Computational Mathematics and Mathematical Physics, Vol. 43, No. 12, 2003, pp. 1812–1822.
Translated from Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, Vol. 43, No. 12, 2003, pp. 1884–1895.
Original Russian Text Copyright © 2003 by Djukova, Inyakin.
English Translation Copyright © 2003 by MAIK "Nauka/Interperiodica" (Russia).

Classification Procedures Based on the Construction of Class Coverings

E. V. Djukova and A. S. Inyakin

Computing Center, Russian Academy of Sciences, ul. Vavilova 40, Moscow, 119991 Russia

e-mail: djukova@ccas.ru, andre_w@mail.ru

Received February 28, 2003

Abstract—An approach to the clustering problem with integer data is described, in which special sets of feature values not contained in the feature descriptions of objects are constructed. The problem is reduced to the construction of irredundant coverings of integer matrices, which can be done by constructing irreducible coverings of Boolean matrices. Based on a geometric interpretation of the concept of covering, a new method is proposed for constructing irreducible and minimal coverings of a Boolean matrix. The metric properties of close-to-minimal coverings of integer matrices are examined.

INTRODUCTION

A frequent task arising in a number of applied classification problems is to partition a set of objects into "homogeneous" groups (classes) in the absence of learning data. Such problems are referred to as taxonomy problems.

Let M be a finite sample of elements, each of which is described by a finite set of features (properties). The task is to partition the sample into homogeneous groups, whose number may be prescribed or unknown. The process of partitioning a set into classes is known as clustering or cluster analysis.

In the statement of the problem, we need to describe the rules observed by objects of a single class and by objects of different classes. Such a rule can be defined as the compactness of objects in a given feature space, i.e., a rule according to which, for example, the "distance" between objects assigned to the same class cannot be greater than a prescribed value. One can also use "good" relative remoteness of classes; for example, the distance between objects assigned to different classes cannot be less than a prescribed value.

In this paper, we consider some approaches to taxonomy problems with integer data. In contrast to classical methods, our methods do not require selecting a distance function, which is sometimes rather difficult to evaluate. Our methods are based on the following arguments.

Suppose that we want to determine the degree of membership of an object S in a group M of objects. If the description of S includes a set of feature values not contained in the description of any object in M, we can say that the union of S and M violates the internal structure of M. Examining various combinations of feature values not contained in the descriptions of objects in M, we can quantify the proximity of S to M. Thus, sets of admissible feature values not contained in the descriptions of all objects in M are regarded as informative while determining the degree of proximity of an object to M.

Based on the arguments above, an algorithm has been designed that considerably improves clustering for some model and actual problems as compared to available algorithms, such as those based on the nearest neighbor, the farthest neighbor, and central-element choice, in which the distance function is defined to be the Hamming distance.

In the clustering algorithm we propose, the evaluation of the proximity of S to M is reduced to a search for the so-called irredundant σ-coverings of the integer matrix made up of the feature descriptions of objects in M. This concept is a generalization of an irreducible covering of a Boolean matrix, a well-known concept in discrete mathematics, and was first introduced in [1, 2] in connection with constructing irredundant representative descriptors in Kora-type algorithms [3, 4] and with examining the metric properties of a set of representative descriptors. It was shown in [5] that the construction of (irredundant) coverings of an integer matrix can be reduced to a search for (irreducible) coverings of a Boolean matrix. The approach to recognition problems that involves the construction of coverings of Boolean and integer matrices was described in the most comprehensive form in [6, 7].

The sets of feature values defined by short σ-coverings are believed to be more informative. That is why not all σ-coverings are usually constructed but rather those whose length does not exceed a prescribed parameter. In this connection, we examine the metric (quantitative) properties of σ-coverings that are close in length to minimal σ-coverings. An asymptotic value of the number of such coverings is obtained in a typical case.

We propose a geometric interpretation of σ-covering and irredundant σ-covering of an m-by-n matrix L with entries in $\{0, 1, \ldots, k-1\}$, $k \geq 2$. Let M_L be the set of such k-ary n-tuples (points of the k-ary n-cube) that are the rows in L. The construction of (irredundant) σ-coverings of L is shown to reduce to the construction of (maximal) subcubes in the set $\overline{M}_L$ consisting of all k-ary n-tuples not contained in L. In the case $k = 2$, we propose a search algorithm for maximal subcubes in $\overline{M}_L$ corresponding to irreducible coverings of a Boolean matrix L and a modification of that algorithm in the case of searching for maximal subcubes in $\overline{M}_L$ corresponding to minimal (in length) coverings of L.

1. BASIC DEFINITIONS AND GEOMETRIC INTERPRETATION OF COVERING AND IRREDUNDANT COVERING OF AN INTEGER MATRIX

We introduce the following notation: E_k^r $(k \geq 2)$ is the set of all k-ary r-tuples with $r \leq n$, $E^r = E_2^r$, and σ is a tuple of the form $(\sigma_1, \ldots, \sigma_r)$ in E_k^r.

A set H of r different columns in a matrix L is called a σ-covering of L if the submatrix L^H of L consisting of the columns of H does not contain the row $(\sigma_1, \ldots, \sigma_r)$.

A set of columns H that is a σ-covering of L is called an irredundant σ-covering of L if for any $p \in \{1, 2, \ldots, r\}$, L^H contains at least one row of the form $(\beta_1, \ldots, \beta_r)$, where $\beta_p \neq \sigma_p$ and $\beta_i = \sigma_i$ for $i = \overline{1, r}$ and $i \neq p$; i.e., L^H contains a submatrix of the from

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

Such a submatrix is called a σ-submatrix.

In the case $k = 2$ and $\sigma = (0, \ldots, 0)$, the concept of an irredundant σ-covering coincides with the well-known concept of an irreducible covering of a Boolean matrix (see [8]), with the σ-submatrix being an identity matrix.

