случая двоичных данных одинаковой размерности, выделяется два подмножества пар данных [7]. К первому относятся пары совпадающих данных $t_{i,j} = t_{k,r}$, для которых стоит задача определения оптимального их распределения между парами для минимизации суммы расстояний между ними. Количество таких пар определяется соотношением $Q_e = n + \min[w(T_i), w(T_k)] - \max[w(T_i), w(T_k)]$, где $w(T_i)$ является весом (количеством единиц) двоичного вектора T_i. Второе подмножество содержит $Q_n = \max[w(T_i), w(T_k)] - \min[w(T_i), w(T_k)]$ произвольных пар несовпадающих данных $t_{i,j} \neq t_{k,r}$, для которых $D(t_{i,j}, t_{k,r}) = \lfloor n / 2 \rfloor$. Для тестовых наборов T_i и T_k двоичных данных $t_{i,j}$ и $t_{k,r}$ справедливо следующее утверждение.

Утверждение 1. Для пары совпадающих данных $t_{i,j} = t_{k,r}$ с расстоянием $D(t_{i,j}, t_{k,r}) \neq 0$ всегда существует пара $t_{i,r} = t_{k,j}$ данных с расстоянием $D(t_{i,r}, t_{k,j}) = D(t_{i,j}, t_{k,r})$. Справедливость данного утверждения следует из того, что если $t_{i,j} = t_{k,r}$ образуют пару с $D(t_{i,j}, t_{k,r}) \neq 0$, то $t_{i,j} \neq t_{k,j}$ и $t_{i,r} \neq t_{k,r}$. Соответственно при $t_{i,j} = t_{k,r} = 0$ существует пара $t_{k,j} = t_{i,r} = 1$, либо, наоборот, при $t_{i,j} = t_{k,r} = 1$ имеем пару $t_{k,j} = t_{i,r} = 0$.

Мера различия $AD(T_i, T_k)$ позволяет оценить степень различия двух тестовых наборов T_i и T_k, которые могут быть неразличимыми при использовании других мер различия. В качестве иллюстрации данного утверждения рассмотрим пример двоичных наборов, второй T_k из которых является инверсией первого T_i. Для подобных наборов T_i и $T_k = \overline{T_i}$ расстояние Хэмминга $HD\left(T_i, \overline{T_i}\right)$ всегда неизменно и равняется n. В то же время характеристика $AD(T_i, \overline{T_i})$ принимает различные значения в зависимости от веса $w(T_i)$ исходного набора T_i, а также от взаимного расположения в нем данных $t_{i,0}, t_{i,1}, \ldots, t_{i,n-1}$, входящих в данный набор. Например, для случая двоичных данных, входящих в набор T_i, и различных значений веса $w(T_i)$ характеристика $AD(T_i, \overline{T_i})$ принимает следующие значения: $AD(10000000, 01111111) = 26$, $AD(11000000, 00111111) = 20$, $AD(11100000, 00011111) = 18$ и $AD(11110000, 00001111) = 16$. В качестве второго примера рассмотрим случай двоичных наборов той же размерности $n = 8$, когда первый набор T_i имеет постоянный вес, предположим, $w(T_i) = 4$. Соответственно имеем: $AD(11110000, 00001111) = 16$, $AD(11101000, 00010111) = 12$ и $AD(11001100, 00110011) = 8$. Расстояние Хэмминга во всех рассмотренных выше примерах равняется 8, что свидетельствует об одинаковом максимальном отличии всех рассмотренных пар наборов $T_i, \overline{T_i}$ в терминах указанной меры различия.

Мера различия $AD(T_i, T_k)$ для инверсных тестовых наборов T_i и $T_k = \overline{T_i}$

