ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2024, № 4, с. 33–39

КОМПЬЮТЕРНЫЕ МЕТОДЫ

УДК 519.7

ЛОГИЧЕСКАЯ КЛАССИФИКАЦИЯ НА ОСНОВЕ ПОИСКА ПРАВИЛЬНЫХ ПРЕДСТАВИТЕЛЬНЫХ ЭЛЕМЕНТАРНЫХ КЛАССИФИКАТОРОВ

© 2024 г. Н. А. Драгунов[a], *, Е. В. Дюкова[a], А. П. Дюкова[a]

[a]ФИЦ ИУ РАН, Москва, Россия

**e-mail: nikitadragunovjob@gmail.com*

Поступила в редакцию 29.01.2024 г.

После доработки 19.03.2024 г.

Принята к публикации 13.05.2024 г.

Рассмотрен подход к задаче классификации по прецедентам, базирующийся на применении аппарата дискретной математики (логических методов анализа данных). Исследована возможность сокращения временных затрат на стадии обучения корректного логического классификатора. Предложены новые модели классификаторов, основанные на поиске в описаниях прецедентов часто встречающихся фрагментов специального вида, названных правильными элементарными классификаторами. Описания моделей классификаторов даны с использованием понятий теории логических функций. Для построения искомых фрагментов авторами разработан и реализован оригинальный алгоритм. Эффективность предлагаемых моделей классификаторов обоснована экспериментально и подтверждена теоретическими оценками сложности их обучения. Получена верхняя асимптотическая оценка типичного числа правильных элементарных классификаторов.

Ключевые слова: задача классификации по прецедентам, корректная классификация, представительный элементарный классификатор, правильный элементарный классификатор, тупиковое покрытие целочисленной матрицы.

DOI: 10.31857/S0002338824040027 **EDN:** UENRUE

LOGICAL CLASSIFICATION BASED ON FINDING REGULAR REPRESENTATIVE ELEMENTARY CLASSIFIERS

N. Dragunov[a], *, E. Djukova[a], A. Djukova[a]

[a]Federal Research Center «Computer Science and Control» of the Russian Academy of Sciences, Moscow, Russia

**e-mail: nikitadragunovjob@gmail.com*

An approach to the supervised classification problem based on the apparatus of discrete mathematics (logical methods of data analysis) is considered. The possibility of time costs reducing at the stage of correct logical classifier training is investigated. New models of classifiers are proposed. These models are based on finding frequently occurring fragments of a special type in the descriptions of precedents — regular elementary classifiers. Descriptions of classifier models are given using the concepts of logical functions theory. To construct sought fragments, the authors have developed and implemented an original algorithm. The effectiveness of proposed classifier models has been experimentally substantiated and confirmed by theoretical estimates of their training complexity. An upper asymptotic estimate of the typical number of regular elementary classifiers is obtained.

Keywords: supervised classification problem, correct classification, representative elementary classifier, regular elementary classifier, irredundant covering of an integer matrix.

Введение. Задача классификации по прецедентам является одной из основных задач интеллектуального анализа данных и формулируется следующим образом. Исследуется некоторое множество объектов M, описываемых в системе числовых признаков $x_1,\ldots,x_n$. Известно, что M редставимо в виде объединения l подмножеств $K_1,\ldots,K_l$, называемых классами. Дан набор объектов из M, о которых известно, каким классам они принадлежат. Это прецеденты или

обучающие объекты. Требуется на базе анализа множества прецедентов построить алгоритм, определяющий класс любого объекта из M.

Дискретный или логический подход к задаче классификации предполагает, что каждый признак имеет ограниченное число допустимых значений, каждое из которых кодируется целым числом. Рассматриваемый подход имеет целью построение корректных моделей классификаторов, обеспечивающих безошибочное распознавание прецедентов.

Одними из известных направлений логической классификации являются LAD (logical analysis of data) и CVP (correct voting procedures). Каждое из направлений базируется на поиске таких фрагментов описаний прецедентов, которые позволяют отличать прецеденты из разных классов. В LAD искомые фрагменты называют логическими закономерностями, а в CVP — представительными элементарными классификаторами. Различным образом определяется понятие информативности фрагмента. В первом случае ищутся «максимальные» логические закономерности и решается сложная в вычислительном плане оптимизационная задача линейного программирования. Во втором случае ищутся «тупиковые» (в некотором смысле минимальные) представительные элементарные классификаторы, при этом возникают труднорешаемые дискретные перечислительные задачи. Направление LAD предложено в [1] и в основном развивается за рубежом. В России это направление представлено работами [2, 3]. Для направления CVP основополагающими являются публикации отечественных ученых [4–10].

