# Методы повышения эффективности логических корректоров*

***Е. В. Дюкова, Ю. И. Журавлев, П. А. Прокофьев***
edjukova@mail.ru, zhuravlev@ccas.ru, p_prok@mail.ru
Вычислительный центр им. А. А. Дородницына РАН, Москва, ул. Вавилова, 42

Рассматривается алгебро-логический подход к корректному распознаванию по прецедентам для задач с целочисленными признаками. Исследуются вопросы повышения распознающей способности и скорости обучения логических корректоров — процедур распознавания, основанных на голосовании по семействам корректных наборов элементарных классификаторов. Вводится понятие корректного набора элементарных классификаторов общего вида, и на этой основе строится модель логического корректора, в которой голосующие семейства наборов элементарных классификаторов формируются итеративно. Рассматривается более широкий, чем в ранее построенных моделях, класс корректирующих функций. Качество работы построенной модели логического корректора тестируется на прикладных задачах.

**Ключевые слова**: *корректное распознавание по прецедентам; логические процедуры распознавания; алгебро-логический подход; логические корректоры; корректный набор элементарных классификаторов; локальный базис; бустинг*

# Methods to improve the effectiveness of logical correctors*

***E. V. Djukova, Yu. I. Zhuravlev, and P. A. Prokofjev***
Dorodnicyn Computing Centre of RAS, Moscow

**Background**: One of the key concepts used to build the correct recognition procedures is the concept of elementary classifier. Elementary classifier is an elementary conjunction defined on integer attributive descriptions of objects. Elementary classifier is *correct* if it highlights only the objects of the same class. Classical correct logical recognition procedures are based upon the construction of correct elementary classifier families. There are challenges that cannot find a sufficient number of correct informative elementary classifiers. One way to solve the problem is to build the recognition procedures based on the construction of the families of the correct sets of elementary classifiers (*logical correctors*). The elementary classifiers of the sets of these families are not necessarily correct.
**Methods**: Some new results concerning the improvement of recognition quality and learning rate of logical correctors are presented. The model of the logical corrector based on a more general concept of the correct set of elementary classifiers is built.
**Results**: New design allows more succinctly describe the patterns in the classes of objects. New logical correctors have a higher quality of recognition in almost all test problems. Learning rate of the logical correctors increases due to the preselection of high-informative elementary classifiers (local basis).
**Concluding Remarks**: The proposed methods allow to apply logical correctors for the large-size problems and well-known logical classifiers. Further refinement of the proposed models can be produced by introducing the partial orders on the sets of feature values.

*Работа частично поддержана грантами РФФИ № 13-01-00787-а, № 14-07-00819-а и грантом президента РФ НШ-4908.2014.1.

**Keywords**: *correct classifier; logic classifier; algebraic-logical approach; logical correctors; correct set of elementary classifiers; local basis; boosting*

## 1 Введение

Рассматривается задача распознавания по прецедентам с множеством объектов $M$, представимым в виде объединения непересекающихся подмножеств $K_1, \ldots, K_l$, называемых классами. Задано обучающее множество объектов $T = \{S_1, \ldots, S_m\}$ из $M$. Каждый объект $S_i \in T$ описан набором значений признаков $x_1, \ldots, x_n$ (числовых характеристик объекта $S_i$), и известен номер класса $y_i \in \{1, \ldots, l\}$, которому принадлежит $S_i$. Объекты из $T$ называются *прецедентами* или *обучающими объектами*. Требуется построить алгоритм $A_T : M \to \{0, 1, \ldots, l\}$, ставящий в соответствие произвольному объекту из $M$, представленному описанием в системе признаков $\{x_1, \ldots, x_n\}$, либо номер класса, которому он принадлежит, либо 0 в случае отказа от распознавания. Алгоритм $A_T$ называется *алгоритмом (процедурой) распознавания*.

Алгоритм распознавания называется *корректным*, если он не ошибается на обучающих объектах. Качество работы алгоритма распознавания на объектах, не являющихся прецедентами, характеризует его *обобщающую способность*. Представляет интерес синтез корректных алгоритмов распознавания с хорошей обобщающей способностью.

Пусть $x_j$ — признак из $\{x_1, \ldots, x_n\}$, $S$ — объект из $M$ и $H = (x_{j_1}, \ldots, x_{j_r})$ — набор признаков. Обозначим через $x_j(S)$ значение признака $x_j$ на объекте $S$ и через $H(S)$ вектор значений признаков $(x_{j_1}(S), \ldots, x_{j_r}(S))$.

В случае, когда множество допустимых значений каждого признака конечно и состоит из целых чисел, задача корректного распознавания успешно решается в рамках *логического* подхода [1–5]. Одним из базовых понятий этого подхода является понятие элементарного классификатора [3].

Пусть $H = (x_{j_1}, \ldots, x_{j_r})$ — набор различных признаков и $\sigma = (\sigma_1, \ldots, \sigma_r)$ — набор, в котором $\sigma_q$ — допустимое значение признака $x_{j_q}$, $q \in \{1, \ldots, r\}$. Пара $(H, \sigma)$ называется *элементарным классификатором* (эл.кл.). Число $r$ называется *рангом* эл.кл. $(H, \sigma)$. Говорят, что эл.кл. $(H, \sigma)$ является фрагментом описания объекта $S$ (выделяет объект $S$), если $H(S) = \sigma$. Элементарный классификатор $(H, \sigma)$ называется *корректным* для класса $K, K \in \{K_1, \ldots, K_l\}$, если не существует двух выделяемых эл.кл. $(H, \sigma)$ прецедентов $S_i$ и $S_t$ таких, что $S_i \in K$, $S_t \notin K$, т.е. множество прецедентов, выделяемых корректным эл.кл. $(H, \sigma)$, является подмножеством либо $T \cap K$, либо $T \setminus K$.

В классических логических процедурах распознавания на этапе обучения для каждого класса $K$ формируется семейство корректных для $K$ эл.кл. При распознавании объекта осуществляется голосование по эл.кл. построенных семейств. Корректность процедуры распознавания обеспечивается за счет корректности каждого эл.кл., участвующего в голосовании. Естественно, что качество работы распознающей процедуры напрямую связано с информативностью использующихся корректных эл.кл. Этап формирования семейств из информативных корректных эл.кл. является наиболее трудоемким в плане вычислительной сложности. Хорошие результаты дает предварительный анализ обучающей выборки, нацеленный на выделение «типичных» обучающих объектов [3].

Довольно часто встречаются задачи, когда почти все корректные эл.кл. имеют большой ранг. Такие задачи являются сложными для рассматриваемых алгоритмов. Несмотря на то что каждый голосующий эл.кл. корректен для некоторого класса $K$, он плохо характеризует класс $K$ в целом (не является информативным для $K$). Возникает эффект

переобучения, связанный с тем, что вместо выявления скрытых закономерностей класса фактически происходит «копирование» прецедентов этого класса по отдельности. Зачастую описанная ситуация возникает в связи с большой значностью признаков (под значностью признака понимается число его различных значений, встречающихся в обучающей выборке).

Одним из способов решения указанной проблемы является корректная перекодировка признаков [5]. Другой способ заключается в построении корректной процедуры распознавания на базе произвольных эл.кл., необязательно корректных, что, как правило, осуществляется методами *алгебро-логического* подхода, объединяющего идеи логического и *алгебраического* подходов.

Алгебраический подход применяется, когда требуется скорректировать работу нескольких различных алгоритмов, каждый из которых безошибочно классифицирует лишь часть обучающих объектов. Цель коррекции — сделать так, чтобы ошибки одних алгоритмов были скомпенсированы другими и качество результирующего алгоритма оказалось лучше, чем каждого из базовых алгоритмов в отдельности (см., например, [6, 7]).

Об алгебро-логическом подходе говорят, когда каждый базовый алгоритм распознавания однозначно определяется некоторым эл.кл. и корректирующие функции являются булевыми функциями. Идея алгебро-логического синтеза корректных логических процедур распознавания предложена в [8]. В указанной работе введено понятие корректного набора эл.кл. Подход развит в работах [9–12], в которых рассмотрены вопросы практического применения различных моделей логических корректоров — корректных процедур распознавания, основанных на голосовании по корректным наборам эл.кл.

Определим понятие корректного набора эл.кл. Пусть имеется упорядоченный набор эл.кл. $U = ((H_1, \sigma_1), \ldots, (H_d, \sigma_d))$. Набор $U$ ставит в соответствие объекту $S$ из $M$ бинарный вектор $U(S) = ([H_1(S) = \sigma_1], \ldots, [H_d(S) = \sigma_d])$, который называется *откликом* набора эл.кл. $U$ на объекте $S$ (здесь и далее через $[p]$ обозначается предикат, принимающий значение 1 в случае, когда выражение $p$ истинно, и 0 — в противном случае). Набор эл.кл. $U$ называется *корректным* для класса $K$, если для любых двух обучающих объектов $S_i$ и $S_t$ таких, что $S_i \in K$ и $S_t \notin K$, отклики $U(S_i)$ и $U(S_t)$ различны. Булевая функция $F(t_1, \ldots, t_d)$ такая, что для любых двух обучающих объектов $S_i \in K$ и $S_t \notin K$ выполняется $F(U(S_i)) \neq F(U(S_t))$, называется *корректирующей* для $U$.

На этапе обучения логического корректора для каждого класса $K$ формируется семейство корректных для $K$ наборов эл.кл. При распознавании объекта осуществляется голосование по построенным семействам. Корректность процедуры распознавания обеспечивается за счет корректности каждого набора эл.кл., участвующего в голосовании.

Наиболее существенным и трудоемким является этап построения семейства корректных наборов эл.кл., в котором каждый набор обладает высокой распознающей способностью. Для эффективного осуществления этого этапа применяются генетические алгоритмы, а также итерационные и стохастические методы предобработки обучающей информации с целью формирования так называемых *локальных базисов классов*. Под локальным базисом класса понимается специальный корректный набор эл.кл., который в дальнейшем используется для построения искомого семейства корректных наборов эл.кл. Наилучшее качество показывают процедуры голосования по наборам эл.кл. с монотонной корректирующей функцией. Подробный обзор результатов, полученных ранее в рассматриваемой области, приведен в следующем разделе.

