ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2003, том 43, № 12, с. 1910–1921

УДК 519.714

О ПРОЦЕДУРАХ КЛАССИФИКАЦИИ, ОСНОВАННЫХ НА ПОСТРОЕНИИ ПОКРЫТИЙ КЛАССОВ[1)]

© 2003 г. Е. В. Дюкова, А. С. Инякин

(119991 Москва, ул. Вавилова, 40, ВЦ РАН)
e-mail: djukova@ccas.ru, andre–w@mail.ru

Поступила в редакцию 28.02.2003 г.

Описывается подход к решению задачи кластеризации с целочисленной информацией, основанный на построении специальных наборов значений признаков, не содержащихся в признаковых описаниях объектов. Задача сводится к построению тупиковых покрытий целочисленных матриц, которое может быть осуществлено на основе построения неприводимых покрытий булевых матриц. Предлагается новый метод построения неприводимых и минимальных покрытий булевой матрицы, использующий геометрическую интерпретацию понятия покрытия. Изучаются метрические свойства близких к минимальным покрытий целочисленных матриц. Библ. 13. Фиг. 5.

ВВЕДЕНИЕ

В ряде прикладных задач классификации зачастую возникает необходимость разбить совокупность объектов на “однородные” группы (классы) при условии отсутствия обучающей информации. Задачи такого типа называют задачами таксономии.

Итак, пусть имеется конечная выборка M, каждый элемент которой описывается конечным набором признаков (свойств). Требуется разбить данную выборку на однородные группы, число которых может быть заранее заданным, а может быть и неизвестным. Процесс деления на классы называют кластеризацией или кластерным анализом.

При постановке задачи необходимо описать правила, которым должны подчиняться объекты из одного класса и объекты из разных классов. В качестве такого правила можно рассматривать компактность объектов в заданном признаковом пространстве, т.е. правило, в соответствии с которым, например, “расстояние” между объектами, отнесенными в один класс, должно быть не более заданного. Может также использоваться “хорошая” взаимная удаленность классов, например, расстояние между объектами, отнесенными в разные классы, должно быть не менее заданного.

В данной работе рассматриваются некоторые подходы к решению задач таксономии с целочисленной информацией. Отличием предлагаемых методов от классических является то, что при решении указанных задач не нужно подбирать функцию расстояния, которая не всегда может быть достаточно простой для вычисления. В их основе лежат следующие соображения.

Рассмотрим ситуацию, когда определяется степень принадлежности объекта S к группе объектов M. Если описание объекта S содержит набор значений признаков, который не присутствует в описании ни одного объекта из M, то можно сказать, что объединение S и M нарушает внутреннюю структуру множества M. Рассматривая различные комбинации значений признаков, не содержащихся в описаниях объектов из M, можно оценить близость объекта S к множеству M. Таким образом, при определении степени близости объекта к множеству M информативными считаются такие наборы из допустимых значений признаков, которые отсутствуют в описаниях всех объектов из класса M.

На основе приведенных выше рассуждений был построен алгоритм, который в ряде случаев позволил значительно улучшить результаты кластеризации на модельных и реальных задачах по сравнению с такими известными алгоритмами, как алгоритмы, основанные на принципах “ближайший сосед”, “дальний сосед” и выбор центрального элемента, в которых в качестве функции расстояния используется расстояние Хемминга.

[1)]Работа выполнена при частичной финансовой поддержке РФФИ (коды проектов 01-01-00575, 00-15-96064) и Программы Президиума РАН “Интеллектуальные компьютерные системы” (проект № 1.1).

В предложенном алгоритме кластеризации задача оценки близости объекта S к множеству M сводится к поиску так называемых тупиковых σ-покрытий целочисленной матрицы, образованной признаковыми описаниями объектов из M. Данное понятие является обобщением хорошо известного в дискретной математике понятия неприводимого покрытия булевой матрицы и первоначально было введено в [1], [2] в связи с задачами построения тупиковых представительных наборов в алгоритмах типа "Кора" (см. [3], [4]) и с изучением метрических свойств множества представительных наборов. В [5] показано, что задача построения (тупиковых) покрытий целочисленных матриц может быть сведена к задаче поиска (неприводимых) покрытий булевой матрицы. Наиболее полно подход к решению задач распознавания, основанный на построении покрытий булевых и целочисленных матриц, описан в [6], [7].

Более информативными считаются наборы значений признаков, порождаемые короткими σ-покрытиями, поэтому, как правило, строят не все σ-покрытия, а только те, длины которых не превосходят некоторой наперед заданной величины. В связи с этим в данной работе изучаются метрические (количественные) свойства σ-покрытий, по длине близких к минимальным σ-покрытиям. Получена асимптотика числа таких покрытий в типичном случае.