Логические классификаторы наиболее эффективны в случае целочисленной информации низкой значности. Их описание может быть дано с использованием аппарата функций k-значной логики ($k \geq 2$). Тогда представительный элементарный классификатор (логическая закономерность) класса K является элементарной конъюнкцией над переменными $x_1,\ldots,x_n$, принимающей значение 1 на описании хотя бы одного прецедента из класса K и значения 0 на описаниях всех прецедентов из других классов [6].

Поиск тупиковых представительных элементарных классификаторов класса K основан, как правило, на первоначальном анализе множества прецедентов из других классов и сводится к решению сложной перечислительной задачи, называемой монотонной дуализацией, или к обобщениям этой задачи [6, 8]. Фактически сначала строятся элементарные конъюнкции над переменными $x_1,\ldots,x_n$, принимающие значение 0 на описаниях тех прецедентов, которые не принадлежат классу K, и теряющие это свойство при удалении хотя бы одного сомножителя. Затем из найденных конъюнкций отбираются те, которые не менее p ($p \geq 1$) раз принимают значение 1 на описаниях прецедентов класса K, т.е. отбираются тупиковые p-представительные элементарные классификаторы класса K (здесь p-настраиваемый параметр). В данной модели классификатора, обозначаемой далее A_0, вычисление оценки принадлежности распознаваемого объекта классу K существляется на основе проведения классической процедуры «голосования» [2], в которой участвуют все отобранные элементарные классификаторы.

В настоящей работе предлагаются и исследуются модели A_1, A_2, A_3 корректных логических классификаторов, обучение которых осуществляется путем поиска для каждого класса K так называемых правильных p-представительных элементарных классификаторов, т.е. таких представительных элементарных классификаторов этого класса, которые имеют ранг p ($p \geq 1$) и не менее p раз принимают значение 1 на описаниях прецедентов класса K. При этом классификатор A_1 действует по схеме классификатора A_0, но в голосовании участвуют только те тупиковые p-представительные элементарные классификаторы класса K, которые имеют ранг p. Классификаторы A_2 и A_3 действуют по иной схеме. Первоначально анализируются описания прецедентов класса K и строятся элементарные конъюнкции, которые не менее p раз принимают значение 1 на описаниях прецедентов этого класса и имеют ранг p. Такие конъюнкции называются правильными элементарными классификаторами. Затем просматриваются прецеденты из других классов и в A_2 из найденных конъюнкций отбираются представительные элементарные классификаторы класса K, а в A_3 отбираются тупиковые представительные элементарные классификаторы класса K. Процедура вычисления оценки принадлежности распознаваемого объекта классу K такая же, как и в алгоритме A_0.

Таким образом, на этапе обучения модель A_1 решает задачу монотонной дуализации, а модели A_2 и A_3 осуществляют поиск правильных элементарных классификаторов, базирующийся на предложенном в работе оригинальном алгоритме. Идея применения методов поиска часто встречающихся фрагментов в данных на этапе обучения логического классификатора была анонсирована авторами в [11].

Экспериментальное сравнение рассматриваемых алгоритмов на реальных и случайных модельных данных свидетельствует о целесообразности (в плане сокращения временных затрат) предлагаемого подхода к построению логических классификаторов. Получены теоретические результаты, характеризующие сложность обучения классификаторов A_2 и A_3 для случая, когда число

прецедентов класса K существенно больше числа признаков n. В экспериментах значение параметра p выбиралось согласно оценке типичного ранга правильного элементарного классификатора.

1. Основные понятия. Описание классификаторов A_1, A_2 и A_3. Пусть E_k^n, $k \geq 2$ – множество наборов вида $(\alpha_1,\ldots,\alpha_n)$, где $\alpha_i \in \{0,1,\ldots,k-1\}$.

