ISSN 1064-2307, Journal of Computer and Systems Sciences International, 2023, Vol. 62, No. 5, pp. 817–826. © Pleiades Publishing, Ltd., 2023.
ISSN 1064-2307, Journal of Computer and Systems Sciences International, 2023. © Pleiades Publishing, Ltd., 2023.
Russian Text © The Author(s), 2023, published in Izvestiya Akademii Nauk, Teoriya i Sistemy Upravleniya, 2023, No. 5, pp. 57–66.

COMPUTER METHODS

Number of Solutions for Some Special Logical Analysis Problems of Integer Data

A. P. Djukova[a] and E. V. Djukova[a,*]

[a] *Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, 119333 Russia*

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

Received April 3, 2023; revised April 28, 2023; accepted June 5, 2023

Abstract—In the class of discrete enumeration problems, an important place belongs to the problems of searching for frequently and infrequently occurring elements in integer data. Questions on the effectiveness of such a search are directly related to the study of the metric (quantitative) properties of sets of frequent and infrequent elements. It is assumed that the initial data are presented in the form of an integer matrix, whose rows are descriptions of the studied objects in the given system of the numerical characteristics of these objects, called attributes. The case is considered when each attribute takes values from the set $\{0,1,\ldots,k-1\}$, $k \geq 2$. Asymptotic estimates for the typical number of special, frequent fragments of object descriptions, called correct fragments, and estimates for the typical length of such a fragment are given. We also present new results concerning the study of the metric properties of the minimal infrequent fragments of descriptions of objects.

DOI: 10.1134/S1064230723050052

INTRODUCTION

The considered problems of the analysis of integer data arise at the stage of training logical classification procedures by precedents. The metric (quantitative) properties of the sets of solutions to these problems need to be studied in order to obtain theoretical estimates of the complexity of the synthesis of logical classifiers and forecast the time costs.

We introduce the basic concepts. The set of M objects are explored. It is known that each object of the set M can be represented as a numerical vector obtained based on the observation or measurement of a number of its characteristics. Such characteristics are called attributes. It is assumed that each attribute has a limited set of valid values, which are encoded as integers.

Assume $X = \{x_1,\ldots,x_n\}$ is the given set of attributes; H is the set from r attributes of the form $H = \{x_{j_1},\ldots,x_{j_r}\}$, $j_1 < \ldots < j_r$; and $\sigma = (\sigma_1,\ldots,\sigma_r)$ the set in which σ_i is an admissible value of attribute x_{j_i}, $i = \overline{1, r}$. The pair (σ, H) is called an elementary fragment (EF) of rank r. The set of all EFs generated by the set of attributes X is denoted through $W(X)$.

We assume $S = (a_1,\ldots,a_n)$ is an object from M (here a_j, $j \in \{1,2,\ldots,n\}$, is the value of attribute x_j for object S). We will consider that S *contains* $\mathrm{EF}(\sigma,H)$, $H = \{x_{j_1},\ldots,x_{j_r}\}$, $\sigma = (\sigma_1,\ldots,\sigma_r)$ if $a_{j_i} = \sigma_i$ at $i = \overline{1, r}$.

The set of objects D from M and number p, $1 \leq p \leq |D|$, where $|D|$ is the number of objects in D, are given. The objects in D are not necessarily different.

EF (σ,H), $(\sigma,H) \in W(X)$ is called (p, D)-*frequent* if at least p objects from D contain (σ, H). EF (σ,H), $(\sigma,H) \in W(X)$, of rank r, $r \leq |D|$ is called *correct* in D if (σ, H) is (r,D)-frequent. EF (σ,H), $(\sigma,H) \in W(X)$, is called *infrequent* in D if no object from D contains (σ, H), and it is called *minimal* infrequent in D if from the condition $\sigma' \subset \sigma$, $H' \subset H$ it follows that $\mathrm{EF}(\sigma', H')$ is not infrequent in D.