В статье предлагается геометрическая интерпретация понятий σ-покрытия и тупикового σ-покрытия матрицы L размера $m \times n$ с элементами из $\{0, 1, \ldots, k-1\}$, $k \geq 2$. Пусть M_L – множество таких k-х наборов длины n (точек k-значного n-мерного куба), которые являются строками матрицы L. Показано, что задача построения (тупиковых) σ-покрытий матрицы L сводится к задаче построения (максимальных) подкубов в множестве $\overline{M}_L$, состоящем из всех k-х наборов длины n, не вошедших в матрицу L. Для случая $k = 2$ предложен алгоритм поиска максимальных подкубов в $\overline{M}_L$, соответствующих неприводимым покрытиям булевой матрицы L, и модификация этого алгоритма на случай поиска максимальных подкубов в $\overline{M}_L$, соответствующих минимальным по длине покрытиям матрицы L.

1. ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ. ГЕОМЕТРИЧЕСКАЯ ИНТЕРПРЕТАЦИЯ ПОНЯТИЙ ПОКРЫТИЯ И ТУПИКОВОГО ПОКРЫТИЯ ЦЕЛОЧИСЛЕННОЙ МАТРИЦЫ

Введем следующие обозначения:

E_k^r, $k \geq 2$, – множество всех k-х наборов длины r, $r \leq n$; $E^r = E_2^r$; σ – набор из E_k^r вида $(\sigma_1, \ldots, \sigma_r)$.

Набор H из r различных столбцов матрицы L назовем σ-покрытием, если подматрица L^H матрицы H, образованная столбцами набора H, не содержит строку $(\sigma_1, \ldots, \sigma_r)$.

Набор столбцов H, являющийся σ-покрытием матрицы L, назовем тупиковым σ-покрытием, если для любого $p \in \{1, 2, \ldots, r\}$ подматрица L^H содержит хотя бы одну строку вида $(\beta_1, \ldots, \beta_r)$, где $\beta_p \neq \sigma_p$ и $\beta_i = \sigma_i$ при $i = \overline{1, r}$ и $i \neq p$, т.е. L^H содержит подматрицу вида

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

Такую подматрицу назовем σ-подматрицей.

В случае $k = 2$ и $\sigma = (0, \ldots, 0)$ понятие тупикового σ-покрытия совпадает с известным понятием неприводимого покрытия булевой матрицы (см. [8]), при этом σ-подматрица является единичной подматрицей.

Обозначим через $C(L, \sigma)$, $B(L, \sigma)$ и $S(L, \sigma)$, соответственно, множество σ-покрытий матрицы L, множество тупиковых σ-покрытий матрицы L и множество σ-подматриц матрицы L. Положим

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

Покажем, что задачи построения множеств $C(L)$ и $B(L)$ можно свести, соответственно, к задачам построения $(0, \ldots, 0)$-покрытий и тупиковых $(0, \ldots, 0)$-покрытий (неприводимых покрытий) булевой матрицы, которая специальным образом строится по L.

Пусть матрица L имеет вид

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

где каждый элемент столбца с номером j может принимать k_j ($k_j \le k$) значений (для дальнейших рассуждений удобно считать, что элементы столбца с номером j принимают значения от 1 до k_j).

Положим

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

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

Построим матрицу L^B, состоящую из m строк, в которой строка с номером i, $i \in \{1, 2, \ldots, m\}$, имеет вид

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

Нетрудно заметить, что столбцу с номером j, $j \in \{1, 2, \ldots, n\}$, исходной матрицы соответствует группа, состоящая из k_j столбцов бинарной матрицы L^B, обозначаемая далее через g_j. В группе g_j столбцы будем нумеровать числами от 1 до k_j.

Обозначим через (j, i) столбец матрицы L^B из группы g_j с номером i, $j \in \{1, 2, \ldots, n\}$, $i \in \{1, 2, \ldots, k_j\}$.

Очевидным является

Утверждение 1. *Набор столбцов матрицы L^B вида $((j_1, i_1), \ldots, (j_r, i_r))$, $r \in 1, 2, \ldots, n$, $j_l \in \{1, 2, \ldots, n\}$, $i_l \in \{1, 2, \ldots, k_{j_l}\}$, $j_{l_1} \ne j_{l_2}$ при $l_1 \ne l_2$, $l_1, l_2 = 1, 2, \ldots, r$, является (неприводимым) покрытием тогда и только тогда, когда набор столбцов с номерами (j_1, j_r) является (тупиковым) (j_1, j_r)-покрытием матрицы L.*

Множество E_k^n можно рассматривать как k-значный n-мерный куб. Обозначим через M_L множество точек куба E_k^n, задаваемых строками матрицы L. Положим $\overline{M}_L = E_k^n \backslash M_L$. Набору столбцов H с номерами $(j_1, \ldots, j_r)$, являющимся σ-покрытием матрицы L, поставим в соответствие множество точек $E^{(\sigma, H)} = \{(\alpha_1, \ldots, \alpha_n) | \alpha \in E_k^n, \alpha_{j_i} = \sigma_i, i = \overline{1, r}\}$. Очевидно, $E^{(\sigma, H)}$ является $(n - r)$-мерным подкубом куба E_k^n и целиком принадлежит множеству $\overline{M}_L$. В случае, когда $r = n$ множество точек $E^{(\sigma, H)}$ является 0-мерным подкубом, состоящим из одной точки σ. Таким образом, множеству $C(L)$ взаимно однозначно соответствует множество всех подкубов, содержащихся в $\overline{M}_L$. Отметим, что $E^{(\sigma_1, H_1)}$ является подкубом $E^{(\sigma_2, H_2)}$ тогда и только тогда, когда $\sigma_2 \subset \sigma_1$, $H_2 \subset H_1$.

