

О метрических (количественных) свойствах логических классификаторов

Н.А. ДРАГУНОВ

Федеральный исследовательский центр «Информатика и управление» Российской академии наук, г. Москва, Россия

Аннотация. Рассматривается логический подход к задаче классификации по прецедентам. Предполагается, что прецеденты описаны в системе числовых признаков, каждый из которых имеет ограниченное число значений и кодируется целыми числами. Этап обучения классификатора заключается в поиске специальных фрагментов в признаковых описаниях прецедентов, позволяющих различать их из разных классов и называемых представительными элементарными классификаторами. Традиционно важным для рассматриваемого подхода является получение оценок количественных характеристик множества искомых элементарных классификаторов. В работе приведена верхняя асимптотическая оценка типичного числа так называемых правильных представительных элементарных классификаторов и оценка типичного «ранга» такого элементарного классификатора для важного в практическом отношении случая, когда число прецедентов существенно больше числа признаков.

Ключевые слова: классификация по прецедентам, правильный представительный элементарный классификатор, правильная подматрица целочисленной матрицы, асимптотические оценки типичного числа и порядка правильных подматриц.

DOI: 10.14357/20790279240402 **EDN:** KADXXE

Введение

Задача классификации по прецедентам является одной из наиболее актуальных решений интеллектуального анализа данных [1]. Важность задачи обусловлена большим числом приложений в различных прикладных областях. Существенным преимуществом логического подхода к классификации является возможность получения результатов в тех случаях, когда формирование большого числа прецедентов требует дорогостоящих и трудоемких работ или, в принципе, невозможно. Например, такая ситуация имеет место в задачах прогнозирования месторождений редких ископаемых и свойств твердых сплавов. Важной отличительной особенностью логических методов классификации является их высокая интерпретируемость. В настоящей работе рассматривается постановка задачи классификации по прецедентам.

Исследуется некоторое множество объектов M . Объекты из M описываются в системе числовых признаков $x_1, \dots, x_n$. Предполагается, что каждый признак x_i принимает значения из $E_k = \{0, 1, \dots, k-1\}$. Известно, что M представимо в виде объединения непересекающихся

подмножеств $K_1, \dots, K_l$ (классов). Даны примеры объектов из каждого класса, называемые прецедентами. Требуется на основе анализа множества прецедентов построить алгоритм, умеющий по признаковому описанию объекта из M определять его класс.

Одним из основных понятий логического подхода к классификации является представительный элементарный классификатор. Пусть $H = \{x_{j_1}, \dots, x_{j_r}\} \subseteq \{x_1, \dots, x_n\}$, $j_1 < \dots < j_r$ – набор из r различных признаков; $\sigma = (\sigma_1, \dots, \sigma_r)$, $\sigma_i \in E_k$, $i = \overline{1, r}$ – набор допустимых значений признаков из H . Пара (σ, H) называется элементарным классификатором (ЭК) ранга r . Объект $S \in M$ содержит ЭК (σ, H) , если подписание объекта S , порожденное признаками из H , совпадает с σ . Пусть $K \in \{K_1, \dots, K_l\}$, $\bar{K} = \{K_1, \dots, K_l\} \setminus \{K\}$. ЭК (σ, H) называется корректным для класса K если не существует пары прецедентов $S' \in K$, $S'' \in \bar{K}$ таких, что S' и S'' содержат (σ, H) . Корректные ЭК позволяют безошибочно классифицировать обучающие объекты. Корректный ЭК класса K называется представительным для класса K , если хотя бы один прецедент из K содержит этот ЭК.

Общая схема обучения логического классификатора заключается в построении для каждого класса K некоторого множества $P(K)$ представительных ЭК. Алгоритмы логической классификации отличаются искомыми множествами $P(K)$ и методами их построения. При поиске $P(K)$ решаются сложные в вычислительном плане задачи [2-4].

