БЕЗОПАСНОСТЬ ЦИФРОВЫХ ТЕХНОЛОГИЙ. – 2022. – № 1 (104). – 61–84

МЕТОДЫ И СИСТЕМЫ ЗАЩИТЫ ИНФОРМАЦИИ, ИНФОРМАЦИОННАЯ БЕЗОПАСНОСТЬ

УДК 004.056 DOI: 10.17212/2782-2230-2022-1-61-84

ВОЗМОЖНОСТИ АНАЛИЗА НОМИНАТИВНЫХ ПРИЗНАКОВ В ЗАДАЧАХ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ*

В.Е. ХИЦЕНКО[1], Н.А. ФЕДОТОВ[2]

[1] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, кандидат технических наук, доцент. E-mail: xicenko@corp.nstu.ru*

[2] *630073, РФ, г. Новосибирск, пр. Карла Маркса, 20, Новосибирский государственный технический университет, магистрант кафедры вычислительной техники. E-mail: nicoss.fd@yandex.ru*

В статье на различных примерах демонстрируются и обсуждаются возможности проверки гипотез и применения информационных мер для выявления и оценки силы связи номинативных признаков в задачах классификации при анализе информационной безопасности. Основной вид представления исходных данных в этой шкале – это таблица сопряженности номинативных признаков или таблица «объект – признак», из которой могут быть получены частоты совпадения категорий признаков и, собственно, таблица сопряженности. По этой таблице несложно проверить гипотезу о независимости или однородности признаков. Рассмотрен альтернативный подход к этому анализу на основе статистики Кульбака, представляющей собой среднюю различающую информацию в пользу гипотезы о зависимости признаков. В частных случаях практический интерес представляет гипотеза о симметрии квадратных таблиц, которая также может быть проверена на основе информационных мер и критериев. Показан пример обработки дихотомических данных типа «да – нет» по критерию Кокрена. В работе обсуждаются пути измерения силы связи признаков и различные информационные характеристики в виде относительного уменьшения энтропии одного признака при известном другом или в виде средневзвешенного количества информации, приходящегося на различные категории признака. Эти меры полезны для сравнительного анализа признаков в задачах принятия решений. Используются показатель информативности Шеннона, дивергенция Кульбака – Лейблера, Дженсена – Шеннона и мера попарного различения классов эффективности защиты по законам распределения соответствующих им категорий признака. Последовательно сопоставляются классические процедуры проверки гипотез и подходы на основе информационных характеристик. Рассмотренные в работе методы и примеры охватывают многие актуальные задачи информационной безопасности, ассоциированные с номинативными признаками.

* Статья получена 10 февраля 2022 г.

Ключевые слова: статистические методы, защита информации, проверка гипотез о сопряженности качественных признаков, критерии Кокрена, критерии Кульбака, проверка симметрии таблиц, меры связи признаков, информационный подход к анализу связи признаков, показатель информативности Шеннона, дивергенция Кульбака – Лейблера и Дженсена – Шеннона

ВВЕДЕНИЕ

Номинативные признаки – это по сути наименования объектов. Например, способы защиты, уровни доступа, должности, номера отдела, типы программного обеспечения, виды вторжений, протоколы, техники обнаружения атак и т. п. Различие между признаками в такой слабой шкале не может быть измерено количественно. Однако связи между двумя и более такими признаками могут быть выявлены с помощью анализа таблицы сопряженности признаков, в ячейках которой фиксируются частоты или факты совпадения категорий (градаций) признаков в процессе наблюдения. В статье сравниваются два подхода к анализу таблиц: классический и информационный. Особый интерес представляют различные меры для оценки силы связи признаков, оценки их информативности и возможности этих мер для классификации и принятия решений в сфере информационной безопасности.

1. КЛАССИЧЕСКИЙ АНАЛИЗ СОПРЯЖЕННОСТИ ПРИЗНАКОВ

Обсудим несколько подходов к измерению сопряженности признаков. Исходные данные представляются в виде таблицы частот совпадения категорий исследуемых признаков.

Таблица 1

Table 1

Таблица сопряженности двух признаков

Contingency table of two features

Признак 1	Признак 2				
	1	*2*	…	*l*	Сумма
1	n_{11}	n_{12}	…	n_{1l}	$n_{1\cdot}$
2	n_{21}	n_{22}	…	n_{2l}	$n_{2\cdot}$
⋮	⋮	⋮	⋮	⋮	⋮
k	n_{k1}	n_{k2}	…	n_{kl}	$n_{k\cdot}$
Сумма	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot l}$	$n_{\cdot\cdot}$

Количество категорий первого признака (число строк) равно k. Второй признак представлен l категориями-столбцами; n_{ij} – частота совпадений i-й строки и j-го столбца; $n_{i.}$ и $n_{.j}$ – суммы частот i-й строки и j-го столбца соответственно; $n_{..} = n$ – объем выборки.

Разные группы людей или объектов могут отличаться и классифицироваться именно по распределению частот. В другом случае мы имеем одну группу объектов и располагаем частотами проявления одного признака в разных условиях или, что характерно для этой шкалы, частотами проявления разных признаков при фиксированных условиях. Однако принадлежность к группе можно рассматривать как еще один номинативный признак, как столбец «переменная группирования» таблицы «объект-признак», строки которой соответствуют разным объектам.

Метод давно используют для выявления связи двух номинативных признаков, представленных таблицами сопряженности [1, 2]. Для этого находят ожидаемые частоты, которые соответствуют полному отсутствию связи между строками и столбцами (гипотеза H_0) и вычисляют некоторую меру отклонения частот от ожидаемых.

Рассмотрим метод на иллюстративном примере. В табл. 2 показаны частоты сопряженности признака 1 (разные группы специалистов) и признака 2 (мнения об уровне защиты). Ранговую шкалу с небольшим количеством категорий (скажем, низкий, средний, высокий) обычно также относят к номинативной.