В настоящей работе введено более общее понятие корректного набора эл.кл. и на его основе построена новая модель логического корректора. Для удобства описания модели

выбран язык предикатов. На этапе обучения строятся семейства предикатов, каждый из которых порождается некоторым корректным набором эл.кл. и зависит от свойств корректирующей функции. На конструкцию предиката влияет характер монотонности корректирующей функции по ее отдельным переменным, что является важным отличием от ранее построенных логических корректоров. Семейства голосующих предикатов формируются итеративно по принципу бустинга. Приведены результаты тестирования построенного логического корректора на прикладных задачах.

## 2 Обзор предыдущих результатов

Классическими логическими распознающими процедурами принято считать тестовый алгоритм (голосование по тестам) [1] и голосование по представительным наборам [2].

*Тестом* называется набор признаков $H$ такой, что для любых двух прецедентов $S_i$ и $S_t$, принадлежащих разным классам, векторы значений признаков $H(S_i)$ и $H(S_t)$ различны. На этапе обучения тестового алгоритма формируется семейство тестов $\mathcal{H}$. Распознавание объекта $S$ осуществляется путем голосования по построенным тестам. В простейшей модификации тестового алгоритма для каждого класса $K$ вычисляются оценки принадлежности объекта $S$ классу $K$, имеющие вид:

$$\Gamma(S,K) = \frac{1}{|\mathcal{H}|}\sum_{H\in\mathcal{H}}\frac{1}{|T\cap K|}\sum_{S_i\in T\cap K}[H(S_i)=H(S)]\,.$$

Объект $S$ относится к тому классу $K$, для которого оценка $\Gamma(S,K)$ имеет наибольшее значение. Если таких классов несколько, то алгоритм отказывается от распознавания.

*Представительным набором* класса $K$ называется корректный для $K$ эл.кл., являющийся признаковым подописанием хотя бы одного прецедента из $K$.

На этапе обучения процедуры голосования по представительным наборам для каждого класса $K$, $K \in \{K_1,\ldots,K_l\}$, строится семейство $C_K$ представительных наборов, которое является некоторым подмножеством всех представительных наборов класса $K$. Распознавание объекта $S$ осуществляется путем взвешенного голосования по построенным представительным наборам. Для каждого класса $K$ вычисляются оценки принадлежности объекта $S$ классу $K$, имеющие вид:

$$\Gamma(S,K) = \sum_{(H,\sigma)\in C_K}\alpha_{(H,\sigma)}[H(S)=\sigma]\,.$$

Вес $\alpha_{(H,\sigma)}$ положителен и, как правило, пропорционален числу прецедентов из $K$, выделяемых представительным набором $(H,\sigma)$. Ясно, что корректность процедуры распознавания обеспечивается за счет корректности каждого представительного набора, участвующего в голосовании.

Заметим, что на практике хорошо себя зарекомендовало голосование по представительным наборам небольшого ранга. При этом, как правило, строятся семейства из так называемых тупиковых представительных наборов. Представительный набор $(H,\sigma)$, $H = (x_{j_1},\ldots,x_{j_r})$, $\sigma=(\sigma_1,\ldots,\sigma_r)$, называется *тупиковым*, если для любого $q\in\{1,\ldots,r\}$ эл.кл. $(H',\sigma')$, где $H'=(x_{j_1},\ldots,x_{j_{q-1}},x_{j_{q+1}},\ldots,x_{j_r})$ и $\sigma'=(\sigma_1,\ldots,\sigma_{q-1},\sigma_{q+1},\ldots,\sigma_r)$, не является корректным. Ясно, что чем больше прецедентов выделяет представительный набор класса $K$, тем лучше он характеризует класс $K$ в целом. Поэтому процедуры голосования по тупиковым представительным наборам или по представительным наборам небольшого ранга, как правило, обладают лучшей обобщающей способностью.

Как было сказано во введении, в рамках алгебро-логического подхода был построен ряд моделей логических корректоров [9–12].

Простейший логический корректор построен в [9]. На этапе обучения логического корректора для каждого класса $K$ строится семейство $W_K$ корректных для $K$ наборов эл.кл. Распознавание объекта $S$ осуществляется путем голосования по наборам эл.кл. построенных семейств. Для каждого класса $K$ вычисляется оценка принадлежности объекта $S$ классу $K$, имеющая вид:

$$\Gamma(S,K) = \frac{1}{|W_K|} \sum_{U \in W_K} \frac{1}{|T \cap K|} \sum_{S_i \in T \cap K} [U(S_i) = U(S)] .$$

Далее используется стандартное решающее правило голосования. Корректность распознающего алгоритма обеспечивается за счет корректности каждого набора эл.кл., участвующего в голосовании. Практика показывает, что качество распознавания может быть улучшено за счет построения семейств из тупиковых корректных наборов эл.кл. Корректный для $K$ набор эл.кл. $U$ называется *тупиковым*, если любое его собственное подмножество не является корректным для $K$ набором эл.кл.

Заметим, что вид оценки принадлежности распознаваемого объекта $S$ классу $K$, вычисляемой по семейству корректных наборов эл.кл., аналогичен виду оценки, вычисляемой в тестовом алгоритме, т. е. понятие корректного набора эл.кл. близко к понятию теста.

Фактически коррекция эл.кл. осуществляется за счет того, что при распознавании объекта $S$ отклик $U(S)$ каждого корректного набора эл.кл. $U$ из семейства $W_K$ сравнивается с откликами $U(S_i)$, $S_i \in T \cap K$. Из корректности набора эл.кл. $U$ следует, что предикат $[U(S_i) = U(S)]$ обращается в 1 только в случае, когда $S \notin T \setminus K$.

В [8] предложено два способа сравнения откликов. Первый способ основан на отношении «равно» и используется в описанной выше процедуре голосования по корректным наборам эл.кл.: распознаваемый объект $S$ близок к прецеденту $S_i$ по набору эл.кл. $U$, если $U(S_i) = U(S)$. Второй — на отношении «меньше или равно» и предполагает, что распознаваемый объект $S$ близок к прецеденту $S_i$ по набору эл.кл. $U$, если каждая координата отклика $U(S_i)$ не превосходит соответствующую координату отклика $U(S)$.

Способ сравнения откликов влияет на свойства корректирующей булевой функции. Отношение «равно», вообще говоря, не накладывает никаких ограничений на ее вид. В случае же использования отношения «меньше или равно» корректирующая функция должна быть монотонной булевой функцией. Корректный набор эл.кл. с монотонной корректирующей функцией называется *монотонным*.

В [9] помимо описанного выше логического корректора построена модель, в которой используются только корректные наборы эл.кл. с монотонной корректирующей функцией (корректор МОН). Счет на прикладных задачах показал, что корректор МОН превосходит по качеству корректор, основанный на голосовании по корректным наборам эл.кл. с произвольной корректирующей функцией.

В [11] показано, что вычисление оценки $\Gamma(S,K)$ принадлежности объекта $S$ классу $K$ по корректным наборам эл.кл. можно осуществлять не только на основании сравнения откликов объекта $S$ с откликами прецедентов из $K$, но также сравнивая их с откликами прецедентов не из $K$. На этом принципе построен корректор АМОН, в котором в качестве корректирующей функции использовалась монотонная булева функция. В случае двух классов корректоры МОН и АМОН эквивалентны. Если же в задаче более двух классов, то в ряде случаев корректор АМОН опережает корректор МОН.

Построение семейств корректных наборов эл.кл. с хорошей распознающей способностью является сложной дискретной задачей [8]. Каждый корректный для $K$ набор эл.кл. однозначно соответствует покрытию булевой матрицы $L_K$, специальным образом построенной по обучающей выборке. Каждому столбцу матрицы $L_K$ соответствует один из эл.кл. Каждая строка $L_K$ образована одной из пар прецедентов $S_i \in T \cap K$ и $S_t \in T \setminus K$. Элемент, находящийся на пересечении строки $(S_i, S_t)$ и столбца $(H, \sigma)$, равен 1, только если эл.кл. $(H, \sigma)$ позволяет различить объекты $S_i$ и $S_t$.

Перечислять все покрытия $L_K$ и выбирать среди них наилучшие очень трудоемко. В [9] для построения семейства $W_K$ используется генетический алгоритм. Кроме этого, временные затраты удается существенно сократить за счет использования только одноранговых эл.кл.

В [10] построены две модели логических корректоров, в которых голосование ведется по корректным наборам эл.кл. произвольного ранга. Для снижения временных затрат при построении голосующих семейств добавлена процедура формирования локальных базисов классов. Под локальным базисом класса $K$ понимается корректный для $K$ набор $\mathcal{U}_K$, состоящий из информативных эл.кл. Семейство $W_K$ формируется из корректных наборов эл.кл., каждый из которых является подмножеством локального базиса $\mathcal{U}_K$.

Вообще говоря, идея применения локального базиса в алгебраическом подходе не нова и впервые встречается в работе К.В. Воронцова [13]. Однако применение локального базиса в логических корректорах из [10] имеет свои особенности, вследствие чего потребовалось разработать специальные алгоритмы формирования локального базиса, лучшим из которых оказался алгоритм, основанный на методе бустинга [14]. Отметим, что один из простейших способов построения локального базиса является его случайный выбор. Этот метод был успешно реализована в [12] при построении стохастического логического корректора МОНС, который опережает по качеству распознавания корректор МОН, в основном благодаря снятию ограничения на ранг эл.кл.

Метод бустинга в [10] используется не только для построения локального базиса, но и для итеративного формирования семейств голосующих наборов эл.кл. Вообще говоря, метод бустинга является универсальным методом построения алгоритмов взвешенного голосования по базовым распознающим алгоритмам произвольного типа. При обучении логического корректора на каждой итерации ищется корректный набор эл.кл. такой, что его добавление в семейство наилучшим образом компенсирует ошибки ранее построенных наборов. Пополнение семейств останавливается при достижении требуемого качества или после выполнения заданного числа итераций. Каждый набор эл.кл. получает «оптимальный» вес, и при распознавании объекта осуществляется взвешенное голосование по построенным наборам. Одним из достоинств бустинга является то, что с его помощью удается построить семейства из «непохожих» наборов эл.кл.