Этап распознавания логического классификатора, как правило, базируется на процедуре «голосования», в которой участвуют все найденные представительные ЭК из $P(K)$. Пусть S – распознаваемый объект; $W_{(\sigma, H)}^K$ – число прецедентов класса K , содержащих (σ, H) ; $\Omega(\sigma, H, S)$ – величина, равная 1, если S содержит (σ, H) , и равная 0 иначе. Для каждого класса K вычисляется оценка $\Gamma(S, K)$ принадлежности S рассматриваемому классу согласно формуле:

$$\Gamma(S, K) = \frac{1}{|P(K)|} \sum_{(\sigma, H) \in P(K)} W_{(\sigma, H)}^K \Omega(\sigma, H, S).$$

Объект S относится к классу с наибольшей оценкой. В случае, если таких классов несколько, то происходит отказ от распознавания.

Одним из ключевых направлений логической классификации являются процедуры корректного голосования (Correct Voting Procedures или CVP). Основополагающий вклад в развитие методов CVP сыграли работы С. В. Яблонского, в которых введено хорошо известное в дискретной математике понятие теста, а также труды Ю. И. Журавлева, М. М. Бонгарда и М. Н. Вайнцвайга. Позднее это направление активно развивалось как отечественными, так и зарубежными учеными. Особенно значительный вклад внесен научной школой академика РАН Ю.И. Журавлева.

В CVP наибольшее внимание уделяется построению тупиковых представительных ЭК класса K . Представительный для класса K ЭК (σ, H) называется тупиковым, если любой ЭК (σ', H') класса K такой, что $\sigma' \subset \sigma, H' \subset H$ не является представительным ЭК класса K . На этапе поиска тупиковых представительных ЭК возникает труднорешаемая перечислительная задача дискретной математики, называемая монотонной дуализацией [3, 4]. Дуализация может быть сформулирована как задача перечисления всех неприводимых покрытий булевой матрицы или как задача построения сокращенной ДНФ монотонной булевой функции по ее КНФ.

В [5, 6] предложен логический классификатор, основанный на поиске правильных представительных ЭК. Представительный ЭК (σ, H) класса K , имеющий ранг r , называется правильным, если не менее r прецедентов из K содержат (σ, H) . По-

строение правильных представительных ЭК для каждого класса K осуществляется в два шага. Сначала строятся правильные ЭК класса K . Такой ЭК имеет заданный ранг r и не менее r прецедентов из K его содержат. Затем из найденных ЭК отбираются те, которые не содержатся в прецедентах из $\bar{K}$. Поиск правильных ЭК класса K осуществляется на основе перечисления так называемых правильных подматриц матрицы, строками которой являются признаковые описания прецедентов класса K . Правильной подматрицей называется квадратная подматрица порядка r , состоящая из одинаковых строк. Для перечисления правильных подматриц в [6] предложен оригинальный алгоритм, базирующийся на методах поиска частых элементов в данных [7]. Рассматриваемая модель на практике работает быстрее классической модели голосования по тупиковым представительным ЭК, не уступает ей по качеству классификации [6] и хорошо себя зарекомендовала при решении важной задачи генной инженерии – задачи предсказания промоторов [8].

Традиционно важным для обоснования вычислительной эффективности логических классификаторов является получение асимптотических оценок типичного числа и оценок типичного ранга искомого ЭК. Для правильных ЭК такие оценки приведены в [9], где исследованы два случая, а именно, когда число признаков существенно больше числа прецедентов и когда число прецедентов не меньше числа признаков.

В настоящей работе получены асимптотическая оценка типичного числа правильных подматриц матрицы с элементами из E_k и оценка типичного порядка такой подматрицы в случае, когда число ее строк существенно больше числа столбцов. Полученная оценка числа правильных подматриц является верхней асимптотической оценкой типичного числа правильных представительных ЭК в случае, когда число прецедентов существенно больше числа признаков.

Проведено экспериментальное сравнение алгоритма голосования по правильным представительным ЭК с такими хорошо известными классификаторами как случайный лес и логистическая регрессия. При выборе ранга искомого ЭК (ранг – параметр модели) используется полученная в работе оценка типичного порядка правильной подматрицы.

