ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2023, № 5, с. 57–66

КОМПЬЮТЕРНЫЕ МЕТОДЫ

УДК 519.7

О ЧИСЛЕ РЕШЕНИЙ НЕКОТОРЫХ СПЕЦИАЛЬНЫХ ЗАДАЧ ЛОГИЧЕСКОГО АНАЛИЗА ЦЕЛОЧИСЛЕННЫХ ДАННЫХ

© 2023 г. А. П. Дюкова[a], Е. В. Дюкова[a,*]

[a]*ФИЦ ИУ РАН, Москва, Россия*

**e-mail: edjukova@mail.ru*

Поступила в редакцию 03.04.2023 г.
После доработки 28.04.2023 г.
Принята к публикации 05.06.2023 г.

В классе дискретных перечислительных задач важное место принадлежит задачам поиска в целочисленных данных часто и нечасто встречающихся элементов. Вопросы эффективности такого поиска напрямую связаны с изучением метрических (количественных) свойств множеств частых и нечастых элементов. Предполагается, что исходные данные представлены в виде целочисленной матрицы, строки которой являются описаниями исследуемых объектов в заданной системе числовых характеристик этих объектов, называемых атрибутами. Рассмотрен случай, когда каждый атрибут принимает значения из множества $\{0,1,\ldots,k-1\}$, $k \geq 2$. Приведены асимптотические оценки типичного числа специальных частых фрагментов описаний объектов, называемых правильными фрагментами, и оценки типичной длины такого фрагмента. Представлены также новые результаты, касающиеся изучения метрических свойств минимальных нечастых фрагментов описаний объектов.

DOI: 10.31857/S0002338823050050, **EDN:** OHCWCE

Введение. Рассматриваемые задачи анализа целочисленных данных возникают на этапе обучения логических процедур классификации по прецедентам. Исследование метрических (количественных) свойств множеств решений этих задач необходимо для получения теоретических оценок сложности синтеза логических классификаторов и прогноза временных затрат.

Введем основные понятия. Исследуется множество объектов M. Известно, что каждый объект множества M может быть представлен в виде числового вектора, полученного на основе наблюдения или измерения ряда его характеристик. Такие характеристики называют атрибутами. Предполагается, что каждый атрибут имеет ограниченное множество допустимых значений, которые кодируются целыми числами.

Пусть $X = \{x_1,\ldots,x_n\}$ – заданное множество атрибутов; H – набор из r атрибутов вида $H = \{x_{j_1},\ldots,x_{j_r}\}$, $j_1 < \ldots < j_r$; $\sigma = (\sigma_1,\ldots,\sigma_r)$ – набор, в котором σ_i – допустимое значение атрибута $x_{j_i}, i = \overline{1,\ r}$. Пару (σ, H) назовем элементарным фрагментом (ЭФ) ранга r. Через $W(X)$ обозначим множество всех ЭФ, порождаемых набором атрибутов X.

Пусть $S = (a_1,\ldots,a_n)$ – объект из M (здесь a_j, $j \in \{1,2,\ldots,n\}$, – значение атрибута x_j для объекта S). Будем говорить, что S *содержит* ЭФ $(\sigma,H), H = \{x_{j_1},\ldots,x_{j_r}\}, \sigma = (\sigma_1,\ldots,\sigma_r)$, если $a_{j_i} = \sigma_i$ при $i = \overline{1,\ r}$.

Дана некоторая совокупность объектов D из M и задано число p, $1 \leq p \leq |D|$, где $|D|$ – число объектов в D. Объекты в D не обязательно различны.

ЭФ (σ, H), $(\sigma,H) \in W(X)$, называется (p, D)-*частым*, если для не менее p объектов из D содержат (σ, H). ЭФ (σ,H), $(\sigma,H) \in W(X)$, ранга r, $r \leq |D|$, называется *правильным* в D, если (σ, H) – – (r, D)-частый. ЭФ (σ,H), $(\sigma,H) \in W(X)$, называется *нечастым в D*, если ни один объект из D не содержит (σ, H), и *минимальным* нечастым в D, если из условия $\sigma' \subset \sigma$, $H' \subset H$ следует, что ЭФ (σ', H') не является нечастым в D.

Логическая классификация целочисленных данных предполагает наличие нескольких непересекающихся выборок $D_1,\dots,D_l$, $l \ge 2$, объектов из M, каждая из которых представляет некоторый класс объектов. Объекты, содержащиеся в этих выборках, называются прецедентами, а атрибуты из X – признаками. На этапе обучения в каждой выборке D_i, $i \in \{1,2,\dots,l\}$, ищутся такие частые ЭФ, которые являются нечастыми в D_j при любом $j \ne i$. Найденные ЭФ позволяют различать прецеденты из разных классов и называются логическими закономерностями или представительными элементарными классификаторами [1–6].

Могут накладываться некоторые дополнительные условия на вид искомых ЭФ (в зависимости от рассматриваемой модели классификатора). Например, ищутся так называемые тупиковые представительные элементарные классификаторы. Элементарный классификатор (σ, H) называется тупиковым представительным для D_i, $i \in \{1,2,\dots,l\}$, если выполнены два условия: 1) $(\sigma, H) - (1, D_i)$-частый ЭФ; 2) (σ, H) – минимальный нечастый в D_j при любом $j \ne i$. В этом случае при поиске минимальных нечастых ЭФ возникает необходимость рассматривать труднорешаемую дискретную задачу построения тупиковых покрытий целочисленной матрицы [3], строками которой являются описания прецедентов, не принадлежащих D_i.