## 3 Логический корректор общего вида

### 3.1 Основные понятия и обозначения

Пусть $K \in \{K_1, \ldots, K_l\}$. Введем обозначения $\overline{K} = M \setminus K$, $\mathbb{K}^+ = \{K_1, \ldots, K_l\}$, $\mathbb{K}^- = \{\overline{K}_1, \ldots, \overline{K}_l\}$ и $\mathbb{K}^\pm = \mathbb{K}^+ \cup \mathbb{K}^-$.

Из соображения удобства перейдем на язык предикатов. Рассмотрим произвольный предикат $B : M \to \{0, 1\}$, заданный на множестве объектов $M$. Будем говорить, что $B$ *корректен для* $K \in \mathbb{K}^\pm$, если множество прецедентов, на которых предикат $B$ равен 1, является подмножеством либо $T \cap K$, либо $T \setminus K$. Корректный для $K$ предикат $B$ будем

называть *представительным для* $K \in \mathbb{K}^{\pm}$, если существует прецедент $S_i \in T \cap K$ такой, что $B(S_i) = 1$.

Понятия корректного эл.кл., представительного набора, теста и корректного набора эл.кл. могут быть переформулированы на языке предикатов.

Элементарный классификатор $(H, \sigma)$ корректен для $K$ (является представительным набором класса $K$) тогда и только тогда, когда предикат $B(S) = [H(S) = \sigma]$ является корректным (представительным) для $K$.

Набор признаков $H$ является тестом тогда и только тогда, когда для любого $K \in \mathbb{K}^{+}$ и любого прецедента $S_i \in T \cap K$ предикат $B_i(S) = [H(S_i) = H(S)]$ корректен для $K$.

Пусть $U = ((H_1, \sigma_1), \ldots, (H_d, \sigma_d))$ — набор эл.кл. и $F(t_1, \ldots, t_d)$ — булева функция от $d$ переменных. Обозначим через $F(U)$ предикат, задаваемый композицией $F(U(S)) = F([H_1(S) = \sigma_1], \ldots, [H_d(S) = \sigma_d]),\ S \in M$.

Набор эл.кл. $U$ корректен для класса $K \in \mathbb{K}^{+}$ тогда и только тогда, когда существует булева функция $F$ такая, что предикаты $F(U)$ и $1 - F(U)$ являются представительными соответственно для $K$ и $\overline{K}$.

Ослабим условия, которым удовлетворяет корректный набор эл.кл. Набор эл.кл. $U$ будем называть *полукорректным* для $K \in \mathbb{K}^{\pm}$, если существует булева функция $F$ такая, что предикат $F(U)$ является представительным для $K$. Функция $F$ называется *корректирующей* для набора $U$ относительно класса $K$. В общем случае корректирующая функция $F$ определена неоднозначно, поскольку ее значения заданы лишь в точках $U(S_i)$, $S_i \in T \backslash K$, в которых $F(U(S_i)) = 0$.

Ясно, что каждый корректный для $K$ набор эл.кл. является полукорректным как для $K$, так и для $\overline{K}$. Набор эл.кл., состоящий из одного представительного набора класса $K$, является полукорректным для $K$.

### 3.2 Информативность предиката

Пусть объект $S_i \in T$ имеет неотрицательный вес $w_i$. Обозначим $\boldsymbol{w} = (w_1, \ldots, w_m)$. Пусть $B$ — предикат на множестве объектов $M$ и $K \in \mathbb{K}^{\pm}$. Введем зависящие от взвешенной выборки $(T, \boldsymbol{w})$ функционалы

$$P(B, K) = \sum_{S_i \in K} w_i B(S_i)\,; \quad N(B, K) = \sum_{S_i \in \overline{K}} w_i B(S_i)\,.$$

Потребуем, чтобы веса объектов из $T$ удовлетворяли дополнительному условию нормировки, $w_1 + \cdots + w_m = 1$. Тогда $\boldsymbol{w}$ можно интерпретировать как распределение вероятностей объектов из $T$, и значение $P(B, K)$ будет равно вероятности того, что случайно выбранный из $T$ объект принадлежит $K$ и выделяется предикатом $B$. Очевидно, $N(B, K) = P(B, \overline{K})$, т. е. $N(B, K)$ — вероятность того, что случайно выбранный из $T$ объект не принадлежит $K$ и выделяется предикатом $B$.

Вероятность того, что предикат $B$ выделяет случайно выбранный из $T$ объект $S_i$ только в случае, когда $S_i$ лежит в $K$, равна

$$\sum_{S_i \in K} w_i B(S_i) + \sum_{S_i \notin K} w_i (1 - B(S_i)) = P(B, K) - N(B, K) + \sum_{S_i \notin K} w_i\,.$$

Вероятность указанного события является естественной характеристикой качества предиката $B$ относительно $K$. Чем она больше, тем лучше предикат $B$ подходит для описания $K$. Разность $P(B, K) - N(B, K)$ будем называть *информативностью* предиката $B$

для $K$ и обозначать через $I(B,K)$. Очевидно, если предикат $B$ представителен для $K$, то $N(B,K)=0$ и $I(B,K)=P(B,K)$.

### 3.3 Корректные предикаты специального вида

Рассмотрим множество бинарных логических операций $\mathcal{O}=\{o(x,y):\{0,1\}^2\to\{0,1\}\}$, которое состоит из 16 элементов (отношений). Пусть $O=(o_1,\ldots,o_d)$ — набор операций из $\mathcal{O}$ и $\alpha=(\alpha_1,\ldots,\alpha_d)$, $\beta=(\beta_1,\ldots,\beta_d)$ — бинарные векторы. Введем обозначение:

$$O(\alpha,\beta)=\bigwedge_{j=1}^{d}o_j(\alpha_j,\beta_j)\,.$$

Пусть $G$ — набор объектов из $M$, $U$ — набор эл.кл., $O$ — набор отношений из $\mathcal{O}$ и длины наборов $U$ и $O$ совпадают. Построим предикат

$$B_{(U,O,G)}(S)=\bigvee_{S'\in G}O(U(S'),U(S)).$$

Выявим условия, при которых предикат $B_{(U,O,G)}(S)$ корректен для $K\in\mathbb{K}^{\pm}$.

Пусть $G_1$ и $G_2$ — множества объектов из $M$, $U$ — набор эл.кл., $O$ — набор операций из $\mathcal{O}$ и длины наборов $U$ и $O$ совпадают. Будем говорить, что набор эл.кл. $U$ *отделяет* объекты из $G_1$ от объектов из $G_2$ с помощью набора бинарных логических операций $O$, если не существует двух объектов $S'\in G_1$ и $S''\in G_2$, для которых выполняется равенство $O(U(S'),U(S''))=1$.

В частности, когда набор эл.кл. $U$ отделяет прецеденты из $T\cap K$ от прецедентов из $T\setminus K$ с помощью набора $O$, состоящего из одинаковых бинарных логических операций, совпадающих с отношением «равно» («меньше или равно»), $U$ является (монотонным) корректным для класса $K$.

**Утверждение 1.** *Пусть $K\in\mathbb{K}^{\pm}$, $G$ — набор объектов из $M$ и набор эл.кл. $U$ отделяет объекты из $G$ от прецедентов из $\overline{K}$ с помощью набора операций $O$.*

*Тогда предикат $B_{(U,O,G)}(S)$ корректен для $K$, и набор эл.кл. $U$ является полукорректным для $K$ с корректирующей функцией*

$$F_{(U,O,G)}(t_1,\ldots,t_d)=\bigvee_{S'\in G}O(U(S'),(t_1,\ldots,t_d))\,.$$

**Доказательство.** Справедливость утверждения 1 очевидно вытекает из конструкции предиката $B_{(U,O,G)}(S)$. ■

Далее рассматривается ряд полезных свойств корректных предикатов вида $B_{(U,O,G)}(S)$. Пусть $B_1$ и $B_2$ — предикаты, корректные для $K$. Будем говорить, что $B_1$ *эквивалентен* $B_2$, если для любого прецедента $S_i$ из $K$ выполняется $B_1(S_i)=B_2(S_i)$. В случае, когда предикаты $B_{(U,O,G)}(S)$ и $B_{(U',O',G')}(S)$ эквивалентны, есть смысл отдавать предпочтение тому предикату, который имеет более простую конструкцию. Докажем несколько свойств, связанных с отношением эквивалентности предикатов.

**Утверждение 2.** *Пусть $U=((H_1,\sigma_1),\ldots,(H_d,\sigma_d))$ — набор эл.кл., $O=(o_1,\ldots,o_d)$ — набор отношений из $\mathcal{O}$, $G$ — набор объектов из $M$ и $B_{(U,O,G)}(S)$ — корректный для $K$ предикат. Тогда выполняются следующие свойства.*

1. *Если для некоторого объекта $S^*\in G$ найдется $j\in\{1,\ldots,d\}$ такой, что $o_j(B_{(H_j,\sigma_j)}(S^*),0)=o_j(B_{(H_j,\sigma_j)}(S^*),1)=0$, то предикаты $B_{(U,O,G)}(S)$ и $B_{(U,O,G\setminus\{S^*\})}(S)$ эквивалентны.*

2. *Если наборы $O'$ и $U'$ получаются путем удаления соответственно из $O$ операции $o_j$ и из $U$ эл.кл. $(H_j, \sigma_j)$ таких, что $\forall S \in G$, $o_j(B_{(H_j,\sigma_j)}(S), 0) = o_j(B_{(H_j,\sigma_j)}(S), 1) = 1$, то предикаты $B_{(U,O,G)}(S)$ и $B_{(U',O',G)}(S)$ эквивалентны.*

**Доказательство.** Первое свойство следует из того, что $\forall S \in M$, $O(U(S^*), U(S)) = 0$, т. е. слагаемое, соответствующее объекту $S^*$, можно не включать в дизъюнкцию, задающую предикат $B_{(U,O,G)}(S)$. Второе свойство вытекает из очевидного тождества $O'(U(S), U(S')) = O(U(S), U(S'))$, $\forall S \in G, \forall S' \in M$. ■

Утверждение 2 фактически дает два правила «упрощения» предикатов вида $B_{(U,O,G)}(S)$. Из наборов $U$, $O$ и $G$ можно удалять элементы, не влияющие на результат применения предиката к прецедентам.