1. Постановка задачи и основные результаты

Исследуется некоторое множество объектов M , описываемых в системе числовых признаков $x_1, \dots, x_n$. Не ограничивая общности можно счи-

тать, что каждый признак принимает значения из $E_k = \{0, 1, \dots, k-1\}$. Множество M представимо в виде объединения непересекающихся подмножеств $K_1, \dots, K_l$, называемых классами.

Пусть $K \in \{K_1, \dots, K_l\}$, $\bar{K} = \{K_1, \dots, K_l\} \setminus \{K\}$. Через $Q(K)$ обозначается совокупность прецедентов из K , $m = |Q(K)|$ (здесь и далее $|X|$ - число объектов в множестве X).

ЭК (σ, H) называется p -частым в $Q(K)$, если не менее p прецедентов из $Q(K)$ содержат (σ, H) . ЭК (σ, H) ранга p называется *правильным* для $Q(K)$, если (σ, H) - p -частый в $Q(K)$. ЭК (σ, H) называется *правильным представительным* для $Q(K)$, если (σ, H) - правильный для $Q(K)$ и не содержится ни в одном прецеденте из $Q(\bar{K})$.

Через L обозначим матрицу, строками которой являются описания объектов из $Q(K)$. Нетрудно видеть, что правильные ЭК порождаются квадратными подматрицами матрицы L , состоящими из одинаковых строк. Такие подматрицы назовём правильными. Ниже получены асимптотические оценки типичных значений числа правильных подматриц целочисленной матрицы L и оценки порядка такой подматрицы в случае большого числа строк матрицы L (Теорема 1). Полученная оценка типичного числа правильных подматриц является верхней асимптотической оценкой типичного числа правильных представительных ЭК в случае, когда число прецедентов из $Q(K)$ существенно больше числа признаков.

Пусть M_{mn}^k - множество всех матриц размера $m \times n$ с элементами из E_k ; W_r^i , $r \leq i$, - множество всех наборов вида $\{j_1, \dots, j_r\}$, где $j_t \in \{1, 2, \dots, i\}$ при $t = 1, 2, \dots, r$ и $j_t \neq j_u$ при $t \neq u$; $S(L)$, $L \in M_{mn}^k$, - множество правильных подматриц в матрице L ; $\phi_k(m, n)$ - интервал $(0, r(k, m, n))$, где $r(k, m, n) = 0.5 \log_k mn - 0.5 \log_k \log_k mn + \log_k \log_k \log_k n$; $b_n \sim c_n$, $n \rightarrow \infty$, означает, что $\lim_{n \rightarrow \infty} b_n/c_n = 1$; $b_n \leq_n c_n$ означает, что $b_n \leq c_n$ при всех достаточно больших n .

Выявление типичной ситуации связано с высказыванием вида «для почти всех матриц L из M_{mn}^k при $n \rightarrow \infty$ выполнено $F_1(L) \sim F_2(L)$ » (здесь $F_1(L)$ и $F_2(L)$ - два функционала, заданные на матрицах из M_{mn}^k). Данное высказывание означает, что существуют две положительные бесконечно малые функции $\alpha(n)$ и $\beta(n)$, такие, что для всех достаточно больших n имеет место:

$$1 - |M|/|M_{mn}^k| \leq \alpha(n),$$

где M - множество таких матриц L в M_{mn}^k , для которых

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

Теорема 1. Если $n^a \leq m \leq k^n$, $a > 1$, то для почти всех матриц L из M_{mn}^k справедливо

$$|S(L)| \sim \sum_{r \in \phi_k} C_n^r C_m^r k^{r-r^2}, \text{ при } n \rightarrow \infty,$$

и порядки почти всех правильных подматриц из $S(L)$ принадлежат интервалу ϕ_k .

Доказательство теоремы опирается на ряд приведенных Лемм 1-7.

Пусть $v \in W_r^m$, $w \in W_r^n$. Матрица $L \in M_{mn}^k$ называется (v, w) -правильной, если подматрица матрицы L , образованная строками с номерами из v и столбцами с номерами из w , является правильной. Через $M(v, w)$ обозначается множество всех (v, w) -правильных матриц из M_{mn}^k . Через $L[v, w]$ обозначается подматрица матрицы L с номерами строк из v и номерами столбцов из w .