Пусть M – некоторое подмножество множества E_k^n. Будем говорить, что подкуб $E' \subset M$ является максимальным в M, если не существует другого подкуба $E'' \subset M$ такого, что $E' \subset E''$. Множество всех максимальных подкубов в M будем обозначать через $E^{\max}(M)$. Нетрудно видеть, что множеству $B(L)$ взаимно однозначно соответствует множество $E^{\max}(\overline{M}_L)$. Следовательно, задачу построения множества $B(L)$ можно свести к задаче построения множества всех максимальных подкубов в $\overline{M}_L$.

2. АЛГОРИТМЫ ПОИСКА НЕПРИВОДИМЫХ И МИНИМАЛЬНЫХ ПОКРЫТИЙ БУЛЕВОЙ МАТРИЦЫ, ИСПОЛЬЗУЮЩИЕ ГЕОМЕТРИЧЕСКУЮ ИНТЕРПРЕТАЦИЮ ПОНЯТИЯ ПОКРЫТИЯ

В [9] построен асимптотически оптимальный алгоритм поиска всех неприводимых покрытий булевой матрицы в случае, когда число строк матрицы существенно меньше числа столбцов. Описываемые в данном разделе алгоритмы предназначены в основном для поиска неприводимых и минимальных по длине покрытий булевой матрицы в случае, когда число строк матрицы существенно больше числа столбцов (в этом случае они работают достаточно быстро).

Пусть $k = 2$. Очевидно, что если точка $\tilde{0} = (0, \ldots, 0)$ принадлежит M_L, то в матрице L нет ни одного неприводимого покрытия, поэтому далее будем полагать, что $\tilde{0} \in \overline{M}_L$.

Легко видеть, что $(0, \ldots, 0)$-покрытию булевой матрицы L соответствует подкуб в $\overline{M}_L$, одной из вершин которого является точка $\tilde{0}$. Неприводимому покрытию матрицы L соответствует максимальный подкуб в $\overline{M}_L$, одной из вершин которого является точка $\tilde{0}$. Таким образом, множеству всех неприводимых покрытий булевой матрицы L соответствует пучок всех максимальных подкубов в $\overline{M}_L$, исходящий из точки $\tilde{0}$.

Будем говорить, что точка $\tilde{\alpha} = (\alpha_1, \ldots, \alpha_n)$, $\tilde{\alpha} \in E^n$, охватывает точку $\tilde{\beta} = (\beta_1, \ldots, \beta_n)$, $\tilde{\beta} \in E^n$, и обозначать это через $\tilde{\alpha} > \tilde{\beta}$, если $\alpha_i \geq \beta_i$, $i = \overline{1, n}$. Через $M(\tilde{\alpha}, \tilde{\beta})$ будем обозначать подкуб минимальной размерности в E^n, содержащий точки $\tilde{\alpha}$ и $\tilde{\beta}$. Будем говорить, что подкуб $M(\tilde{\alpha}, \tilde{\beta})$ является покрывающим для M_L, если $M(\tilde{\alpha}, \tilde{\beta}) \subseteq \overline{M}_L$.

Очевидным является

Утверждение 2. *Если $\tilde{\alpha} \in \overline{M}_L$, то $M(\tilde{\alpha}, \tilde{0})$ не содержится в $\overline{M}_L$ тогда и только тогда, когда в M_L можно указать точку $\tilde{\beta}$ такую, что $\tilde{\alpha} > \tilde{\beta}$.*

Алгоритм построения всех неприводимых покрытий матрицы L

Шаг 1. Из $\overline{M}_L$ удаляются все точки $\tilde{\alpha}$ такие, что в M_L, найдется точка $\tilde{\beta}$ для которой $\tilde{\alpha} > \tilde{\beta}$.

Шаг 2. Пусть $\tilde{\alpha}' \in \overline{M}_L$ и $\rho(\tilde{\alpha}', \tilde{0}) = \max\limits_{\tilde{\alpha} \in \overline{M}_L} \rho(\tilde{\alpha}, \tilde{0})$ (здесь и далее $\rho(\tilde{\alpha}, \tilde{\beta})$ – расстояние Хемминга между точками $\tilde{\alpha}$ и $\tilde{\beta}$). Очевидно, $M(\tilde{\alpha}', \tilde{0})$ – максимальный покрывающий подкуб для M_L. Пусть $j_1, \ldots, j_r$ – координаты, в которых точки $\tilde{\alpha}'$ и $\tilde{0}$ совпадают, тогда набор столбцов с номерами $j_1, \ldots, j_r$ является неприводимым покрытием матрицы L, соответствующим максимальному подкубу $M(\tilde{\alpha}', \tilde{0})$. Положим $\overline{M}_L = \overline{M}_L \backslash M(\tilde{\alpha}', \tilde{0})$.