Возможен другой путь упрощения предиката $B_{(U,O,G)}(S)$. Рассмотрим множество операций $\mathcal{O}^* = \{[x \leqslant y], [x \geqslant y], [x \vee y], [\neg x \vee \neg y]\}$. Каждая операция из $\mathcal{O}^*$ принимает нулевое значение всего лишь на одной паре значений аргументов. Например, $[x \leqslant y] = 0$, только если $x = 1$ и $y = 0$. Легко убедиться, что любую операцию $o$ из $\mathcal{O}$ можно представить в виде $o(x, y) = o_1(x, y) \wedge \cdots \wedge o_u(x, y)$, $\{o_1, \ldots, o_u\} \subseteq \mathcal{O}^*$.

**Утверждение 3.** *Пусть $U$ — набор эл.кл., $O$ — набор операций из $\mathcal{O}$, $G$ — набор объектов из $M$ и $B_{(U,O,G)}(S)$ — корректный для $K$ предикат.*

*Тогда существуют набор отношений $O'$ из $\mathcal{O}^*$ и набор эл.кл. $U'$ такие, что предикаты $B_{(U,O,G)}(S)$ и $B_{(U',O',G)}(S)$ эквивалентны.*

**Доказательство.** Построим требуемые наборы $O'$ и $U'$ соответственно из наборов $O$ и $U$ по следующему правилу. Каждую операцию $o_j$ в $O$, не принадлежащую $\mathcal{O}^*$, заменим на набор операций $\{o'_1, \ldots, o'_u\} \subseteq \mathcal{O}^*$ таких, что $o_j(x, y) = o'_1(x, y) \wedge \cdots \wedge o'_u(x, y)$, и каждой операции $o'_v, v \in \{1, \ldots, u\}$ сопоставим эл.кл. $(H'_v, \sigma'_v)$, совпадающий с эл.кл. $(H_j, \sigma_j)$. Полученные в результате замен наборы $O'$ и $U'$ и будут определять предикат $B_{(U',O',G)}(S)$, эквивалентный предикату $B_{(U,O,G)}(S)$. ■

Утверждение 3 позволяет при построении предикатов вида $B_{(U,O,G)}(S)$ не использовать отношения из $\mathcal{O} \setminus \mathcal{O}^*$.

**Утверждение 4.** *Пусть $U = ((H_1, \sigma_1), \ldots, (H_d, \sigma_d))$ — набор эл.кл., $O = (o_1, \ldots, o_d)$ — набор отношений из $\mathcal{O}$, $G$ — набор объектов из $M$ и $B_{(U,O,G)}(S)$ — корректный для $K$ предикат. Тогда выполняются следующие свойства.*

1. *Для любого подмножества $G' \subset G$ предикат $B_{(U,O,G')}(S)$ корректен для $K$.*
2. *Для любого эл.кл. $(H', \sigma')$ и любого отношения $o'$ из $\mathcal{O}^*$ предикат $B_{(U',O',G)}(S)$, $U' = ((H_1, \sigma_1), \ldots, (H_d, \sigma_d), (H', \sigma'))$, $O' = (o_1, \ldots, o_d, o')$, корректен для $K$.*

**Доказательство.** Доказательство основывается на том, что предикат $B_{(U,O,G)}(S)$ выделяет все объекты, выделяемые и предикатом $B_{(U,O,G')}(S)$, и предикатом $B_{(U',O',G)}(S)$. ■

Пусть $U = ((H_1, \sigma_1), \ldots, (H_d, \sigma_d))$ — набор эл.кл., $O = (o_1, \ldots, o_d)$ — набор отношений из $\mathcal{O}$, $G$ и $G^*$ — наборы объектов из $M$ такие, что $G \subseteq G^*$. Корректный для $K$ предикат $B_{(U,O,G)}(S)$ будем называть *тупиковым относительно* $G^*$, если выполняются два условия:

1) для любого объекта $S^* \in G^* \setminus G$ предикат $B_{(U,O,G\cup\{S^*\})}(S)$ не является корректным для $K$;
2) для любого $j \in \{1, \ldots, d\}$ предикат $B_{(U',O',G)}(S)$, порожденный наборами $U' = ((H_1, \sigma_1), \ldots, (H_{j-1}, \sigma_{j-1}), (H_{j+1}, \sigma_{j+1}), \ldots (H_d, \sigma_d))$, $O' = (o_1, \ldots, o_{j-1}, o_{j+1}, \ldots, o_d)$, не является корректным для $K$.

Тупиковый корректный для $K$ относительно $T \cap K$ предикат будем просто называть тупиковым. Пусть $G_1 \subseteq G_2 \subseteq M$. Обозначим через $\mathcal{P}_K(G_1, G_2)$ множество всех корректных для $K$ предикатов вида $B_{(U,O,G)}(S)$, для которых $G_1 \subseteq G \subseteq G_2$. Будем оценивать информативность предикатов из $\mathcal{P}_K(G_1, G_2)$, полагая, что контрольная выборка совпадает с основной. Нетрудно видеть, что при $T^* = T$ функция информативности $I(B_{(U,O,G)}, K)$ совпадает с функцией $P(B_{(U,O,G)}, K)$, которая на множестве $\mathcal{P}_K(G_1, G_2)$ достигает локального максимума в каждом тупиковом относительно $G_2$ предикате. Обозначим через $\mathcal{P}^*_K(G_1, G_2)$ множество тупиковых относительно $G_2$ предикатов из $\mathcal{P}_K(G_1, G_2)$.

Пусть предикат $B_{(U,O,G)}(S)$ корректен для $K$. Исследуем свойства корректирующей функции $F_{(U,O,G)}$ полукорректного набора эл.кл. $U$. В частности, выясним условия, при которых $F_{(U,O,G)}$ является монотонной или поляризуемой булевой функцией (булева функция $F(t_1, \ldots, t_d)$ называется *поляризуемой*, если для некоторого бинарного вектора $(\alpha_1, \ldots, \alpha_d)$ функция $F(t_1 \oplus \alpha_1, \ldots, t_d \oplus \alpha_d)$ монотонна). Разделим множество отношений $\mathcal{O}^*$ на два подмножества $\mathcal{O}^*_0 = \{[x \geqslant y], [\neg x \vee \neg y]\}$ и $\mathcal{O}^*_1 = \{[x \leqslant y], [x \vee y]\}$.

**Утверждение 5.** *Пусть $B_{(U,O,G)}(S)$ — корректный для $K$ предикат. Тогда имеют место следующие два критерия.*

1. *Корректирующая функция $F_{(O,U,G)}$ является монотонной тогда и только тогда, когда предикат $B_{(U,O,G)}(S)$ эквивалентен предикату $B_{(U',O',G)}(S)$ такому, что каждое отношение из набора $O'$ принадлежит $\mathcal{O}^*_1$.*
2. *Корректирующая функция $F_{(O,U,G)}$ является поляризуемой тогда и только тогда, когда предикат $B_{(U,O,G)}(S)$ эквивалентен предикату $B_{(U',O',G)}(S)$, $U' = ((H'_1, \sigma'_1), \ldots, (H'_u, \sigma'_u))$, $O' = (o'_1, \ldots, o'_u)$, такому, что каждое отношение из набора $O'$ принадлежит $\mathcal{O}^*$, и в наборе $U'$ не существует двух одинаковых эл.кл. $(H'_i, \sigma'_i)$ и $(H'_t, \sigma'_t)$, для которых $o'_i \in \mathcal{O}^*_0$ и $o'_t \in \mathcal{O}^*_1$.*

**Доказательство.** Заметим, что $[x \leqslant y] = [\neg x \vee y]$ и $[x \geqslant y] = [x \vee \neg y]$.

Докажем первый критерий. Использование отношений из $\mathcal{O}^*_1$ гарантирует, что в дизъюнктивную нормальную форму корректирующей функция $F_{(U',O',G)}$ не будут входить переменные с отрицанием, что эквивалентно монотонности $F_{(U',O',G)}$.

Второй критерий справедлив, поскольку переменные корректирующей функции $F_{(U',O',G)}$, соответствующие одинаковым эл.кл. набора $U'$, либо входят в $F_{(U',O',G)}$ только с отрицанием (им соответствуют операции из $\mathcal{O}^*_0$), либо входят только без отрицания (им соответствуют операции из $\mathcal{O}^*_1$). ■

Проиллюстрируем на примере с модельной задачей распознавания преимущества предикатов вида $B_{(U,O,G)}$.

**Пример 1.** Рассмотрим задачу распознавания с двумя классами и двумя признаками, изображенную на рис. 1. Построим следующие наборы эл.кл.:

1. Набор $U_1 = ([x_2 = 0],\ [x_1 = 0],\ [x_1 = 1],\ [x_1 = 2],\ [x_1 = 3],\ [x_1 = 4])$ принадлежит семейству монотонных корректным для класса $K_1$ наборов эл.кл.
2. Набор $U_2 = ([x_1 = 5],\ [x_1 = 0],\ [x_2 = 0])$ принадлежит семейству корректных для класса $K_1$ наборов эл.кл. и не является монотонным.
3. Набор $U_3 = ([x_1 = 5],\ [x_2 = 0],\ [x_1 = 0])$ отделяет прецеденты из $K_1$ от прецедентов из $K_2$ с помощью набора отношений $O = ([x \geqslant y], [x \leqslant y], [x \leqslant y])$.