Лемма 1. Если $v \in W_r^m$, $w \in W_r^n$, то

$$|M(v, w)| = k^{mn+r-r^2}.$$

Доказательство. Правильная подматрица $L[v, w]$ определяется своей первой строкой. Поэтому элементы правильной подматрицы $L[v, w]$ могут быть выбраны k^r способами. Остальные $mn - r^2$ элементов матрицы L могут быть выбраны k^{mn-r^2} способами.

Лемма 2. Если $v_1 \in W_r^m$, $w_1 \in W_r^n$, $v_2 \in W_l^m$, $w_2 \in W_l^n$ и наборы v_1, v_2 пересекаются по a ($a \geq 0$) элементам, a наборы w_1, w_2 пересекаются по b ($b \geq 0$) элементам, то

$$|M(v_1, w_1) \cap M(v_2, w_2)| = k^{mn+r-r^2+l-l^2+ab-a}.$$

Доказательство. Положим

$$M = M(v_1, w_1) \cap M(v_2, w_2).$$

Рассмотрим матрицу $L \in M$. Подматрицы $L[v_1, w_1]$ и $L[v_2, w_2]$ являются правильными и содержат r^2 и l^2 элементов, соответственно. Подматрицы $L[v_1, w_1]$ и $L[v_2, w_2]$ пересекаются по ab элементам. Правильные подматрицы $L[v_1, w_1]$ и $L[v_2, w_2]$ определяются своими первыми строками. Они могут быть выбраны k^{r+l-a} способами. Остальные $mn - r^2 - l^2 + ab$ элементы матрицы L могут быть выбраны $k^{mn-r^2-l^2+ab}$ способами.

Лемма 3. Если $n^a \leq m \leq k^n$, $a > 1$, то при $n \rightarrow \infty$ имеет место:

$$\sum_{r=1}^n C_n^r C_m^r k^{r-r^2} \sim \sum_{r \in \phi_k(m, n)} C_n^r C_m^r k^{r-r^2}.$$

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

Пусть $r \geq p + q - 1$, тогда

$$\frac{a_{r+1}}{a_r} \leq \frac{(m-r)(n-r)}{r^2} k^{-2r} \leq$$

$$\leq \frac{mn}{p^2} k^{-2p-2q+2} \leq_n k^{-2q+2}.$$

Следовательно,

$$\sum_{r \in [p+q, n]} a_r = o(\sum_{r \in \phi_k(m, n)} a_r) \text{ при } n \rightarrow \infty.$$

Лемма 4. Если $n^a \leq m$, $a > 1$ и $r < \log_k mn$, $l < \log_k mn$, то имеет место

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

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

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

Утверждение леммы совпадает с утверждением леммы 3.3.3 из работы [10] в случае $k = 2$ и доказывается аналогично.

Лемма 5 [11]. Пусть для случайных величин $X_1(L)$ и $X_2(L)$, определенных на M_{mn}^k , выполнено $X_1(L) \geq X_2(L) \geq 0$ и при $n \rightarrow \infty$ верно $MX_1(L) \sim MX_2(L)$, $DX_2(L)/(MX_2(L))^2 \rightarrow 0$. Тогда для почти всех матриц L из M_{mn} имеет место $X_1(L) \sim X_2(L) \sim MX_2(L)$, $n \rightarrow \infty$.

На множестве элементарных событий $M_{mn}^k = \{L\}$ зададим случайную величину $\zeta_{(v, w)}(L)$, принимающую значение 1, если $L \in M(v, w)$, и значение 0, иначе. Положим

$$\zeta_1(L) = \sum_{r=1}^n \sum_{v \in W_r^m, w \in W_r^n} \zeta_{(v, w)}(L),$$

$$\zeta_2(L) = \sum_{r \in \phi_k(m, n)} \sum_{v \in W_r^m, w \in W_r^n} \zeta_{(v, w)}(L).$$

Нетрудно видеть, что $|S(L)| = \zeta_1(L)$, а $\zeta_2(L)$ — это число таких правильных подматриц в $S(L)$, порядки которых принадлежат интервалу $\phi_k(m, n)$.