Let $C(L, \sigma)$ be the set of σ-coverings of L, $B(L, \sigma)$ be the set of irredundant σ-coverings of L, and $S(L, \sigma)$ be the set of σ-submatrices of L. Define

$$C(L) = \bigcup_{r=1}^{n} \bigcup_{\sigma \in E_k^r} C(L, \sigma), \quad B(L) = \bigcup_{r=1}^{n} \bigcup_{\sigma \in E_k^r} B(L, \sigma), \quad S(L) = \bigcup_{r=1}^{n} \bigcup_{\sigma \in E_k^r} S(L, \sigma).$$

We show that the construction of $C(L)$ and $B(L)$ can be reduced to the construction of $(0, \ldots, 0)$-coverings and irredundant $(0, \ldots, 0)$-coverings (irreducible coverings), respectively, of a specific Boolean matrix constructed from L.

Suppose that L has the form

$$\begin{bmatrix} x_{11} & \ldots & x_{1n} \\ \ldots & \ldots & \ldots \\ x_{1m} & \ldots & x_{mn} \end{bmatrix},$$

where each entry in the column with index j can take k_j values with $k_j \leq k$ (for further analysis, it is convenient to assume that the entries of the column with index j range from 1 to k_j).

Define

$$\delta(x_{ij}, a) = \begin{cases} 1, & x_{ij} \neq a, \\ 0, & x_{ij} = a, \end{cases}$$

where $i = \overline{1, m}$ and $j = \overline{1, n}$.

We construct a matrix L^B consisting of m rows such that the row with index i, $i \in \{1, 2, \ldots, m\}$ has the form

$$(\delta(x_{i1}, 1), \ldots, \delta(x_{i1}, k_1), \delta(x_{i2}, 1), \ldots, \delta(x_{i2}, k_2), \ldots, \delta(x_{in}, 1), \ldots, \delta(x_{in}, k_n)).$$

It is easy to see that the column with index j ($j \in \{1, 2, \ldots, n\}$) in the original matrix corresponds to a group of k_j columns in L^B, hereafter denoted by g_j. The columns in g_j will be indexed by numbers ranging from 1 to k_j.

Denote by (j, i) the column of L^B with index i in g_j, where $j \in \{1, 2, \ldots, n\}$ and $i \in \{1, 2, \ldots, k_j\}$.

The following result is straightforward.

Proposition 1. *A set of columns of the form* $((j_1, i_1), \ldots, (j_r, i_r))$ *in* L^B (*where* $r \in 1, 2, \ldots, n$, $j_l \in \{1, 2, \ldots, n\}$, $i_l \in \{1, 2, \ldots, k_{j_l}\}$, $j_{l_1} \neq j_{l_2}$ if $l_1 \neq l_2$, $l_1, l_2 = 1, 2, \ldots, r$) *is an* (*irreducible*) *covering if and only if the set of columns indexed by* $(j_1, \ldots, j_r)$ *is an* (*irredundant*) $(j_1, \ldots, j_r)$*-covering of* L.

The set E_k^n can be regarded as a k-ary n-cube. Denote by M_L the set of points in E_k^n that correspond to the rows of L. Let $\overline{M}_L = E_k^n \backslash M_L$. The set H of columns indexed by $(j_1, \ldots, j_r)$, which is a σ-covering of L, is assigned the set of points $E^{(\sigma, H)} = \{(\alpha_1, \ldots, \alpha_n) | \alpha \in E_k^n, \alpha_{j_i} = \sigma_i, i = \overline{1, r}\}$. Obviously, $E^{(\sigma, H)}$ is an $(n-r)$-subcube of E_k^n and entirely belongs to $\overline{M}_L$. When $r = n$, $E^{(\sigma, H)}$ is a 0-subcube consisting of a single point σ. Thus, $C(L)$ is in one-to-one correspondence with the set of all subcubes in $\overline{M}_L$. Note that $E^{(\sigma_1, H_1)}$ is a subcube of $E^{(\sigma_2, H_2)}$ if and only if $\sigma_2 \subset \sigma_1$ and $H_2 \subset H_1$.

Let M be a subset of E_k^n. A subcube $E' \subset M$ is said to be maximal in M if there does not exist any other subcube $E'' \subset M$ such that $E' \subset E''$. The set of all maximal subcubes in M is denoted by $E^{\max}(M)$. It is easy to see that $B(L)$ is in one-to-one correspondence with $E^{\max}(\overline{M}_L)$. Consequently, the construction of $B(L)$ can be reduced to the construction of the set of all maximal subcubes in $\overline{M}_L$.

2. SEARCH ALGORITHMS FOR IRREDUCIBLE AND MINIMAL COVERINGS OF A BOOLEAN MATRIX

In [9] an asymptotically optimal algorithm for search for all irreducible coverings of a Boolean matrix was designed in the case where the number of rows in the matrix is much less than the number of its columns. The algorithms described in this section are intended primarily for search for irreducible and minimal (in length) coverings of a Boolean matrix when the number of rows in the matrix is much greater than the number of its columns (in which case the algorithms are sufficiently fast).

Let $k = 2$. Obviously, if the point $\tilde{0} = (0, \ldots, 0)$ is in M_L, then there is no irreducible covering in L, so we assume in what follows that $\tilde{0} \in \overline{M}_L$.

It is easy to see that the $(0, \ldots, 0)$-covering of a Boolean matrix L corresponds to a subcube in $\overline{M}_L$, with $\tilde{0}$ being one of its vertices. An irreducible covering of L corresponds to a maximal subcube in $\overline{M}_L$, with $\tilde{0}$ being one of its vertices. Thus, the set of all irreducible coverings of L corresponds to the stencil of all maximal subcubes in $\overline{M}_L$ originating in $\tilde{0}$.