В [6] предложена модель логического классификатора, базирующаяся на первоначальном поиске в каждой выборке D_i, $i \in \{1,2,\dots,l\}$, правильных ЭФ и последующем отборе среди них тех, которые не содержатся в описаниях прецедентов из других классов. Данная модель демонстрирует существенное преимущество по скорости счета перед классической моделью, основанной на построении тупиковых представительных элементарных классификаторов, не уступая последней в качестве классификации.

Представляет интерес получение асимптотических оценок (при $n \to \infty$) типичного числа правильных ЭФ и оценок типичной длины правильного ЭФ. В [7] требуемые оценки получены для случая, когда число объектов в D существенно меньше числа атрибутов и каждый атрибут принимает значения из множества $\{0,1,\dots,k-1\}, k \ge 2$.

Полученные в работе новые результаты в основном касаются изучения метрических свойств множества правильных ЭФ в случае $n \le |D|$. Следует отметить, что аналогичные свойства множества минимальных нечастых ЭФ ранее изучались в ряде публикаций (например, [3, 8, 9]), в которых в том числе рассматривался случай $n \le |D|$. Приводимые в статье оценки числа минимальных нечастых ЭФ имеют вид, позволяющий сравнивать их с соответствующими оценками правильных ЭФ. Результат сравнения свидетельствует о целесообразности (в плане сокращения временных затрат) применения методов поиска частых ЭФ для синтеза логических классификаторов и согласуется с полученными в [6] результатами экспериментов на случайных модельных данных.

В разд. 1 дана постановка задачи. Исходные данные представлены в виде целочисленной матрицы, строками которой являются описания объектов из D. Приведены формулировки двух основных теорем о числе правильных ЭФ. Доказательства этих теорем содержатся в разд. 2. В разд. 3 приведены полученные ранее и новые оценки типичных значений числа минимальных нечастых ЭФ и длины минимального нечастого ЭФ.

1. Постановка задачи и формулировки основных результатов. Пусть L, $L = (a_{ij}), i = \overline{1,m}, j = \overline{1,n}$, – матрица с элементами из $\{0,1,\dots,k-1\}$, $k \ge 2$; E_k^r, $r \le n$, $k \ge 2$, – множество наборов $(\sigma_1, \dots, \sigma_r)$, $\sigma_i \in \{0,1,\dots,k-1\}$, $i = \overline{1, r}$; W_r^n, $r \le n$, – множество всех наборов вида $\{j_1,\dots,j_r\}$, где $j_t \in \{1,2,\dots,n\}$ при $t = \overline{1, r}$ и $j_1 < \dots < j_r$; V_r^m, $r \le m$, – множество всех упорядоченных наборов вида $(i_1,\dots,i_r)$, где $i_t \ne i_l$ при $t,l = \overline{1, r}$.

Положим $\sigma \in E_k^r$, $\sigma = (\sigma_1,\dots,\sigma_r)$, $w \in W_r^n$, $w = \{j_1,\dots,j_r\}$. Число r назовем *длиной* набора w.

Набор w назовем σ-*допустимым* для L, если можно указать набор $v = (i_1,\dots,i_r), v \in V_r^m$, такой, что $a_{i_t j_t} = \sigma_t$ при $t = \overline{1, r}$. Будем говорить, что σ-допустимый набор w *порождается* набором σ.

Нетрудно видеть, что в случае, когда в качестве строк матрицы L берутся описания объектов из выборки D, то набор $w \in W_r^n$, $w = \{j_1,\dots,j_r\}$, является σ-допустимым для L тогда и только тогда, когда ЭФ (σ, H), $H = \{x_{j_1},\dots,x_{j_r}\}$ – правильный в D.

Введем обозначения: $\mathfrak{M}_{mn}^k$ – множество всех матриц размера $m\times n$ с элементами из $\{0,1,\ldots,k-1\}$, $k\ge 2$; $U(L,\sigma)$, $L\in\mathfrak{M}_{mn}^k$, $\sigma\in E_k^r$, – множество всех σ-допустимых наборов для матрицы L; $U_r(L,\sigma)$ – множество всех наборов в $U(L,\sigma)$ длины r; $U(L)$, $L\in\mathfrak{M}_{mn}^k$, – совокупность всех допустимых для матрицы L наборов, в которой каждый набор встречается столько раз, сколькими наборами из E_k^r он порождается; $|N|$ – мощность множества N;

$$|U_r(L)|=\sum_{\sigma\in E_k^r}|U_r(L,\sigma)|;$$

$$|U(L)|=\sum_{r=1}^{n}\sum_{\sigma\in E_k^r}|U_r(L,\sigma)|;$$

$r_1=[0.5\log_k mn-0.5\log_k\log_k^2 mn-\log_k\log_k\log_k n]$, здесь и далее $[q]$ – целая часть от числа q; $r_2=\left]0.5\log_k mn-0.5\log_k\log_k^2 mn+\log_k\log_k\log_k n\right[$, здесь и далее $]q[$ – наименьшее целое, превосходящее q; ϕ_1 – интервал $[r_1,r_2]$; $r_3=]\log_k m+\log_k\log_k m[$; ϕ_2 – интервал $[1,r_3]$; $b_n\approx c_n$, $n\to\infty$, означает, что $\lim_{n\to\infty}b_n/c_n=1$; $b_n\preceq c_n$, $n\to\infty$, означает, что $\lim_{n\to\infty}b_n/c_n\le 1$.