Через $P(\zeta_{(v, w)}(L) = 1)$ обозначается вероятность события $\zeta_{(v, w)}(L) = 1$. Очевидно, в силу Леммы 1

$$P(\zeta_{(v, w)}(L) = 1) = \frac{|M(v, w)|}{|M_{mn}|} =$$

$$= \frac{k^{mn+r-r^2}}{k^{mn}} = k^{r-r^2}.$$

Лемма 6. Если $n^a \leq m \leq k^n$, $a > 1$, то имеет место

$$M\zeta_1(L) \sim M\zeta_2(L) \sim \sum_{r \in \phi_k(m, n)} C_n^r C_m^r k^{r-r^2}, \text{ при } n \rightarrow \infty.$$

Доказательство. Действительно,

$$M\zeta_1(L) = \sum_{r=1}^n \sum_{v \in W_r^m, w \in W_r^n} P(\zeta_{(v, w)}(L) = 1) =$$

$$= \sum_{r=1}^n C_n^r C_m^r k^{r-r^2}$$

$$M\zeta_2(L) = \sum_{r \in \phi_k(m, n)} \sum_{v \in W_r^m, w \in W_r^n} P(\zeta_{(v, w)}(L) = 1) =$$

$$= \sum_{r \in \phi_k(m, n)} C_n^r C_m^r k^{r-r^2}.$$

Отсюда и из Леммы 3 следует утверждение Леммы 6.

Лемма 7. Если $n^a \leq m \leq k^n$, $a > 1$, то имеет место

$$D\zeta_2(L)/(M\zeta_2(L))^2 \rightarrow 0, \text{ при } n \rightarrow \infty.$$

Доказательство. Действительно,

$$D\zeta_2(L) = M(\zeta_2(L))^2 - (M\zeta_2(L))^2. \quad (1)$$

Нетрудно видеть, что

$$M(\zeta_2(L))^2 = \sum_{r, l \in \phi_k(m, n)} \sum_{v_1 \in W_r^m, w_1 \in W_r^n, v_2 \in W_l^m, w_2 \in W_l^n} \frac{|M|}{k^{mn}},$$

где $M = M(v_1, w_1) \cap M(v_2, w_2)$.

В силу Лемм 2 и 4

$$M(\zeta_2(L))^2 \leq \sum_{r, l \in \phi_k(m, n)} \sum_{b=0}^{\min(r, l)} C_n^r C_r^b C_{n-r}^{l-b} C_m^l C_m^l k^{r-r^2+l-l^2+lb-l} <$$

$$< \sum_{r, l \in \phi_k(m, n)} C_n^r C_n^l C_m^r C_m^l k^{r-r^2+l-l^2} (1 + \delta(n)). \quad (2)$$

В силу Леммы 6

$$(M\zeta_2(L))^2 = \sum_{r, l \in \phi_k(m, n)} C_n^r C_m^r C_n^l C_m^l k^{r-r^2+l-l^2}. \quad (3)$$

Из утверждений (1) – (3) следует утверждение доказываемой леммы.

Таким образом, из Лемм 6 и 7 следует, что для случайных величин $X_1 = \zeta_1(L)$ и $X_2 = \zeta_2(L)$ выполнены все условия Леммы 5. Следовательно, утверждение Теоремы 1 полностью доказано.

Так как правильные ЭК порождаются правильными подматрицами из $S(L)$, то приведенная в теореме 1 оценка типичного порядка правильной подматрицы косвенно свидетельствует о том, что в случае, когда число прецедентов m в $Q(K)$ существенно больше числа признаков n , типичный ранг правильного представительного ЭК в $Q(K)$ не превосходит $r(k, m, n)$.

2. Экспериментальное исследование

Проведено экспериментальное сравнение качества работы алгоритма голосования по правильным представительным ЭК (REC) с такими хорошо известными алгоритмами, как случайный лес (RF) и логистическая регрессия (LR). Результаты счета на реальных данных с целочисленными признаками приведены в табл. Задачи взяты из репозитория UCI [archive.ics.uci.edu] и ВЦ ФИЦ ИУ РАН. Дополнительная настройка алгоритмов не производилась.