Отметим, что наборы $U_1$ и $U_2$ являются наименее «громоздкими» представителями своих семейств. Выпишем предикаты, которые порождаются наборами $U_1, U_2, U_3$ и преце-

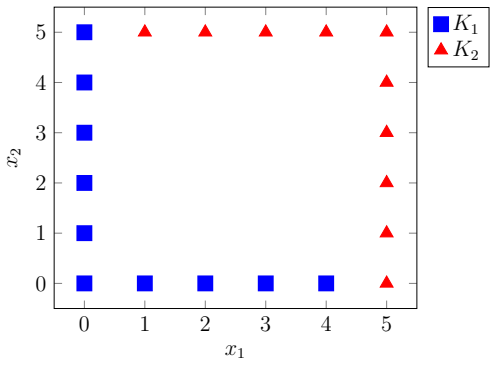



**Рис. 1** Модельная задача распознавания с двумя классами и двумя признаками из примера 1

дентами класса $K_1$. Оценим информативность этих предикатов по обучающей выборке $T$, полагая, что вес каждого объекта равен $1/20$.

1. Набор $U_1$ и прецеденты из $K_1$ порождают следующие предикаты: $[x_1 = 0]$, $[x_1 = 0 \wedge x_2 = 0]$, $[x_1 = 1 \wedge x_2 = 0]$, $[x_1 = 2 \wedge x_2 = 0]$, $[x_1 = 3 \wedge x_2 = 0]$, $[x_1 = 4 \wedge x_2 = 0]$. Первый предикат имеет информативность 0,3, информативности остальных равны 0,05.
2. Набор $U_2$ и прецеденты из $K_1$ порождают предикаты $[x_1 \neq 5 \wedge x_1 \neq 0 \wedge x_2 = 0]$, $[x_1 = 0 \wedge x_2 \neq 0]$ и $[x_1 = 0 \wedge x_2 = 0]$, информативности которых соответственно равны 0,2, 0,25 и 0,05.
3. Набор $U_3$ и прецеденты из $K_1$ порождают предикаты $[x_1 \neq 5 \wedge x_2 = 0]$ и $[x_1 = 0]$, информативности которых соответственно равны 0,25 и 0,3. Отметим, что приведены все предикаты $B_{(U_3,O,\{S_i\})}, S_i \in K_1$.

Видно, что набор $U_3$ порождает более лаконичные предикаты с высокой информативностью. Это становится возможным благодаря тому, что с помощью расширенного набора отношений удается одним предикатом проверить как наличие, так и отсутствие некоторого признакового подописания у распознаваемого объекта.

### 3.4 Голосование по представительным предикатам

Опишем логический корректор, основанный на голосовании по предикатам вида $B_{(U,O,G)}(S)$.

На *этапе обучения* для каждого класса $K \in \mathbb{K}^+$ строятся два семейства $Z_K$ и $Z_{\overline{K}}$ предикатов на множестве объектов $M$. Каждый предикат семейства $Z_K, K \in \mathbb{K}^{\pm}$, является представительным для $K$. Для каждого предиката $B \in Z_K$ задается положительный вес $\alpha_B$.

Семейства предикатов $Z_K$, $K \in \mathbb{K}^{\pm}$, формируются итеративно. При инициализации берутся $Z_K := \varnothing$, $K \in \mathbb{K}^{\pm}$. На итерации $t \geqslant 1$ по некоторому правилу выбираются $K \in \mathbb{K}^{\pm}$ и подмножества прецедентов $G_1 \subseteq G_2 \subseteq T \cap K$. Далее осуществляется поиск одного или нескольких предикатов из $\mathcal{P}^*_K(G_1, G_2)$ с высокой информативностью. Каждый найденный предикат $B$ получает вес $\alpha_B$, вычисляемый по определенному правилу, и добавляется в семейство $Z_K$. Если не выполнен критерий останова, то происходит переход к следующей итерации.

Этап обучения имеет несколько параметров:

1) правило выбора $K \in \mathbb{K}^{\pm}$ и подмножеств прецедентов $G_1 \subseteq G_2 \subseteq T \cap K$ на каждой итерации;
2) алгоритм поиска корректных предикатов с высокой информативностью;

3) правило вычисления весов предикатов;
4) критерий останова обучения.

*При распознавании* осуществляется взвешенное голосование по предикатам, построенным на этапе обучения. Возможны два режима распознавания: базовый и аддитивный.

1. В *базовом режиме* для распознаваемого объекта $S$ вычисляются оценки $\Gamma(S,K)$ принадлежности объекта $S$ классу $K \in \mathbb{K}^+$, имеющие вид:

$$\Gamma(S,K) = \sum_{B \in Z_K} \alpha_B B(S) - \sum_{B \in Z_{\overline{K}}} \alpha_B B(S)\,.$$

2. В *аддитивном режиме* для распознаваемого объекта $S$ и каждого предиката $B_{(U,O,G)} \in \in Z_K$, $K \in \mathbb{K}^{\pm}$, вычисляется оценка

$$\gamma(S, B_{(U,O,G)}) = \frac{1}{|G|} \sum_{S_i \in G} O(U(S_i), U(S))\,.$$

Затем для каждого класса $K \in \mathbb{K}^+$ вычисляется оценка $\Gamma(S,K)$ принадлежности объекта $S$ классу $K$, имеющая вид:

$$\Gamma(S,K) = \sum_{B \in Z_K} \alpha_B \gamma(S,B) - \sum_{B \in Z_{\overline{K}}} \alpha_B \gamma(S,B)\,.$$

Описанный распознающий алгоритм будем называть *логическим корректором общего вида*. Для обеспечения его корректности достаточно, чтобы было справедливо

**Утверждение 6.** *Пусть $A$ — логический корректор общего вида и $\{Z_K,\ K \in \mathbb{K}^{\pm}\}$ — семейства предикатов, по которым осуществляется голосование при распознавании объектов. Алгоритм $A$ корректен, если для любого класса $K \in \mathbb{K}^+$ и любого прецедента $S_i \in K$ выполняется одно из двух условий:*

*1) в семействе $Z_K$ найдется предикат, выделяющий $S_i$;*
*2) для каждого $\overline{K'} \neq \overline{K}$, $\overline{K'} \in \mathbb{K}^-$, в семействе $Z_{\overline{K'}}$ найдется предикат, выделяющий $S_i$.*

**Доказательство.** Пусть $P$ — семейство предикатов на множестве объектов $M$ и $S$ — объект из $M$. Введем обозначение $b(P,S) = \{B \in P : B(S) = 1\}$.

Зафиксируем класс $K \in \mathbb{K}^+$ и объект $S_i \in K$.

1) Если $b(P_K, S_i) \neq \varnothing$, то $\Gamma(S_i, K) > 0$. Поскольку $\forall K' \in \mathbb{K}^+ \setminus \{K\}$, $\Gamma(S_i, K') \leqslant 0$, отступ $\Delta(S_i, K) > 0$, и объект $S_i$ распознается алгоритмом $A$ правильно.

2) Если $\forall \overline{K'} \in \mathbb{K}^- \setminus \{\overline{K}\}$, $b(Z_{\overline{K'}}, S_i) \neq \varnothing$, то $\forall K'' \in \mathbb{K}^+ \setminus \{K\}$, $\Gamma(S_i, K'') < 0$. Поскольку $\Gamma(S_i, K) \geqslant 0$, отступ $\Delta(S_i, K) > 0$, и объект $S_i$ распознается алгоритмом $A$ правильно. ■

### 3.5 Построение корректных предикатов

В работе [8] построение тупиковых корректных наборов эл.кл. сводится к поиску неприводимых покрытий булевой матрицы, построенной специальным образом по обучающей выборке. В данном подразделе выполняется аналогичное сведение построения корректного для $K$ предиката вида $B_{(U,O,G)}$ к поиску покрытия булевой матрицы. Отдельно рассматривается вопрос поиска в семействах $\mathcal{P}_K(G_1,G_2)$ и $\mathcal{P}^*_K(G_1,G_2)$ предикатов с наибольшей информативностью.

Пусть $L = \|a_{ij}\|$ — булева матрица размера $m \times n$. Говорят, что столбец с номером $j$ покрывает строку с номером $i$ булевой матрицы $L$, если $a_{ij} = 1$. Обозначим через $R_0(L,J)$

набор строк матрицы $L$, непокрытых ни одним столбцом из $J$. *Покрытием* булевой матрицы $L$ называется набор столбцов $J$ такой, что каждую строку матрицы $L$ покрывает хотя бы один столбец из $J$, т. е. $R_0(L,J)=\varnothing$. Обозначим через $\mathcal{C}(L)$ набор покрытий булевой матрицы $L$. Покрытие $J$ матрицы $L$ называется *неприводимым*, если любое его собственное подмножество не является покрытием матрицы $L$. Обозначим через $\mathcal{P}(L)$ набор неприводимых покрытий булевой матрицы $L$.

Пусть $K\in\mathbb{K}^{\pm}$, $G_1\subseteq G_2\subseteq T\cap K$. Покажем, что каждый тупиковый корректный для $K$ предикат из $\mathcal{P}_K(G_1,G_2)$ однозначно соответствует неприводимому покрытию булевой матрицы, построенной по прецедентной информации и зависящей от $G_1$ и $G_2$.

Обозначим множество всех эл.кл. через $\mathcal{U}^*$. Построим булеву матрицу $L_{T\setminus K}(G_1,G_2)$ по следующему правилу. Каждой строке матрицы $L_{T\setminus K}(G_1,G_2)$ сопоставим пару обучающих объектов $(S_i,S_t)$ таких, что $S_i\in G_2$ и $S_t\in T\setminus K$. Столбцы матрицы $L_K(G_1,G_2)$ будут иметь один из двух типов. Каждому столбцу первого типа сопоставим тройку $(H,\sigma,o)$, где $(H,\sigma)$ — эл.кл. из $\mathcal{U}^*$ и $o$ — отношение из $\mathcal{O}^*$. Каждому столбцу второго типа — прецедент $S_j$ из $G_2\setminus G_1$. Элемент матрицы $L_{T\setminus K}(G_1,G_2)$, расположенный на пересечении строки $(S_i,S_t)$ и столбца первого типа $(H,\sigma,o)$, равен $1-o(B_{(H,\sigma)}(S_i),B_{(H,\sigma)}(S_t))$. Элемент матрицы $L_{T\setminus K}(G_1,G_2)$, расположенный на пересечении строки $(S_i,S_t)$ и столбца второго типа $S_j$, равен $[i=j]$. Матрицу, построенную по указанному правилу, принято называть *матрицей сравнения*.