Ниже приводятся асимптотические оценки типичного значения величины $|U(L)|$ и оценки типичной длины допустимого для L набора при различных значениях m и n.

Выявление типичной ситуации связано с высказыванием типа “для почти всех матриц L из $\mathfrak{M}_{mn}^k$ при $n\to\infty$ выполнено $F_1(L)\approx F_2(L)$” (здесь $F_1(L)$ и $F_2(L)$ – два функционала, заданные на матрицах из $\mathfrak{M}_{mn}^k$). Данное высказывание означает, что существуют две положительные бесконечно убывающие функции $\alpha(n)$ и $\beta(n)$, такие, что для всех достаточно больших n имеет место

$$1-|\mathfrak{M}|/|\mathfrak{M}_{mn}^k|\le\alpha(n),$$

где $\mathfrak{M}$ – множество таких матриц L в $\mathfrak{M}_{mn}^k$, для которых

$$1-\beta(n)<|F_1(L)|/|F_2(L)|<1+\beta(n).$$

Справедливы приведенные ниже теоремы 1 и 2.

Т е о р е м а 1. Если $m^a\le n\le k^{m^\beta}, a>1,\beta<1,k\ge 2$, то при $n\to\infty$ для почти всех матриц L из $\mathfrak{M}_{mn}^k$ имеет место

$$\sum_{r\le r_1}|U_r(L)|\approx|U_{r_1}(L)|\approx C_n^{r_1}C_m^{r_1}k^{r_1-r_1^2},$$

$$\sum_{r\ge r_2}|U_r(L)|\approx|U_{r_2}(L)|\approx C_n^{r_2}C_m^{r_2}k^{r_2-r_2^2},$$

$$|U(L)|\approx\sum_{r\in\phi_1}|U_r(L)|\approx\sum_{r\in\phi_1}C_n^rC_m^rk^{r-r^2}$$

и длины почти всех наборов из $U(L)$ принадлежат интервалу ϕ_1.

Т е о р е м а 2. Если $n\le m\le k^{n^\beta}, \beta<1/2, k\ge 2$, то при $n\to\infty$ для почти всех матриц L из $\mathfrak{M}_{mn}^k$ имеет место

$$\sum_{r\ge r_3}|U_r(L)|\approx|U_{r_3}(L)|\approx C_n^{r_3}C_m^{r_3}k^{r_3-r_3^2},$$

$$|U(L)|\precsim\sum_{r\in\phi_2}C_n^rC_m^rk^{r-r^2}$$

и длины почти всех наборов из $U(L)$ принадлежат интервалу ϕ_2.

Доказательства теорем 1 и 2 опираются на ряд лемм, приводимых в разд. 2.

2. Доказательства теорем 1 и 2. Пусть $v \in V_r^m, v = (i_1,\dots,i_r)$; $\sigma \in E_k^r, \sigma = (\sigma_1,\dots,\sigma_r)$; $w \in W_r^n$, $w = \{j_1,\dots,j_r\}$. Матрица $L = (a_{ij})$, $i = \overline{1,m}$, $j = \overline{1,n}$, $L \in \mathfrak{M}_{mn}^k$, называется (v,σ,w)-матрицей, если $a_{i_t j_t} = \sigma_t$ при $t = \overline{1,r}$. Обозначим через $N_{(v,\sigma,w)}$ совокупность (v,σ,w) -матриц в $\mathfrak{M}_{mn}^k$, через $N_{(v,\sigma,w)}^*$ – совокупность всех матриц L в $N_{(v,\sigma,w)}$, таких, что $L \notin N_{(v_1,\sigma,w)}$ при $v_1 \in V_r^m, v_1 \neq v$.

Лемма 1. Если $v \in V_r^m, w \in W_r^n$, $\sigma \in E_k^r$, то

$$\left|N_{(v,\sigma,w)}\right| = k^{mn-r^2}.$$

Доказательство. Оценим, сколькими способами можно построить матрицу L из $N_{(v,\sigma,w)}$. Однозначным образом определяются те элементы матрицы L, которые расположены на пересечении строк с номерами из v и столбцов с номерами из w. Остальные элементы этой матрицы могут быть выбраны произвольным образом (k^{mn-r^2} способов). Отсюда получаем требуемую оценку. Лемма 1 доказана.

Лемма 2. Если $v \in V_r^m, w \in W_r^n, \sigma \in E_k^r$, то

$$\left|N_{(v,\sigma,w)}^*\right| = (1-k^{-r})^{m-r} k^{mn-r^2}.$$

Доказательство. Оценим, сколькими способами можно построить матрицу L из $N_{(v,\sigma,w)}^*$. Элементы этой матрицы, расположенные в столбцах с номерами не входящими в w, могут быть выбраны произвольным образом ($k^{m(n-r)}$ способов). Отсюда, учитывая, что строки в подматрице матрицы L, образованной столбцами с номерами из w, можно выбрать $(k^r-1)^{m-r}$ способами, получаем требуемую оценку. Лемма 2 доказана.