Элементарной конъюнкцией над переменными $x_1,\ldots,x_n$ называется функция вида $x_{j_1}^{\sigma_1} \& \ldots \& x_{j_r}^{\sigma_r}$, где $\sigma_i \in \{0,1,\ldots k-1\}$, $x_{j_i} \in \{x_1,\ldots,x_n\}$ при $i = \overline{1,r}$, и при $r \geq 2$ выполнено $x_{j_q} \neq x_{j_t}$, $t = \overline{1,r}$, $q = \overline{1,r}$, $t \neq q$. Для краткости знак & опускается. Конъюнкция $B = x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$ обращается в 1 на тех наборах $(\alpha_1,\ldots,\alpha_n)$ из E_k^n, в которых $\alpha_{j_i} = \sigma_i$, $i = \overline{1,r}$. Множество наборов из E_k^n, на которых B принимает значение 1, обозначается через N_B, а через $\mathcal{B}(n,k)$ – множество всех элементарных конъюнкций рассматриваемого вида. Не ограничивая общности, можно считать, что объекты из исследуемого множества M описаны признаками, каждый из которых принимает значения из множества $\{0,1,\ldots,k-1\}$.

Пусть $K \in \{K_1,\ldots K_l\}$. Зададим на множестве прецедентов двузначную частичную (не всюду определенную) функцию $f_K(x_1,\ldots,x_n)$, которая принимает значение 1 на наборах, являющихся описаниями прецедентов класса K, и значение 0 на наборах, описывающих остальные обучающие объекты. Функция $f_K(x_1,\ldots,x_n)$ называется характеристической функцией класса K. Решение задачи классификации заключается в доопределении f_K на наборах, не входящих в обучающую выборку.

Далее U_K и Z_K обозначают соответственно множества прецедентов, на которых функция f_K равна 1 и 0. Положим $|U_K| = m_1$, $|Z_K| = m_2$, $1 \leq p \leq m_1$ (здесь и далее $|W|$ – мощность множества W).

Элементарным классификатором (*ЭК*) *ранга r* называется элементарная конъюнкция из $\mathcal{B}(n,k)$, зависящая от r переменных. ЭК B называется покрытием для Z_K, если $N_B \cap Z_K = \varnothing$. ЭК B, являющийся покрытием для Z_K, называется тупиковым покрытием для Z_K, если не существует покрытия B' для Z_K, такого, что $N_B \subset N_{B'}$.

Пусть $p \in \{1,2\ldots,m_1\}$. ЭК B называется p-частым в U_K, если $|N_B \cap U_K| \geq p$. ЭК B называется *p-представительным* для класса K, если B – p-частый в U_K и B – покрытие для Z_K. ЭК B называется *тупиковым p-представительным* для класса K, если B – p-частый в U_K и B – тупиковое покрытие для Z_K.

ЭК B ранга p называется *правильным* для U_K, если B – p-частый в U_K. ЭК B ранга p называется *правильным p-представительным* для класса K, если B – p-частый в U_K и B – покрытие для Z_K.

Приведем подробное описание моделей корректных классификаторов A_1, A_2 и A_3, о которых говорилось во Введении. Пусть $T_1(p, K)$ – множество всех тупиковых правильных p-представительных ЭК для класса K; $T_2(p, K)$ – множество всех правильных p-представительных ЭК для класса K; $T_3(p, K) = T_1(p, K)$; P_B^i, $B \in \mathrm{T}_i(p,K)$, $i \in \{1,2,3\}$, – число объектов S в U_K, таких, что $S \in N_B$.

На стадии обучения классификатор A_i, $i \in \{1,2,3\}$, строит некоторое множество ЭК из $\mathrm{T}_i(p,K)$. На следующей стадии (стадии распознавания) каждый найденный ЭК B участвует в процедуре голосования, заключающейся в вычислении величин P_B^i и $\Omega(B,S)$, где S – распознаваемый объект и $\Omega(B, S) = 1$, если $S \in N_B$, иначе $\Omega(B, S) = 0$. В результате получается оценка $\Gamma_i(S,K)$ принадлежности объекта S классу K, имеющая вид

$$\Gamma_i(S,K) = \frac{1}{|T_i(p,K)|} \sum_{B \in T_i(p,K)} P_B^i \ \Omega(B,S).$$

Объект S относится к классу с наибольшей оценкой. Если таких классов несколько, то объект относится к классу с наибольшим числом прецедентов.