**Утверждение 7.** *Пусть $K\in\mathbb{K}^{\pm}$, $G_1\subseteq G\subseteq G_2\subseteq T\cap K$, $U=((H_1,\sigma_1),\ldots,(H_d,\sigma_d))$ — набор эл.кл. и $O=(o_1,\ldots,o_d)$ — набор отношений из $\mathcal{O}^*$.*

*Предикат $B_{(U,O,G)}$ является (тупиковым относительно $G_2$) корректным для $K$ тогда и только тогда, когда набор столбцов матрицы $L_{T\setminus K}(G_1,G_2)$, соответствующих прецедентам из $G_2\setminus G$ и тройкам $(H_1,\sigma_1,o_1),\ldots,(H_d,\sigma_d,o_d)$, является (неприводимым) покрытием матрицы $L_{T\setminus K}(G_1,G_2)$.*

**Доказательство.** Обозначим $J=(G_2\setminus G)\cup\{(H_1,\sigma_1,o_1),\ldots,(H_d,\sigma_d,o_d)\}$. Корректность предиката $B_{(U,O,G)}$ для $K$ по определению означает, что выполняется $B_{(U,O,G)}(S_t)=0$, $\forall S_t\in T\setminus K$. Поскольку верны тождества

$$B_{(U,O,G)}(S)=\bigvee_{S_i\in G}O(U(S_i),U(S))=\bigvee_{S_j\in G_2}[S_j\notin G_2\setminus G]\,O(U(S_j),U(S))\,,$$

корректность предиката $B_{(U,O,G)}$ эквивалентна условию

$$\bigvee_{S_j\in G_2}\bigvee_{S_t\in T\cap K}[S_j\notin G_2\setminus G]\,O(U(S_j),U(S_t))=0\,. \tag{1}$$

Отрицая левую и правую часть равенства (1), получаем условие

$$\bigwedge_{S_j\in G_2}\bigwedge_{S_t\in T\cap K}[S_j\in G_2\setminus G]\vee\neg o_1(B_{(H_1,\sigma_1)}(S_j),B_{(H_1,\sigma_1)}(S_t))\vee\cdots\vee\neg o_d(B_{(H_d,\sigma_d)}(S_j),B_{(H_d,\sigma_d)}(S_t))=$$
$$=1,$$

которое равносильно тому, что набор столбцов $J$ покрывает матрицу $L_{T\setminus K}(G_1,G_2)$.

Из определения тупикового корректного предиката легко выводится, что корректный для $K$ предикат $B_{(U,O,G)}$ является тупиковым относительно $G_2$ тогда и только тогда, когда при удалении любого столбца из $J$ получается набор, не являющийся покрытием $L_{T\setminus K}(G_1,G_2)$, т. е. $J$ — неприводимое покрытие матрицы $L_{T\setminus K}(G_1,G_2)$. ■

Пусть $K \in \mathbb{K}^{\pm}$, $G_1 \subseteq G_2 \subseteq T \cap K$. Рассмотрим задачу построения предиката $B_{(U,O,G)}$ из $\mathcal{P}_K(G_1, G_2)$, обладающего максимальной информативностью.

В случае, когда логический корректор используется в базовом режиме распознавания, информативность корректного для $K$ предиката $B_{(U,O,G)}$ будем оценивать значением $I(B_{(U,O,G)}, K)$. Ставится оптимизационная

**Задача 1.**

$$I(B_{(U,O,G)}) \underset{B_{(U,O,G)} \in \mathcal{P}_K(G_1,G_2)}{\to} \max .$$

В аддитивном режиме более адекватную оценку информативности предиката $B_{(U,O,G)}$ дает функционал

$$\hat{I}(B_{(U,O,G)}, K) = \hat{P}(B_{(U,O,G)}, K) - \hat{N}(B_{(U,O,G)}, K) ,$$

где

$$\hat{P}(B_{(U,O,G)}, K) = \sum_{S \in G} P(B_{(U,O,\{S\})}, K) , \quad \hat{N}(B_{(U,O,G)}, K) = \sum_{S \in G} N(B_{(U,O,\{S\})}, K) .$$

**Задача 2.**

$$\hat{I}(B_{(U,O,G)}) \underset{B_{(U,O,G)} \in \mathcal{P}_K(G_1,G_2)}{\to} \max .$$

Задачи 1 и 2 могут быть рассмотрены в варианте, когда в качестве области поиска предикатов вместо $\mathcal{P}_K(G_1, G_2)$ берется его подмножество $\mathcal{P}^*_K(G_1, G_2)$, состоящее из тупиковых относительно $G_2$ предикатов.

Сформулируем две дискретные оптимизационные задачи, являющиеся специальными разновидностями задачи о поиске покрытий булевой матрицы.

**Задача 3 (поиск набора столбцов, покрывающего оптимальную комбинацию матриц).** Пусть даны булевы матрицы $L_1, \ldots, L_d$ и их веса $\alpha_1, \ldots, \alpha_d$. Каждая матрица имеет $n$ столбцов. Вес $\alpha_i$ матрицы $L_i$ либо является рациональным числом, либо равен $+\infty$. Требуется найти набор столбцов $J \subseteq \{1, \ldots, n\}$ такой, что сумма весов матриц, непокрытых набором $J$, минимальна, т. е.

$$\sum_{i=1}^{d} \alpha_i [J \notin \mathcal{C}(L_i)] \underset{J \subseteq \{1,\ldots,n\}}{\to} \min .$$

Очевидно, не теряя общности, можно считать, что веса всех матриц отличны от нуля, число матриц с весом $+\infty$ не превосходит единицы и ни одна из матриц не содержит нулевой строки. Случай, когда все веса положительны, тривиален, так как одним из решений всегда будет набор, состоящий из всех столбцов $\{1, \ldots, n\}$. Возможен вариант постановки задачи, когда решение должно являться неприводимым покрытием матрицы с весом $+\infty$.

**Задача 4 (поиск набора столбцов, покрывающего оптимальную комбинацию строк).** Пусть дана булева матрица $L$ размера $m \times n$. Для каждой строки $i \in \{1, \ldots, m\}$ задан вес $\beta_i$, который либо является рациональным числом, либо равен $+\infty$. Требуется найти набор столбцов $J \subseteq \{1, \ldots, n\}$ такой, что сумма весов строк, непокрытых набором $J$, минимальна, т. е.

$$\sum_{i=1}^{m} \beta_i [i \notin R_0(L, J)] \underset{J \subseteq \{1,\ldots,n\}}{\to} \min .$$

Снова, не теряя общности, можно считать, что веса всех строк отличны от нуля, и в $L$ нет нулевой строки. В случае, когда веса всех строк положительны, описанная задача тривиальна, поскольку набор столбцов $\{1,\ldots,n\}$ является решением. Возможен вариант постановки задачи, когда решение должно являться неприводимым покрытием подматрицы, составленной из строк с весом $+\infty$.

Покажем, что задачи 1 и 2 сводятся соответсвенно к задачам 3 и 4. Для каждого объекта $S_i^* \in T^*$ построим матрицу сравнения $L_{\{S_i^*\}}(G_1, G_2)$. Справедливо

**Утверждение 8.** *Пусть $K \in \mathbb{K}^{\pm}$, $U = ((H_1,\sigma_1),\ldots,(H_d,\sigma_d))$ — набор эл.кл., $O = (o_1,\ldots,o_d)$ — набор отношений из $\mathcal{O}^*$, $G_1 \subseteq G \subseteq G_2 \subseteq T \cap K$ и $J = (G_2 \setminus G) \cup \cup\{(H_1,\sigma_1,o_1),\ldots,(H_d,\sigma_d,o_d)\}$ — покрытие матрицы $L_{T\setminus K}(G_1,G_2)$. Тогда*

$$\begin{aligned} P(B_{(U,O,G)},K) &= \sum_{S_i^* \in K} w_i[J \notin \mathcal{C}(L_{\{S_i^*\}}(G_1,G_2))]\,; \\ N(B_{(U,O,G)},K) &= \sum_{S_i^* \notin K} w_i[J \notin \mathcal{C}(L_{\{S_i^*\}}(G_1,G_2))]\,. \end{aligned}$$

**Доказательство.** Первого равенства следует из простой цепочки тождеств:

$$\begin{aligned} P(B_{(U,O,G)},K) &= \sum_{S_i^* \in K} w_i B_{(U,O,G)}(S_i^*) = \sum_{S_i^* \in K} w_i \bigvee_{S_j \in G} O(U(S_j),U(S_i^*)) = \\ &= \sum_{S_i^* \in K} w_i \bigvee_{S_j \in G_2} [S_j \notin G_2 \setminus G]\, O(U(S_j),U(S_i^*)) = \\ &= \sum_{S_i^* \in K} w_i \Big(1 - \bigwedge_{S_j \in G_2} [S_j \in G_2 \setminus G] \vee \neg O(U(S_j),U(S_i^*))\Big) = \\ &= \sum_{S_i^* \in K} w_i [J \notin \mathcal{C}(L_{\{S_i^*\}}(G_1,G_2))]\,. \end{aligned}$$

Равенство для $N(B_{(U,O,G)},K)$ доказывается аналогично. ■

Построим матрицу сравнения $L_{(T\setminus K)\cup T^*}(G_1,G_2)$. Аналогично утверждению 8 доказывается