The logical classification of integer data assumes the presence of several nonoverlapping samples $D_1,\ldots,D_l$, $l \geq 2$, of objects from M, each of which represents a certain class of objects. The objects contained in these samples are called precedents, and the attributes from X are called features. At the training

stage in each sample D_i, $i \in \{1,2,\ldots,l\}$, we search for those frequent EFs that are infrequent in D_j for any $j \neq i$. The found EFs make it possible to distinguish precedents from different classes and are called logical patterns or representative elementary classifiers [1–6].

Some additional conditions may be imposed on the type of the desired EF (depending on the classifier model under consideration). For example, the so-called irredundant representative elementary classifiers are sought. An elementary classifier (σ, H) is called a irredundant representative for D_i, $i \in \{1,2,\ldots,l\}$ if two conditions are met: (1) (σ, H) is an $(1, D_i)$-frequent EF; and (2) (σ, H) is minimal infrequent in D_j for any $j \neq i$. In this case, when searching minimal infrequent EFs we need to consider the intractable discrete problem of constructing the irredundant covers of an integer matrix [3], whose rows are descriptions of precedents that do not belong to D_i.

In [6], a model of a logical classifier is proposed, based on the initial search in each sample D_i, $i \in \{1,2,\ldots,l\}$, of the correct EFs and the subsequent selection among them of those that are not contained in the descriptions of precedents from other classes. This model demonstrates a significant advantage in terms of counting speed over the classical model based on the construction of irredundant representative elementary classifiers, which are not inferior to the latter in terms of classification.

It is of interest to obtain asymptotic estimates (for $n \to \infty$) of the typical number of correct EFs and estimates of the typical length of the correct EF. In [7], the required estimates are obtained for the case when the number of objects in D is significantly less than the number of attributes and each attribute takes values from the set $\{0,1,\ldots,k-1\}$, $k \geq 2$.

The new results obtained in this paper mainly concern research on the metric properties of the set of correct EFs in the case $n \leq |D|$. It should be noted that similar properties of the set of minimal infrequent EFs were previously studied in a number of publications (for example, [3, 8, 9]), in which, among other things, the case $n \leq |D|$ is considered. The estimates of the number of minimal infrequent EFs given in the article have a form that allows us to compare them with the corresponding estimates of the correct EFs. The result of the comparison indicates the expediency (in terms of reducing the time costs) of using methods for searching for frequent EFs for the synthesis of logical classifiers and agrees with the experimental results obtained in [6] on random model data.

In Section 1 the problem statement is given. The initial data are presented as an integer matrix, whose rows are descriptions of the objects from D. Statements of the two main theorems on the number of correct EFs are given. The proofs of these theorems are contained in Section 2. The previously obtained and new estimates of the typical values for the number of minimal infrequent EFs and the length of the minimum infrequent EF are shown in Section 3.

1. STATEMENT OF THE PROBLEM AND FORMULATION OF THE MAIN RESULTS

We assume L, $L = (a_{ij})$, $i = \overline{1,m}$, $j = \overline{1,n}$, is a matrix with elements from $\{0,1,\ldots,k-1\}$, $k \geq 2$; $E_k^r, r \leq n$, $k \geq 2$, is the set of sets $(\sigma_1,\ldots,\sigma_r)$, $\sigma_i \in \{0,1,\ldots,k-1\}$, $i = \overline{1, r}$; W_r^n, $r \leq n$, is the set of all sets of the form $\{j_1,\ldots,j_r\}$, where $j_t \in \{1,2,\ldots,n\}$ at $t = \overline{1, r}$ and $j_1 < \ldots < j_r$; V_r^m, $r \leq m$, is the set of all ordered sets of the form $(i_1,\ldots,i_r)$, where $i_t \neq i_l$ at $t,l = \overline{1, r}$.

We put $\sigma \in E_k^r$, $\sigma = (\sigma_1,\ldots,\sigma_r)$, $w \in W_r^n$, $w = \{j_1,\ldots,j_r\}$. We will call number r the *length* of set w.