We say that $\tilde{\alpha} = (\alpha_1, \ldots, \alpha_n)$ ($\tilde{\alpha} \in E^n$) encloses than $\tilde{\beta} = (\beta_1, \ldots, \beta_n)$ ($\tilde{\beta} \in E^n$) (and write this as $\tilde{\alpha} > \tilde{\beta}$) if $\alpha_i \geq \beta_i$, $i = \overline{1, n}$. Denote by $M(\tilde{\alpha}, \tilde{\beta})$ the subcube of minimal dimension in E^n that contains $\tilde{\alpha}$ and $\tilde{\beta}$. The subcube $M(\tilde{\alpha}, \tilde{\beta})$ is said to be covering for M_L if $M(\tilde{\alpha}, \tilde{\beta}) \subseteq \overline{M}_L$.

The following result is straightforward.

Proposition 2. *If* $\tilde{\alpha} \in \overline{M}_L$, *then* $M(\tilde{\alpha}, \tilde{0})$ *is not contained in* $\overline{M}_L$ *if and only if there exists* $\tilde{\beta}$ *in* M_L *such that* $\tilde{\alpha} > \tilde{\beta}$.

Algorithm for Constructing All Irreducible Coverings of L

Step 1. Delete from $\overline{M}_L$ all points $\tilde{\alpha}$ such that there exists $\tilde{\beta}$ in M_L such that $\tilde{\alpha} > \tilde{\beta}$.

Step 2. Let $\tilde{\alpha}' \in \overline{M}_L$ and $\rho(\tilde{\alpha}', \tilde{0}) = \max\limits_{\tilde{\alpha} \in \overline{M}_L} \rho(\tilde{\alpha}, \tilde{0})$ (here and below, $\rho(\tilde{\alpha}, \tilde{\beta})$ is the Hamming distance between $\tilde{\alpha}$ and $\tilde{\beta}$). Obviously, $M(\tilde{\alpha}', \tilde{0})$ is a maximal covering subcube for M_L. Let the coordinates of $\tilde{\alpha}'$ indexed by $j_1, \ldots, j_r$ be zero. Then the set of columns indexed by $j_1, \ldots, j_r$ is the irreducible covering of L that corresponds to $M(\tilde{\alpha}', \tilde{0})$. Set $\overline{M}_L = \overline{M}_L \backslash M(\tilde{\alpha}', \tilde{0})$.

Step 3. If $\overline{M}_L \neq \varnothing$, then return to Step 2; otherwise, stop.

Algorithm for Constructing All Minimal Coverings of L

Step 1. Delete from $\overline{M}_L$ all points $\tilde{\alpha}$ such that there exists $\tilde{\beta}$ in M_L such that $\tilde{\alpha} > \tilde{\beta}$.

Step 2. Let $\tilde{\alpha}' \in \overline{M}_L$ and $\rho(\tilde{\alpha}', \tilde{0}) = \max\limits_{\tilde{\alpha} \in \overline{M}_L} \rho(\tilde{\alpha}, \tilde{0})$. Set

$$Q_L = \{\tilde{\alpha} \mid \tilde{\alpha} \in \overline{M}_L, \rho(\tilde{\alpha}, \tilde{0}) = \rho(\tilde{\alpha}', \tilde{0})\}.$$

Step 3. Let $\tilde{\alpha}' \in Q_L$, and let the coordinates of α' indexed by $j_1, \ldots, j_r$ be zero. Then the set of columns indexed by $j_1, \ldots, j_r$ is the minimal covering of L that corresponds to $M(\tilde{\alpha}', \tilde{0})$. Set $Q_L = Q_L \backslash \tilde{\alpha}'$.

Step 4. If $Q_L \neq \varnothing$, then return to Step 3; otherwise, stop.

Note that the algorithm for search for irreducible coverings could be so modified that it would search for irredundant σ-coverings.

3. CLUSTERING ALGORITHM BASED ON THE CONSTRUCTION OF IRREDUNDANT σ-COVERINGS

The measure of the similarity of $M'' \subset E^n$ and $M' \subset E^n$ is defined as

$$P(M', M'') = \frac{|E^{\max}(E^n \backslash M') \cap E^{\max}(E^n \backslash (M' \cup M''))|}{|E^{\max}(E^n \backslash M')|}.$$

Here, $|A|$ is the cardinality of A.

Let p be a given number such that $p \in (0, 1)$. A set M'' is said to be similar to M' in terms of a threshold p if $P(M', M'') \geq p$.

Suppose that the clustering problem is to be solved for a set of objects $M = \{S_1, \ldots, S_m\}$.

The first class M_1 is constructed as follows.

Step 1. Set $M_1 = S$, where S is an arbitrary element in M. If $M \backslash M_1 = \varnothing$, then stop; otherwise, go to Step 2.

Step 2. If there exists an element S' in $M \backslash M_1$ such that

$$P(M_1, \{S'\}) = \max_{S \in \Omega \backslash M_1} P(M_1, \{S\}) \quad \text{and} \quad P(M_1, \{S'\}) \geq p,$$

set $M_1 = M_1 \cup \{S'\}$ and repeat Step 2. Otherwise, begin to construct the next class.

Suppose that the classes $M_1, \dots, M_{i-1}$ $(i \geq 2)$ have been constructed. Let

$$\tilde{M}_i = M \setminus \bigcup_{j=1}^{i-1} M_j.$$

If $\tilde{M}_i = \emptyset$, then stop; otherwise, construct M_i. The procedure for constructing M_i is entirely analogous to that for constructing M_1, with the initial set being $\tilde{M}_i$ rather than M.

In this algorithm, the similarity of objects or groups of objects is defined by using the concept of a maximal subcube of a set of points, in which case the distance function between objects is not required.

Experience gained from solving practical problems has shown that p should be chosen in the range (0.5, 0.9). The greater the value of p, the more the number of classes formed in clustering.

A description of the algorithm without using geometric language can be found in [5].