**Утверждение 9.** *Пусть $K \in \mathbb{K}^{\pm}$, $G_1 \subseteq G \subseteq G_2 \subseteq T \cap K$, $U = ((H_1,\sigma_1),\ldots,(H_d,\sigma_d))$ — набор эл.кл., $O = (o_1,\ldots,o_d)$ — набор отношений из $\mathcal{O}^*$ и $J = (G_2 \setminus G) \cup \cup\{(H_1,\sigma_1,o_1),\ldots,(H_d,\sigma_d,o_d)\}$ — покрытие матрицы $L_{T\setminus K}(G_1,G_2)$. Найдем множество строк матрицы $L_{(T\setminus K)\cup T^*}(G_1,G_2)$, непокрытых набором столбцов $J$. Обозначим $R_0 = R_0(L_{(T\setminus K)\cup T^*}(G_1,G_2),J)$. Тогда*

$$\begin{aligned} \hat{P}(B_{(U,O,G)},K) &= \sum_{(S,S_i^*) \notin R_0} w_i[S_i^* \in K]\,; \\ \hat{N}(B_{(U,O,G)},K) &= \sum_{(S,S_i^*) \notin R_0} w_i[S_i^* \notin K]\,. \end{aligned}$$

Каждой матрице $L_{\{S_i^*\}}(G_1,G_2)$, $S_i^* \in T^*$, и каждой строке $(S,S_i^*)$, $S \in G_2$, $S_i^* \in T^*$, матрицы $L_{(T\setminus K)\cup T^*}(G_1,G_2)$ припишем вес

$$\begin{cases} -w_i, & S_i^* \in K\,; \\ w_i, & S_i^* \notin K\,, \end{cases}$$

а матрице $L_{T\setminus K}(G_1, G_2)$ и каждой строке $(S, S_i)$, $S \in G_2$, $S_i \in T \cap K$, матрицы $L_{(T\setminus K)\cup T^*}(G_1, G_2)$ — вес, равный $+\infty$.

Из утверждений 7 и 8 следует, что набор столбцов, покрывающий оптимальную комбинацию взвешенных матриц $L_{T\setminus K}(G_1, G_2), L_{\{S_1^*\}}(G_1, G_2), \ldots, L_{\{S_p^*\}}(G_1, G_2)$, является решением задачи 1.

Аналогично из утверждений 7 и 9 следует, что набор столбцов, покрывающий оптимальную комбинацию взвешенных строк матрицы $L_{(T\setminus K)\cup T^*}(G_1, G_2)$, является решением задачи 2.

Аналоги задачи 3 авторам неизвестны. Очевидно, что если матрицы с числовыми весами однострочны, то задачи 3 и 4 эквивалентны. Задача 4 обобщает ряд известных задач, однако ее исследования в приведенной постановке не проводились.

**Задача 5 (Red-Blue Set Cover Problem (RBSC)).** Простой вариант задачи RBSC формулируется следующим образом [15, 16]. Входом являются множество «красных» элементов $R$, множество «синих» элементов $B$ и набор $\mathcal{D}$ подмножеств множества $R \cup B$. Говорят, что элемент $e \in R \cup B$ покрыт набором $\mathcal{D}' \subseteq \mathcal{D}$, если $e$ принадлежит хотя бы одному множеству из $\mathcal{D}'$. Обозначим через $\mathcal{C}(\mathcal{D}')$ множество элементов, покрытых набором $\mathcal{D}'$. Требуется найти подмножество $\mathcal{D}'$ множества $\mathcal{D}$, которое покрывает все синие элементы и как можно меньше красных элементов, т. е.

$$|R \cap \mathcal{C}(\mathcal{D})| \underset{\mathcal{D}' \subseteq \mathcal{D}: B \subseteq \mathcal{C}(\mathcal{D}')}{\to} \min.$$

В [16] также рассматривается «взвешенный» вариант RBSC. Во взвешенном варианте RBSC каждому красному элементу присваивается положительный вес и требуется минимизировать сумму весов покрытых красных элементов.

В случае, когда часть строк матрицы $L$ имеет вес $+\infty$, а остальные строки имеют отрицательные веса, задача 4 эквивалентна RBSC, в которой синими элементами являются строки матрицы $L$, имеющие вес $+\infty$, красными — остальные строки $L$ и вес каждого красного элемента равен весу соответствующей строки, взятому с противоположным знаком.

**Задача 6 (Positive–Negative Partial Set Cover Problem (±PSC)).** Входом, аналогично RBSC, являются множество «красных» (отрицательных) элементов $R$, множество «синих» (положительных) элементов $B$ и набор $\mathcal{D}$ подмножеств множества $R \cup B$. Требуется найти подмножество $\mathcal{D}'$ множества $\mathcal{D}$, которое покрывает как можно больше синих элементов и как можно меньше красных, т. е.

$$|R \cap \mathcal{C}(\mathcal{D})| - |B \cap \mathcal{C}(\mathcal{D})| \underset{\mathcal{D}' \subseteq \mathcal{D}: B \subseteq \mathcal{C}(\mathcal{D}')}{\to} \min.$$

Задача ±PSC изучается в [17].

В случае, когда каждая строка матрицы $L$ имеет вес $\pm 1$, задача 4 эквивалентна ±PSC, в которой красными элементами являются строки матрицы $L$, имеющие вес 1, синими — остальные строки $L$.

В настоящей работе для решения задач 3 и 4 использовался метод ветвей и границ на базе алгоритмов дуализации из [18]. Сложность такого варианта решения не исследовалась. Однако очевидно, что она существенно зависит от размеров входных матриц. Например, даже для сравнительно небольших прикладных задач матрица сравнения

$L_{T\setminus K}(G_1, G_2)$ может иметь достаточно большой размер, поскольку у нее $|T \setminus K||G_2|$ строк и $|\mathcal{U}^*||\mathcal{O}^*|+|G_2\setminus G_1|$ столбцов. Далее рассматриваются методики, позволяющие строить логические корректоры без использовании всех строк и столбцов матриц сравнения, работая только с их подматрицами.

# 4 Повышение эффективности логических корректоров

## 4.1 Построение семейств предикатов методом бустинга

В данном подразделе предлагается и исследуется бустинг-алгоритм построения логического корректора. Применяя бустинг для обучения логического корректора, можно одновременно решить две проблемы: сократить временные затраты и повысить качество распознавания. Временные затраты снижаются благодаря тому, что при поиске предикатов, добавляемых в семейство $Z_K$, вместо всей матрицы сравнения $L_{T\setminus K}(G_1, G_2)$ используется лишь часть ее строк. Качество распознавания логического корректора улучшается за счет настройки весов предикатов, а также построения семейств предикатов с высоким уровнем диверсификации. Под диверсификацией семейства $Z_K$ подразумевается различность входящих в него предикатов. Чем разнообразнее наборы прецедентов, выделяемые различными предикатами семейства, тем лучше распознающая способность алгоритма в целом.

Обозначим через $A_t$ логический корректор, голосующий по предикатам, построенным за $t$, $t \geqslant 0$, итераций. Пусть $S_i \in T$, $y_i$ — номер класса, которому принадлежит $S_i$, и $K \in \mathbb{K}^+$. Обозначим через $\Gamma_t(S_i, K)$ оценку за отнесение объекта $S_i$ к классу $K$, вычисляемую по семействам предикатов $Z_K$ и $Z_{\overline{K}}$ логического корректора $A_t$. Далее понадобится обозначение $M_t(S_i, K) = \Gamma_t(S_i, K_{y_i}) - \Gamma_t(S_i, K)$.

Для числа ошибок и отказов алгоритма $A_t$ на обучении справедливо неравенство

$$Q(A_t) = \sum_{i=1}^{m}[A_t(S_i) \neq y_i] \leqslant \sum_{i=1}^{m} \sum_{K'\neq K_{y_i}, K'\in\mathbb{K}^+} [M_t(S_i, K') \leqslant 0]\,,$$

которое в случае двух классов ($\mathbb{K}^+ = \{K_1, K_2\}$) обращается в равенство.

Построим логический корректор $A_{t+1}$, не меняя предикаты и их веса, найденные на итерациях $1, \ldots, t$. На итерации $t+1$ по некоторому правилу выберем класс $K \in \mathbb{K}^{\pm}$ и сформируем предикат $B$. Добавим $B$ в семейство $Z_K$ с весом $\alpha_B > 0$. Семейства $Z_{K'}$, $K' \neq K$, $K' \in \mathbb{K}^{\pm}$, оставим без изменений.

Предикат $B$ и его вес $\alpha_B$ целесообразно выбирать так, чтобы суммарные потери $Q(A_{t+1})$ были минимальными. Однако решать оптимизационную задачу $Q(A_{t+1}) \to \min$ неудобно. В методе бустинга предлагается вместо нее решать задачу

$$\hat{Q}(A_{t+1}) = \sum_{i=1}^{m} \sum_{K'\neq K_{y_i}, K'\in\mathbb{K}^+} d(M_{t+1}(S_i, K')) \to \min\,,$$

где $d(x)$ — монотонно убывающая, дифференцируемая на $\mathbb{R}$ функция, ограничивающая сверху функцию $f(x) = [x \leqslant 0]$. В этом есть смысл, поскольку верно неравенство $Q(A_{t+1}) \leqslant \hat{Q}(A_{t+1})$.

Рассмотрим случай, когда логический корректор используется в базовом режиме. Для аддитивного режима применимы все приводимые ниже рассуждения с незначительными изменениями.