Исследуем характеристику $AD(T_i, \overline{T_i})$ в зависимости от значения веса $w(T_i)$, представляющего собой количество единичных данных в исходном наборе T_i. Как иллюстрируют приведенные выше примеры, взаимное расположение данных в сильной мере влияет на величину характеристики $AD(T_i, \overline{T_i})$. Первоначально рассмотрим исходный набор T_i с весом $w = w(T_i) \leq \lfloor n/2 \rfloor$, в котором все w единиц сгруппированы в виде одной серии. В общем виде набор T_i принимает вид $T_i = t_{i,0}, \ldots, t_{i,l-1}, t_{i,l}, t_{i,l+1}, t_{i,l+2}, \ldots, t_{i,l+w-3}, t_{i,l+w-2}, t_{i,l+w-1}, t_{i,l+w}, t_{i,l+w+1}, t_{i,l+w+2}, \ldots, t_{i,n-1} =$ $= 0 \ldots 0\,0\,1\,1 \ldots 1\,1\,1\,1\,0\,0 \ldots 0$ и, соответственно, $T_k = \overline{T_i} = t_{k,0}, \ldots, t_{k,l-1}, t_{k,l}, t_{k,l+1}, t_{k,l+2}, \ldots, t_{k,l+w-3}, t_{k,l+w-2}, t_{k,l+w-1}, t_{k,l+w}, t_{k,l+w+1}, t_{k,l+w+2}, \ldots, t_{k,n-1} = 1 \ldots 1\,1\,0\,0 \ldots 0\,0\,0\,0\,1\,1 \ldots 1$. В соответствии с Определением 1 оптимальное сочетание пар совпадающих данных наборов T_i и $T_k = \overline{T_i}$ для вычисления величины характеристики $AD(T_i, T_k)$ имеет вид $\{(t_{i,l+1}, t_{k,l}), (t_{i,l}, t_{k,l+1}), (t_{i,l+w}, t_{k,l+w+1}), (t_{i,l+w+1}, t_{k,l+w}), (t_{i,l+2}, t_{k,l-1}), (t_{i,l-1}, t_{k,l+2}), (t_{i,l+w-1}, t_{k,l+w+2}), (t_{i,l+w+2}, t_{k,l+w-1}), \ldots\}$ и представлено на рис. 1. Графически пары совпадающих данных показаны сплошными линиями с двусторонними стрелками, а несовпадающие пары – пунктирными линиями (рис. 1).

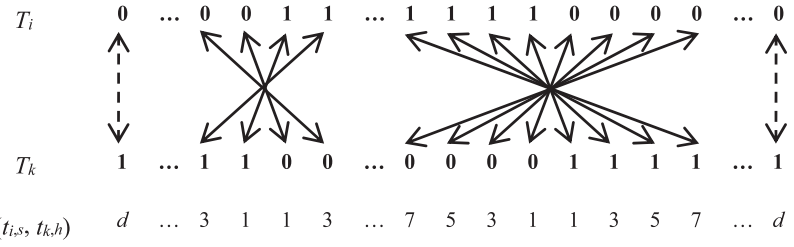

Рис. 1. Оптимальное сочетание пар совпадающих данных наборов T_i и $T_k = \overline{T_i}$ для одной серии из единиц
Fig. 1. Optimal combination of matching data set pairs T_i and $T_k = \overline{T_i}$ for single series of ones values case

Общее количество пар совпадающих и несовпадающих двоичных данных в наборах T_i и T_k определяется величинами Q_e и Q_n [7]. В случае, когда $T_k = \overline{T_i}$, учитывается свойство симметричности ($AD(T_i, \overline{T_i}) = AD(\overline{T_i}, T_i)$) данной метрики и принимается ограничение $w = w(T_i) \leq \lfloor n/2 \rfloor$, количество совпадающих пар данных равняется $2w$, а несовпадающих – $(n - 2w)$.

Регулярность серии из w единиц в наборе T_i и соответствующей ей серии из нулей в наборе $T_k = \overline{T_i}$, как это видно из рис. 1, позволяет формализовать вычисление характеристики $AD(T_i, T_k)$.

Слагаемыми для вычисления данной метрики являются расстояния между совпадающими данными, которые принимают следующие значения: $D(t_{i,l+1}, t_{k,l}) = D(t_{i,l}, t_{k,l+1}) = D(t_{i,l+w}, t_{k,l+w+1}) = D(t_{i,l+w+1}, t_{k,l+w}) = 1$, $D(t_{i,l+2}, t_{k,l-1}) = D(t_{i,l-1}, t_{k,l+2}) = D(t_{i,l+w-1}, t_{k,l+w+2}) = D(t_{i,l+w+2}, t_{k,l+w-1}) = 3$ и так далее для последующих пар совпадающих значений данных в наборах. Значения расстояний $D(t_{i,s}, t_{k,h})$ приведены на рис. 1 для $s, h \in \{0, 1, 2, \ldots, n-1\}$. В зависимости от величины веса w максимальное значение расстояния для серии, состоящей из единиц, принимают расстояния $D(t_{i,l+w/2}, t_{k,l-w/2+1}) = D(t_{i,l-w/2+1}, t_{k,\,l+w/2}) = D(t_{i,l+w/2+1}, t_{k,l+w+w/2}) = D(t_{i,l+w+w/2}, t_{k,l+w/2+1}) = w-1$ для четырех пар данных при четных w, и двух пар $D(t_{i,l+(w+1)/2}, t_{k,l-(w+1)/2+1}) = D(t_{i,l-(w+1)/2+1}, t_{k,l+(w+1)/2}) = w$ в противном случае. Для остальных $(n-2w)$ пар несовпадающих данных в наборах T_i и $T_k = \overline{T}_i$, согласно Определению 1, расстояние принимается равным $\lfloor n/2 \rfloor$, которое на рис. 1 обозначено символом d [7]. Окончательно выражение для характеристики $AD(T_i, \overline{T}_i)$, где $w = w(T_i) \le \lfloor n/2 \rfloor$, а все единичные данные набора T_i расположены последовательно в виде серии из w единиц, принимает вид