Лемма 3. Пусть $v_1 \in V_r^m$, $v_2 \in V_l^m$, $w_1 \in W_r^n$, $w_2 \in W_l^n$, $\sigma' \in E_k^r$, $\sigma'' \in E_k^l$ и наборы v_1 и v_2 пересекаются по $a (a \geq 0)$ элементам, а наборы w_1 и w_2 пересекаются по $b (b \geq 0)$ элементам. Тогда

$$\left|N_{(v_1,\sigma',w_1)} \cap N_{(v_2,\sigma'',w_2)}\right| \leq k^{mn-r^2-l^2+ab}.$$

Доказательство леммы 3 не приводится в силу ее очевидности.

При доказательстве приводимых ниже лемм 4–6 используется выражение $b_n \leq_n c_n$, которое означает, что $b_n \leq c_n$ при всех достаточно больших n.

Лемма 4. 1. Если $m \leq n \leq k^{m^\beta}, \beta < 1$, то

$$\sum_{r \leq r_1} C_n^r C_m^r k^{r-r^2} \precsim C_n^{r_1} C_m^{r_1} k^{r_1-r_1^2}, \quad n \to \infty.$$

2. Имеет место

$$\sum_{r \geq r_2} C_n^r C_m^r k^{r-r^2} \precsim C_n^{r_2} C_m^{r_2} k^{r_2-r_2^2}, \quad n \to \infty.$$

3. Если $n \leq m$, то

$$\sum_{r \geq r_3} C_n^r C_m^r k^{r-r^2} \precsim C_n^{r_3} C_m^{r_3} k^{r_3-r_3^2}, \quad n \to \infty.$$

Доказательство. Положим $a_r = C_n^r C_m^r k^{r-r^2}$, $q = 0.5\log_k mn - 0.5\log_k \log_k^2 mn$, $t = \log_k \log_k \log_k n$.

1. Пусть $m \leq n \leq k^{m^\beta}, \beta < 1$, и $r \leq r_1 + 1$. Тогда, пользуясь тем, что $q \leq 0.5\log_k mn$, $k^{2q} = mn/\log_k^2 mn$ и $(n-q) \geq_n 0.5n$ при $m \leq n$, $(m-q) \geq_n 0.5m$ при $n \leq 2^{m^\beta}$, получаем

$$\frac{a_{r-1}}{a_r} = \frac{r^2 k^{2r-2}}{(n-r+1)(m-r+1)} \leq \frac{q^2 k^{2q-2t}}{(n-q)(m-q)} \leq_n k^{-2t}.$$

2. При $r \geq r_2 - 1$ получаем

$$\frac{a_{r+1}}{a_r} \leq \frac{mn}{r^2} k^{-2r} \leq_n \frac{mn}{q^2} k^{-2q-2t+2} \leq_n k^{-2t}.$$

3. При $n \leq m,\ r \geq r_3 - 1$ получаем

$$\frac{a_{r+1}}{a_r} \leq \frac{mn}{r^2} k^{-2r} \leq_n \frac{1}{(\log_k n)^2}.$$

Таким образом, $a_{r-1} = o(a_r),\ n \to \infty$, в случае 1 и $a_{r+1} = o(a_r),\ n \to \infty$, в каждом из случаев 2 и 3. Лемма 4 доказана.

Л е м м а 5. Если $m \leq n$ и $r,l \leq r_2$, то имеет место

$$\sum_{b=0}^{\min(r,l)} k^{lb} C_n^r C_r^b C_{n-r}^{l-b} \leq C_n^r C_n^l (1 + \delta(n)),$$

где $\delta(n) \to 0$ при $n \to \infty$.

Д о к а з а т е л ь с т в о. Обозначим $\lambda_b = k^{lb} C_n^r C_r^b C_{n-r}^{l-b} / C_n^r C_{n-r}^l$. Так как

$$\frac{C_r^b C_{n-r}^{l-b}}{C_{n-r}^l} \leq \left(\frac{rl}{n-r-l}\right)^b$$

и по условию $r,l \leq_n 0.5\log_k mn \leq \log_k n, (r+l)/n \leq_n 0.5$, то

$$\lambda_b \leq_n \left(\frac{2\log_k^2 n}{n}\right)^b.$$

Следовательно, оцениваемая сумма не превосходит $C_n^r C_{n-r}^l (1 + \delta(n))$, где $\delta(n) \to 0$ при $n \to \infty$. Отсюда, пользуясь неравенством $C_{n-r}^l \leq C_n^l$, получаем утверждением леммы. Лемма 5 доказана.

Л е м м а 6. Если $m \leq k^{n^\beta}, \beta < 1/2$, и $r,l \leq r_3$, то имеет место

$$\sum_{b=0}^{\min(r,l)} k^{lb} C_n^r C_r^b C_{n-r}^{l-b} < C_n^r C_n^l (1 + \delta(n)),$$

где $\delta(n) \to 0$ при $n \to \infty$.

Доказательство леммы 6 аналогично доказательству леммы 5 (в этом случае $r,\ l \leq 2n^\beta$ и $\lambda_b \leq_n (8n^{2\beta-1})^b$).

Будем считать $\mathfrak{M}_{mn}^k = \{L\}$ пространством элементарных событий, в котором каждое событие L происходит с вероятностью $1/\left|\mathfrak{M}_{mn}^k\right|$. Математическое ожидание случайной величины $X(L)$, определенной на множестве $\mathfrak{M}_{mn}^k$, будем обозначать через $\mathbf{M}X(L)$, дисперсию – через $\mathbf{D}X(L)$.