4. TESTING BASED ON MODEL AND ACTUAL PROBLEMS

Model and actual problems were used to test the clustering algorithms based on the farthest neighbor (algorithm A_1), nearest neighbor (algorithm A_2), central-element choice (algorithm A_3), and the construction of irredundant σ-coverings (algorithm A_4).

Algorithms A_1–A_3 are based on hierarchical clustering methods, which can be described as follows. Consider a sequence of partitions of m objects into groups. The first step in the sequence is a partition into m groups, each of which contains only one object. The next step is a partition into $m - 1$ groups; then, into $m - 2$ groups, and so on up to the mth step, at which all objects constitute a single group. Thus, the kth step corresponds to a partition into $m + 1 - k$ groups, where $k = 1, 2, \dots, m$. A sequence of partitions is called hierarchical grouping if any two objects, S_1 and S_2, assigned to the same group at the kth step stay in a single group at all subsequent steps.

Suppose that $S_1, \dots, S_m$ are m objects to be partitioned into l classes. The basic steps in hierarchical grouping are described by the following procedure.

1. Let $\tilde{l} = m$ and $\Omega_i = \{S_i\}$, $i = \overline{1, m}$.

2. Stop if $\tilde{l} \leq l$.

3. Find two groups Ω_i and Ω_j such that

$$\rho(\Omega_i, \Omega_j) = \min_{u \neq v} \rho(\Omega_u, \Omega_v),$$

where $\rho(\Omega_u, \Omega_v)$ is the distance between Ω_u and Ω_v (see below).

4. Unite Ω_i and Ω_j, delete Ω_i, and decrement $\tilde{l}$ by 1. Return to Step 2.

The procedure halts when a prescribed number of groups is achieved, i.e., when $\tilde{l} = l$.

When the number of groups to partition a set is not specified, various stopping criteria can be applied, in particular, the condition that the smallest distance between the groups at the ith step is more than p times greater than the smallest distance at the preceding step ($p > 1$ is specified in advance; usually, $p > 1.5$).

Note that the results of the procedure depend strongly on the distance function between groups used. In our tests, we used the following functions.

1. The nearest neighbor distance

$$\rho(\Omega_u, \Omega_v) = \min_{S_i \in \Omega_u, S_j \in \Omega_v} \rho(S_i, S_j), \quad (u, v) = \overline{1, l}.$$

2. The farthest neighbor distance

$$\rho(\Omega_u, \Omega_v) = \max_{S_i \in \Omega_u, S_j \in \Omega_v} \rho(S_i, S_j), \quad (u, v) = \overline{1, l}.$$

3. The distance between the central elements of classes:

$$\rho(\Omega_u, \Omega_v) = \rho(\bar{S}_u, \bar{S}_v),$$

where $\bar{S}_u$ is the central element of Ω_u and $\bar{S}_v$ is the central element of Ω_v. In particular, the central element of a class can be defined to be the center of gravity of that class.

In the formulas above, the distance $\rho(S_i, S_j)$ between $S_i = (x_{i1}, \ldots, x_{in})$ and $S_j = (x_{j1}, \ldots, x_{jn})$ was calculated as

$$\rho(S_i, S_j) = \sum_{k=1}^{n} |x_{ik} - x_{jk}|.$$

Algorithm A_4 was compared with A_1–A_3 by using the model sets of data displayed in Fig. 1.

The results obtained with A_1 are shown in Fig. 2. Note that A_1 gives a right (i.e., intuitively most natural) partition of the set displayed in Fig. 1a for an arbitrary order of objects appearing in the original sample (in general, the results produced by the algorithms depend on this order).

Figure 3 shows the results obtained with A_2. This algorithm considerably improves the partition of the set displayed in Fig. 1b but degrades the partition of the sets displayed in Figs. 1a and 1c.

Figure 4 shows the results obtained with A_3. The algorithm based on central-element choice exhibits an average of the nearest neighbor and farthest neighbor methods.

Figure 5 shows the results obtained with A_4. The algorithm based on the construction of irredundant σ-coverings produces the best results as compared to the previous algorithms.

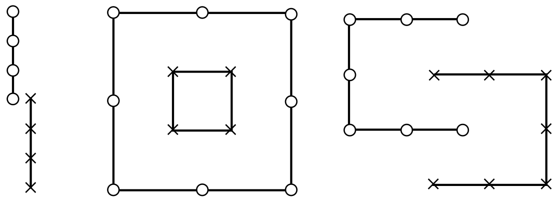
Clustering algorithms A_3 and A_4 were also tested against poll results. The questionnaire consisted of 21 questions, with 2 to 15 variants of answers to each. Some 799 persons were asked to respond to the questionnaire. The set of respondents was partitioned into classes according to their answers to the key question of the questionnaire: “What is your attitude toward party A?” The alternative answers were favor, disinterested, disfavor, and undecided. Thus, we obtained four classes. For every experiment, a small group of respondents (4 to 20 persons) was selected at random in each class. No data on the membership in some class were available. Algorithm A_4 (based on the construction of irredundant σ-coverings) and algorithm A_3 (based on the calculation of the Hamming distance with central-element choice) were used to partition the set into clusters. We conducted 32 experiments. The partitions produced by the algorithms were compared with the original partition into classes. The performance of an algorithm was characterized by the ratio of the number of “correctly” classified pairs of objects to the total number of pairs. A classification was treated as correct if one of the following conditions was fulfilled: (a) a pair of objects belonging to the same class in the original partition remained in a single class after clustering; and (b) a pair objects belonging to different classes in the original partition remained in different classes after clustering.

The accuracy of clustering was found to range between 60 and 80% for A_3 and between 70 and 90% for A_4.