$$AD(T_i,\overline{T}_i) = \lfloor n/2 \rfloor \cdot (n-2w) + 2w\left(w - 2\lfloor w/2 \rfloor\right) + 4\sum_{v=1}^{\lfloor w/2 \rfloor}(2v-1) = \\ = w^2 + \left(w - 2\lfloor w/2 \rfloor\right) + \lfloor n/2 \rfloor \cdot (n-2w). \quad (1)$$

Выражение (1) справедливо только для набора T_i веса w, когда все w единичные данные расположены в виде одной серии. Для оценки меры различия $AD(T_i, \overline{T}_i)$ при отсутствии ограничений на вид произвольного набора T_i приведем следующее утверждение.

Утверждение 2. Максимальное значение $\max AD(T_i, \overline{T}_i)$ меры различия $AD(T_i, \overline{T}_i)$, где $w = w(T_i)$, равняется $w^2 + \left(w - 2\lfloor w/2 \rfloor\right) + \lfloor n/2 \rfloor \cdot (n-2w)$. Учитывая, что для меры различия $AD(T_i, \overline{T}_i)$ тестовых наборов T_i и $T_k = \overline{T}_i$ выполняется требование симметричности ($AD(T_i, \overline{T}_i) = AD(\overline{T}_i, T_i)$), будем рассматривать случай, когда $w = w(T_i) \le \lfloor n/2 \rfloor$. Как было показано ранее для T_i с весом $w = w(T_i) \le \lfloor n/2 \rfloor$, в котором w единиц представлены в виде серии, $AD(T_i, \overline{T}_i) = w^2 + \left(w - 2\lfloor w/2 \rfloor\right) + \lfloor n/2 \rfloor \cdot (n-2w)$. Рассмотрим процедуру перехода от набора T_i, в котором все w единиц сгруппированы в виде одной серии, к набору с тем же весом w, но с большим количеством единичных серий. Первоначально исследуем T_i, в котором w единичных значений сгруппированы в виде двух серий, разделенных одним нулевым значением, как это показано на рис. 2.

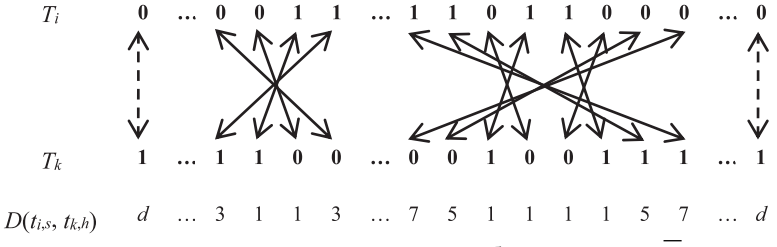

Рис. 2. Оптимальное сочетание пар совпадающих данных наборов T_i и $T_k = \overline{T}_i$ для двух серий из единиц
Fig. 2. Optimal combination of matching data set pairs T_i and $T_k = \overline{T}_i$ for two series of ones values

В силу симметрии пар совпадающих данных $t_{i,j} = t_{k,r}$ рассмотрим только единичные значения $t_{i,j}$, так как аналогичные рассуждения будут верны и для $t_{i,j} = 0$, что следует из Утверждения 1. Построение двух серий из последовательных единичных значений $t_{i,j} = 1$ на основании одной серии, состоящей из w единиц, заключается во внесении в эту серию нулевого значения $t_{i,j}$, как это показано на рис. 2. Соответственно в ту же позицию в набор $T_k = \overline{T}_i$ вносится единичное значение. Для примера, приведенного на рис. 2, имеем T_i = 0 … 0 0 1 1 … 1 1 0 1 1 0 … 0 и T_k = 1 … 1 1 0 0 … 0 0 1 0 0 1 … 1. В результате пара $(t_{i,l+w-1}, t_{k,l+w+2})$ идентичных данных $t_{i,l+w-1} = t_{k,l+w+2} = 1$, для которых $D(t_{i,l+w-1}, t_{k,l+w+2}) = 3$, преобразовалась в пару $(t_{i,l+w}, t_{k,l+w-1})$, которая имеет расстояние $D(t_{i,l+w}, t_{k,l+w-1}) = 1$, т. е. в пару с меньшим расстоянием. Остальные пары единичных данных остались без изменений, так же, как и соответствующие им расстояния. В результате значение характеристики $AD(T_i, \overline{T}_i)$ уменьшилось. Таким образом, можно заключить,