Л е м м а 7 [10]. Пусть для случайных величин $X_1(L)$ и $X_2(L)$, определенных на $\mathfrak{M}_{mn}^k$, выполнено $X_1(L) \geq X_2(L) \geq 0$ и при $n \to \infty$ верно $\mathbf{M}X_1(L) \approx \mathbf{M}X_2(L), \mathbf{D}X_2(L)/(\mathbf{M}X_2(L))^2 \to 0$. Тогда для почти всех матриц L из $\mathfrak{M}_{mn}^k$ имеет место $X_1(L) \approx X_2(L) \approx \mathbf{M}X_2(L), n \to \infty$.

Пусть $\sigma \in E_k^r,\ w \in W_r^n$. На $\mathfrak{M}_{mn}^k = \{L\}$ рассмотрим случайную величину $\zeta_{(\sigma,w)}(L)$, равную 1, если w – σ-допустимый набор для матрицы L, и равную 0 в противном случае. Положим

$$\mu_r(L) = \sum_{w \in W_r^n} \sum_{\sigma \in E_k^r} \zeta_{(\sigma,w)}(L), \quad \zeta(L) = \sum_{r=1}^{\min(m,n)} \mu_r(L), \quad \zeta_i(L) = \sum_{r \in \phi_i} \mu_r(L), \quad i \in \{1,2\}.$$

Нетрудно видеть, что $\mu_r(L) = \left|U_r(L)\right|$ (число наборов в $U(L)$ длины r, $\zeta(L) = \left|U(L)\right|$ и $\zeta_i(L), i \in \{1,2\}$, – число тех наборов в $U(L)$, длины которых принадлежат интервалу ϕ_i.

Оценим вероятность события $\zeta_{(\sigma,w)}(L)=1$, $\sigma\in E_k^r$, $w\in W_r^n$, обозначаемую далее через $P(\zeta_{(\sigma,w)}(L)=1)$. Очевидно, в силу леммы 1

$$P(\zeta_{(\sigma,w)}(L)=1)\le\sum_{v\in V_r^m}\left|N_{(v,\sigma,w)}\right|/|\mathfrak{M}_{mn}^k|=C_m^r k^{-r^2}. \tag{2.1}$$

С другой стороны, в силу леммы 2 имеем

$$P(\zeta_{(\sigma,w)}(L)=1)\ge\sum_{v\in V_r^m}|N_{(v,\sigma,w)}^*|/|\mathfrak{M}_{mn}^k|=C_m^r(1-k^{-r})^{m-r}k^{-r^2}. \tag{2.2}$$

Из (2.1), а также леммы 4 сразу вытекает следующая лемма.

Л е м м а 8. Если $m\le n\le k^{m^\beta}$, $\beta<1$, то имеет место

$$\mathbf{M}\mu_{r_1}(L)\le C_n^{r_1}C_m^{r_1}k^{r_1-r_1^2},\quad n\to\infty,$$

$$\sum_{r\le r_1}\mathbf{M}\mu_r(L)\precsim C_n^{r_1}C_m^{r_1}k^{r_1-r_1^2},\quad n\to\infty.$$

Л е м м а 9. Если $m^a\le n, a>1$, то

$$\mathbf{M}\mu_{r_1}(L)\succsim C_n^{r_1}C_m^{r_1}k^{r_1-r_1^2},\quad n\to\infty.$$

$$\sum_{r\le r_1}\mathbf{M}\mu_r(L)\succsim C_n^{r_1}C_m^{r_1}k^{r_1-r_1^2},\quad n\to\infty.$$

Д о к а з а т е л ь с т в о. Имеем

$$\sum_{r\le r_1}\mathbf{M}\mu_r(L)\ge\mathbf{M}\mu_{r_1}(L).$$

Так как $mk^{-r_1}\to 0$, $n\to\infty$, то $(1-k^{-r_1})^{m-r_1}\to 1$, $n\to\infty$. Откуда, пользуясь (2.2), получаем

$$\mathbf{M}\mu_{r_1}(L)\succsim C_n^{r_1}C_m^{r_1}k^{r_1-r_1^2},\quad n\to\infty.$$

Лемма 9 доказана.

Из лемм 8 и 9 сразу вытекает следующая лемма.

Л е м м а 10. Если $m^a\le n\le k^{m^\beta}$, $a>1$, $\beta<1$, то

$$\sum_{r\le r_1}\mathbf{M}\mu_r(L)\approx\mathbf{M}\mu_{r_1}(L)\approx C_n^{r_1}C_m^{r_1}k^{r_1-r_1^2},\quad n\to\infty.$$

Доказательства представленных ниже лемм 11–13 не приводятся, поскольку они полностью аналогичны доказательству леммы 10.