В модели A_1 множество $T_1(p,K)$ строится в два этапа. Сначала анализируется множество Z_K и строятся тупиковые покрытия для Z_K ранга p. При этом решается задача монотонной дуализациии, которая относится к труднорешаемым дискретным перечислительным задачам. Затем из найденных ЭК отбираются те, которые являются p-частыми в U_K. Основная вычислительная сложность в этой модели заключается в необходимости решать задачу монотонной дуализации. Эффективность перечислительных задач принято оценивать сложностью нахождения нового решения (сложностью одного шага). В настоящее время для монотонной дуализации не построен алгоритм с полиномиальным шагом (алгоритм с полиномиальной задержкой [12]). Наиболее

эффективными в практическом отношении для этой задачи являются асимптотически оптимальные алгоритмы [10].

В моделях A_2 и A_3 множества $\mathrm{T}_2(p,K)$ и $\mathrm{T}_3(p,K)$ строятся также в два этапа. Однако, в отличие от модели A_1, сначала вместо анализа множества Z_K проводится анализ множества U_K, которое обычно меньше по мощности, чем Z_K, в случае, если число классов больше двух. В результате такого анализа строится множество правильных ЭК для U_K ранга p. На втором этапе в моделях A_2 и A_3 из найденных ЭК отбираются соответственно покрытия для Z_K и тупиковые покрытия для Z_K.

В настоящей работе при реализации классификаторов A_1, A_2 и A_3 к исходным данным применялась известная процедура one-hot кодирования [9]. В результате классификаторы работали с бинарными описаниями объектов. Для поиска правильных ЭК в бинарных данных разработан алгоритм ADR, описание которого приведено в разд. 2.

2. Алгоритм ADR поиска правильных ЭК. Типичное число правильных ЭК. Обозначим через L матрицу, строками которой являются бинарные описания объектов класса K, полученные с помощью one-hot кодирования.

Пусть Q – набор различных столбцов матрицы L, L^Q – подматрица матрицы L образованная набором Q. Набор столбцов Q называется p-частым, если L^Q содержит не менее p строк, все элементы которых равны 1. Набор столбцов Q называется p-правильным, если он p-частый и его мощность равна p. Несложно видеть, что поиск всех правильных ЭК ранга p эквивалентен поиску всех p-правильных наборов столбцов матрицы L.

Обозначим через $R(L,p)$ множество всех столбцов матрицы L, имеющих не менее p элементов, равных 1. Пронумеруем столбцы матрицы L слева направо, начиная с 1. Пусть $e_1(R)$ и $e_2(R)$ – столбцы соответственно с наименьшим номером и наибольшим номером из R, $R \subseteq R(L,p)$. Через $U_p(L)$ обозначим множество всех p-частых наборов столбцов матрицы L, мощность которых не превосходит p. Алгоритм ADR строит множество всех p-правильных наборов столбцов матрицы L, перечисляя с полиномиальной задержкой наборы из $U_p(L)$.

Определим порядок, в котором происходит перечисление наборов из $U_p(L)$. На первом шаге рассматривается набор $Q = \{e_1(R(L,p))\}$.

Пусть на шаге $i\,(i \geq 1)$ построен набор $Q \in U_p(L)$, состоящий из столбцов с номерами $j_1,\ldots,j_r$, $j_1 < \ldots < j_r$, $r \leq p$. Если $Q = \{e_2(R(L,p))\}$, то алгоритм заканчивает работу. Если же $Q \neq \{e_2(R(L,p))\}$, то на шаге $i+1$ алгоритм ADR строит новый набор ΔQ из $U_p(L)$. При этом возможны два случая: $r < p$ и $r = p$. В первом случае алгоритм строит ΔQ согласно приведенным ниже правилам $1-4$. Во втором случае алгоритм строит ΔQ по правилам $2-4$.

Для описания правил построения ΔQ введем обозначения: Q_t, $t = \overline{1,r}$, – набор столбцов матрицы L с номерами $j_1,\ldots,j_t$; R_t, $t = \overline{1,r}$, – множество столбцов в $R(L,p)$, номера которых больше j_t; G_t, $t = \overline{1,r}$, $r < p$, – множество столбцов из R_t, каждый из которых в объединении со столбцами из Q_t образует набор из $U_p(L)$. Положим $G_r = \varnothing$ в случае $r = p$.