что преобразование набора T_i, для которого $w = w(T_i)$ и все w единиц сгруппированы в виде одной серии, к набору с тем же весом w, но уже с двумя единичными сериями, приводит к тому, что значение $AD(T_i, \overline{T_i})$ не увеличивается. Продолжая подобную процедуру разбиения на серии из единичных значений, можно отметить, что характеристика $AD(T_i, \overline{T_i})$ полученных таким образом наборов не будет увеличиваться. Отсюда следует, что $\max AD(T_i, \overline{T_i})$ вычисляется согласно (1) и равняется величине, определяемой выражением $w^2 + \left(w - 2\lfloor w/2 \rfloor\right) + \lfloor n/2 \rfloor \cdot \left(n - 2w\right)$.

Утверждение 3. Минимальное значение $\min AD(T_i, \overline{T_i})$ меры различия $AD(T_i, \overline{T_i})$, где $T_k = \overline{T_i}$ и $w = w(T_i)$, равняется $2w + \lfloor n/2 \rfloor \cdot (n - 2w)$. Для получения минимальной оценки меры различия $AD(T_i, \overline{T_i})$ рассмотрим минимальные значения слагаемых, участвующих в ее вычислении, т. е. расстояний $D(t_{i,j}, t_{k,r})$ для наборов T_i и $\overline{T_i}$. В силу того, что набор T_k является инверсным по отношению к T_i, не существует пар идентичных данных $t_{i,j} = t_{k,r}$, для которых $D(t_{i,j}, t_{k,r}) = 0$. Минимально возможное значение расстояния $D(t_{i,j}, t_{k,r})$ равняется 1, таким образом, w пар единичных данных и столько же нулевых (Утверждение 1) данных будут принимать минимальные значения расстояний, равные 1, а остальные $(n - 2w)$ несовпадающие пары имеют расстояние $\lfloor n/2 \rfloor$. Окончательно получим $\min AD(T_i, \overline{T_i}) = 2w + \lfloor n/2 \rfloor \cdot (n - 2w)$.

Мера различия $AD(T_i, T_k)$ для произвольных двоичных тестовых наборов T_i и T_k

Для произвольных двоичных тестовых наборов T_i и T_k минимальное значение меры различия $\min AD(T_i, T_k) = 0$, которое достигается при $T_i = T_k$ и вытекает из ее свойства тождественности [7]. Максимальное различие достигается, только когда наборы состоят из различных несовпадающих данных, т. е. $T_i = 0\ 0\ \ldots\ 0\ 0\ 0$, а $T_k = 1\ 1\ \ldots\ 1\ 1\ 1$, либо, наоборот, $T_i = 1\ 1\ \ldots\ 1\ 1\ 1$, а $T_k = 0\ 0\ \ldots\ 0\ 0\ 0$. В результате $\max AD(T_i, T_k) = n\lfloor n/2 \rfloor$.

В общем случае двоичные тестовые наборы различаются по их весу $w_i = w(T_i)$ и $w_k = w(T_k)$ и позволяют определять количество Q_e совпадающих пар данных $t_{i,j} = t_{k,r}$ и количество Q_n пар несовпадающих данных $t_{i,j} \neq t_{k,r}$. Конкретные соотношения весов $w_i = w(T_i)$ и $w_k = w(T_k)$ и их значения позволяют уточнить оценки максимального и минимального значений меры различия $AD(T_i, T_k)$. Величины этих оценок могут уменьшить вычислительную сложность определения $AD(T_i, T_k)$ либо вообще исключить необходимость ее вычисления. Поясним это утверждение на простейшем примере. Имеем два набора, для которых, например, $w(T_i) = 1$ и $w(T_k) = n - 1$, и, соответственно, $\min AD(T_i, T_k) = (n-2) \cdot \lfloor n/2 \rfloor$, а $\max AD(T_i, T_k) = 2 + (n-2) \cdot \lfloor n/2 \rfloor$. Как видно, в данном случае $\min AD(T_i, T_k)$ и $\max AD(T_i, T_k)$ практически не отличаются и имеют большие значения. Это означает, что независимо от распределения данных в обоих наборах T_i и T_k с весами $w(T_i) = 1$ и $w(T_k) = n - 1$ указанные наборы имеют большое различие и не требуют точного вычисления $AD(T_i, T_k)$. Первоначально оценим $\min AD(T_i, T_k)$.