Л е м м а 11. Если $m^a\le n, a>1$, то имеет место

$$\sum_{r\ge r_2}\mathbf{M}\mu_r(L)\approx\mathbf{M}\mu_{r_2}(L)\approx C_n^{r_2}C_m^{r_2}k^{r_2-r_2^2},\quad n\to\infty.$$

Л е м м а 12. Если $n\le m$, то

$$\sum_{r\ge r_3}\mathbf{M}\mu_r(L)\approx\mathbf{M}\mu_{r_3}(L)\approx C_n^{r_3}C_m^{r_3}k^{r_3-r_3^2},\quad n\to\infty.$$

Л е м м а 13. Если $m^a\le n\le k^{m^\beta}$, $a>1, \beta<1$, то

$$\mathbf{M}\zeta(L)\approx\mathbf{M}\zeta_1(L)\approx\sum_{r\in\phi_1}C_n^rC_m^rk^{r-r^2},\quad n\to\infty.$$

Л е м м а 14. Если $m^a\le n\le k^{m^\beta}$, $a>1, \beta<1$, то имеет место

$$\mathbf{D}\zeta_1(L)/(\mathbf{M}\zeta_1(L))^2\to 0,\quad n\to\infty.$$

Доказательство. Имеем

$$\mathbf{D}\zeta_1(L) = \mathbf{M}(\zeta_1(L))^2 - (\mathbf{M}\zeta_1(L))^2. \tag{2.3}$$

Нетрудно видеть, что

$$\mathbf{M}(\zeta_1(L))^2 \le \sum_{r,l\in\phi_1} \sum_{\substack{v_1\in V_r^m, v_2\in V_l^m \\ w_1\in W_r^n, w_2\in W_l^n}} \sum_{\substack{\sigma'\in E_k^r \\ \sigma''\in E_k^l}} |N| / k^{mn},$$

где $N = N_{(v_1,\sigma',w_1)} \cap N_{(v_2,\sigma'',w_2)}$. Отсюда, пользуясь леммами 3 и 5, получаем

$$\mathbf{M}(\zeta_1(L))^2 \le \sum_{r,l\in\phi_1} \sum_{b=0}^{\min(r,l)} k^{r+l}k^{-r^2-l^2+lb} C_n^r C_r^b C_{n-r}^{l-b} C_m^r C_m^l \le$$

$$\le \sum_{r,l\in\phi_1} C_n^r C_n^l C_m^r C_m^l k^{r+l}k^{-r^2-l^2}(1+\delta(n)), \tag{2.4}$$

где $\delta(n) \to 0$ при $n \to \infty$.

С другой стороны, в силу леммы 13

$$(\mathbf{M}\zeta_1(L))^2 \approx \sum_{r,l\in\phi_1} C_n^r C_n^l C_m^r C_m^l k^{r+l}k^{-r^2-l^2}, \quad n\to\infty. \tag{2.5}$$

Из (2.3)–(2.5) следует утверждение доказываемой леммы. Лемма 14 доказана.

Аналогично лемме 14 доказываются приводимые ниже леммы 15–17.

Лемма 15. Если $m^a \le n \le k^{m^\beta}$, $a > 1$, $\beta < 1$, то

$$\mathbf{D}\mu_{r_1}(L) / (\mathbf{M}\mu_{r_1}(L))^2 \to 0, \quad n\to\infty.$$

Лемма 16. Если $m^a \le n$, $a > 1$, то

$$\mathbf{D}\mu_{r_2}(L)/(\mathbf{M}\mu_{r_2}(L))^2 \to 0, \quad n\to\infty.$$

Лемма 17. Если $n \le m$, то

$$\mathbf{D}\mu_{r_3}(L)/(\mathbf{M}\mu_{r_3}(L))^2 \to 0, \quad n\to\infty.$$

Пусть $v \in V_r^m, \sigma \in E_k^r, w \in W_r^n$. На $\mathfrak{M}_{mn}^k = \{L\}$ рассмотрим случайную величину $\xi_{(v,\sigma,w)}(L)$, равную 1, если $L \in N_{(v,\sigma,w)}$, и равную 0 в противном случае. Положим

$$\xi(L) = \sum_{r=1}^{\min(m,n)} \sum_{v\in V_r^m, w\in W_r^n} \sum_{\sigma\in E_k^r} \xi_{(v,\sigma,w)}(L),$$

$$\xi_1(L) = \sum_{r\in\phi_2} \sum_{v\in V_r^m, w\in W_r^n} \sum_{\sigma\in E_k^r} \xi_{(v,\sigma,w)}(L).$$

Лемма 18. Если $n \le m \le k^{n^\beta}$, $\beta < 1/2$, то при $n \to \infty$ для почти всех матриц L из $\mathfrak{M}_{mn}^k$ имеет место

$$\xi(L) \approx \xi_1(L) \approx \sum_{r\in\phi_2} C_n^r C_m^r k^{r-r^2}.$$

Доказательство. Оценим вероятность события, $\xi_{(v,\sigma,w)}(L) = 1, v \in V_r^m, \sigma \in E_k^r, w \in W_r^n$, обозначаемую далее через $P(\xi_{(v,\sigma,w)}(L) = 1)$. В силу леммы 1

$$P(\xi_{(v,\sigma,w)}(L) = 1) = |N_{(v,\sigma,w)}|/|\mathfrak{M}_{mn}^k| = k^{-r^2}.$$

Следовательно, согласно лемме 4,

$$\mathbf{M}\xi(L) \approx \mathbf{M}\xi_1(L) \approx \sum_{r\in\phi_2} C_n^r C_m^r k^{r-r^2}, \quad n\to\infty. \tag{2.6}$$

Из (2.6) и леммы 6, используя схему доказательства леммы 14, получаем

$$\mathbf{D}\xi_1(L)/(\mathbf{M}\xi_1(L))^2, \quad n\to\infty. \tag{2.7}$$

Из (2.6), (2.7) и леммы 7 следует утверждение доказываемой леммы. Лемма 18 доказана.