We will call the set w σ-*admissible* for L if we can specify a set $v = (i_1,\ldots,i_r), v \in V_r^m$ such that $a_{i_t j_t} = \sigma_t$ at $t = \overline{1, r}$. We will consider that the σ-admissible set w *is generated* by the set σ.

It is easy to see that in the case when the matrix L takes descriptions of objects from the sample D as its rows, the set $w \in W_r^n$, $w = \{j_1,\ldots,j_r\}$, is σ-admissible for L if and only if the EF (σ, H), $H = \{x_{j_1},\ldots,x_{j_r}\}$ is correct in D.

Let us introduce the following notation: $\mathfrak{M}_{mn}^k$ is the set of all matrices of size $m \times n$ with elements from $\{0,1,\ldots,k-1\}$, $k \geq 2$; $U(L,\sigma)$, $L \in \mathfrak{M}_{mn}^k$, $\sigma \in E_k^r$, is the set of all σ-admissible sets for matrices L; $U_r(L,\sigma)$ is the set of all sets in $U(L,\sigma)$ of length r; $U(L)$, $L \in \mathfrak{M}_{mn}^k$, is the aggregate of all admissible sets

for matrices L in which each set occurs as many times as the number of sets it generates from E_k^r; $|N|$ is the cardinality of the set N;

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

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

$r_1 = [0.5\log_k mn - 0.5\log_k \log_k^2 mn - \log_k \log_k \log_k n]$; hear and further, $[q]$ is the integer part of the number q; $r_2 =]0.5\log_k mn - 0.5\log_k \log_k^2 mn + \log_k \log_k \log_k n[$; hear and further, $]q[$ is the smallest integer greater than q; ϕ_1 is the interval $[r_1, r_2]$; $r_3 =]\log_k m + \log_k \log_k m[$; ϕ_2 is the interval $[1, r_3]$; $b_n \approx c_n$, $n \to \infty$ means that $\lim_{n\to\infty} b_n/c_n = 1$ and $b_n \preceq c_n$, $n \to \infty$ means that $\lim_{n\to\infty} b_n/c_n \le 1$.

Below we present the asymptotic estimates for the typical value of $|U(L)|$ and an estimate of the typical length admissible set for L for different values of m and n.

The identification of the typical situation is connected with a statement of the type "for almost all matrices L from $\mathfrak{M}_{mn}^k$ at $n \to \infty$ $F_1(L) \approx F_2(L)$ is satisfied" (here $F_1(L)$ and $F_2(L)$ are two functionals defined on matrices from $\mathfrak{M}_{mn}^k$). This statement means that there are two positive infinitely decreasing functions $\alpha(n)$ and $\beta(n)$ such that for all sufficiently large n

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

where $\mathfrak{M}$ is the set of such matrices L in $\mathfrak{M}_{mn}^k$ for which

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

is fulfilled.

Theorems 1 and 2 below are valid.

Theorem 1. If $m^a \le n \le k^{m^\beta}$, $a > 1$, $\beta < 1$, $k \ge 2$, *then at* $n \to \infty$ *for almost all matrices* L *from* $\mathfrak{M}_{mn}^k$,

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

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

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

are fulfilled and the lengths of almost all sets from $U(L)$ *belong to the interval* ϕ_1.

Theorem 2. *If* $n \le m \le k^{n^\beta}$, $\beta < 1/2$, $k \ge 2$, *then at* $n \to \infty$ *for almost all matrices* L *from* $\mathfrak{M}_{mn}^k$,

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

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

are valid and the lengths of almost all sets from $U(L)$ *belong to the interval* ϕ_2.

The proofs of Theorems 1 and 2 are based on a number of lemmas given in Section 2.