Утверждение 4. Минимальное значение $\min AD(T_i, T_k)$ меры различия $AD(T_i, T_k)$ запишется как $Q_n \lfloor n/2 \rfloor = \max(w_i, w_k) \cdot \lfloor n/2 \rfloor - \min(w_i, w_k) \cdot \lfloor n/2 \rfloor$. Значение $\min AD(T_i, T_k)$ достигается в том случае, когда для всех $Q_e = n + \min(w_i, w_k) - \max(w_i, w_k)$ пар совпадающих данных $D(t_{i,j}, t_{k,r}) = 0$. Тогда $\min AD(T_i, T_k) = Q_e \cdot 0 + Q_n \lfloor n/2 \rfloor$. Для ранее рассмотренного примера наборов T_i и T_k с весами $w(T_i) = 1$ и $w(T_k) = n - 1$ при $n = 8$ указанное $\min AD(T_i, T_k) = 24$ достигается в случае, когда, например, $T_i = 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0$, а $T_k = 1\ 1\ 1\ 1\ 1\ 1\ 1\ 0$.

Утверждение 5. Максимальное значение $\max AD(T_i, T_k)$ меры различия $AD(T_i, T_k)$ запишется в виде $(w_k - w_i) \cdot \lfloor n/2 \rfloor + 2\lceil (n - w_k)/2 \rceil \cdot \lceil w_i/2 \rceil + 2\lfloor (n - w_k)/2 \rfloor \cdot \lfloor w_i/2 \rfloor$, где $w_i = w(T_i) \le w_k = w(T_k)$ и $w_i \le n - w_k$. Первоначально сформулируем условия и ограничения для получения оценки максимального значения $AD(T_i, T_k)$ в случае произвольных наборов T_i и T_k с весами $w_i = w(T_i)$ и $w_k = w(T_k)$. Отметим, что, основываясь на Утверждении 1, задачу вычисления характеристики $AD(T_i, T_k)$ для T_i и T_k можно рассматривать как задачу вычисления этой же характеристики для инверсных значений наборов T_i и T_k. Кроме того, принимая во внимание свойство симметрии $AD(T_i, T_k)$, будем рассматривать случай, когда $w_i \le w_k$. Таким образом, при любом сочетании значений весов w_i и w_k наборов T_i и T_k вычисление значения $AD(T_i, T_k)$ можно свести к задаче вычисления данной характеристики для случаев, когда $w_i \le w_k$ и $w_i \le n - w_k$.

Первоначально рассмотрим наборы T_i и T_k, когда все w_i единиц в T_i и все w_k единиц в T_k расположены в виде одной серии. Предположим, что первые w_i значений данных $t_{i,j}$ набора T_i и первые w_k

данных $t_{k,r}$ набора T_k принимают единичные значения, подобно как для случаев T_i = 1 1 0 0 0 0 0 0 и T_k = 1 1 1 1 0 0 0 0. Взаимное расположение единичных серий в наборах T_i и T_k будет определять значение характеристики $AD(T_i, T_k)$. Так, для T_i = 1 1 0 0 0 0 0 0 и T_k = 1 1 1 1 0 0 0 0 имеем $AD(T_i, T_k) = 8$. Такое же значение будет для T_i и T_k наборов: {0 1 1 0 0 0 0 0, 1 1 1 1 0 0 0 0}; {0 0 1 1 0 0 0 0, 1 1 1 1 0 0 0 0}. В то же время максимальное значение $AD(T_i, T_k) = 16$ достигается для T_i = 1 1 1 1 0 0 0 0 и T_k = 0 0 0 0 0 1 1 0. В общем случае существует n взаимных расположений единичных серий наборов T_i и T_k. В качестве примера все возможные значения величин $AD(T_i, T_k)$ для различного взаимного расположения единичных серий наборов T_i = 1 1 0 0 0 0 0 0 и T_k = 1 1 0 0 0 0 1 1 приведены в табл. 1.