The results of comparison based on model and actual problems suggest that the algorithms with Hamming distance calculation are characterized by a high speed of computation. These methods are used primarily for solving problems with binary data. They can be applied to problems with features of higher arity, but in that case they poorly perform on specific initial datasets, for example, on those shown in Figs. 1–3. Algorithm A_4 is more laborious, but it performs well on nonbinary data and yields better results in that case.

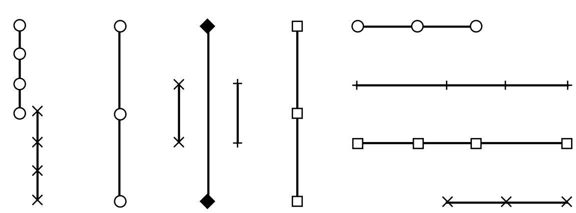
5. METRIC PROPERTIES OF CLOSE-TO-MINIMAL COVERINGS OF INTEGER MATRICES

Traditionally, issues related to higher performance and speedup of recognition algorithms based on the construction of coverings of Boolean and integer matrices are associated with asymptotic estimates obtained for typical values of important quantitative characteristics of this set. Such characteristics are, for example,



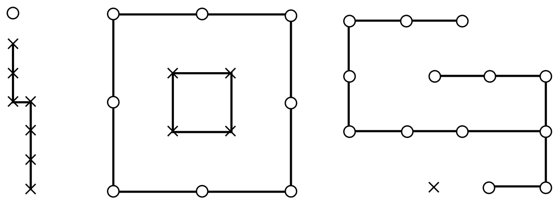
(a) (b) (c)

Fig. 1.



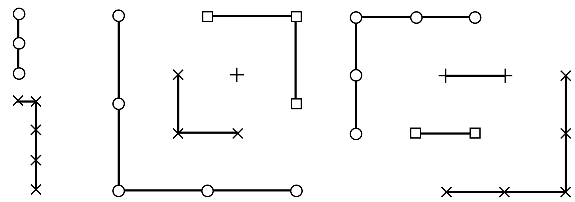
(a) (b) (c)

Fig. 2.



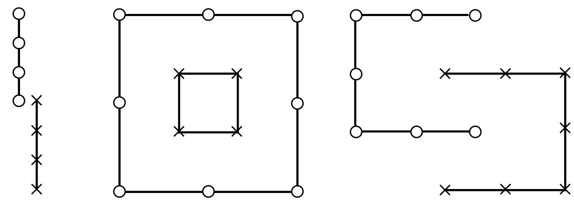
(a) (b) (c)

Fig. 3.



(a) (b) (c)

Fig. 4.



(a) (b) (c)

Fig. 5.

the number of coverings and the length of a covering. The metric properties of irredundant coverings are of special interest but present the greatest difficulties in their study.

In [1, 2, 6, 7] asymptotic values of the number of coverings in $B(L)$ and the length of a covering in $B(L)$ were obtained for almost all m-by-n matrices with entries in $\{0, 1, \ldots, k-1\}$, $k \ge 2$, as $n \longrightarrow \infty$ with $m^{\alpha} \le n \le k^{m^{\beta}}$, $\alpha > 1$, and $\beta < 1$. For almost all such matrices, the number of coverings in $B(L)$ was shown to be asymptotically equal to the number of all submatrices in $S(L)$ and smaller in order than the number of coverings in $C(L)$.

An opposite case (namely, $n^{\alpha} \le m \le k^{n^{\beta}}$ with $\alpha > 1$ and $\beta < 1$) was considered in [11]. Asymptotic estimates were obtained for the typical value of $S(L)$ and the typical order of a submatrix in $S(L)$. In the case $n^{\alpha} \le m \le k^{n^{\beta}}$ with $\alpha > 1$ and $\beta < 1/2$, the number of submatrices in $S(L)$ was shown to be almost always greater in order than the number of coverings in $B(L)$. In a practically important case, asymptotic estimates were obtained for the typical number of coverings in $C(L)$ and for the typical length of a covering in $C(L)$. Moreover, for $r \le \log_k m - \log_k(\log_k m \times \ln n)$ and $n \longrightarrow \infty$, coverings of length r in $C(L)$ were shown to be absent in almost all m-by-n matrices L.

An open question is under which conditions the number of irredundant coverings is asymptotically equal to the number of coverings. An answer is given in Theorem 1 below.

Suppose that $r_1 = [\log_k m - \log_k \ln \log_k m - 1]$ (here, $[x]$ is the integer part of x), φ denotes the interval $(\log_k m - \log_k(\log_k m \times \ln n), \log_k m - \log_k \ln \log_k m - 1]$, $a_n \approx b_n$ means that $\lim\limits_{n \to \infty} a_n/b_n = 1$, $a_n \underset{n}{\ge} b_n$ means that $a_n \ge b_n$ for all sufficiently large n, M_{mn}^k $(k \ge 2)$ is the set of all m-by-n matrices with entries in $\{0, 1, \ldots, k-1\}$, $C_{\varphi}(L)$ is the set of coverings in $C(L)$ with lengths ranging over φ, and $B(L)$ is the set of coverings in $B(L)$ with lengths ranging over φ, $C_{r_1}(L)$ is the set of coverings of length r_1 in $C(L)$, and $B_{r_1}(L)$ is the set of coverings of length r_1 in $B(L)$.

Theorem 1. *If* $m \le k^{n^{\beta}}$ *and* $\beta < 1/2$, *then*

$$|B_{\varphi}(L)| \approx |C_{\varphi}(L)| \approx |C_{r_1}(L)| \approx |B_{r_1}(L)| \approx k^{r_1} C_n^{r_1} (1 - k^{-r_1})^m$$

as $n \longrightarrow \infty$ *for almost all matrices* L *in* M_{mn}^k.

The proof of the theorem is based on Lemmas 1–6 below.

Denote by W_r^n $(r \le n)$ the set of all ordered sets of the form $(j_1, \ldots, j_r)$, where $j_l \in \{1, 2, \ldots, n\}$ for $l = \overline{1, r}$ and $j_1 < \ldots < j_r$.