Таблица 2

Table 2

Анализ различия мнений сотрудников об уровне защиты

The analysis of distinction of opinions of employees about protection level

Признак 1	Признак 2			
	Низкий	Средний	Высокий	$n_{i.}$
Группа 1	24	7	7	38
Группа 2	76	38	70	184
Группа 3	69	32	82	183
Группа 4	27	9	55	91
$n_{.j}$	196	86	214	$n_{..} = 496$

Найдем ожидаемые частоты, которые получились бы при полном отсутствии связи этих двух признаков, то есть зависимости мнения от принадлежности к группе (гипотеза H_0). Эти частоты находятся по формуле

$$n_{ij}^0 = \frac{n_{i.} n_{.j}}{n_{..}}. \quad (1)$$

При справедливости H_0 статистика (2) распределена по закону χ-квадрат с параметром $(k-1)(l-1)$:

$$\chi^2_{\text{эмп}} = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{\left(n_{ij}^0 - n_{ij}\right)^2}{n_{ij}^0}. \quad (2)$$

В данном примере имеем $\chi^2_{\text{эмп}} = 24{,}93$. Достигнутая значимость составляет менее 0,0004. Это вероятность получить такое или даже большее значение $\chi^2_{\text{эмп}}$ при справедливости H_0. Эту гипотезу отклоняем. Признаки явно сопряжены. Группы радикально отличаются во мнениях. Мы не можем считать это случайным, и выводы из анализа таблицы будут объективны и полезны. Примечательно, что гипотеза H_0 предполагает еще и однородность признака 1 относительно признака 2. Это означает, что все частоты принадлежности к группам при различных мнениях (признак 2) относятся к одной генеральной совокупности.

В сущности, та же статистика (2) традиционно используется для сравнения законов распределения двух признаков. Эта ситуация возникает, когда таблица сопряженности имеет две строки (например, филиал 1 и филиал 2). Столбцы определяют частоты предпочтения шести способов защиты (табл. 3).

Таблица 3

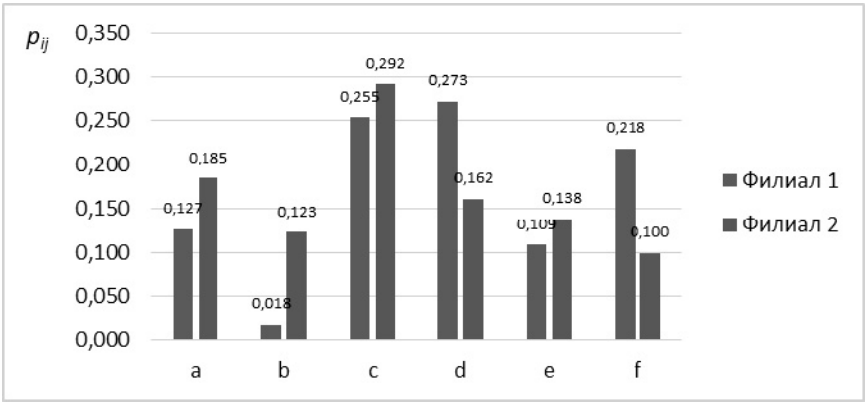
Table 3

Частоты предпочтения видов защиты в разных филиалах

Frequencies of preference for types of protection in different branches

Филиалы	Виды защиты						$n_{i.}$
	a	*b*	*c*	*d*	*e*	*f*	
Филиал 1	7	1	14	15	6	12	55
Филиал 2	24	16	38	21	18	13	130
$n_{.j}$	31	17	52	36	24	25	$n_{..} = 185$

Переходя к относительным частотам предпочтений $p_{ij} = n_{ij} / n_{i\cdot}$, построим гистограммы (рисунок).



Гистограммы предпочтений способов защиты

Histograms of protection methods preferences

Визуальное сопоставление гистограмм не дает уверенности в существенном их различии. Однако проверка гипотезы H_0 о равенстве законов распределения предпочтений дает $\chi^2_{\text{эмп}} = 12,29$. Достигнутая значимость составляет 0,031. Гипотезу H_0 следует отклонить с риском ошибки около 3 % и исследовать причины различия предпочтений. Этот анализ послужит математической основой организационных или технических решений и выводов.

2. ИНФОРМАЦИОННЫЙ КРИТЕРИЙ СОПРЯЖЕННОСТИ

При этом подходе к анализу используют те же таблицы сопряженности признаков (табл. 1), однако затем применяют информационные меры [3], которые являются более гибким инструментом анализа по сравнению со статистикой (2).

Обозначим $p_{ij} = n_{ij} / n_{\cdot\cdot}$, $p_{i\cdot} = n_{i\cdot} / n_{\cdot\cdot}$, $p_{\cdot j} = n_{\cdot j} / n_{\cdot\cdot}$ эмпирические вероятности попадания в клетки таблицы. Для проверки гипотезы о независимости признаков H_0: «$p_{ij} = p_{i\cdot}p_{\cdot j}$ для всех клеток» против двусторонней альтернативы H_1: «$p_{ij} \neq p_{i\cdot}p_{\cdot j}$ хотя бы для одной клетки» вычисляют статистику,

представляющую собой удвоенную среднюю различающую информацию в пользу H_1 против H_0:

$$2I(H_1 : H_0) = 2\sum_{i=1}^{k}\sum_{j=1}^{l} p_{ij} \ln \frac{p_{ij}}{p_{i.}p_{.j}} = 2\sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} \ln \frac{n_{..}n_{ij}}{n_{i.}n_{.j}}. \quad (3)$$

Статистика $I(H_1 : H_0)$ может быть представлена в виде (4), соответствующем разности логарифмов наблюдаемых и ожидаемых частот при справедливости H_0:

$$I(H_1 : H_0) = \sum_{i=1}^{k}\sum_{j=1}^{l} n_{ij} \ln n_{ij} - \sum_{i=1}^{k} n_{i.} \ln n_{i.} - \sum_{j=1}^{l} n_{.j} \ln n_{.j} + n \ln n. \quad (4)$$

Доказано, что при выполнении H_0 статистика (3) имеет в асимптотике то же χ^2-распределение с $(k - 1)(l - 1)$ степенями свободы. Если встречаются несколько равных нулю частот, то соответствующее слагаемое в (3) берем равным нулю и из полученной статистики вычитаем число таких слагаемых [1, с. 445].

Был реализован эксперимент по сравнению предпочтений альтернативных служебных клавиш. Исследование проводилось в рамках формирования клавиатурных портретов пользователей для задач идентификации и аутентификации [4]. В табл. 4 показаны сопряженности частот для пяти пользователей.