2. PROOFS OF THEOREMS 1 AND 2

Assume $v \in V_r^m, v = (i_1, \ldots, i_r)$; $\sigma \in E_k^r, \sigma = (\sigma_1, \ldots, \sigma_r)$; and $w \in W_r^n$, $w = \{j_1, \ldots, j_r\}$. Matrix $L = (a_{ij})$, $i = \overline{1,m}$, $j = \overline{1,n}$, $L \in \mathfrak{M}_{mn}^k$, is called (v, σ, w)-matrix if $a_{i_t j_t} = \sigma_t$ at $t = \overline{1,r}$. We denote by $N_{(v,\sigma,w)}$ the set of

(v,σ,w)-matrices in $\mathfrak{M}_{mn}^{k}$; and through $N^{*}_{(v,\sigma,w)}$, the set of all matrices L in $N_{(v,\sigma,w)}$ such that $L \notin N_{(v_1,\sigma,w)}$ at $v_1 \in V_r^m, v_1 \neq v$.

Lemma 1. *If* $v \in V_r^m$, $w \in W_r^n$, $\sigma \in E_k^r$, *then*

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

Proof. We estimate in how many ways it is possible to construct the matrix L from $N_{(v,\sigma,w)}$. Those elements of matrix L that are located at the intersection of rows with numbers from v and columns with numbers from w are uniquely determined. The remaining elements of this matrix can be chosen arbitrarily (k^{mn-r^2} ways). From this we obtain the required estimate. Lemma 1 is proved.

Lemma 2. *If* $v \in V_r^m$, $w \in W_r^n$, $\sigma \in E_k^r$, *then*

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

Proof. We estimate in how many ways it is possible to construct the matrix L from $N^{*}_{(v,\sigma,w)}$. The elements of this matrix, located in columns with numbers not included in w, can be chosen arbitrarily (in $k^{m(n-r)}$ ways). Hence, given that the rows in the submatrix of matrix L formed by columns with numbers from w, can be chosen by $(k^r-1)^{m-r}$ methods, we obtain the required estimate. Lemma 2 is proved.

Lemma 3. *We assume* $v_1 \in V_r^m$, $v_2 \in V_l^m$, $w_1 \in W_r^n$, $w_2 \in W_l^n$, $\sigma' \in E_k^r$, $\sigma'' \in E_k^l$; *sets* v_1 *and* v_2 *intersect along a* $(a \geq 0)$ *elements*; *and sets* w_1 *and* w_2 *intersect along b* $(b \geq 0)$ *elements. Then*

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

The proof of Lemma 3 is not given due to its obviousness.

Lemmas 4–6 below are proved using the expression $b_n \leq_n c_n$, which means that $b_n \leq c_n$ for all sufficiently large n.

Lemma 4. 1. *If* $m \leq n \leq k^{m^{\beta}}$, $\beta < 1$, *then*

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

2. *The following relation is valid:*

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

3. *If* $n \leq m$, *then*

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

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

1. We assume $m \leq n \leq k^{m^{\beta}}$, $\beta < 1$, and $r \leq r_1 + 1$. Then, using the fact that $q \leq 0.5\log_k mn$, $k^{2q} = mn/\log_k^2 mn$ and $(n-q) \geq_n 0.5n$ at $m \leq n$, $(m-q) \geq_n 0.5m$, at $n \leq 2^{m^{\beta}}$, we get

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

2. At $r \geq r_2 - 1$, we get

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

3. At $n \le m,\ r \ge r_3 - 1$, we get

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

Thus, $a_{r-1} = o(a_r),\ n \to \infty$, in case 1 and $a_{r+1} = o(a_r),\ n \to \infty$, in each of cases 2 and 3. Lemma 4 is proved.

Lemma 5. *If* $m \le n$ and $r,\ l \le r_2$, *then*

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

where $\delta(n) \to 0$ *at* $n \to \infty$.