Заметим, что в случае $r < p$ для построения G_t в L нужно оставить только те столбцы, номера которых больше j_t и которые имеют не менее p элементов, равных 1 в подматрице, полученной после удаления из L строк, дающих 0 в пересечении со столбцами с номерами $j_1,\ldots,j_t$.

Положим $Q_0 = \varnothing$ и $G_0 = \varnothing$. Перечислим возможные случаи и в каждом из них укажем правила построения ΔQ:

1) $G_r \neq \varnothing : \Delta Q = Q_r \cup \{e_1(G_r)\}$;

2) $G_r = \varnothing,\ G_{r-1} \cap R_r \neq \varnothing : \Delta Q = Q_{r-1} \cup \{e_1(G_{r-1} \cap R_r)\}$;

3) $G_r = \varnothing,\ G_{r-1} \cap R_r = \varnothing,\ r = 1 : \Delta Q = \{e_1(R_r)\}$;

4) $G_r = \varnothing,\ G_{r-1} \cap R_r = \varnothing,\ r > 1 : \Delta Q = Q_{r-2} \cup \{e_1(G_{r-2} \cap R_{r-1})\}$.

Заметим, что $R_r \neq \varnothing$ при $r = 1$, так как $Q \neq \{e_2(R(L,p))\}$, и $G_{r-2} \cap R_{r-1} \neq \varnothing$ при $\mathrm{T}_2(p,K)$, так как столбец с номером j_r принадлежит этому множеству.

Из описания работы алгоритма ADR видно, что в его основе лежит процесс ветвления, который удобно представить в виде обхода дерева решений в глубину. Вершинами этого дерева являются наборы из $U_p(L)$, причем p-правильные наборы столбцов находятся среди висячих вершин. Через L_K обозначим матрицу, строками которой являются описания прецедентов класса K. Правильные ЭК порождаются квадратными подматрицами матрицы L_K, состоящими из одинаковых строк. Такие подматрицы назовем правильными.

Ниже приведены асимптотические оценки типичных значений числа правильных подматриц целочисленной матрицы L_K и порядка такой подматрицы в случае большого числа строк матрицы L_K. Пусть M_{mn}^k – множество всех целочисленных матриц размера $m \times n$ с элементами из $\{0,1,\ldots,k-1\}$; $S(L)$, $L \in M_{mn}^k$, – множество правильных подматриц в матрице L; $\phi_k(m,n)$ –

интервал $(0,\ r(k,m,n))$, где $r(k,m,n) = 0.5\log_k mn - 0.5\log_k\log_k mn + \log_k\log_k\log_k n$; $b_n \sim c_n$, $n\to\infty$ означает, что $lim_{n\to\infty} b_n / c_n = 1$.

Т е о р е м а. Если $n^\alpha \le m \le k^n$, $\alpha > 1$, то при $n\to\infty$ для почти всех матриц L из M^k_{mn} справедливо

$$|S(L)| \sim \sum_{r\in\phi_k(m,n)} C^r_n C^r_m k^{r-r^2}.$$

и порядки почти всех подматриц из $S(L)$ принадлежат интервалу $\phi_k(m,n)$.

Доказательство теоремы аналогично доказательству теоремы 3 из [13], в которой при тех же ограничениях на m и n получена асимптотическая оценка типичного числа так называемых σ -подматриц матрицы L, служащая верхней оценкой числа тупиковых покрытий для Z_K при условии, что $|Z_K| = m$.

Приведенная в теореме оценка типичного порядка подматрицы из $S(L)$ косвенно свидетельствуют о том, что в случае, когда число прецедентов m_1 класса K существенно больше числа признаков n, типичный ранг правильного ЭК в U_K не превосходит $r(k,m_1,n)$.

З а м е ч а н и е 1. В работе [14] получены асимптотические оценки типичного числа правильных ЭК в U_K для двух случаев: 1) $m_1{}^a \le n \le k^{m_1\beta}$, $a>1$, $\beta < 1$; 2) $n \le m_1 \le k^{n^\beta}$, $\beta < 1/2$. Авторами показано, что типичный ранг правильного ЭК в U_K в случаях 1) и 2) соответственно принадлежит интервалу $\phi_k(m_1,n)$ и не превосходит $\log_k m_1 + \log_k\log_k m_1$.