Let $w \in W_r^n$ and $\omega \in E_k^r$. Denote by $M_{(w,\sigma)}$ the set of matrices L in M_{mn}^k such that the submatrix of L made up of the columns of w does not contain σ.

Suppose that $M_{mn}^k = \{L\}$ is the space of elementary events in which every event L has the probability $\left|M_{mn}^k\right|^{-1}$. The expectation and variance of a random variable $X(L)$ are denoted by $\mathbf{M}X(L)$ and $\mathbf{D}X(L)$, respectively.

Lemma 1 (see [12]). *Let random variables $X_1(L)$ and $X_2(L)$ defined on M_{mn}^k be such that $X_1(L) \ge X_2(L) \ge 0$ and $\mathbf{M}X_1(L) \approx \mathbf{M}X_2(L)$ and $\mathbf{D}X_2(L)/[\mathbf{M}X_2(L)]^2 \longrightarrow 0$ as $n \longrightarrow \infty$. Then, $X_2(L) \approx X_1(L) \approx \mathbf{M}X_2(L)$ as $n \longrightarrow \infty$ for almost all matrices L* in M_{mn}^k.

On $M_{mn}^k = \{L\}$ consider the random variables

$$\xi_{w,\sigma}(L) = \begin{cases} 1, & \text{if the set of columns with indices in } w \text{ is a } \sigma\text{-sovering of } L, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$\eta_{w,\sigma}(L) = \begin{cases} 1, & \text{if the set of columns with indices in } w \text{ is a} \\ & \sigma\text{-covering of } L, \\ 0 & \text{otherwise.} \end{cases}$$

Define

$$\xi_r(L) = \sum_{w \in W_r^n} \sum_{\sigma \in E_k^r} \xi_{w,\sigma}(L), \quad \xi_\varphi(L) = \sum_{r \in \varphi} \sum_{w \in W_r^n} \sum_{\sigma \in E_k^r} \xi_{w,\sigma}(L),$$

$$\eta_r(L) = \sum_{w \in W_r^n} \sum_{\sigma \in E_k^r} \eta_{w,\sigma}(L), \quad \eta_\varphi(L) = \sum_{r \in \varphi} \sum_{w \in W_r^n} \sum_{\sigma \in E_k^r} \eta_{w,\sigma}(L).$$

Note that $\eta_\varphi(L) = B_\varphi(L)$ and $\xi_\varphi(L) = C_\varphi(L)$.

Let $\mathbf{P}(\xi_{w,\sigma}(L) = 1)$ be the probability that $\xi_{w,\sigma}(L) = 1$ and $\mathbf{P}(\eta_{w,\sigma}(L) = 1)$ be the probability that $\eta_{w,\sigma}(L) = 1$.

It is obvious that

$$\begin{aligned} \mathbf{P}(\xi_{w,\sigma}(L) = 1) &= |M_{(w,\sigma)}|\left|M_{mn}^k\right|^{-1} = (1 - k^{-r})^m, \\ \mathbf{P}(\eta_{w,\sigma}(L) = 1) &\le \mathbf{P}(\xi_{w,\sigma}(L) = 1) = (1 - k^{-r})^m. \end{aligned} \tag{5.1}$$

According to the inclusion–exclusion formula, we have

$$\mathbf{P}(\eta_{w,\sigma}(L) = 1) = (1 - k^{-r})^m - C_r^1(1 - 2k^{-r})^m + C_r^2(1 - 3k^{-r})^m - \ldots + (-1)^r C_r^r[1 - (r+1)k^{-r}]^m. \tag{5.2}$$

Therefore, for $r \le r_1$,

$$\mathbf{P}(\eta_{w,\sigma}(L) = 1) \ge (1 - k^{-r})^m[1 - r\exp(-2mk^{-r})] \ge (1 - k^{-r})^m(1 - 1/\log_k^{k-1} m).$$

Lemma 2. *It holds that*

$$\mathbf{M}\xi_{r_1}(1 - 1/\log_k^{k-1} m) \le \mathbf{M}\eta_{r_1} \le \mathbf{M}\xi_{r_1}.$$

Proof. The upper estimate for $\mathbf{M}\eta_{r_1}$ is obvious. Let us prove the lower estimate. According to (5.1) and

(5.2), we have

$$\mathbf{M}\xi_r(L) = \sum_{w \in W_r^n} \sum_{\sigma \in E_k^r} \mathbf{P}(\xi_{w,\sigma}(L) = 1) = C_n^r k^r (1 - k^{-r})^m,$$

$$\mathbf{M}\eta_r(L) = \sum_{w \in W_r^n} \sum_{\sigma \in E_k^r} \mathbf{P}(\eta_{w,\sigma}(L) = 1) = C_n^r k^r \sum_{i=0}^{r} (-1)^i C_r^i [1 - (i+1)k^{-r}]^m.$$

Consequently,

$$\mathbf{M}\xi_r - \mathbf{M}\eta_r = C_n^r k^r \sum_{i=1}^{r} (-1)^{i-1} C_r^i [1 - (i+1)k^{-r}]^m \le C_n^r k^r r(1 - 2k^{-r})^m \le r(1 - k^{-r})^m \mathbf{M}\xi_r.$$

Setting $r = r_1$ gives

$$\mathbf{M}\xi_{r_1} - \mathbf{M}\eta_{r_1} \le r_1 \exp(-mk^{-r_1}) \mathbf{M}\xi_{r_1} \le \mathbf{M}\xi_{r_1} / \log_k^{k-1} m,$$

which completes the proof.