Таблица 4

Table 4

Частоты использования служебных клавиш для пяти пользователей

Service keys usage frequencies for five users

№ п/п	Левый Shift	Правый Shift	CapsLock	Левый Ctrl	Правый Ctrl	Backspace	Delete	$n_{i.}$
1	485	273	8	2	0	520	0	1288
2	1035	16	0	14	6	1639	0	2710
3	383	225	6	13	17	373	39	1056
4	338	372	0	5	11	27	304	1057
5	470	1	518	7	6	8	84	1094
$n_{.j}$	2711	887	532	41	40	2567	427	7205

Статистика (3) с учетом поправки оказалась огромной и составила 5757,4. Напомним, что это средняя информация в пользу H_1. Вероятность получить такое значение при справедливости H_0 равна нулю. Эти пять пользователей радикально различны по манере работы на клавиатуре.

Понятно, что при обычной проверке сопряженности признаков по критерию (2) результат получился бы таким же бесспорным: очевидно различие почерка.

Даже если повторить проверку для визуально более близких по почерку пользователей 1, 3 и 4, результат не изменится. Статистика (3) уменьшилась до $2I$ = 1152,7, но достигнутая значимость практически нулевая. Статистика (2), естественно, тоже очень велика, $\chi^2_{\text{эмп}}$ = 970,3.

Преимущество информационного подхода в возможности анализировать не только таблицы с небольшим числом нулевых ячеек, но и таблицы с тремя и более входами, то есть позволяющие исследовать условные независимости и взаимодействие более двух признаков. Это непараметрический, не предполагающий нормальность признаков дисперсионный анализ на основе частот сопряженности.

3. КРИТЕРИЙ КОКРЕНА (COCHRAN)

Этот метод применяется для анализа связи признаков, если мера сопряженности имеет двузначный характер типа «да – нет», «за – против» и измеряется на одной группе объектов в разных экспериментальных условиях или на разных, но одинаковых по численности группах [5]. Такая наиболее слабая среди номинативных шкала называется дихотомической. По сути, это выявление неоднородности наборов двоичных данных.

Рассмотрим метод на примере. Пять сотрудников проверяли на компетентность с помощью шести тестовых заданий. Результаты проверки в табл. 5. Выполнение отмечалось знаком «1», невыполнение – «0». Имеем гипотезу H_0: «все тесты одинаково трудны» против H_1: «есть различия».

Суммы элементов строк и столбцов обозначены здесь ΣX_R и ΣX_C соответственно. Находим *Q*-статистику [5, 6] по формуле

$$Q = \frac{(k-1)\left[k\sum\left(\sum X_C\right)^2 - \left(\sum\left(\sum X_C\right)\right)^2\right]}{k\sum\left(\sum X_C\right) - \sum\left(\sum X_C\right)^2} = \frac{(6-1)[6\cdot 57 - 17^2]}{6\cdot 17 - 59} = 6{,}163, \quad (5)$$

где k – число тестов (столбцов).

Таблица 5

Table 5

Иллюстрация метода Кокрена

Cochran's method illustration

Сотрудники	Тест 1	Тест 2	Тест 3	Тест 4	Тест 5	Тест 6	ΣX_R	$(\Sigma X_R)^2$
1	1	1	0	0	1	1	4	16
2	1	0	0	1	0	1	3	9
3	0	0	1	0	1	1	3	9
4	1	1	1	0	1	0	4	16
5	1	0	0	0	1	1	3	9
ΣX_C	4	2	2	1	4	4	17	59
$(\Sigma X_C)^2$	16	4	4	1	16	16	57	

Статистика Q при справедливости H_0 асимптотически распределена по закону χ^2 с параметром $k-1$. Достигнутая значимость 0,29 не так уж мала. Нет оснований считать, что тесты различаются и результаты тестирования необъективны. Если в результате проверки мы вынуждены отклонить H_0, встает задача совершенствования тестов.

После этого мы можем аналогично проверить гипотезу о близости компетентности сотрудников, просто поменяв местами строки и столбцы в табл. 3. В случае отклонения этой гипотезы мы обоснованно выявим наиболее и наименее компетентных сотрудников.

В результате такой проверки по транспонированной табл. 5 было получено $Q = 3{,}20$ со значимостью 0,525. Различий компетентности не обнаружено.

К сожалению, при ограниченных размерах таблиц χ^2-аппроксимация распределения статистики Q ненадежна. В таких ситуациях применяют статистическое моделирование огромного числа случайно сформированных таблиц. Затем находят долю таблиц, у которых Q достигает полученную в эксперименте. Это и есть оценка Монте-Карло достигнутой значимости, равная 0,273 для этого примера. Современные программы статистической обработки данных [6] предусматривают такое моделирование.

4. ПРОВЕРКА СИММЕТРИИ ТАБЛИЦ

Проверка симметрии таблиц возникает и может быть реализована в квадратных таблицах $k \times k$ произвольного размера. Ясно, что k – это число категорий одного и того же признака, но зафиксированного в разных условиях (например, до и после какого-то события, стажировки, изменения правил, законодательства, уровня защиты).

Гипотеза симметрии H_0 есть предположение о том, что клетки, симметричные главной диагонали, содержат равные частоты. Это значит, что число людей, изменивших мнение с категории *i* на категорию *j*, равно числу людей, поступивших прямо противоположно. Следовательно, H_0: «$p_{ij} = p_{ji}$» против H_1: «$p_{ij} \neq p_{ji}$ хотя бы для одной пары (i, j)».

Для проверки мы должны просуммировать квадраты разностей пар частот, симметричных относительно этой диагонали, деленные на сумму этих же частот. Желательно, чтобы число клеток с частотами менее трех не превышало пятой части всех клеток таблицы [1, 6].

Статистика $\chi^2_{\text{сим}}$ при справедливости гипотезы H_0 о симметрии распределена по закону χ^2 с $k(k-1)/2$ степенями свободы. То есть число степеней свободы равно половине числа недиагональных элементов таблицы.