Таблица 1. Значения характеристики $AD(T_i, T_k)$ для T_i = 1 1 0 0 0 0 0 0 и T_k = 1 1 0 0 0 0 1 1
Table 1. Values of characteristic $AD(T_i, T_k)$ for T_i = 1 1 0 0 0 0 0 0 and T_k = 1 1 0 0 0 0 1 1

$AD(T_i, T_k)$	8	12	16	16	16	12	8	8
T_i	11000000	01100000	00110000	00011000	00001100	00000110	00000011	10000001
T_k	11000011	11000011	11000011	11000011	11000011	11000011	11000011	11000011

Рассматривая наборы T_i и T_k с принятыми ранее ограничениями как циклические наборы данных, можно отметить наличие в них по одной серии из единиц и одной серии из нулей. В наборе T_i имеем серию из w_i единиц и серию из $(n - w_i)$ нулей, а в наборе T_k – серию из w_k единиц и серию из $(n - w_k)$ нулей соответственно. Взаимное расположение этих серий и определяет величину $AD(T_i, T_k)$. Анализ рассмотренного выше примера и данных, приведенных в табл. 1, показывает, что максимальное значение характеристики $AD(T_i, T_k)$ для наборов T_i и T_k достигается при максимальном удалении единичной серии набора T_i от единичной серии набора T_k. Либо, что есть то же самое, при симметричном расположении единичной серии из w_i единиц набора T_i по отношению к нулевой серии из $(n - w_k)$ нулей набора T_k. Отметим, что, согласно принятым ограничениям, $w_i \le (n - w_k)$. Для четных значений $w_i = 2$ и $(n - w_k) = 4$ в примере, представленном в табл. 1, $AD(T_i, T_k)$ принимает максимальное значение 16. Для произвольных значений величин n, w_i и $(n - w_k)$ схема оптимального сочетания пар для вычисления $AD(T_i, T_k)$ в соответствии с Определением 1 приведена на рис. 3.

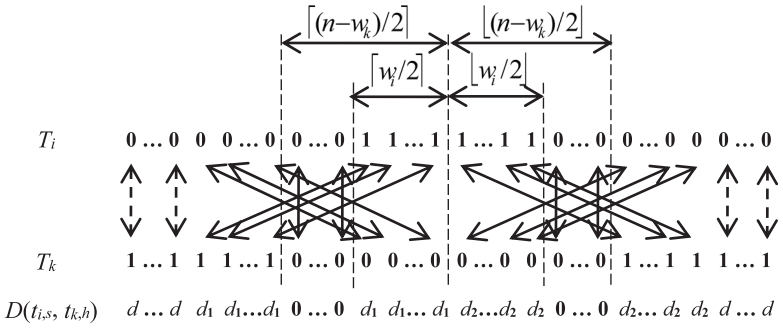

Рис. 3. Оптимальное сочетание пар совпадающих данных наборов T_i и T_k
Fig. 3. Optimal combination of matching data set pairs T_i and T_k

Аналогично, как и на рис. 1, пары совпадающих данных показаны сплошными линиями с двусторонними стрелками, а несовпадающие пары – пунктирными линиями. Общая схема сочетаний пар данных, представленных на рис. 3, позволяет получить выражение для вычисления $\max AD(T_i, T_k)$. Количество пар несовпадающих двоичных данных в наборах T_i и T_k, как было показано ранее, запишется как $w(T_k) - w(T_i) = w_k - w_i$, для которых, согласно Определению 1, $D(t_{i,s}, t_{k,h}) = \lfloor n/2 \rfloor$. Пары совпадающих данных могут иметь три возможных значения. Во-первых, $(n - w_k - w_i)$ пар будут иметь нулевое расстояние $D(t_{i,s}, t_{k,h}) = 0$. Далее, как видно из схемы пар данных на рис. 3, расстояние $D(t_{i,s}, t_{k,h}) = d_1 = \lceil (n - w_k)/2 \rceil$ имеет $2\lceil w_i/2 \rceil$ пары идентичных данных, а расстояние $D(t_{i,s}, t_{k,h}) = d_2 = \lfloor (n - w_k)/2 \rfloor$ соответствует $2\lfloor w_i/2 \rfloor$ парам. Окончательно $\max AD(T_i,T_k) = (w_k - w_i)\cdot\lfloor n/2 \rfloor + 2\lceil (n - w_k)/2 \rceil \cdot \lceil w_i/2 \rceil + 2\lfloor (n - w_k)/2 \rfloor \cdot \lfloor w_i/2 \rfloor$. Отметим, что для вычисления $\max AD(T_i, T_k)$ возможно альтернативное оптимальное сочета-

ние пар совпадающих данных, аналогичное сочетанию, представленному на рис. 1. В обоих случаях результат одинаков. Для примера наборов T_i и T_k с $w(T_i) = 1$ и $w(T_k) = n - 1$ при $n = 8$ значение $\max AD(T_i, T_k) = 26$ достигается в случае, когда, например, T_i = 0 0 0 0 1 0 0 0, а T_k = 1 1 1 1 0 1 1 1. Аналогично, как и для случая инверсных наборов, можно показать, что преобразование наборов T_i и T_k, для которых все w_i и w_k единиц сгруппированы в виде серии, к наборам с теми же весами w_i и w_k, но уже с большим числом единичных серий, приводит к тому, что значение $AD(T_i, T_k)$ не увеличивается. Соответственно получим, что $\max AD(T_i, T_k) = (w_k - w_i)\cdot\lfloor n/2\rfloor + 2\lceil (n-w_k)/2\rceil\cdot\lceil w_i/2\rceil + 2\lfloor (n-w_k)/2\rfloor\cdot\lfloor w_i/2\rfloor$.