Утверждения теоремы 1 следуют непосредственно из лемм 7, 10, 11, 13, 14–16, а утверждения теоремы 2 следуют непосредственно из лемм 7, 12, 17, 18 и неравенства $\zeta(L) \le \xi(L)$.

3. Оценки типичных значений числа минимальных нечастых ЭФ и длины минимального нечастого ЭФ. Положим $L \in \mathfrak{M}_{mn}^k, L = (a_{ij}),\ i = 1,\ldots,m,\ j = 1,\ldots,n;\ \sigma \in E_k^r, \sigma = (\sigma_1,\ldots,\sigma_r); w \in W_r^n, w = \{j_1,\ldots,j_r\}$.

Набор w называется σ-*покрытием матрицы L длины r*, если для любого $i \in \{1,2,\ldots,m\}$ найдется $j \in \{j_1,\ldots,j_r\}$, такое, что $a_{ij} \ne \sigma_j$. Будем говорить, что σ-покрытие w *порождается* набором σ.

Набор w, являющийся σ-покрытием матрицы L, называется тупиковым, если при любом $t \in \{1,2,\ldots,\ r\}$ набор $w\backslash\{j_t\}$ не является γ_t-покрытием матрицы L, где $\gamma_t = (\sigma_1,\ldots,\sigma_{t-1},\sigma_{t+1},\ldots,\sigma_r)$. Если w – тупиковое σ-покрытие матрицы L, то нетрудно видеть, что столбцы матрицы L с номерами из w содержат подматрицу, имеющую с точностью до перестановки строк вид

$$\begin{pmatrix} \beta_1\sigma_2\sigma_3\ldots\sigma_{r-1}\sigma_r \\ \sigma_1\beta_2\sigma_3\ldots\sigma_{r-1}\sigma_r \\ \ldots \\ \sigma_1\sigma_2\sigma_3\ldots\sigma_{r-1}\beta_r \end{pmatrix},$$

где $\beta_p \ne \sigma_p$ при $p = 1,2,\ldots,r$. Такая подматрица называется σ-*подматрицей*.

Отметим, что в случае, когда в качестве строк матрицы L берутся описания объектов из выборки D, то набор $w \in W_r^n,\ w = \{j_1,\ldots,j_r\}$, является тупиковым σ-покрытием матрицы L тогда и только тогда, когда ЭФ $(\sigma, H),\ H = \{x_{j_1},\ldots,x_{j_r}\}$ – минимальный нечастый в D.

Введем обозначения: $B(L,\sigma),\ L \in \mathfrak{M}_{mn}^k,\ \sigma \in E_k^r,$ – множество всех тупиковых σ-покрытий матрицы L; $S(L,\sigma),\ L \in \mathfrak{M}_{mn}^k,\ \sigma \in E_k^r,$ – множество всех σ-подматриц матрицы L; $B_r(L,\ \sigma),\ L \in \mathfrak{M}_{mn}^k, \sigma \in E_k^r,$ – множество всех наборов в $B(L,\ \sigma)$ длины r; $S_r(L,\sigma),\ L \in \mathfrak{M}_{mn}^k,\ \sigma \in E_k^r,$ – множество всех подматриц в $S(L,\sigma)$ порядка r; $B(L),\ L \in \mathfrak{M}_{mn}^k,$ – совокупность всех тупиковых σ-покрытий матрицы L, в которой каждое покрытие встречается столько раз, сколькими наборами из E_k^r оно порождается; $S(L),\ L \in \mathfrak{M}_{mn}^k,$ – совокупность всех σ-подматриц матрицы L для всех σ из E_k^r;

$$|B(L)| = \sum_{r=1}^{n} \sum_{\sigma\in E_k^r} |B_r(L,\sigma)|;$$

$$|S(L)| = \sum_{r=1}^{n} \sum_{\sigma\in E_k^r} |S_r(L,\sigma)|;$$

$r_3 =]\log_k m + \log_k\log_k m[$; ϕ_2 – интервал $[1,\ r_3]$; $r_4 = [0.5\log_k mn - 0.5\log_k\log_k mn - \log_k\log_k\log_k n]$; $r_5 =]0.5\log_k mn - 0.5\log_k\log_k mn + \log_k\log_k\log_k n[$; ϕ_3 – интервал $[r_4,\ r_5]$; $r_6 = [\log_k m + \log_k\log_k m + \log_k\log_k\log_k n[$; ϕ_4 – интервал $[1,\ r_6]$.

Т е о р е м а 3 [3]. Если $m^a \le n \le k^m, a > 1,\ k \ge 2$, то при $n\to\infty$ для почти всех матриц L из $\mathfrak{M}_{mn}^k$ имеет место

$$\sum_{r\le r_4} |B_r(L)| \approx |B_{r_4}(L)| \approx C_n^{r_4} C_m^{r_4} r!(k-1)^{r_4} k^{r_4-r_4^2},$$

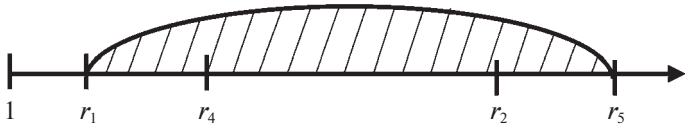

Рис. 1. Типичные значения длин наборов из $U(L)$ (см. разд. 1) и $B(L)$ в случае $m^a \le n \le k^{m^\beta}, a > 1, \beta < 1$.