Результаты усреднены по 10 случайным независимым разбиениям прецедентов, из которых 80% использовалось для обучения моделей, а оставшиеся 20% — для оценки качества классификации. В каждом из разбиений распределение прецедентов по классам сохранялось неизменным. Функционалом качества выбрана сбалансиро-

рованная точность классификации, вычисляемая по формуле:

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

где q_i — доля верно классифицированных объектов класса K_i ; $i = 1, l$. Данный функционал хорошо себя зарекомендовал при несбалансированных классах. В случае равномоощных классов сбалансированная точность совпадает с долей верно классифицированных объектов.

Как видно из результатов счета, приведенных в табл. 1, алгоритм REC в среднем не уступает по качеству классификации алгоритмам RF и LR. Классификатор REC демонстрирует наилучший результат по качеству классификации на трех задачах из семи, RF — на двух задачах, LR — на двух задачах.

Табл. 1

Качество работы классификаторов

Задача	REC	RF	LR
Остеосаркома	0.570	0.545	0.578
Инсульт	0.623	0.542	0.553
Манелис	0.735	0.742	0.774
Молекулярная Биология 2 (UCI)	0.794	0.946	0.831
Молекулярная Биология 1 (UCI)	0.971	0.960	0.922
Шахматы (UCI)	0.974	0.988	0.956
Крестики-нолики (UCI)	0.976	0.939	0.639

Заключение

В рамках логического подхода к классификации по прецедентам рассматриваются вопросы, связанные со сложностью поиска правильных представительных ЭК. Поиск таких ЭК осуществляется на основе перечисления правильных подматриц целочисленной матрицы $L \in M_{mn}^k$. Получены точная асимптотическая оценка типичного числа правильных подматриц и оценка типичного порядка правильной подматрицы матрицы L в случае большого числа строк, а именно, при $n^a \leq m \leq k^n$, $a > 1$. Оценка типичного числа правильных подматриц матрицы L является верхней асимптотической оценкой типичного числа правильных представительных ЭК для случая, когда число прецедентов существенно больше числа признаков. Оценка типичного порядка правильной подматрицы косвенно свидетельствует о том, что в рассматриваемом случае ранг правильного представительного ЭК почти всегда не превосходит величину $r(k, m, n) = 0.5 \log_k mn - 0.5 \log_k \log_k mn + \log_k \log_k \log_k n$. В ходе экспериментального исследования на ре-

альных целочисленных данных показано, что алгоритм голосования по правильным представительным ЭК в среднем не уступает по качеству классификации таким алгоритмам как случайный лес и логистическая регрессия.

Хочу выразить благодарность Дюковой Елене Всеволодовне, под руководством которой выполнена эта работа.

Литература

1. Бонгард М.М. Проблема узнавания. М.Ж. Физматгиз. 1967. 321 с.
2. Crata, Y., Hammer P.L., Ibaraki T. Cause-effect relationships and partially defined Boolean functions // Ann. Oper. Res. 1988. Vol. 16. Iss. 1. P. 299–325.
3. Дюкова Е.В., Журавлёв Ю.И. Задача монотонной дуализации и её обобщения: асимптотические оценки числа решений // Ж. вычисл. матем. и матем. физ. 2018. Т. 58. № 12. С. 2153–2168.
4. Хачиян Л. Г. Избранные труды. М.: МЦНМО. 2009. 519 с.
5. Dragunov N., Djukova E., Djukova A. Supervised classification and finding frequent elements in data // 8th Conference (International) on Information Technology and Nanotechnology Proceedings. NJ: IEEE. 2022. P. 5.
6. Драгунов Н.А., Дюкова Е.В., Дюкова А.П. Логическая классификация на основе поиска правильных представительных элементарных классификаторов // Известия РАН. Теория и системы управления. 2024. № 3 (в печати).
7. Charu C. Aggarwal, Jiawei Han. Frequent Pattern Mining. Springer International Publishing. 2014. 471 p.
8. Дюкова Е.В., Дюкова А.П. О методах машинного обучения в задаче предсказания промотеров // Доклады Международной конференции «Математическая биология и биоинформатика». 2024. Т 10. № 42. С. 1–4.
9. Дюкова Е.В., Дюкова А.П. О числе решений некоторых специальных задач логического анализа целочисленных данных // Известия РАН. Теория и системы управления. 2023. № 5. С. 57–66.
10. 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.
11. Носков В.Н., Слепая В.А. О числе тупиковых тестов для одного класса таблиц // Кибернетика. 1972. № 1. С. 60–65.