Экспериментальные результаты

Для проверки правильности полученных теоретических результатов провели эксперименты для вычисления меры различия $AD(T_i, T_k)$ согласно Определению 1. В качестве метода вычисления данной характеристики использовали Венгерский алгоритм, применяемый для решения задачи о назначениях [7]. Результаты, показывающие корректность Утверждений 2 и 3, приведены в табл. 2, где представлены максимальные и минимальные значения данной меры $AD(T_i, \overline{T_i})$, полученные для большого числа наборов T_i с различными значениями w для разных n.

Таблица 2. Значения $\min AD(T_i, \overline{T_i})$ и $\max AD(T_i, \overline{T_i})$ для различных n
Table 2. Values of $\min AD(T_i, \overline{T_i})$ and $\max AD(T_i, \overline{T_i})$ for various n

n	w	$\max AD(T_i, \overline{T_i})$	$\min AD(T_i, \overline{T_i})$	Количество наборов T_i
8	3	18	14	56
	4	16	8	70
9	3	22	18	84
	4	20	12	126
16	3	90	86	560
	4	80	72	1820
	7	66	30	11 440
	8	64	16	12 870
32	3	426	422	4960
	4	400	392	30 000
	10	292	212	100 000
	11	282	182	100 000

Справедливость теоретических результатов, сформулированных в Утверждениях 4 и 5 для произвольных тестовых наборов T_i и T_k, подтверждают данные, приведенные в табл. 3.

Таблица 3. Значения меры различия $AD(T_i, T_k)$ для произвольных наборов T_i и T_k
Table 3. Values of dissimilarity measure $AD(T_i, T_k)$ for arbitrary patterns T_i and T_k

n	w_i	w_k	$\max AD(T_i, T_k)$	$\min AD(T_i, T_k)$	$\mathrm{av} AD(T_i, T_k)$	Количество пар наборов T_i и T_k с весами w_i и w_k
16	1	3	30	16	19.1513	8960
	2	4	40	16	22.2669	30 000
	2	12	88	80	81.0568	30 000
	2	14	100	96	96.3393	14 400
32	2	4	88	32	45.5338	30 000

Как видно из приведенных в табл. 2 данных, для небольших значений w по отношению к n значения $\min AD(T_i, \overline{T_i})$ и $\max AD(T_i, \overline{T_i})$ практически не отличаются, что позволяет использовать значение $\min AD(T_i, \overline{T_i})$, полученное в соответствии с Утверждением 4, в качестве оценки меры различия $AD(T_i, \overline{T_i})$. Анализ данных, полученных для произвольных тестовых наборов T_i и T_k и частично представленных в табл. 3, показывает тенденцию к близости среднего значения меры различия $\mathrm{av} AD(T_i, T_k)$ к ее минимальному значению $\min AD(T_i, T_k)$. Приведенные практические результаты полностью подтверждают граничные оценки значений для мер различия $AD(T_i, \overline{T_i})$ и $AD(T_i, T_k)$, сформулированные в виде Утверждений 2–5. Их численные значения в ряде случаев

могут быть использованы в качестве самих мер различия $AD(T_i, \overline{T_i})$ и $AD(T_i, T_k)$, что позволит избежать трудоемких вычислений их точных значений [7].

Выводы

Получены оценки максимальных и минимальных значений мер различия $AD(T_i, \overline{T_i})$ и $AD(T_i, T_k)$ для произвольного случая двоичных тестовых наборов T_i и T_k. Экспериментально подтверждены их корректность и возможность применения в качестве оценочных значений меры различия, сформулированной в Определении 1. Дальнейшие исследования целесообразно расширить в части свойств новой меры отличия для различных сочетаний весов тестовых наборов и их размерности, а также применимости данной меры различия для других прикладных задач.