3. Результаты экспериментов. Результаты счета на реальных целочисленных задачах приведены в таблице. Задачи взяты из репозитория UCI [archive.ics.uci.edu] и репозитория ВЦ ФИЦ ИУ РАН. Описанные выше алгоритмы A_1, A_2, A_3 оценивались по качеству классификации и по времени обучения. Алгоритмы реализованы на языке программирования C++. В тестировании на качество классификации также участвовали такие известные алгоритмы, как случайный лес (RF) и логистическая регрессия (LR). Дополнительная настройка алгоритмов RF и LR не производилась.

Результаты счета усреднялись по 10 случайным независимым разбиениям прецедентов, 80% которых использовалось для обучения моделей, а 20% — для оценки качества классификации. В каждом из разбиений распределение прецедентов по классам сохранялось неизменным.

Таблица 1.

m, n_1, l	Время, мс			Качество			
$(p_1, \dots, p_1)$	A_1	A_2	A_3	A_1, A_3	A_2	RF	LR
144, 379, 2 (3, 3)	512.1	**47.0**	48.6	0.691	0.735	0.742	**0.774**
267, 566, 2 (3, 4)	289.2	**18.3**	18.4	0.560	0.570	0.545	**0.578**
957, 27, 2 (3, 3)	71.7	**1.0**	**1.0**	**0.976**	**0.976**	0.939	0.639
79, 160, 2 (2, 3)	238.4	**140.0**	150.0	0.614	**0.623**	0.542	0.553
3195, 73, 2 (4, 4)	5294.0	**903.7**	1061.9	0.903	0.974	**0.988**	0.956
1532, 284, 2 (5, 5)	2763106	**59265**	69387	0.960	**0.971**	0.960	0.922
2056, 83, 3 (4, 4, 4)	35471	**8.3**	9.4	0.641	0.770	**0.905**	0.790
3190, 287, 3 (5, 5, 5)	10487213	**235045**	315275	0.793	0.794	**0.946**	0.831

В таблице последовательно указаны результаты счета для следующих задач: Манелис, Остеосаркома, Крестики-нолики (UCI), Инсульт, Шахматы (UCI), Молекулярная Биология 1 (UCI), Задача 5, Молекулярная Биология 2 (UCI). Для каждой задачи указаны число прецедентов m, число признаков n_1 полученное после one-hot перекодировки, число классов l и ранг p_i,

$i \in \{1,2,\ldots,l\}$, голосующих ЭК класса K_i. Время работы алгоритмов указано в миллисекундах. Функционалом качества выбрана сбалансированная точность классификации, вычисляемая по формуле

$$\psi = \sum_{i=1}^{l} q_i \,/\, l,$$

где q_i — доля верно классифицированных объектов класса K_i. Данный функционал хорошо себя зарекомендовал при несбалансированных классах. В случае равномощных классов сбалансированная точность совпадает с долей верно классифицированных объектов.

Как видно из таблицы, модель A_2 превосходит по качеству и времени работы модели A_1 и A_3 на всех рассмотренных данных, кроме задачи Крестики-нолики, и в среднем не уступает по качеству ни случайному лесу, ни логистической регрессии. На трех задачах (Крестики-нолики, Инсульт, Молекулярная Биология 1) модель A_2 превосходит все модели.

Модель A_1 работает существенно медленнее модели A_3 при том, что оба алгоритма строят множество всех тупиковых p-представительных ЭК ранга p. Однако модель A_1 на первом этапе обучения ищет тупиковые покрытия для Z_K ранга p, а модель A_3 перечисляет правильные ЭК ранга p для U_K. Стоит отметить, что на шести задачах (Манелис, Остеосаркома, Крестики-нолики, Инсульт, Шахматы, Задача 5) модели A_2 и A_3 обучались менее чем за 1 с, что свидетельствует об их высокой вычислительной эффективности.