Шаг 3. Если $\overline{M}_L \neq \varnothing$, то перейти к шагу 2, иначе конец работы алгоритма.

Алгоритм построения всех минимальных покрытий матрицы L

Шаг 1. Из $\overline{M}_L$ удаляются все точки $\tilde{\alpha}$ такие, что в M_L найдется точка $\tilde{\beta}$ для которой $\tilde{\alpha} > \tilde{\beta}$.

Шаг 2. Пусть $\tilde{\alpha}' \in \overline{M}_L$ и $\rho(\tilde{\alpha}', \tilde{0}) = \max\limits_{\tilde{\alpha} \in \overline{M}_L} \rho(\tilde{\alpha}, \tilde{0})$. Положим

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

Шаг 3. Пусть $\tilde{\alpha}' \in Q_L$, $j_1, \ldots, j_r$ – координаты, в которых точки α' и 0 совпадают, тогда набор столбцов с номерами $j_1, \ldots, j_r$ является минимальным покрытием матрицы L, соответствующим максимальному подкубу $M(\tilde{\alpha}', \tilde{0})$. Положим $Q_L = Q_L \backslash \tilde{\alpha}'$.

Шаг 4. Если $Q_L \neq \varnothing$, то перейти к шагу 3, иначе – конец работы алгоритма.

Заметим, что описанный алгоритм поиска неприводимых покрытий может быть модифицирован на случай поиска тупиковых σ-покрытий.

3. АЛГОРИТМ КЛАСТЕРИЗАЦИИ, ОСНОВАННЫЙ НА ПОСТРОЕНИИ ТУПИКОВЫХ σ-ПОКРЫТИЙ

Меру близости множества $M'' \subset E^n$ к множеству $M' \subset E^n$ определим следующей величиной:

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

здесь $|A|$ – мощность множества A.

Зададим некоторое число p, $p \in (0, 1)$. Будем говорить, что множество M'' близко к M' по порогу p, если $P(M', M'') \geq p$.

Пусть задача кластеризации решается для множества объектов $M = \{S_1, \ldots, S_m\}$.

Формирование первого класса M_1 осуществляется следующим образом.

Шаг 1. Положить $M_1 = S$, где S – произвольный элемент из множества M. Если $M \backslash M_1 = \varnothing$, то алгоритм заканчивает работу. В противном случае перейти к шагу 2.

Шаг 2. Если в $M \backslash M_1$ существует элемент S' такой, что

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

то положить $M_1 = M_1 \cup \{S'\}$и повторить шаг 2. В противном случае перейти к формированию следующего класса.

Пусть построены классы $M_1, \ldots, M_{i-1}$, $i \geq 2$, и пусть

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

Если $\tilde{M}_i = \varnothing$, то алгоритм заканчивает работу, в противном случае осуществляется формирование класса M_i. Процедура построения класса M_i полностью аналогична процедуре построения класса M_1 при условии, что за исходное множество вместо M берется $\tilde{M}_i$.

В данном алгоритме для определения близости объектов или групп объектов используется понятие максимального подкуба множества точек, что дает возможность не вводить функцию расстояния между объектами.

Решение практических задач показало, что число p целесообразно выбирать из интервала (0.5, 0.9). Чем больше величина p, тем, соответственно, большее число классов образуется при кластеризации.

Описание алгоритма без использования геометрического языка приведено в [5].

4. РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ НА МОДЕЛЬНЫХ И РЕАЛЬНЫХ ЗАДАЧАХ

Приведем результаты тестирования на модельных и реальных задачах алгоритмов кластеризации, основанных на принципах дальний сосед (алгоритм A_1), ближайший сосед (алгоритм A_2), выбора центрального элемента (алгоритм A_3) и алгоритма кластеризации, основанного на построении тупиковых σ-покрытий (алгоритм A_4).

Алгоритмы A_1, A_2 и A_3 относятся к классу так называемых иерархических методов группировки, суть которых состоит в следующем. Рассматривается последовательность разделений m объектов на группы. Первый шаг последовательности – разделение на m групп таких, что каждая группа содержит в точности по одному объекту. Следующий шаг – разделение на $m - 1$ группу, затем на $m - 2$ групп и т.д., до m-го шага, в котором все объекты образуют одну группу. Таким образом, k-му шагу соответствует разделение на $m + 1 - k$ групп, где $k = 1, 2, \ldots, m$. Последовательность разделений называется иерархической группировкой, если любые два объекта S_1 и S_2, объединенные в одну группу на k-м шаге, содержатся в одной группе и на всех последующих шагах.

Пусть требуется разделить m объектов S_1 и S_m на l классов. Основные этапы иерархической группировки описываются следующей процедурой.