Список литературы

1. A Survey on Adaptive Random Testing / R. Huang [et al.] // IEEE Transactions on Software Engineering. 2021. Vol. 47, No 10. P. 2052–2083. DOI: 10.1109/tse.2019.2942921.
2. 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.
3. Ярмолик, В. Н. Многократные управляемые вероятностные тесты / В. Н. Ярмолик, В. А. Леванцевич, И. Мрозек // Информатика. 2015. № 2. С. 63–76.
4. Sadovsky, M. G. Comparison of Symbol Sequences: no Editing, no Alignment / M. G. Sadovsky // Open Systems & Information Dynamics. 2002. Vol. 9, No 1. P. 19–36. https://doi.org/10.1023/A:1014278811727.
5. Ярмолик, В. Н. Мера отличия для управляемых вероятностных тестов / В. Н. Ярмолик, Н. А. Шевченко, В. В. Петровская // Доклады БГУИР. 2022. Т. 20, № 6. С. 52–60. http://dx.doi.org/10.35596/1729-7648-2022-20-6-52-60.
6. О мерах сходства расположения компонентов в массивах естественно упорядоченных данных / А. С. Гуменюк [и др.] // Труды СПИИРАН. 2019. Т. 18, № 2. С. 471–503. https://doi.org/10.15622/sp.18.2.471-503.
7. 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.
8. Гайдамакин, Н. А. Мера сходства последовательностей одинаковой размерности / Н. А. Гайдамакин // Математические структуры и моделирование. 2016. Т. 40, № 4. С. 5–16.

References

1. Huang R., Sun W., Xu Y., Chen H., Towey D., Xia X. (2021) A Survey on Adaptive Random Testing. *IEEE Transactions on Software Engineering*. 47 (10), 2052–2083. DOI: 10.1109/tse.2019.2942921.
2. Wu H., Nie C., Petke Y., Jia Y., Harman M. (2020) An Empirical Comparison of Combinatorial Testing, Random Testing and Adaptive Random Testing. *IEEE Transactions on Software Engineering*. 46 (3), 302–320.
3. Yarmolik V. N., Levantsevich B. A., Mrozek I. (2015) Multiple Controlled Random Tests. *Informatics*. (2), 63–76 (in Russian).
4. Sadovsky M. G. (2002) Comparison of Symbol Sequences: no Editing, no Alignment. *Open Systems & Information Dynamics.* 9 (1), 19–36. https://doi.org/10.1023/A:1014278811727.
5. Yarmolik V. N., Shauchenka M. A., Petrovskaya V. V. (2022) Distance Measure for Controlled Random Tests. *Doklady BGUIR*. 20 (6), 52–60. http://dx.doi.org/10.35596/1729-7648-2022-20-6-52-60 (in Russian).
6. Gumenjuk A. S, Skiba A. A., Pozdnichenko N. N., Shpunov S. N. (2019) On the Measures of Similarity of the Arrangement of Components in Arrays of Naturally Ordered Data. *Proc. SPIIRAS.* 18 (2), 471–503 (in Russian).
7. Yarmolik V. N., Petrovskaya V. V., Mrozek I. (2022) A Measure of the Difference between Test Sets for Generating Controlled Random Tests. *Informatics*. 19 (4), 7–26 (in Russian).
8. Gaydamakin N. A. (2016) Measures of Similarity Among Finite Sequences. *Mathematical Structures and Simulation.* 40 (4), 5–16 (in Russian).

Вклад авторов

Ярмолик В. Н. предложил меру отличия для управляемых вероятностных тестов.
Петровская В. В. участвовала в обобщении результатов и проведении экспериментов.
Иванюк А. А. принял участие в анализе результатов и проведении экспериментов.

Authors' contribution

Yarmolik V. N. proposed a distance measure for controlled random tests.
Petrovskaya V. V. took part in the generalization of the results and conduct of experiment.
Ivaniuk A. A. took part in the analysis of the results and experiments.

Сведения об авторах

Ярмолик В. Н., д. т. н., профессор Белорусского государственного университета информатики и радиоэлектроники

Петровская В. В., магистр т. н. Белорусского государственного университета информатики и радиоэлектроники

Иванюк А. А., д. т. н., доцент, профессор кафедры информатики, заведующий совместной учебной лабораторией «СК хайникс мемори солюшнс Восточная Европа» Белорусского государственного университета информатики и радиоэлектроники

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. (Eng.), Professor at the Belarusian State University of Informatics and Radioelectronics

Petrovskaya V. V., M. of Sci. at the Belarusian State University of Informatics and Radioelectronics

Ivaniuk A. A., Dr. of Sci. (Eng.), Associate Professor, Professor at the Computer Science Department, Head of the Joint Educational Laboratory "SK Hynix Memory Solutions Eastern Europe" of the Belarusian State University of Informatics and Radioelectronics

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