З а м е ч а н и е 2. В экспериментах ранг p_i, $i \in \{1,2,\ldots,l\}$, голосующих ЭК класса K_i брался равным числу $0.5\log_2 m_i n_1 - 0.5\log_2\log_2 m_i n_1 - \log_2\log_2\log_2 n_1$, где m_i — число прецедентов класса K_i. Обучение с таким рангом в среднем показывало лучшее качество по сравнению с обучением с другими значениями ранга p_i, также принадлежащими интервалу $\phi_2(m_i, n_1)$.

На рис. 1, 2 приведено время обучения моделей A_1 и A_2 на случайных модельных данных из равномерного распределения при $l = 2$, $k = 2$, m_1 — число прецедентов в каждом классе, n_1 — число признаков. Результаты счета усреднены по 20 независимым запускам. Время работы алгоритмов указано в секундах. Время счета модели A_3 не приводится на графиках, так как в рассматриваемых примерах оно практически совпадает с временем работы A_2.

На рис. 1 показан экспоненциальный рост временных затрат на этапе обучения классификаторов A_1 и A_2 при $m_1 = 250$ в зависимости от числа признаков n_1. Видно, что при относительно небольшом n_1 разрыв во времени счета для A_1 и A_2 незначителен. При $n_1 \geq 150$ алгоритм A_1 работает значительно медленнее алгоритма A_2. Например, A_1 обучается примерно в 1.3 раза медленнее A_2 при $n_1 = 150$, а при $n_1 = 250$ — в 1.7 раз медленнее.

На рис. 2 продемонстрирован линейный рост временных затрат на этапе обучения классификаторов A_1 и A_2 при $n_1 = 100$ в зависимости от числа прецедентов m_1. Видно, что время работы A_1 растет быстрее по сравнению с временем работы A_2. Например, A_1 обучается примерно в 1.2 раза медленнее A_2 при $m_1 = 100$ и почти в 2 раз медленнее при $m_1 = 700$.

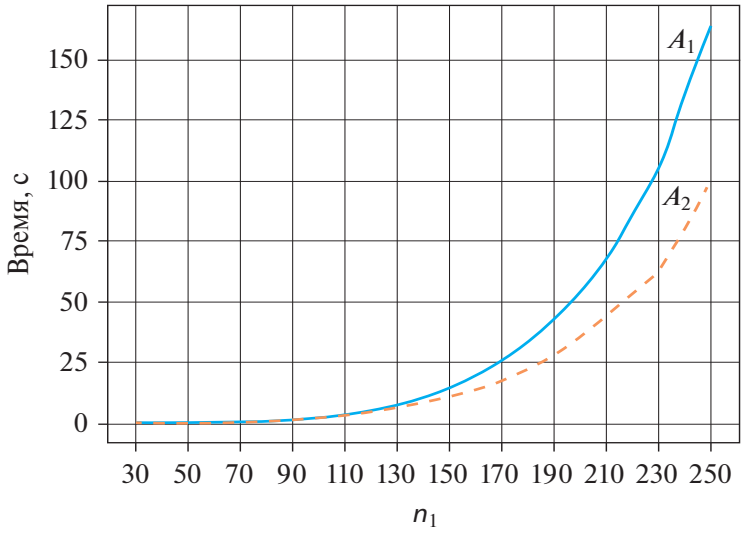

Рис. 1. Зависимость времени обучения моделей A_1 и A_2 от числа признаков при $m_1 = 250$

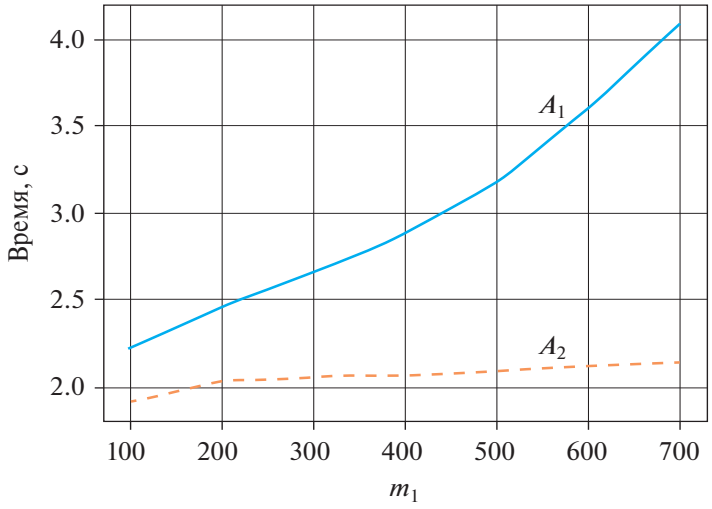

Рис. 2. Зависимость времени обучения моделей A_1 и A_2 от числа прецедентов при $n_1 = 100$