Lemma 3. *If* $m < k^{n^\beta}$ *and* $\beta < 1$, *then*

$$\mathbf{M}\xi_\varphi \approx \mathbf{M}\eta_\varphi \approx \mathbf{M}\xi_{r_1} \approx \mathbf{M}\eta_{r_1} \approx k^{r_1} C_n^{r_1} (1 - k^{-r_1})^m, \quad n \longrightarrow \infty.$$

Proof. We have

$$\mathbf{M}\xi_\varphi(L) = \sum_{r \in \varphi} \sum_{w \in W_r^n} \sum_{\sigma \in E_k^r} \mathbf{P}(\xi_{w,\sigma}(L) = 1) = \sum_{r \in \varphi} C_n^r k^r (1 - k^{-r_1})^m.$$

Define $A_r = C_n^r k^r (1 - k^{-r})^m$. Let $r \le r_1$. Then

$$\frac{A_{r-1}}{A_r} = \frac{C_n^{r-1} k^{r-1} (1 - k^{-r+1})^m}{C_n^r k^r (1 - k^{-r})^m} \le \frac{r}{k(n - r + 1)} \le \frac{\log_k m - \log_k \ln \log_k m + 1}{n - (\log_k m - \log_k \ln \log_k m)} \le \frac{\log_k m}{n - \log_k m} \le o(1).$$

Consequently,

$$\mathbf{M}\xi_\varphi \approx \mathbf{M}\xi_{r_1}, \quad n \longrightarrow \infty. \tag{5.3}$$

Applying Lemma 2 and taking into account (5.3) and $\mathbf{M}\eta_{r_1} \le \mathbf{M}\eta_\varphi \le \mathbf{M}\xi_\varphi$, we derive the required relation.

Lemma 4. *If* $m < k^{n^\beta}$ *and* $\beta < 1/2$, *then*

$$\mathbf{D}\xi_{r_1}(L) / [\mathbf{M}\xi_{r_1}(L)]^2 \longrightarrow 0, \quad n \longrightarrow \infty.$$

Proof. Let $r = r_1$. We have

$$\mathbf{D}\xi_r(L) = \mathbf{M}\xi_r^2(L) - [\mathbf{M}\xi_r(L)]^2.$$

It is obvious that

$$\mathbf{M}\xi_r^2(L) = \sum_{w_1, w_2 \in W_r^n} \sum_{\sigma_1, \sigma_2 \in E_k^r} \left| M_{(w_1, \sigma_1)} \cap M_{(w_2, \sigma_2)} \right|.$$

Suppose that $w_1$and w_2 intersect in $r - t$ elements. Then

$$\left| M_{(w_1, \sigma_1)} \cap M_{(w_2, \sigma_2)} \right| \le (k^r - 1)^m (k^t - 1)^m k^{m(n - r - t)} = (k^{r+t} - k^t - k^r + 1)^m k^{m(n - r - t)}$$
$$= (k^r - 1 - k^{r-t} + k^{-t})^m k^{m(n-r)} \le (k^r - 2 + k^{-t})^m k^{m(n-r)}.$$

Hence,

$$\mathbf{M}\xi_r^2(L) \le C_n^r k^r k^{-mr} \sum_{t=0}^{\min(r, n-r)} C_{n-r}^t C_r^t k^t (k^r - 2 + k^{-t})^m.$$

Since

$$(k^r - 2 + k^{-t})^m = (k^r - 2)^m \left(1 + \frac{1}{(k^r - 2)k^t}\right)^m \le (k^r - 2)^m \left(1 + \frac{1}{k^{r+t-1}}\right)^m \quad \text{for} \quad r \ge 2,$$

we have

$$\mathbf{M}\xi_r^2(L) \le C_n^r k^r k^{-mr} \sum_{t=0}^{\min(r,\, n-r)} C_{n-r}^t C_r^t k^t (k^r - 2)^m \exp\left(\frac{m}{k^{r+t-1}}\right).$$

Let $a_t = C_{n-r}^t C_r^t k^t \exp(m/k^{r+t-1})$ and $t_0 = [\log_k \log_k m + \log_k \ln \log_k m]$.

For $t \le t_0$, we have

$$a_t \le (knr)^{t_0} \exp(m/k^{r-1}) \le (knr)^{t_0} \log_k^{k^2} m. \tag{5.4}$$

For $t > t_0$, we have

$$a_t \le k^t C_{n-r}^t C_r^t \exp(k^4/\log_k m). \tag{5.5}$$

Note that

$$\sum_{t=0}^{\min(r,\, n-r)} C_{n-r}^t C_r^t \le C_n^r. \tag{5.6}$$

It follows from (5.4)–(5.6) that

$$\sum_{t=0}^{\min(r,\, n-r)} a_t \le C_n^r k^r \left[\frac{(t_0 + 1)(rn)^{t_0} \log_k^{k^2} m}{C_n^r} + \exp\left(\frac{k^4}{\log_k m}\right)\right]. \tag{5.7}$$

By using the Stirling formula, it is easy to show that

$$C_n^r = \frac{n!}{(n-r)!\,r!} \approx \sqrt{\frac{n}{n-r}} \times \frac{n^n}{(n-r)^{n-r} r^r \sqrt{2\pi r}}, \quad n \longrightarrow \infty.$$

Since $\sqrt{n/n-r} \longrightarrow 1$ for $m < k^{n^\beta}$, $\beta < 1$, $n \longrightarrow \infty$, we obtain

$$\frac{(t_0 + 1)(rn)^{t_0} \log_k^{k^2} m}{C_n^r} \lessapprox \frac{(n-r)^{n-r} r^r \sqrt{2\pi r}(t_0 + 1)(rn)^{t_0} \log_k^{k^2} m}{n^n}, \quad n \longrightarrow \infty. \tag{5.8}$$

We estimate

$$\begin{aligned} (n-r)^{n-r} n^{-n} &= n^{n-r}\left(1 - \frac{r}{n}\right)^{n-r} n^{-n} \le \exp\left(-\frac{r(n-r)}{n}\right) n^{-r} \\ &= \exp\left(-r + \frac{r^2}{n}\right) n^{-r} \approx \frac{1}{\exp(r) n^r} \quad \text{for} \quad m \le k^{n^\beta}, \quad \beta < 1/2, \quad n \longrightarrow \infty. \end{aligned} \tag{5.9}$$