Например, мнения относительно наибольшей эффективности одного из четырех типов защиты у 100 опрошенных лиц поменялось после стажировки согласно табл. 6. Сумма частот в главной диагонали – это число людей, не изменивших мнения. Проверим симметрию относительно главной диагонали, выделенной серым цветом. Вычислим статистику:

$$\chi^2_{\text{сим}} = \frac{(10-4)^2}{10+4} + \frac{(16-12)^2}{16+12} + \frac{(1-6)^2}{1+6} +$$

$$+ \frac{(8-4)^2}{8+4} + \frac{(4-12)^2}{4+12} + \frac{(7-2)^2}{7+2} = 14,82. \qquad (6)$$

Достигнутая значимость составляет 0,022. Слишком маловероятно получить такое отклонение от симметрии. Следовательно, на уровне значимости 0,02 отклоняем гипотезу симметрии и приступаем к анализу изменений мнений. Значимая асимметрия относительно главной диагонали говорит о направленной деформации мнений в результате стажировки.

Таблица 6

Table 6

Деформация мнений об опасности атак

Changing attitudes about the dangers of attacks

До стажировки	Тип атаки	После стажировки			
		Тип 1	Тип 2	Тип 3	Тип 4
	Тип 1	7	10	16	1
	Тип 2	4	3	8	4
	Тип 3	12	4	4	7
	Тип 4	6	12	2	11

Результаты исследования сравним с информационным подходом к этой же задаче. Средняя информация в пользу H_1 против H_0 может быть найдена согласно [3] по формуле

$$2I(H_1 : H_0) = 2\sum_{i \neq j} n_{ij} \ln \frac{2n_{ij}}{n_{ij} + n_{ji}}. \qquad (7)$$

Достигнутая значимость при использовании этой информационной статистики, которая в данном примере равна 14,144, составляет 0,028. Следовательно, гипотезу H_0 отклоняем.

Понятно, что вместо опасностей атак могли быть изменения других убеждений, симпатий после каких-то мероприятий. Вообще говоря, любая квадратная таблица сопряженности при анализе симметрии может дать нетривиальную информацию.

5. МЕРЫ СВЯЗИ НОМИНАТИВНЫХ ПРИЗНАКОВ

Если признаки оказались взаимосвязаны, то есть гипотеза об их независимости была проверена и отвергнута, возможна оценка силы доказанной связи. Ее хочется видеть в привычном интервале величин, например, от –1 до +1 с нулевым значением при отсутствии связи или от нуля до единицы, если отрицательные значения бессмысленны.

Для измерения силы связи на основе частотных таблиц предложены и используются в статистике около десятка формул [6], которые можно свести к трем основным группам: а) традиционные коэффициенты связи, использующие статистику χ^2; б) меры связи направленного типа; в) коэффициенты на основе информационных характеристик влияния признаков.

Известны многочисленные коэффициенты связи двух признаков, представленных таблицей сопряженности 2×2, т. е. оба признака представлены в двух категориях значений. Это коэффициент ассоциации Пирсона, коэффициент контингенции, частотный коэффициент детерминации, критерий Мак-Нимара и др. [6–8].

Рассмотрим **частотный коэффициент детерминации**, используемый для квадратных таблиц произвольного размера и тесно связанный с симметрией. Он вычисляется по формуле

$$\eta^2 = \frac{\sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{ii}^0}{\sum_{i=1}^{k}\sum_{j=1}^{k} n_{ij} - \sum_{i=1}^{k} n_{ii}^0}, \qquad (8)$$

где n_{ii}^0 – ожидаемые частоты на главной диагонали, вычисляемые по формуле (2). Значимость найденного коэффициента η^2 проверяем так же, как при проверке симметрии, предполагая асимптотическое распределение χ^2 с $k\,(k-1)/2$ степенями свободы.

Пример. Сотрудники организации оценивали в четырехбалльной шкале надежность режима секретности до и после неких модификаций. Изменение мнений представлено в табл. 7.

Таблица 7

Table 7

Пример изменения мнений о надежности защиты

Example of changing opinions about the reliability of protection

	Оценки после модификаций					
		2	**3**	**4**	**5**	Итого
Оценки до модификаций	**2**	3	3	5	2	13
	3	1	2	6	1	10
	4	6	6	2	1	15
	5	1	1	2	1	5
	Итого	11	12	15	5	43

Визуальное изучение таблицы сложно. Вычисление по формуле (8) дает $\eta^2 = -0{,}126$. Связь мнений «до» и «после» мала. Отрицательное значение говорит о том, что большинство респондентов всё же поменяли мнение о надежности. Однако асимптотическая значимость, равная 0,199, не позволяет признать эту асимметрию таблицы существенной. Это может быть случайным.

Рассмотрим **меры, использующие статистику** $\chi^2_{\text{эмп}}$ (2), позволяющие оценить связь двух признаков, измеренных в шкалах с различным числом категорий. То есть таблицы сопряженности признаков могут и не быть квадратными.

Коэффициент сопряженности Пирсона вычисляется по формуле

$$P = \sqrt{\frac{\chi^2_{\text{эмп}}}{\chi^2_{\text{эмп}} + n}}, \tag{9}$$

где $\chi^2_{\text{эмп}}$ находим по формуле (2). Минимальное значение P равно нулю, а максимальное не достигает единицы и зависит от размерности таблицы. Так что сопоставлять по этому коэффициенту можно только таблицы с одинаковым числом клеток. Коэффициент P можно улучшить [6, 1] делением на

$$P_{\max} = \sqrt{\frac{\min\{k,l\}-1}{\min\{k,l\}}}, \tag{10}$$

и тогда его значение попадает в интервал [0,1].

Показатель Чупрова считается более объективным и вычисляется по формуле

$$T = \sqrt{\frac{\chi^2_{\text{эмп}}}{n\sqrt{(k-1)(l-1)}}}. \tag{11}$$

Показатель T достигает единицы, когда все частоты лежат на одной из самых длинных диагоналей таблицы. В квадратных таблицах ($k = l$) максимум T равен единице.

Для неквадратных таблиц часто используют **показатель Крамера**, который вычисляют по формуле

$$C = \sqrt{\frac{\chi^2_{\text{эмп}}}{\min\{k,l\}[(k-1)(l-1)]}}\,. \qquad (12)$$

Он всегда достигает единицы при полной связи и совпадает с T в квадратной таблице.

Для примера, показанного в табл. 8, значения этих мер следующие: $P = 0{,}419$; $T = 0{,}294$; $C = 0{,}326$.