Заключение. Исследованы актуальные вопросы снижения временных затрат, возникающие при логическом анализе данных в задачах классификации на основе прецедентов. Предложены новые модели корректного голосования, базирующиеся на поиске в описаниях прецедентов каждого класса правильных ЭК ранга p (p – настраиваемый параметр модели). Разработан эффективный алгоритм для перечисления искомых правильных ЭК. Получена верхняя асимптотическая оценка типичного числа правильных ЭК для случая, когда число прецедентов существенно больше числа признаков. При этом указан типичный ранг правильного ЭК, который использован в экспериментах для выбора параметра p. Теоретические выводы подтверждены результатами экспериментального исследования на реальных и случайных модельных данных. А именно показано, что время обучения модели A_1, базирующейся на решении задачи монотонной дуализации, растет быстрее времени обучения модели A_2, основанной поиске правильных ЭК.

СПИСОК ЛИТЕРАТУРЫ

1. *Crama Y., Hammer P.L., Ibaraki T.* Cause-effect Relationships and Partially Defined Boolean Functions // Ann. Oper. Res. 1988. V. 16. Iss. 1. P. 299–325.
2. *Журавлёв Ю.И., Рязанов В.В., Сенько О.В.* Распознавание. Математические методы. Программная система. Практические применения. М.: ФАЗИС, 2006. 159 с.
3. *Масич И.С.* Метод оптимальных логических решающих правил для задач распознавания и прогнозирования // Системы управления и информационные технологии. 2019. Т. 75. № 1. С. 31–37.
4. *Бонгард М.М., Вайнцвайг М.Н., Губерман Ш.А., Извекова М.Л., Смирнов М.С.* Использование обучающейся программы для выявления нефтеносных пластов // Геология и геофизика. 1966. № 6.
5. *Баскакова Л.В., Журавлёв Ю.И.* Модель распознающих алгоритмов с представительными наборами и системами опорных множеств // ЖВМ и МФ. 1981. Т. 21. № 5. С. 1264–1275.
6. *Дюкова Е.В., Журавлёв Ю.И.* Дискретный анализ признаковых описаний в задачах распознавания большой размерности // ЖВМ и МФ. 2000. Т. 40. №8. С. 1264–1278.
7. *Яблонский С.В., Чегис И.А.* О тестах для электрических схем // УМН. 1955. Т. 10. Вып. 4(66). С. 182–184.
8. *Дюкова Е.В., Журавлёв Ю.И.* Задача монотонной дуализации и ее обобщения: асимптотические оценки числа решений // ЖВМ и МФ. 2018. Т. 58. № 12. С. 2153–2168.
9. *Дюкова Е.В., Инякин С.А.* Об асимптотически оптимальном построении тупиковых покрытий целочисленной матрицы // Математические вопросы кибернетики. 2008. № 17. С. 247–262.
10. *Дюкова Е.В., Прокофьев П.А.* Об асимптотически оптимальных алгоритмах дуализации // ЖВМ и МФ. 2015. Т. 55. № 5. С. 895–910.
11. *Dragunov N., Djukova E., Djukova. A.* Supervised Classification and Finding Frequent Elements in Data // 8th Intern. Conf. on Information Technology and Nanotechnology Proceedings. N.J.: IEEE, 2022. P. 5.
12. *Johnson D.S., Yannakakis M., Papadimitriou C.H.* On Generating All Maximal Independent Sets // Information Processing Letters. 1988. V. 27. Iss. 3.
13. *Дюкова Е.В., Песков Н.В.* Поиск информативных фрагментов описаний объектов в дискретных процедурах распознавания // ЖВМ и МФ. 2002. Т. 42. № 5. С. 741–753.
14. *Дюкова Е. В., Дюкова А. П.* О числе решений некоторых специальных задач логического анализа целочисленных данных // Изв. РАН. ТиСУ. 2023. № 5. С. 57–66.