Драгунов Никита Аркадьевич. Федеральный исследовательский центр «Информатика и управление» Российской академии наук, г. Москва, Россия. Аспирант Область научных интересов: машинное обучение, искусственный интеллект, задача классификации по прецедентам, логический анализ данных, дискретная математика, перечислительные труднорешаемые задачи. E-mail: nikitadragunovjob@gmail.com

On metric (numerical) properties of logical classifiers

N. A. Dragunov

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

Abstract. A logical approach to the problem of supervised classification is considered. The precedents are assumed to be described in a system of numerical features, each of which has a limited number of values that are encoded by integers. The stage of classifier training consists in finding special fragments of precedents features descriptions that allow distinguishing precedents from different classes and are called representative elementary classifiers. Traditionally, it is crucial for the adopted approach to accurately estimate the quantitative characteristics of required set of elementary classifiers. The paper provides an upper asymptotic estimate of the typical number of so-called regular representative elementary classifiers and an estimate of the typical “rank” of such an elementary classifier for an important practical case when the number of precedents is significantly greater than the number of features.

Keywords: supervised classification, regular representative elementary classifier, regular submatrix of an integer matrix, asymptotic estimates of the typical number and order of regular submatrices.

DOI: 10.14357/20790279240402 **EDN:** KADXXE

References

1. Bongard M. The problem of recognition. M.J. Phisimatgiz. 1967. 321 p. (In Russ).
2. Crama Y., Hammer P.L., Ibaraki T. Cause–effect relationships and partially defined Boolean functions. Ann. Oper. Res. 1988;16 (1):299–325.
3. Djukova E.V., Zhuravlev Yu.I. Monotone dualization problem and its generalizations: Asymptotic estimates of the number of solutions. Comput. Math. Math. Phys. 2018;58(12):2064–2077.
4. Khachiyan L. Selected works. MCNMO; 2009. 519 p. (In Russ).
5. Dragunov N., Djukova E., Djukova A. Supervised classification and finding frequent elements in data. 8th Int. Conf. on Information Technology and Nanotechnology Proceedings (IEEE, New Jersey). 2022. 5 p.
6. Dragunov N., Djukova E., Djukova A. Logical classification based on finding regular representative elementary classifiers. J. Comput. Syst. Sci. Int. 2024;3 (в печати).
7. Charu C. Aggarwal, Jiawei Han. Frequent Pattern Mining. Springer International Publishing. 2014. 471 p.
8. Djukova E.V., Djukova A.P. About Machine Learning Methods in the Gene Promoter Prediction Problem. Proceedings of the International Conference “Mathematical Biology and Bioinformatics” 2024;10(e42):1–4.
9. Djukova E.V., Djukova A.P. Number of solutions for some special logical analysis problems of integer data. J. Comput. Syst. Sci. Int. 2023.62(5):817–826.
10. Dyukova E.V., Peskov N.V. Search for informative fragments in descriptions of objects in discrete recognition procedures. Comput. Math. Math. Phys. 2002.42(5):711–723.
11. Noskov V., Slepyan V. On number of irredundant tests for one class of tables. Kibernetika. 1972.1:60–65 (In Russ).

Dragunov Nikita A. Postgraduate student of the Federal Research Center «Computer Science and Control» of the Russian Academy of Sciences. Research areas: machine learning, artificial intelligence, supervised classification, logical data analysis, discrete mathematics, enumerative intractable problems. E-mail: nikitadragunovjob@gmail.com