1. Пусть $\tilde{l} = m$ и $\Omega_i = \{S_i\}$, $i = \overline{1, m}$.

2. Если $\tilde{l} \le l$, то выход из процедуры.

3. Найти такие две группы Ω_i, Ω_j, для которых справедливо

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

где $\rho(\Omega_u, \Omega_v)$ – расстояние между группами Ω_u и Ω_v (см. ниже).

4. Объединить Ω_i и Ω_j, уничтожить Ω_i и уменьшить $\tilde{l}$ на единицу. Перейти к шагу 2.

Процедура заканчивает работу, когда достигнуто заданное число групп, т.е. при $\tilde{l} = l$.

Заметим, что в случае, когда число групп, на которое нужно разбить множество, не задано, применяются различные условия окончания процедуры группировки, в частности если наименьшее расстояние между группами на i-м шаге более чем в p раз превосходит наименьшее расстояние на $(i-1)$-м шаге (число $p > 1$ задано заранее, обычно $p > 1.5$).

Заметим также, что результат работы описанной выше процедуры сильно зависит от выбора используемой функции расстояния между группами. Для тестирования использовались перечисленные ниже функции.

1. Расстояние, измеряемое по принципу ближайший сосед:

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

2. Расстояние, измеряемое по принципу дальний сосед:

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

3. Расстояние между центральными элементами классов:

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

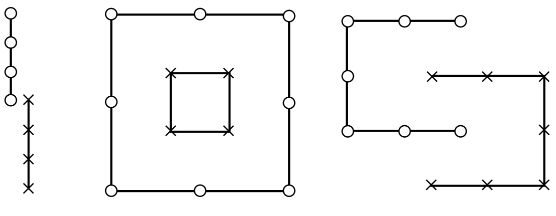
где $\bar{S}_u$ – центральный элемент класса Ω_u, а $\bar{S}_v$ – центральный элемент класса Ω_v. В частности, в качестве центрального элемента класса можно выбрать центр тяжести класса.

В перечисленных выше формулах расстояние $\rho(S_i, S_j)$ между объектами $S_i = (x_{i1}, \ldots, x_{in})$ и $S_j = (x_{j1}, \ldots, x_{jn})$ вычислялось по формуле

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

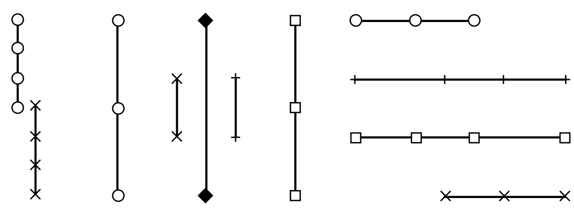
Сравнение алгоритма A_4 с алгоритмами A_1, A_2 и A_3 проводилось на модельных множествах данных, изображенных на фиг. 1.

Результаты работы алгоритма A_1 изображены на фиг. 2. Заметим, что алгоритм A_1 дает "правильное", т.е. наиболее естественное на интуитивном уровне восприятия разбиение множества, изображенного на фиг. 1а, при произвольном порядке следования объектов в исходной выборке (вообще говоря, результаты работы рассматриваемых алгоритмов зависят от этого порядка).



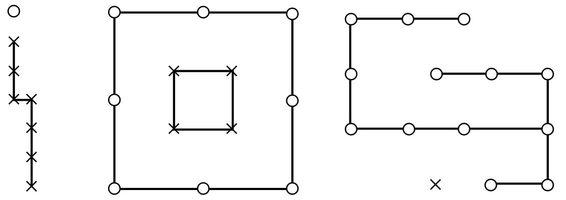
(а) (б) (в)

Фиг. 1.



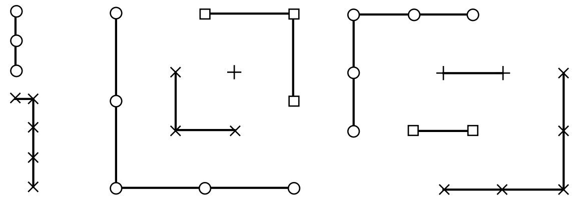
(а) (б) (в)

Фиг. 2.



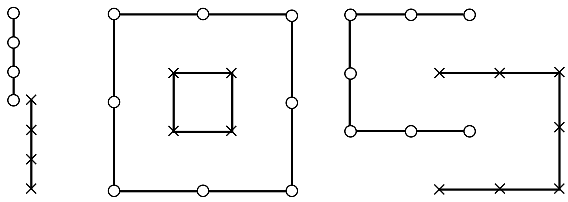
(а) (б) (в)

Фиг. 3.



(а) (б) (в)

Фиг. 4.



(а) (б) (в)

Фиг. 5.

На фиг. 3 изображены результаты работы алгоритма A_2. Применение алгоритма A_2 существенно “улучшает” разбиение множества, изображенного на фиг. 1б, но “ухудшает” разбиение множеств, изображенных на фиг. 1а и 1в.