Таблица 8

Table 8

Пример для мер на основе χ^2

Example for measures based on χ^2

Категории признака 1	Категории признака 2				Итоги
	Плохо	Удовлетв.	Хорошо	Отлично	$n_{i.}$
Низкая	19	12	9	2	42
Средняя	7	16	14	5	42
Высокая	4	11	26	12	53
$n_{.j}$	30	39	49	19	137

Достигнутая значимость при $\chi^2_{\text{эмп}} = 29{,}104$, найденная по (2), равна 0,00006, так что все эти меры, являющиеся функциями от χ^2, можно считать объективными. Эти меры всегда выражаются неотрицательными числами (заключение о знаке связи здесь лишено смысла), поэтому выяснение характера зависимости, ее специфических черт должно определяться непосредственно по таблице сопряженности.

Во многих ситуациях один из признаков явно зависит от другого. Скажем, потенциальная опасность, риск определяется типом вируса, но не наоборот. В таких случаях нужно использовать направленные меры.

Логично потребовать, чтобы мера связи таких признаков отражала бы возможность прогноза зависимого признака по значениям другого.

В табл. 9 показаны результаты проверки эффективности способа защиты от хакерских атак трех типов.

Таблица 9

Table 9

Пример для направленных мер

Example for targeted measures

Результат	Тип атаки			
	A	*B*	*C*	$n_{i.}$
Успех	13	16	7	36
Неудача	9	2	17	28
$n_{.j}$	22	18	24	$n_{..} = 64$

Число отраженных атак составляет 36 из 64, т. е. 56,25 %. Однако в 43,75 % случаев мы ошибемся, предсказывая успех. Если учесть тип атаки, тогда, предсказывая успех или неудачу по большинству экспериментальных результатов, ошибаемся в 9 + 2 + 7 = 18 случаев из 64 и процент ошибок предсказаний достигнет 28,13 %. Относительное уменьшение процента ошибок при учете типа атаки называют **показателем λ**. В этом примере получаем $\lambda = (43{,}75 - 28{,}13) / 43{,}75 = 0{,}357$.

Если учет типа атаки сведет процент ошибок к нулю, то λ станет равной единице и зависимость результата от типа атаки будет абсолютной, не случайной. Если учет типа атаки не улучшит прогноз результата, λ будет равна нулю, следовательно, зависимости мы не обнаружили.

Рассмотрим **показатель τ Гудмена – Крускала**. Если частоты результатов 36 и 28 взять с учетом процентов, то грубый прогноз даст $36 \cdot 0{,}5625 + 28 \cdot 0{,}4375 = 32{,}53$ угадываний из 64 и соответственно 31,48 ошибок. Это составляет 49,22 % от 64. Уточненный прогноз с учетом типа атаки и процентов успехов и неудач при каждом конкретном типе составит $13 \cdot 0{,}5909 + 9 \cdot 0{,}4091 + 16 \cdot 0{,}8889 + 2 \cdot 0{,}1111 + 7 \cdot 0{,}2917 + 17 \cdot 0{,}7083 = 39{,}89$ точных предсказаний и 64 – 39,89 = 24,11 ошибок, что даст 37,67 % от 64. Относительное уменьшение процента ошибок при учете типа атаки при таком подсчете получится равным $\tau = (49{,}22 - 37{,}67) / 49{,}22 = 0{,}235$.

При проверке сопряженности таблицы по критерию (2) достигнутая значимость равна 0,001 и результаты не оставляют сомнений.

6. ИНФОРМАЦИОННЫЕ МЕРЫ СВЯЗИ

Рассмотрим меры связи между признаками (строками X и столбцами Y) таблицы в виде взаимной информации, т. е. уменьшения энтропии одного признака при известном другом.

Вернемся к табл. 1, где сведены частоты сопряжения категорий двух признаков. Если перейти к относительным частотам $p_{ij} = n_{ij} / n..$, то оценки энтропии признаков можно получить по формуле Шеннона

$$H(X) = -\sum_{i=1}^{k} p_{i.} \log p_{i.} ; \tag{13}$$

$$H(Y) = -\sum_{j=1}^{l} p_{.j} \log p_{.j} . \tag{14}$$

Оценка энтропии исследуемой системы двух признаков (X, Y) равна

$$H(X,Y) = -\sum_{i=1}^{k} \sum_{j=1}^{l} p_{ij} \log p_{ij} . \tag{15}$$

Взаимная информация об одном признаке при известном состоянии другого будет равна уменьшению суммарной энтропии вследствие учета фактической зависимости признаков, проявляющейся в таблице:

$$I_{X \leftrightarrow Y} = H(X) + H(Y) - H(X,Y) . \tag{16}$$

В анализе используют информационные коэффициенты, получаемые путем следующей нормировки:

$$R_{X \leftrightarrow Y} = \frac{I_{X \leftrightarrow Y}}{H(X,Y)} , \tag{17}$$

а также направленные коэффициенты связи признаков, то есть качества прогноза одного признака по другому:

$$R_{X \to Y} = \frac{I_{X \leftrightarrow Y}}{H(Y)} ; \tag{18}$$

$$R_{Y \to X} = \frac{I_{X \leftrightarrow Y}}{H(X)} . \tag{19}$$

Допустим, в результате наблюдений эффективности способов защиты информации была составлена следующая таблица (табл. 10).

Таблица 10

Table 10

Пример для расчета информативности и дивергентности

Example for calculating of informativeness and divergence

		Эффективность защиты			Итоги
		Низкая (1)	Средняя (2)	Высокая (3)	$n_{i.}$
Способы обнаружения	*a*	24	7	7	38
	b	76	38	70	184
	c	69	32	82	183
	d	27	9	55	91
	$n_{.j}$	196	86	214	$n = 496$

Переходим к относительным частотам и получаем оценки: $H(X) = 1{,}7942$; $H(Y) = 1{,}4910$; $H(X, Y) = 3{,}2478$; $I_{X \leftrightarrow Y} = 0{,}0374$; $R_{Y \to X} = 0{,}0251$; $R_{X \to Y} = 0{,}0208$; $R_{X \leftrightarrow Y} = 0{,}0115$. Сопряженность признаков по критерию χ^2 достаточно высока, значимость достигает 0,00035. Однако их взаимная информация и коэффициенты связи очень малы, и их трудно интерпретировать в контексте анализа безопасности.