It follows from (5.7)–(5.9) that

$$\sum_{t=0}^{\min(r,\, n-r)} a_t \le C_n^r k^r [1 + \delta(n)], \quad \text{where} \quad \delta(n) \longrightarrow 0, \quad n \longrightarrow \infty.$$

Thus,

$$\mathbf{M}\xi_r^2(L) \le C_n^r k^r k^{-mr} (k^r - 2)^m C_n^r k^r [1 + \delta(n)].$$

We have

$$[\mathbf{M}\xi_r(L)]^2 = (C_n^r k^r)^2 (k^r - 1)^{2m} \frac{1}{k^{2mr}} = (C_n^r k^r)^2 (k^{2r} - 2k^r + 1)^m \frac{1}{k^{2mr}}$$

$$= (C_n^r k^r)^2 k^{rm} \left(k^r - 2 + \frac{1}{k^r}\right)^m \frac{1}{k^{2mr}} \geq (C_n^r k^r)^2 (k^r - 2)^m k^{-mr},$$

which completes the proof of Lemma 4.

Lemma 5. *If* $m < k^{n^\beta}$ *and* $\beta < 1/2$, *then*

$$\mathbf{D}\eta_{r_1}/(\mathbf{M}\eta_{r_1})^2 \longrightarrow 0, \quad n \longrightarrow \infty.$$

Proof. We have

$$\frac{\mathbf{D}\eta_{r_1}}{(\mathbf{M}\eta_{r_1})^2} = \frac{\mathbf{M}\eta_{r_1}^2}{(\mathbf{M}\eta_{r_1})^2} - 1 \leqslant \frac{\mathbf{M}\xi_{r_1}^2}{(\mathbf{M}\xi_{r_1})^2} - 1.$$

The lemma is proved.

Theorem 1 follows directly from Lemmas 1 and 3–5.

Remark. Lemmas 2 and 4 were proved by the respective techniques applied in the proof of Proposition 5 and Lemma 3 in [13], where the number of maximal intervals of a partial Boolean function was estimated and asymptotics of the logarithm of the number of maximal intervals were obtained.

ACKNOWLEDGMENTS

This work was supported in part by the Russian Foundation for Basic Research (project nos. 01-01-00575) and the program "Leading Scientific Schools" (project no. 1721.2003.1).

REFERENCES

1. Djukova, E.V., On the Complexity of Implementation of Some Recognition Procedures, *Zh. Vychisl. Mat. Mat. Fiz.*, 1987, vol. 27, no. 1, pp. 114–127.
2. Djukova, E.V., Recognition Algorithms of the Kora Type: Complexity of Implementation and Metric Properties, *Raspoznavanie, klassifikatsiya, prognoz (matematicheskie metody i ikh primenenie)* (Recognition, Classification, Forecast: Mathematical Methods and Their Applications), Moscow: Nauka, 1989, no. 2, pp. 99–125.
3. Vaintsvaig, M.N., Kora: A Learning Pattern Recognition Algorithm, in *Algoritmy obucheniya raspoznavaniyu obrazov* (Learning Algorithms in Pattern Recognition), Moscow: Sov. Radio, 1973, pp. 82–91.
4. Baskakova, L.V. and Zhuravlev, Yu.I., A Model of Recognition Algorithms with Representative Descriptors and Systems of Reference Sets, *Zh. Vychisl. Mat. Mat. Fiz.*, 1981, vol. 21, no. 5, pp. 1264–1275.
5. Djukova, E.V. and Inyakin, A.S., Taxonomy Problem and Irredundant Coverings of an Integer Matrix, in *Soobshcheniya po prikladnoi matematike* (Reports on Applied Mathematics), Moscow: Vychisl. Tsentr Ross. Akad. Nauk, 2001.
6. Djukova, E.V. and Zhuravlev, Yu.I., Discrete Analysis of Feature Descriptions in Recognition Problems of High Dimensionality, *Zh. Vychisl. Mat. Mat. Fiz.,* 2000, vol. 40, no. 8, pp. 1264–1278.
7. Djukova, E.V. and Zhuravlev, Yu.I., Discrete Methods of Information Analysis and Algorithm Synthesis, *J. Pattern Recogn. Image Anal.*, 1997, vol. 7, no. 2, pp. 192–207.
8. Yablonskii, S.V., *Vvedenie v diskretnuyu matematiku* (Introduction to Discrete Mathematics), Moscow: Nauka, 1986.
9. Djukova, E.V., Asymptotically Optimal Test Algorithms in Recognition Problems, *Problemy kibernetiki* (Problems in Cybernetics), Moscow: Nauka, 1982, vol. 39, pp. 165–199.
10. Djukova, E.V., Asymptotic Estimates for Characteristics of the Set of Representative Descriptors and Stability Problem, *Zh. Vychisl. Mat. Mat. Fiz.*, 1995, vol. 35, no. 1, pp. 122–134.
11. Djukova, E.V. and Peskov, N.V., Search for Informative Fragments in Descriptions of Objects in Discrete Recognition Procedures, *Zh. Vychisl. Mat. Mat. Fiz.*, 2002, vol. 42, no. 5, pp. 741–753.
12. Noskov, V.N. and Slepyan, V.A., On the Number of Irredundant Tests for a Class of Tables, *Kibernetika,* 1972, no. 1, pp. 60–65.
13. Sapozhenko, A.A., An Estimate of the Length and Number of Irredundant Disjunctive Normal Forms for Almost All Incompletely Defined Boolean Functions, *Mat. Zametki*, 1980, vol. 28, no. 2, pp. 279–300.