На фиг. 4 изображены результаты работы алгоритма A_3. Алгоритм, основанный на выборе центрального элемента, демонстрирует некое усреднение методов ближайший сосед и дальний сосед.

На фиг. 5 изображены результаты работы алгоритма A_4. Алгоритм, основанный на построении тупиковых σ-покрытий, в отличие от предыдущих алгоритмов показывает лучшие результаты.

Алгоритмы кластеризации A_3 и A_4 были также протестированы на реальной информации социологического опроса. Анкета социологического опроса состояла из 21 вопроса, каждый содержал от 2 до 15 вариантов ответов. Всего в опросе приняло участие 799 респондентов. Множество респондентов было разбито на классы по ключевому вопросу анкеты: “Ваше отношение к партии A”. На этот вопрос было предложено четыре варианта ответа: симпатия, равнодушие, неприязнь и трудно сказать. Таким образом, было выделено 4 класса. Для каждого эксперимента случайным образом выбиралась небольшая группа (от 4 до 20) респондентов из каждого класса. Информация о принадлежности к какому-либо классу исключалась. Затем на полученном множестве проводилась кластеризация алгоритмом, основанным на построении тупиковых σ-покрытий (алгоритм A_4), и алгоритмом, основанным на вычислении расстояния Хемминга с выбором центрального элемента (алгоритм A_3). Всего было проведено 32 эксперимента. Разбиения, полученные в результате работы указанных алгоритмов, сравнивались с исходным разбиением на классы. Качество работы алгоритма характеризовалось отношением числа “правильно” классифицированных пар объектов к общему числу пар. При этом классификация считалась правильной, если выполнялось одно из условий: пара объектов, принадлежавших одному классу в исходном разбиении, после кластеризации также оказывалась в одном классе; пара объектов, принадлежавших различным классам в исходном разбиении, после кластеризации также не попадала в один класс.

Точность кластеризации, показанная алгоритмом A_3 при проведении экспериментов, лежала в пределах от 60% до 80%. Точность кластеризации, показанная алгоритмом A_4, лежала в пределах от 70% до 90%.

Результаты сравнения на модельных и реальных задачах свидетельствуют о том, что алгоритмы, основанные на вычислении расстояния Хемминга, отличаются высокой скоростью счета. Эти методы используются в основном для решения задач с бинарной информацией. Их можно применять и в задачах с признаками большей значности, но в этом случае они плохо работают со специфическими множествами исходных данных, например с множествами данных, изображенных на фиг. 1–3. Алгоритм A_4 обладает большей трудоемкостью, но он не ограничен бинарными признаками и показывает в этом случае более качественные результаты.

5. МЕТРИЧЕСКИЕ СВОЙСТВА БЛИЗКИХ К МИНИМАЛЬНЫМ ПОКРЫТИЙ ЦЕЛОЧИСЛЕННЫХ МАТРИЦ

Традиционно рассмотрение вопросов повышения качества и быстродействия алгоритмов распознавания, основанных на построении покрытий булевых и целочисленных матриц, связано с получением асимптотических оценок типичных значений важнейших количественных характеристик этого множества. Такими характеристиками являются, например, число покрытий и длина покрытия. Особый интерес и наибольшую трудность представляет изучение метрических свойств тупиковых покрытий.

В [1], [2], [6] и [7] были получены асимптотики числа покрытий из $B(L)$ и длины покрытия из $B(L)$ для почти всех матриц L размера $m \times n$ с элементами из $\{0, 1, \ldots, k-1\}$, $k \geq 2$ при условиях $n \longrightarrow \infty$ и $m^{\alpha} \leq n \leq k^{m^{\beta}}$, $\alpha > 1$, $\beta < 1$. Показано, что в данном случае число покрытий из $B(L)$ почти всегда асимптотически совпадает с числом всех подматриц из $S(L)$ и по порядку меньше числа покрытий из $C(L)$.

В [11] был рассмотрен прямо противоположный случай, а именно, когда $n^{\alpha} \leq m \leq k^{n^{\beta}}$, $\alpha > 1$, $\beta < 1$. Получены асимптотики типичных значений величины $S(L)$ и порядка подматрицы из $S(L)$ и установлено, что в случае, когда $n^{\alpha} \leq m \leq k^{n^{\beta}}$, $\alpha > 1$, $\beta < 1/2$, почти всегда число подматриц из $S(L)$ по порядку больше числа покрытий из $B(L)$. Кроме того, для практически общего случая в [11] получены асимптотики типичных значений числа покрытий из $C(L)$ и длины покрытия из $C(L)$ и показано, что при $r \leq \log_k m - (\log_k m \times \ln n)$ и $n \longrightarrow \infty$ у почти всех матриц L размера $m \times n$ отсутствуют покрытия из $C(L)$ длины r.

Остался открытым вопрос: при каких условиях число тупиковых покрытий асимптотически совпадает с числом покрытий. Одним из ответов на этот вопрос является теорема 1, приводимая ниже.