Для сравнительного анализа номинативных признаков в задачах классификации и принятия решений предлагается [3, 9] оценивать информативность признака как средневзвешенное количество информации, приходящееся на различные его категории. При этом категории другого признака воспринимаются как классы, принадлежность к которым отражает этот показатель информативности. Используя обозначения табл. 1, можно записать показатель информативности Шеннона [9–14] признака 1 для прогноза признака 2 как снятую неопределенность этой задачи классификации эффективности:

$$IS = 1 + \sum_{i=1}^{k} p_i \sum_{j=1}^{l} q_{ji} \log_k q_{ji} \,, \tag{20}$$

где l – количество классов; k – количество категорий признака; q_{ji} – оценка условной вероятности попадания признака эффективности в j-й класс

(при условии, что используем *i*-й способ обнаружения), равная $q_{ji} = n_{ij} / n_i.$; p_i – оценка вероятности попадания признака в *i*-ю категорию, равная $p_i = n_i. / n..$

Для табл. 10 получаем $IS = 0{,}2732$. Таким образом, информативность признака 1 (способ защиты) для предсказания признака 2 (эффективность) заметно больше аналогичного информационного коэффициента $R_{X \to Y} = 0{,}0208$. Но, главное, диапазон значений этого показателя информативности составляет [0, 1]. Эта нормировка позволяет сравнить и объективно выбрать наилучший признак из принципиально возможных типа: топология сети, протокол, программа.

Теория информации предлагает еще один инструмент для принятия решений – это расхождение (дивергенция).

Допустим, что распределение вероятностей некоторого признака X по n категориям согласно двум гипотезам H_j, $j = 1, 2$, таково: $p_{1j}, p_{2j}, \ldots, p_{kj}$. Следуя [3], можно найти среднюю информацию в пользу гипотезы H_1 против гипотезы H_2 при наблюдении X в условии справедливости H_1, равную

$$I_{KL}(1:2) = \sum_{i=1}^{k} p_{i1} \ln \frac{p_{i1}}{p_{i2}}. \qquad (21)$$

Мы получили известную дивергенцию Кульбака – Лейблера [15], которую используют для оценки информационного выигрыша при замене распределения одного признака распределением другого, имеющую вид математического ожидания разности логарифмов вероятностей попадания в категории значений при H_1 и при H_2 при справедливости H_1.

В противоположном случае при справедливости H_2 средняя информация в пользу H_2 равна

$$I_{KL}(2:1) = \sum_{i=1}^{k} p_{i2} \ln \frac{p_{i2}}{p_{i1}} = -\sum_{k=1}^{k} p_{i2} \ln \frac{p_{i1}}{p_{i2}} \qquad (22)$$

и отличается от $I_{KL}(1:2)$. То есть дивергенция Кульбака – Лейблера не является симметричной и не может служить оценкой различия распределений. Это направленная мера.

Например, для табл. 10 попарные сравнения распределений способов защиты при разных эффективностях дают такие дивергенции:

$I_{KL}(1:2) = 0,01772 \neq I_{KL}(2:1) = 0,016326;$

$I_{KL}(1:3) = 0,111845 \neq I_{KL}(3:1) = 0,093937;$

$I_{KL}(2:3) = 0,102129 \neq I_{KL}(3:2) = 0,113980.$

В работах [16, 17] предлагается симметричная мера в виде дивергенции Дженсена – Шеннона:

$$I_{JS}(1:2) = I_{JS}(2:1) = \frac{1}{2}\sum_{i=1}^{k} p_{i1} \ln \frac{2p_{i1}}{p_{i1} + p_{i2}} + \frac{1}{2}\sum_{i=1}^{k} p_{i2} \ln \frac{2p_{i2}}{p_{i1} + p_{i2}}. \qquad (23)$$

Для того же примера имеем $I_{JS}(1:2) = 0,004237;$ $I_{JS}(1:3) = 0,024584;$ $I_{JS}(1:3) = 0,026260.$

В статье [18] показано применение этой меры для распознавания аномальных изменений законов распределения информационных состояний ресурсов беспилотных транспортных средств, таких как канал связи, процессор, память. Предположительно, эти аномалии являются следствием атак.

Информационные подходы на этой основе позволяют избежать ошибок проверки гипотезы об отсутствии изменений и ориентированы на проверку набора из нескольких рабочих гипотез [19, 20]. В частности, показано, что статистика (2) для проверки гипотез уступает данному подходу. Затем несложно ранжировать эти гипотезы от лучших к худшим и масштабировать, чтобы использовать как вес правдоподобия каждой гипотезы.

Вместе с тем в работе [3] вводится другая симметричная оценка, которую можно интерпретировать как меру трудности различения двух гипотез H_1 и H_2 или двух классов (например, низкой и средней эффективности (табл. 10)) по законам распределения соответствующих им категорий признака, т. е. способов защиты:

$$J(1:2) = I_{KL}(1:2) + I_{KL}(2:1) = \sum_{i=1}^{k} (p_{i1} - p_{i2}) \ln \frac{p_{i1}}{p_{i2}}, \qquad (24)$$

где k – число категорий признака; p_{i1} и p_{i2} – оценки условных вероятностей низкой и средней эффективности при применении i-го способа защиты.

Другими словами, если мала дивергенция J на данной паре классов, то наш эксперимент не позволяет уверенно выбрать эффективный способ защиты.

Расчет для трех пар классов эффективности: низкая, средняя и высокая по табл.10 дает $J(1:2) = 0{,}03404$, $J(1:3) = 0{,}205782$ и $J(2:3) = 0{,}216109$. Можно заключить, что хуже всего по этому признаку различаются низкая и средняя эффективности защиты. Понятно, что этот результат явно зависит от наших предварительных правил классификации эффективности при заполнении таблицы, но теперь у нас появляется критерий для этих правил в виде максимума (23), либо мы в итоге убедимся, что исследуемые способы защиты фактически не отличаются по эффективности. Наконец, можно сравнивать различные признаки по показателям (21), (23) и (24) и обоснованно выбирать наиболее информативные.