Proof. We denote $\lambda_b = k^{lb} C_n^r C_r^b C_{n-r}^{l-b} / C_n^r C_{n-r}^l$. Since

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

and on the condition $r,l \le_n 0.5\log_k mn \le \log_k n, (r+l)/n \le_n 0.5$, then

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

Therefore, the estimated amount does not exceed $C_n^r C_{n-r}^l (1 + \delta(n))$, where $\delta(n) \to 0$ at $n \to \infty$. Hence, using the inequality $C_{n-r}^l \le C_n^l$, we obtain the assertion of the lemma. Lemma 5 is proved.

Lemma 6. *If* $m \le k^{n^\beta}$, $\beta < 1/2$, *and* $r,\ l \le r_3$, *then*

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

where $\delta(n) \to 0$ *at* $n \to \infty$.

The proof of Lemma 6 is similar to the proof of Lemma 5 (in this case $r,\ l \le 2n^\beta$ and $\lambda_b \le_n (8n^{2\beta-1})^b$).

We consider $\mathfrak{M}_{mn}^k = \{L\}$ to be the space of elementary events in which each event L happens with probability $1 / \left|\mathfrak{M}_{mn}^k\right|$. The mathematical expectation of a random variable $X(L)$ defined on the set $\mathfrak{M}_{mn}^k$ will be denoted by $\mathbf{M}X(L)$; and dispersion, through $\mathbf{D}X(L)$.

Lemma 7 [10]**.** *We assume that for random variables* $X_1(L)$ *and* $X_2(L)$ *defined on* $\mathfrak{M}_{mn}^k$, $X_1(L) \ge X_2(L) \ge 0$ *is fulfilled*; *and at* $n \to \infty$, $\mathbf{M}X_1(L) \approx \mathbf{M}X_2(L)$ *and* $\mathbf{D}X_2(L)/(\mathbf{M}X_2(L))^2 \to 0$ *are valid. Then for almost all matrices* L *from* $\mathfrak{M}_{mn}^k$, $X_1(L) \approx X_2(L) \approx \mathbf{M}X_2(L),\ n \to \infty$, *is valid.*

Assume $\sigma \in E_k^r,\ w \in W_r^n$. On $\mathfrak{M}_{mn}^k = \{L\}$ we consider a random variable $\zeta_{(\sigma,w)}(L)$, equal to 1 if w is the σ-admissible set for matrix L and equal to 0 otherwise. We put

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

It is easy to see that $\mu_r(L) = |U_r(L)|$ (number of sets in $U(L)$ of length r, $\zeta(L) = |U(L)|$, and $\zeta_i(L),\ i \in \{1,2\}$, is the number of those sets in $U(L)$ whose lengths belong to the interval ϕ_i.

We estimate the probability of an event $\zeta_{(\sigma,w)}(L) = 1, \sigma \in E_k^r, w \in W_r^n$, denoted below by $P(\zeta_{(\sigma,w)}(L) = 1)$. Obviously, by Lemma 1

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

However, by Lemma 2 we have

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

The following lemma immediately follows from (2.1) and Lemma 4.

Lemma 8. *If* $m \leq n \leq k^{m^\beta}$, $\beta < 1$, *then the following relations are valid*:

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

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

Lemma 9. *If* $m^a \leq n$, $a>1$, *then*

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

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

Proof. We have

$$\sum_{r\leq r_1}\mathbf{M}\mu_r(L) \geq \mathbf{M}\mu_{r_1}(L).$$

Since $mk^{-r_1}\to 0$, $n\to\infty$, then $(1-k^{-r_1})^{m-r_1}\to 1$, $n\to\infty$. From this, using (2.2), we obtain

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

Lemma 9 is proved.

Lemmas 8 and 9 immediately imply the following lemma.

Lemma 10. *If* $m^a \leq n \leq k^{m^\beta}$, $a>1$, $\beta<1$, *then*

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

The proofs of Lemmas 11–13 presented below are not given, since they are completely analogous to the proof of Lemma 10.