Пусть $r_1 = [\log_k m - \log_k \ln \log_k m - 1]$ (здесь $[x]$ – целая часть числа x); φ – интервал $(\log_k m - \log_k(\log_k m \times \ln n), \log_k m - \log_k \ln \log_k m - 1]$; $a_n \approx b_n$ означает, что $\lim\limits_{n \to \infty} a_n/b_n = 1$; $a_n \underset{n}{\geq} b_n$ означает, что $a_n \geq b_n$ при всех достаточно больших n; M_{mn}^k, $k \geq 2$, – множество всех матриц размера $m \times n$ с элементами из $\{0, 1, \ldots, k-1\}$; $C_{\varphi}(L)$ и $B_{\varphi}(L)$ – соответственно, совокупность покрытий из $C(L)$, длины которых принадлежат интервалу φ, и совокупность покрытий из $B(L)$, длины которых принадлежат интервалу φ, $C_{r_1}(L)$ и $B_{r_1}(L)$ – соответственно, совокупность покрытий из $C(L)$ длины r_1 и совокупность покрытий из $B(L)$ длины r_1.

Теорема 1. *Если $m \leq k^{n^{\beta}}$, $\beta < 1/2$, то для почти всех матриц L из M_{mn}^k при $n \longrightarrow \infty$ справедливо соотношение*

$$|B_{\varphi}(L)| \approx |C_{\varphi}(L)| \approx |C_{r_1}(L)| \approx |B_{r_1}(l)| \approx k^{r_1} C_n^{r_1}(1 - k^{-k_1})^m.$$

Доказательство теоремы опирается на приводимые ниже леммы 1–6.

Обозначим через W_r^n, $r \leq n$, – множество всех упорядоченных наборов вида $(j_1, \ldots, j_r)$, где $j_l \in \{1, 2, \ldots, n\}$, при $l = \overline{1, r}$ и $j_1 < \ldots < j_r$.

Пусть $w \in W_r^n$, $\omega \in E_k^r$. Обозначим через $M_{(w, \sigma)}$ множество матриц L из M_{mn}^k таких, что подматрица матрицы L, образованная столбцами набора w, не содержит строку σ.

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

Лемма 1 (см. [12]). *Пусть для случайных величин* $X_1(L)$ *и* $X_2(L)$, *определенных на* M_{mn}^k, *выполнено неравенство* $X_1(L) \geq X_2(L) \geq 0$ *и при* $n \longrightarrow \infty$ *верно* $\mathbf{M}X_1(L) \approx \mathbf{M}X_2(L)$, $\mathbf{D}X_2(L)/(\mathbf{M}X_2(L))^2 \longrightarrow 0$. *Тогда для почти всех матриц* L *из* M_{mn}^k *имеет место соотношение* $X_2(L) \approx X_1(L) \approx \mathbf{M}X_2(L)$, $n \longrightarrow \infty$.

На множестве элементарных событий $M_{mn}^k = \{L\}$ рассмотрим случайные величины:

$$\xi_{w,\sigma}(L) = \begin{cases} 1, & \text{если набор столбцов с номерами из } w \text{ является } \sigma\text{-покрытием матрицы } L, \\ 0 & \text{в противном случае,} \end{cases}$$

и

$$\eta_{w,\sigma}(L) =$$

$$= \begin{cases} 1, & \text{если набор столбцов с номерами из } w \text{ является тупиковым } \sigma\text{-покрытием матрицы } L, \\ 0 & \text{в противном случае.} \end{cases}$$

Положим

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

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

Заметим, что $\eta_\varphi(L) = B_\varphi(L)$ и $\xi_\varphi(L) = C_\varphi(L)$.

Пусть $\mathbf{P}(\xi_{w,\sigma}(L) = 1)$ – вероятность того, что $\xi_{w,\sigma}(L) = 1$, $\mathbf{P}(\eta_{w,\sigma}(L) = 1)$ – вероятность того, что $\eta_{w,\sigma}(L) = 1$.

Очевидно, что

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

Согласно формуле включений и исключений имеем

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

Отсюда при $r \leq r_1$ имеем

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

Лемма 2. *Имеет место неравенство*

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

Доказательство. Верхняя оценка величины $\mathbf{M}\eta_{r_1}$ очевидна. Покажем справедливость нижней оценки. Согласно (5.1) и (5.2) имеем

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

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

Следовательно,

$$\mathbf{M}\xi_r - \mathbf{M}\eta_r = C_n^r k^r \sum_{i=0}^{r} (-1)^{i-1} C_r^i [1-(i+1)k^{-r}]^m \le C_n^r k^r r (1-2k^{-r})^m \le r(1-k^{-r})^m \mathbf{M}\xi_r.$$

Полагая $r = r_1$, получаем

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

Лемма доказана.

Лемма 3. *При* $m < k^{n^\beta}$, $\beta < 1$, *имеет место соотношение*

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

Доказательство. Имеем

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

Положим $A_r = C_n^r k^r (1 - k^{-r})^m$. Пусть $r \le r_1$. Тогда

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

Следовательно,

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

Из леммы 2, оценки (5.3) и неравенства $\mathbf{M}\eta_{r_1} \le \mathbf{M}\eta_\varphi \le \mathbf{M}\xi_\varphi$ следует утверждение доказываемой леммы.