ЗАКЛЮЧЕНИЕ

В статье показан подход к анализу ситуаций на основе наблюдений номинативных признаков. Примеры на основе методов проверки гипотез относительно связи признаков и способов оценки степени этой связи и информативности признаков ориентированы на проблематику защиты информации, в основном на стратегические, а не на оперативные задачи безопасности. Показано, что подход с позиций теории информации является сильным конкурентом классическим процедурам проверки гипотез, особенно в задачах классификации и принятия решений на ее основе.

Возможно, что на небольших таблицах сопряженности связи признаков могут быть легко замечены визуально. Однако в сложных случаях в задачах большого масштаба при сравнении более двух признаков без математического подхода, ограничиваясь интуицией и здравым смыслом, получить обоснованные выводы нереально.

Понятно, что некоторые постановки задач информационной безопасности в условиях номинативных признаков не сводятся к анализу таблиц сопряженности. Авторы будут признательны за конструктивные замечания и советы такого рода.

СПИСОК ЛИТЕРАТУРЫ

1. *Закс Л.* Статистическое оценивание. – М.: Статистика, 1976. – 599 с.

2. *Кендалл М., Стьюарт А.* Многомерный статистический анализ и временные ряды. – М.: Наука, 1976. – 736 с.

3. *Кульбак С.* Теория информации и статистика. – М.: Наука, 1967. – 408 с.

4. *Крутохвостов Д.С., Хиценко В.Е.* Парольная и непрерывная аутентификация по клавиатурному почерку средствами математической статистики // Вопросы кибербезопасности. – 2017. – № 5 (24). – С. 91–99.

5. *Рунион Р.* Справочник по непараметрической статистике. – М.: Финансы и статистика, 1982. – 200 с.

6. *Бююль А., Цёфель П.* SPSS: искусство обработки информации. – СПб.: Диасофт, 2002. – 602 с.

7. *Миркин Б.Г.* Анализ качественных признаков и структур. – М.: Статистика, 1980. – 320 с.

8. *Хиценко В.Е.* Математическая статистика для мониторинга информационной безопасности. Непараметрические методы статистики в примерах и задачах. – Saarbrücken: Lap Lambert Academic Publishing, 2013. – 208 с.

9. *Гублер Е.В.* Вычислительные методы анализа и распознавания патологических процессов. – Л.: Медицина, 1978. – 294 с.

10. *Колесникова С.И.* Методы анализа информативности разнотипных признаков // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. – 2009. – № 1 (6). – С. 69–80.

11. *Салахутдинова К.И., Лебедева И.С., Кривцова И.Е.* Подход к выбору информативного признака в задаче идентификации программного обеспечения // Научно-технический вестник информационных технологий, механики и оптики. – 2018. – Т. 18, № 2. – С. 278–285.

12. *Коржук В.М.* Модель и метод идентификации атак сетевого уровня на беспроводные сенсорные сети на основе поведенческого анализа: дис. … канд. техн. наук: 05.13.19. – СПб., 2019. – 206 с.

13. *Быкова В.В., Катаева А.В.* Методы и средства анализа информативности признаков при обработке медицинских данных // Программные продукты и системы. – 2016. – № 2. – С. 172–178.

14. Informativeness of genetic markers for inference of ancestry / N.A. Rosenberg, L.M. Li, R. Ward, J.K. Pritchard // American Journal of Human Genetics. – 2003. – Vol. 73 (6). – P. 1402–1422.

15. *Kullback S., Leibler R.A.* On information and sufficiency // Annals of Mathematical Statistics. – 1951. – Vol. 22 (1). – P. 79–86.

16. *Lin J.* Divergence measures based on the Shannon entropy // IEEE Transactions on Information Theory. – 1991. – Vol. 37, N 1. – P. 145–151.

17. *Nielsen F., Nock R.* Total Jensen divergences: definition, properties and clustering // 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). – South Brisbane, QLD, Australia, 2015. – P. 2016–2020.

18. *Брюховецкий А.А.* Модель обнаружения аномальных данных на основе информационного критерия // Дневник науки. – 2021. – № 4. – URL: www.dnevniknauki.ru/images/publications/2021/4/technics/Bryukhovetskiy.pdf (дата обращения: 09.03.2022).

19. *Burnham K.P., Anderson D.R.* Kullback–Leibler information as a basis for strong inference in ecological studies // Wildlife Research. – 2001. – Vol. 28 (2). – P. 111–119.

20. *Do M.N.* Fast approximation of Kullback–Leibler distance for dependence trees and hidden Markov models // IEEE Signal Processing Letters. – 2003. – Vol. 10, N 4. – P. 115–118.

Хищенко Владимир Евгеньевич, кандидат технических наук, доцент кафедры защиты информации Новосибирского государственного технического университета. E-mail: xicenko@corp.nstu.ru

Федотов Никита Андреевич, магистрант кафедры вычислительной техники Новосибирского государственного технического университета. E-mail: nicoss.fd@yandex.ru

DOI: 10.17212/2782-2230-2022-1-61-84

Possibilities of analysis of nominative signs in tasks of information security *

V.E. Khitsenko[1], N.A. Fedotov[2]

[1] *Novosibirsk State Technical University, 20 Karl Marx Avenue, Novosibirsk, 630073, Russian Federation, candidate of technical sciences, associate Professor of the Department of Information Security. E-mail: xicenko@corp.nstu.ru*

[2] *Novosibirsk State Technical University, 20 Karl Marx Avenue, Novosibirsk, 630073, Russian Federation, computer science student. E-mail: nicoss.fd@yandex.ru*

Using various examples, the article demonstrates and discusses the possibilities of testing hypotheses and applying information measures to identify and assess the strength of the connection of nominative features in classification problems in the analysis of information security. The main type of presentation of the initial data in this scale is a contingency table of nominative features or an "object-feature" table, from which frequencies of coincidence of feature categories and a contingency table can be obtained. Using this table, it is easy to test the hypothesis of independence or homogeneity of features. An alternative approach to this analysis is considered based on the Kullback statistics, which is the average discriminating information in favor of the hypothesis of the dependence of features. In particular cases, the hypothesis of the symmetry of square tables is of practical interest, which can also be tested on the basis of information measures and criteria. An example of the processing of dichotomous data of the "yes-no" type according to the Cochran test is shown. The paper discusses ways to measure the strength of the connection of features. Illustrative examples of calculating measures based on chi-square statistics and directed measures are considered. The possibilities of various information characteristics are discussed in the form of a relative decrease in the entropy of one feature

* *Received 10 February 2022.*

with a known other, or in the form of a weighted average amount of information falling on different categories of a feature. These measures are useful for comparative analysis of nominative features in decision-making problems. Shannon's informativeness index, Kullback-Leibler divergence, and a measure of pairwise differentiation of protection efficiency classes according to the laws of distribution of the corresponding categories of a feature are used. The classical procedures for testing hypotheses and approaches based on information characteristics are consistently compared. The methods and examples considered in the work cover many urgent problems of information security associated with nominative features.