Lemma 11. *If* $m^a \leq n$, $a>1$, *then*

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

Lemma 12. *If* $n\leq m$, *then*

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

Lemma 13. *If* $m^a \leq n \leq k^{m^\beta}$, $a>1$, $\beta<1$, *then*

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

Lemma 14. *If* $m^a \leq n \leq k^{m^\beta}$, $a>1$, $\beta<1$, *then*

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

Proof. We have

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

It is easy to see that

$$\mathbf{M}(\xi_1(L))^2 \le \sum_{r,l\in\phi_1} \sum_{\substack{v_1\in V_r^m, v_2\in V_l^m \\ w_1\in W_r^n, w_2\in W_l^n}} \sum_{\substack{\sigma'\in E_k^r \\ \sigma''\in E_k^l}} |N|/k^{mn},$$

where $N = N_{(v_1,\sigma,w_1)} \cap N_{(v_2,\sigma'',w_2)}$. Hence, using Lemmas 3 and 5, we obtain

$$\mathbf{M}(\xi_1(L))^2 \le \sum_{r,l\in\phi_1} \sum_{b=0}^{\min(r,l)} k^{r+l}k^{-r^2-l^2+lb} C_n^r C_r^b C_{n-r}^{l-b} C_m^r C_m^l$$

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

where $\delta(n) \to 0$ at $n \to \infty$.

However, by Lemma 13

$$(\mathbf{M}\xi_1(L))^2 \approx \sum_{r,l\in\phi_1} C_n^r C_n^l C_m^r C_m^l k^{r+l}k^{-r^2-l^2}, \quad n\to\infty. \tag{2.5}$$

From (2.3)–(2.5) the assertion of the lemma being proved follows. Lemma 14 is proved.

Lemmas 15–17 below are proved similarly to Lemma 14.

Lemma 15. *If* $m^a \le n \le k^{m^\beta}$, $a > 1$, $\beta < 1$, *then*

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

Lemma 16. *If* $m^a \le n$, $a > 1$, *then*

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

Lemma 17. *If* $n \le m$, *then*

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

We assume $v \in V_r^m$, $\sigma \in E_k^r$, $w \in W_r^n$. On $\mathfrak{M}_{mn}^k = \{L\}$, we consider a random variable $\xi_{(v,\sigma,w)}(L)$, equal to 1 if $L \in N_{(v,\sigma,w)}$, and equal to 0 otherwise. We put

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

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

Lemma 18. *If* $n \le m \le k^{n^\beta}$, $\beta < 1/2$, *then at* $n\to\infty$, *the following relation is fulfilled for almost all matrices* L *from* $\mathfrak{M}_{mn}^k$:

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

Proof. We estimate the probability of an event $\xi_{(v,\sigma,w)}(L) = 1$, $v \in V_r^m$, $\sigma \in E_k^r$, $w \in W_r^n$, denoted below by $P(\xi_{(v,\sigma,w)}(L) = 1)$. By Lemma 1

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

Therefore, according to Lemma 4,

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

From (2.6) and Lemma 6, using the scheme of the proof of Lemma 14, we obtain

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

From (2.6), (2.7), and Lemma 7, the assertion of the lemma to be proved follows. Lemma 18 is proved.

The assertions of Theorem 1 follow directly from Lemmas 7, 10, 11, 13, and 14–16, while the assertions of Theorem 2 follow directly from Lemmas 7, 12, 17, 18, and the inequality $\zeta(L) \le \xi(L)$.

3. ESTIMATES OF THE TYPICAL VALUES OF THE NUMBER OF MINIMAL INFREQUENT EFS AND THE LENGTH OF THE MINIMUM INFREQUENT EFS

We put $L \in \mathfrak{M}_{mn}^k, L = (a_{ij}), i = 1,\ldots,m, j = 1,\ldots,n; \sigma \in E_k^r, \sigma = (\sigma_1,\ldots,\sigma_r); w \in W_r^n, w = \{j_1,\ldots,j_r\}$.