Лемма 4. *При* $m < k^{n^\beta}$, $\beta < 1/2$, *имеет место соотношение*

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

Доказательство. Пусть $r = r_1$. Имеем

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

Очевидно, что

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

Пусть наборы w_1 и w_2 пересекаются по $r - t$ элементам. Тогда

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

Следовательно,

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

Так как

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

то

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

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

При $t \le t_0$ имеем

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

При $t > t_0$ имеем

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

Отметим, что

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

Из (5.4)–(5.6) следует

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

Пользуясь формулой Стирлинга, нетрудно показать, что

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

Так как $\sqrt{n/n-r} \longrightarrow 1$ при $m < k^{n^\beta}$, $\beta < 1$, $n \longrightarrow \infty$, то

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

Оценим

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

Из (5.7)–(5.9) следует

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

Таким образом,

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

Имеем

$$\begin{aligned}[\mathbf{M}\xi_r(L)]^2 &= (C_n^r k^r)^2 (k^r-1)^{2m} \frac{1}{k^{2mr}} = (C_n^r k^r)^2 (k^{2r} - 2k^r + 1)^m \frac{1}{k^{2mr}} = \\ &= (C_n^r k^r)^2 k^{rm} \left(k^r - 2 + \frac{1}{k^r}\right)^m \frac{1}{k^{1mr}} \ge (C_n^r k^r)^2 (k^r-2)^m \frac{1}{k^{1mr}} \ge (C_n^r k^r)^2 (k^r-2)^m k^{-mr}.\end{aligned}$$

Утверждение леммы доказано.

Лемма 5. *При* $m < k^{n^\beta}$, $\beta < 1/2$, *справедливо соотношение*

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

Доказательство. Имеем

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

Лемма доказана.

Из лемм 1, 3–5 сразу следует утверждение теоремы 1.

Замечание. При доказательстве лемм 2 и 4 использована техника доказательств соответственно утверждения 5 и леммы 3 из работы [13], в которой рассматривалась задача оценки числа максимальных интервалов частичной булевой функции и получена асимптотика логарифма числа максимальных интервалов.

СПИСОК ЛИТЕРАТУРЫ

1. *Дюкова Е.В.* О сложности реализации некоторых процедур распознавания // Ж. вычисл. матем. и матем. физ. 1987. Т. 27. № 1. С. 114–127.
2. *Дюкова Е.В.* Алгоритмы распознавания типа “Кора”: сложность реализации и метрические свойства // Распознавание, классификация, прогноз (матем. методы и их применение). М.: Наука, 1989. Вып. 2. С. 99–125.
3. *Вайнцвайг М.Н.* Алгоритм обучения распознаванию образов “Кора” // Алгоритмы обучения распознаванию образов. М.: Сов. радио, 1973. С. 82–91.
4. *Баскакова Л.В., Журавлёв Ю.И.* Модель распознающих алгоритмов с представительными наборами и системами опорных множеств // Ж. вычисл. матем. и матем. физ. 1981. Т. 21. № 5. С. 1264–1275.
5. *Дюкова Е.В., Инякин А.С.* Задача таксономии и тупиковые покрытия целочисленной матрицы // Сообщ. по прикл. матем. М.: ВЦ РАН, 2001.
6. *Дюкова Е.В., Журавлёв Ю.И.* Дискретный анализ признаковых описаний в задачах распознавания большой размерности // Ж. вычисл. матем. и матем. физ. 2000. Т. 40. № 8. С. 1264–1278.
7. *Djukova E.V., Zhuravlev Yu.L.* Diskrete methods of information analysis and algorithm synthesis // J. Pattern Recognition and Image Analys. 1997. V. 7. № 2. P. 192–207.
8. *Яблонский С.В.* Введение в дискретную математику // М.: Наука, 1986.
9. *Дюкова Е. В.* Асимптотически оптимальные тестовые алгоритмы в задачах распознавания // Пробл. кибернетики. Вып. 39. М.: Наука, 1982. С. 165–199.
10. *Дюкова Е.В.* Асимптотические оценки некоторых характеристик множества представительных наборов и задача об устойчивости // Ж. вычисл. матем. и матем. физ. 1995. Т. 35. № 1. С. 122–134.
11. *Дюкова Е.В., Песков Н.В.* Поиск информативных фрагментов описаний объектов в дискретных процедурах распознавания // Ж. вычисл. матем. и матем. физ. 2002. Т. 42. № 5. С. 741–753.
12. *Носков В. Н., Слепян В. А.* О числе тупиковых тестов для одного класса таблиц // Кибернетика. 1972. № 1. С. 60–65.
13. *Сапоженко А.А.* Оценка длины и числа тупиковых д. н. ф. для почти всех не всюду определенных булевых функций // Матем. заметки. 1980. Т. 28. № 2. С. 279–300.