Key words: statistical methods, information protection, testing hypotheses about the conjugation of qualitative features, Cochran's and Kullback's criteria, testing the symmetry of tables, measures of the relationship of features, an informational approach to analyzing the relationship of features, Shannon's informativeness index, Kullback-Leibner and Jenssen-Shannon divergence

REFERENCES

1. Sachs L. *Statisticheskoe otsenivanie* [Statistical estimation]. Moscow, Statistika Publ., 1976. 599 p. (In Russian).

2. Kendall M., Stuart A. *Mnogomernyi statisticheskii analiz i vremennye ryady* [Multivariate statistical analysis and time series]. Moscow, Nauka Publ., 1976. 736 p. (In Russian).

3. Kullback S. *Teoriya informatsii i statistika* [Information theory and statistics]. Moscow, Nauka Publ., 1967. 408 p. (In Russian).

4. Krutohvostov D.S., Khitsenko V.E. Parol'naya i nepreryvnaya autentifikatsiya po klaviaturnomu pocherku sredstvami matematicheskoi statistiki [Password authentication and continuous authentication by keystroke dynamics using mathematical statistics]. *Voprosy kiberbezopasnosti = Cybersecurity Issues*, 2017, no. 5 (24), pp. 91–99.

5. Runyon R. *Spravochnik po neparametricheskoi statistike* [Nonparametric statistics]. Moscow, Finansy i statistika Publ., 1982. 200 p. (In Russian).

6. Bühl A., Zöfel P. *SPSS: iskusstvo obrabotki informatsii* [*SPSS:* The art of information processing]. St. Petersburg, Diasoft Publ., 2002. 602 p. (In Russian).

7. Mirkin B.G. *Analiz kachestvennykh priznakov i struktur* [Analysis of qualitative features and structures]. Moscow, Statistika Publ., 1980. 320 p.

8. Khitsenko V.E. *Matematicheskaya statistika dlya monitoringa informatsionnoi bezopasnosti. Neparametricheskie metody statistiki v primerakh i zadachakh* [Mathematical statistics for information security monitoring. Nonparametric methods of statistics in examples and problems]. Saarbrücken, Lap Lambert Academic Publishing, 2013. 208 p.

9. Gubler E.V. *Vychislitel'nye metody analiza i raspoznavaniya patologicheskikh protsessov* [Computational methods of analysis and recognition of pathological processes]. Leningrad, Meditsina Publ., 1978. 294 p.

10. Kolesnikova S.I. Metody analiza informativnosti raznotipnykh priznakov [Methods of analysis of different-type features informativity]. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naya tekhnika i informatika = Tomsk State University Journal of Control and Computer Science*, 2009, no. 1 (6), pp. 69–80.

11. Salakhutdinova K.I., Lebedeva I.S., Krivtsova I.E. Podkhod k vyboru informativnogo priznaka v zadache identifikatsii programmnogo obespecheniya [Informative feature selection in software identification task]. *Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki = Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2018, vol. 18, no. 2, pp. 278–285.

12. Korzhuk V.M. *Model' i metod identifikatsii atak setevogo urovnya na besprovodnye sensornye seti na osnove povedencheskogo analiza*. Diss. kand. tekhn. nauk [Model and method of identification of network layer attacks on wireless sensor networks based on behavioral analysis. PhD eng. sci. diss.]. St. Petersburg, 2019. 206 p.

13. Bykova V.V., Kataeva A.V. Metody i sredstva analiza informativnosti priznakov pri obrabotke meditsinskikh dannykh [Methods and tools for analysing informative features when processing medical data]. *Programmnye produkty i sistemy = Software and Systems*, 2016, no. 2, pp. 172–178.

14. Rosenberg N.A., Li L.M., Ward R., Pritchard J.K. Informativeness of genetic markers for inference of ancestry. *American Journal of Human Genetics*, 2003, vol. 73 (6), pp. 1402–1422.

15. Kullback S., Leibler R.A. On information and sufficiency. *Annals of Mathematical Statistics*, 1951, vol. 22 (1), pp. 79–86.

16. Lin J. Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, 1991, vol. 37, no. 1, pp. 145–151.

17. Nielsen F., Nock R. Total Jensen divergences: definition, properties and clustering. *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, South Brisbane, QLD, Australia, 2015, pp. 2016–2020.

18. Bryukhovetskiy A.A. Model' obnaruzheniya anomal'nykh dannykh na osnove informatsionnogo kriteriya [Anomalous data detection model based on information criterion]. *Dnevnik nauki*, 2021, no. 4. (In Russian). Available at: www.dnevniknauki.ru/images/publications/2021/4/technics/Bryukhovetskiy.pdf (accessed 09.03.2022).

19. Burnham K.P., Anderson D.R. Kullback-Leibler information as a basis for strong inference in ecological studies. *Wildlife Research*, 2001, vol. 28 (2), pp. 111–119.

20. Do M.N. Fast approximation of Kullback–Leibler distance for dependence trees and hidden Markov models. *IEEE Signal Processing Letters*, 2003, vol. 10, no. 4, pp. 115–118.

Для цитирования:

Хищенко В.Е., Федотов Н.А. Возможности анализа номинативных признаков в задачах информационной безопасности // Безопасность цифровых технологий. – 2022. – № 1 (104). – С. 61–84. – DOI: 10.17212/2782-2230-2022-1-61-84.

For citation:

Khitsenko V.E., Fedotov N.A. Vozmozhnosti analiza nominativnykh priznakov v zadachakh informatsionnoi bezopasnosti [Possibilities of analysis of nominative signs in tasks of information security]. *Bezopasnost' tsifrovykh tekhnologii = Digital Technology Security*, 2022, no. 1 (104), pp. 61–84. DOI: 10.17212/2782-2230-2022-1-61-84.