The set w is called a σ-*covering for matric L of length r* if for any $i \in \{1,2,\ldots,m\}$ there are $j \in \{j_1,\ldots,j_r\}$ such that $a_{ij} \ne \sigma_j$. We will consider that the σ-covering w *is generated* by the set σ.

The set w, which is a σ-covering for matric L is called an irredundant if for any $t \in \{1,2,\ldots, r\}$ the set $w\backslash\{j_t\}$ is not a γ_t-covering for matrix L, where $\gamma_t = (\sigma_1,\ldots,\sigma_{t-1},\sigma_{t+1},\ldots,\sigma_r)$. If w is an irredundant σ-covering for matric L, it is easy to see that the columns of matrix L with numbers from w contain a submatrix that, up to row permutation, has the form

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

where $\beta_p \ne \sigma_p$ at $p = 1,2,\ldots,r$. Such a submatrix is called a σ-*submatrix*.

Note that in the case when the descriptions of objects from the sample D are taken as the rows of matrix L, then the set $w \in W_r^n$, $w = \{j_1,\ldots,j_r\}$, is an irredundant σ-covering for matric L if and only if the EF (σ, H), $H = \{x_{j_1},\ldots,x_{j_r}\}$, is minimal infrequent in D.

We introduce the following notation: $B(L,\sigma), L \in \mathfrak{M}_{mn}^k, \sigma \in E_k^r$, is the set of all irredundant of the σ-covering for matrix L; $S(L,\sigma)$, $L \in \mathfrak{M}_{mn}^k$, $\sigma \in E_k^r$, is the set of all σ-matrix submatrices L; $B_r(L, \sigma)$, $L \in \mathfrak{M}_{mn}^k, \sigma \in E_k^r$, is the set of all sets in $B(L, \sigma)$ of length r; $S_r(L,\sigma)$, $L \in \mathfrak{M}_{mn}^k$, $\sigma \in E_k^r$, is the set of all submatrices in $S(L,\sigma)$ of order r; $B(L)$, $L \in \mathfrak{M}_{mn}^k$, is the set of all irredundant σ-covering for matric L, in which each covering occurs as many times as the number of sets of E_k^r it generates; $S(L)$, $L \in \mathfrak{M}_{mn}^k$, is the set of all σ-submatrices of matrix L for all σ from 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 – interval $[1, r_3]$; $r_4 = [0.5\log_k mn - 0.5\log_k\log_k mn - \log_k\log_k\log_k n]$; $r_5 =]0.5\log_k mn - 0.5\log_k\log_k mn + \log_k\log_k\log_k n[$; ϕ_3 – interval $[r_4, r_5]$; $r_6 =]\log_k m + \log_k\log_k m + \log_k\log_k\log_k n[$; ϕ_4 is the interval $[1, r_6]$.

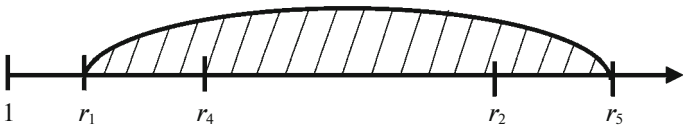

Fig. 1. Typical values for the lengths of sets from $U(L)$ (see Section 1) and $B(L)$ when $m^a \le n \le k^{m^\beta}$, $a > 1$, $\beta < 1$.

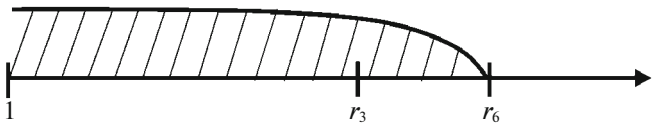

Fig. 2. Typical values for the lengths of sets from $U(L)$ (see Section 1) and $B(L)$ when $n \le m \le k^{n^\beta}$, $\beta < 1/2$.

Theorem 3 [3]. *If $m^a \le n \le k^m$, $a > 1$, $k \ge 2$, then the following relations are valid at $n \to \infty$ for almost all L matrices from $\mathfrak{M}_{mn}^k$:*

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

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

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

and the lengths of almost all sets from $B(L)$ belong to the interval ϕ_3.