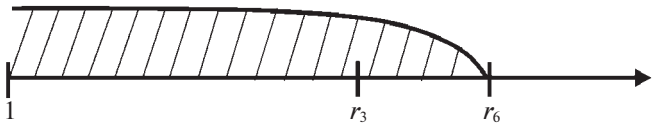

Рис. 2. Типичные значения длин наборов из $U(L)$ (см. разд. 1) и $B(L)$ в случае $n \le m \le k^{n^\beta}, \beta < 1/2$.

$$\sum_{r \ge r_5} |B_r(L)| \approx |B_{r_5}(L)| \approx C_n^{r_5} C_m^{r_5} r!(k-1)^{r_5} k^{r_5 - r_5^2},$$

$$|B(L)| \approx |S(L)| \approx \sum_{r \in \phi_3} C_n^r C_m^r r!(k-1)^r k^{r-r^2}$$

и длины почти всех наборов из $B(L)$ принадлежат интервалу ϕ_3.

Т е о р е м а 4. Если $n \le m \le k^{n^\beta}, \beta < 1/2, k \ge 2$, то при $n \to \infty$ для почти всех матриц L из $\mathfrak{M}_{mn}^k$ имеет место

$$\sum_{r \ge r_6} |B_r(L)| \approx |B_{r_6}(L)| \approx C_n^{r_6} C_m^{r_6} r!(k-1)^{r_6} k^{r_6 - r_6^2},$$

$$|B(L)| \le |S(L)| \approx \sum_{r \in \phi_2} C_n^r C_m^r r!(k-1)^r k^{r-r^2}$$

и длины почти всех наборов из $B(L)$ принадлежат интервалу ϕ_4.

Схема доказательства теоремы 4 аналогична схеме доказательства теоремы 2.

Таким образом, в каждом из двух рассмотренных случаев почти всегда типичная длина набора из $U(L)$ и типичная длина набора из $B(L)$ принадлежат одному интервалу. Результаты теорем 1, 3 и теорем 2, 4 проиллюстрированы ниже соответственно на рис. 1, 2.

Заключение. Рассмотрены актуальные вопросы логического анализа целочисленных данных, касающиеся исследования метрических (количественных) свойств множеств частых и нечастых элементов таких данных. Усовершенствована техника получения оценок для типичных значений основных числовых характеристик указанных множеств и найдены новые оценки таких характеристик. Приведено теоретическое обоснование целесообразности (в плане сокращения временных затрат) применения методов поиска частых элементов на этапе обучения классификаторов, базирующихся на логическом анализе обучающей выборки.

Результаты проведенного в работе исследования важны и для ряда других прикладных областей, среди которых следует выделить нахождение в данных ассоциативных правил. В этом случае D называют базой данных, а каждый объект базы D – транзакцией. Ассоциативное правило устанавливает зависимость между двумя частыми ЭФ, согласно которой один частый ЭФ (посылка) с некоторой “достоверностью” влечет другой частый ЭФ (следствие). При этом посылка и следствие порождаются одним общим частым ЭФ. Вопросы синтеза ассоциативных правил возникли в связи с анализом потребительской корзины [11].

СПИСОК ЛИТЕРАТУРЫ

1. *Баскакова Л.В., Журавлев Ю.И.* Модель распознающих алгоритмов с представительными наборами и системами опорных множеств // ЖВМ и МФ. 1981. Т. 21. № 5. С. 1264–1275.

2. *Hammer P.L.* Partially Defined Boolean Functions and Cause-effect Relationships // Lecture at the Intern. Conf. on Multi-Attrubute Decision Making Via ORBased Expert Systems. Passau, Germany: University of Passau, 1986.
3. *Дюкова Е.В., Журавлев Ю.И.* Дискретный анализ признаковых описаний в задачах распознавания большой размерности // ЖВМ и МФ. 2000. Т. 40. № 8. С. 1264–1278.
4. *Дюкова Е.В., Песков Н.В.* Поиск информативных фрагментов описаний объектов в дискретных процедурах распознавания // ЖВМ и МФ. 2002. Т. 42. № 5. С. 741–753.
5. *Журавлев Ю.И., Рязанов В.В., Сенько О.В.* Распознавание. Математические методы. Программная система. Практические применения. М.: ФАЗИС, 2006. 159 с.
6. *Dragunov N., Djukova E. and Djukova A.* Supervised Classification and Finding Frequent Elements in Data // VIII Intern. Conf. on Information Technology and Nanotechnology (ITNT-2022). Samara, Russian Federation: IEEE, 2022. P. 1–5.
7. *Дюкова Е.В., Дюкова А.П.* О сложности обучения логических процедур классификации // Информатика и ее применения. 2022. Т. 16. Вып. 4. С. 57–62.
8. *Андреев А.Е.* Об асимптотическом поведении числа тупиковых тестов и длины минимального теста для почти всех таблиц // Пробл. кибернетики. 1984. Вып. 41. С. 117–142.
9. *Дюкова Е.В., Сотнезов Р.М.* Асимптотические оценки числа решений задачи дуализации и ее обобщений // ЖВМ и МФ. 2011. Т. 51. № 8. С. 1431–1440.
10. *Носков В.Н., Слепян В.А.* О числе тупиковых тестов для одного класса таблиц // Кибернетика. 1972. № 1. С. 60–65.
11. *Aggarwal Charu C.* Frequent Pattern Mining. N. Y.: Springer International Publishing, 2014. 469 p. (https://www.charuaggarwal.net/freqbook.pdf).