Theorem 4. *If $n \le m \le k^{n^\beta}$, $\beta < 1/2$, $k \ge 2$, then the following relations are valid at $n \to \infty$ for almost all matrices L from $\mathfrak{M}_{mn}^k$:*

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

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

and the lengths of almost all sets from $B(L)$ belong to the interval ϕ_4.

The scheme of the proof of Theorem 4 is similar to that of the proof of Theorem 2.

Thus, in each of the two cases considered, the typical length of a set of $U(L)$ and the typical length of a set of $B(L)$ belong to the same interval. The results of Theorems 1, 3 and Theorems 2, 4 are illustrated, respectively, in Figs. 1 and 2.

CONCLUSIONS

Topical issues of logical analysis of integer data concerning the research on the metric (quantitative) properties of sets of frequent and infrequent elements of such data are considered. The technique for obtaining estimates for the typical values of the main numerical characteristics of the specified sets has been improved and new estimates for such characteristics have been found. A theoretical substantiation of the expediency (in terms of reducing time costs) of using methods for searching for frequent elements at the stage of training classifiers based on a logical analysis of the training sample is given.

The results of the study carried out in this paper are also important for a number of other applied areas, among which it is worth highlighting the searching for associative rules in data. In this case D is called a database, and each object of the database D is a transaction. The associative rule establishes a relationship between two frequent EFs, according to which one frequent EF (premise) with some "certainty" entails another frequent EF (consequence). In this case, the premise and the consequence are generated by one

common frequent EF. Questions of the synthesis of associative rules arose in connection with the analysis of the consumer basket [11].

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

1. L. V. Baskakova and Yu. I. Zhuravlev, "Model of recognition algorithms with representative sets and support set systems," Zh. Vychisl. Mat. Mat. Fiz. **21** (5), 1264–1275 (1981).
2. P. L. Hammer, "Partially defined Boolean functions and cause-effect relationships," in *Lectures at the Int. Conf. on Multi-Attribute Decision Making via O.R.-based Expert Systems* (University of Passau, Passau, Germany, 1986).
3. E. V. Dyukova and Yu. I. Zhuravlev, "Discrete analysis of feature descriptions in recognition problems of high dimensionality," Comput. Math. Math. Phys. **40** (8), 1214–1227 (2000).
4. E. V. Dyukova and N. V. Peskov, "Search for informative fragments in descriptions of objects in discrete recognition procedures," Comput. Math. Math. Phys. **42** (5), 711–723 (2002).
5. Yu. I. Zhuravlev, V. V. Ryazanov, and O. V. Sen'ko, *Recognition. Mathematical Methods. Software System. Practical Applications* (FAZIS, Moscow, 2006) [in Russian].
6. N. Dragunov, E. Djukova, and A. Djukova, "Supervised classification and finding frequent elements in data," in *VIII Int. Conf. on Information Technology and Nanotechnology (ITNT)* (IEEE, Samara, 2022).
7. E. V. Djukova and A. P. Djukova, "On the complexity of learning logical classification procedures," "Informatics and Applications" **16** (4), 57–62 (2022).
8. A. E. Andreev, "On the asymptotic behavior of the number of dead-end tests and the length of the minimum test for almost all tables," Probl. Kibern., No. 41, pp. 117–142 (1984) [in Russian].
9. E. V. Djukova and R. M. Sotnezov, "Asymptotic estimates for the number of solutions of the dualization problem and its generalizations," Comput. Math. Math. Phys. **51** (8), 1431–1440 (2011).
10. V. N. Noskov and V. A. Slepyan, "On the number of dead-end tests for a class of tables," Kibernetika, No. 1, 60–65 (1972) [in Russian].
11. C. Aggarwal Charu, *Frequent Pattern Mining* (Springer, New York, 2014). https://www.charuaggarwal.net/freqbook.pdf.