Введем вспомогательные обозначения:

$$\begin{aligned} D &= \{(S_i, K') \in T \times \mathbb{K}^+ \colon S_i \notin K'\}; \\ D(K) &= \{(S_i, K') \in D \colon K' = K \text{ или } S_i \in K\}, \quad K \in \mathbb{K}^+; \\ D(K) &= D(\overline{K}), \quad K \in \mathbb{K}^-; \\ z_i(K) &= 2[S_i \in K] - 1. \end{aligned}$$

Нетрудно показать, что если на итерации $t+1$ в семейство $Z_K$ добавляется предикат $B$ с весом $\alpha_B$, то выполняется равенство:

$$\hat{Q}(A_{t+1}) = \sum_{(S_i,K')\in D\setminus D(K)} d(M_t(S_i, K')) + \sum_{(S_i,K')\in D(K)} d(M_t(S_i, K') + \alpha_B z_i(K) B(S_i)).$$

По теореме Лагранжа для функции $g(\alpha_B) = d(M_t(S_i, K') + \alpha_B z_i(K) B(S_i))$ при $\alpha_B > 0$ имеет место представление:

$$g(\alpha_B) = g(0) + \alpha_B g'(\xi) = d(M_t(S_i, K')) + \alpha_B z_i(K) B(S_i) d'(M_t(S_i, K') + \xi z_i(K) B(S_i)),$$

где $\xi \in (0, \alpha_B)$. Воспользуемся аппроксимацией

$$g(\alpha_B) \approx d(M_t(S_i, K')) + +\alpha_B z_i(K) B(S_i) d'(M_t(S_i, K'))$$

и вместо $\hat{Q}(A_{t+1})$ будем минимизировать

$$\sum_{(S_i,K')\in D\setminus D(K)} d(M_t(S_i, K')) + \sum_{(S_i,K')\in D(K)} d(M_t(S_i, K')) + \alpha_B z_i(K) B(S_i) d'(M_t(S_i, K')) =$$
$$= \hat{Q}(A_t) - \alpha_B \sum_{(S_i,K')\in D(K)} z_i(K) B(S_i)(-d'(M_t(S_i, K'))). \quad (2)$$

Зафиксируем вес предиката $\alpha_B$ и сопоставим каждому прецеденту $S_i$ вес

$$\tilde{w}_t(S_i, K) = \sum_{(S_i,K')\in D(K)} (-d'(M_t(S_i, K'))).$$

Пусть $G \subseteq T$. Введем обозначение $\tilde{W}_t(G, K) = \sum_{S_i \in G} \tilde{w}_t(S_i, K)$. Нормируем веса объектов:

$$w_t(S_i, K) = \frac{\tilde{w}_t(S_i, K)}{\tilde{W}_t(T, K)}.$$

Для сумм нормированных весов будем использовать аналогичное обозначение: $W_t(G, K) = \sum_{S_i \in G} w_t(S_i, K)$.

Чтобы подчеркнуть, что качество предиката $B$ оценивается на итерации $t$, обозначим:

$$\begin{aligned} P_t(B, K) &= \sum_{S_i \in T \cap K} w_t(S_i, K) B(S_i); \\ N_t(B, K) &= \sum_{S_i \in T \setminus K} w_t(S_i, K) B(S_i); \\ I_t(B, K) &= P_t(B, K) - N_t(B, K). \end{aligned}$$

Поскольку верно равенство

$$\sum_{(S_i,K')\in D(K)} z_i(K)B(S_i)(-d'(M_t(S_i,K'))) = \tilde{W}_t(T,K)I_t(B,K),$$

значение (2) минимально при максимальном значении информативности $I_t(B,K)$. Найдем и зафиксируем предикат $B$ с максимальной информативностью, а затем определим значение веса $\alpha_B$, доставляющее минимум $\hat{Q}(A_{t+1})$.

Наиболее простые выкладки получаются при $d(x) = e^{-x}$. Модель бустинг-алгоритмов с такой функцией потерь носит название AdaBoost. Рассмотрим эту модель более подробно.

В результате несложных преобразований получаем:

$$\hat{Q}(A_{t+1}) = \hat{Q}(A_t) + \tilde{W}_t(T,K)\left((e^{-\alpha_B}-1)P_t(B,K) + (e^{\alpha_B}-1)N_t(B,K)\right). \tag{3}$$

При условии, что $P_t(B,K) > N_t(B,K) > 0$, минимальное значение $\hat{Q}(A_{t+1})$ по $\alpha_B$ достигается в точке

$$\alpha_B = \frac{1}{2}\ln\frac{P_t(B,K)}{N_t(B,K)}.$$

Однако если предикат $B$ корректен для $K$, то $N_t(B,K) = 0$. Чтобы избежать появление неопределенностей, будем вычислять вес предиката $\alpha_B$ по другой формуле, для ввода которой потребуются следующие обозначения:

$$\begin{aligned} w_t^*(G,K) &= \frac{1}{2}\min_{S_i\in G} w_t(S_i,K); \\ N_t^*(B,K) &= \max\{N_t(B,K), w_t^*(T,K)\}. \end{aligned}$$

Если выполнятся неравенство $P_t(B,K) > N_t^*(B,K)$, то вес

$$\alpha_B = \frac{1}{2}\ln\frac{P_t(B,K)}{N_t^*(B,K)} \tag{4}$$

определен и положителен.

Введем вспомогательное обозначение:

$$J_t(B,K) = \frac{\tilde{W}_t(T,K)}{\hat{Q}(A_t)}\left(\sqrt{P_t(B,K)} - \sqrt{N_t^*(B,K)}\right)^2.$$

**Утверждение 10.** *Пусть после $t_0 \geqslant 0$ итераций построен логический корректор $A_{t_0}$ и после $t > t_0$ итераций построен логический корректор $A_t$.*

*Если при построении $A_t$ на каждой итерации $i$, $t_0 < i \leqslant t$, в некоторое семейство $Z_K$ добавлялся предикат $B$ с весом $\alpha_B$, найденным по формуле* (4)*, и при этом всякий раз выполнялось неравенство*

$$J_{i-1}(B,K) > \frac{\ln\hat{Q}(A_{t_0})}{t-t_0},$$

*то распознающий алгоритм $A_t$ корректен.*

**Доказательство.** Подставив (4) в (3), можно убедиться, что верно неравенство

$$\hat{Q}(A_i) \leqslant \hat{Q}(A_{i-1})(1 - J_{i-1}(B,K))\,. \tag{5}$$

Обозначим $\delta = \ln(\hat{Q}(A_{t_0}))/(t - t_0)$. Из (5) получаем цепочку неравенств:

$$Q(A_t) \leqslant \hat{Q}(A_t) < \hat{Q}(A_{t_0})(1-\delta)^{t-t_0} \leqslant \hat{Q}(A_{t_0})e^{-\delta(t-t_0)} = 1.$$

Значение $Q(A_t)$ должно быть целым числом; следовательно, $Q(A_t) = 0$. Доказательство закончено. ■

**Следствие 1.** *Пусть после $t > 0$ итераций построен логический корректор $A_t$. Если при построении $A_t$ на каждой итерации $i$, $1 \leqslant i \leqslant t$, в некоторое семейство $Z_K$ добавлялся предикат $B$ с весом $\alpha_B$, найденным по формуле* (4)*, и при этом всякий раз выполнялось неравенство $J_{i-1}(B,K) > (\ln m)/t$, то распознающий алгоритм $A_t$ корректен.*

Утверждение 10 и его следствие позволяют заменить требование корректности всех используемых при голосовании предикатов другим условием, как правило, более мягким. Далее будет показано, что это зачастую сокращает временные затраты обучения логического корректора за счет использования при поиске предиката подматрицы матрицы сравнения, составленной из относительно небольшой части ее строк.

Пусть $G \subseteq T$, $K \in \mathbb{K}^{\pm}$, $G^+ \subseteq T \cap K$ и $G^- \subseteq T \setminus K$. Введем обозначения:

$$\begin{aligned} W_t^*(G,K) &= \max\{W_t(G,K), w_t^*(G,K)\}\,; \\ \delta_t(G^+,G^-,K) &= \frac{\tilde{W}_t(T,K)}{\hat{Q}(A_t)}\left(\sqrt{W_t(G^+,K)} - \sqrt{W_t^*(T\setminus K\setminus G^-,K)}\right)^2. \end{aligned}$$

**Утверждение 11.** *Пусть $K \in \mathbb{K}^{\pm}$, набор эл.кл. $U$ отделяет набор обучающих объектов $G \subseteq T \cap K$ от набора обучающих объектов $G^- \subseteq T \setminus K$ с помощью набора отношений $O$ и предикат $B = B_{(U,O,G)}$. Тогда верно неравенство $J_t(B,K) \geqslant \delta_t(G,G^-,K)$.*

**Доказательство.** Из условия утверждения и конструкции предиката $B_{(U,O,G)}$ следуют оценки $P_t(B_{(U,O,G)},K) \geqslant W_t(G,K)$ и $N_t^*(B_{(U,O,G)},K) \leqslant W_t^*(T\setminus K\setminus G^-,K)$, которых достаточно для завершения доказательства. ■

Несложно убедиться, что набор эл.кл. $U = ((H_1,\sigma_1),\ldots,(H_d,\sigma_d))$ отделяет обучающие объекты из $G \subseteq T\cap K$ от обучающих объектов из $G^- \subseteq T\setminus K$ с помощью набора отношений $O = (o_1,\ldots,o_d)$ тогда и только тогда, когда набор $((H_1,\sigma_1,o_1),\ldots,(H_d,\sigma_d,o_d))$ покрывает подматрицу $L_{G^-}(G_1^+,G_2^+)$, $G_1^+ \subseteq G \subseteq G_2^+$, матрицы сравнения $L_{T\setminus K}(G_1^+,G_2^+)$. Подматрица $L_{G^-}(G_1^+,G_2^+)$ имеет $|G_2^+||G^-|$ строк.

Процедура `BuildPrettyGoodPredicate`, представленная на схеме алгоритма 1, использует «жадную» стратегию поиска таких $K \in \mathbb{K}^{\pm}$, $G_1^+ \subseteq G_2^+ \subseteq T \cap K$, $G^- \subseteq T \setminus K$, для которых $\delta_t(G_1^+,G^-,K) > \delta$, значение $|G_2^+||G^-|$ близко к наименьшему. Затем по матрице сравнения $L_{G^-}(G_1^+,G_2^+)$ ищется предикат $B = B_{(U,O,G)}$ с наибольшей информативностью $I_t(B,K)$ такой, что $G_1^+ \subseteq G \subseteq G_2^+$ и набор эл.кл. $U$ отделяет прецеденты из $G$ от прецедентов из $G^-$ с помощью набора отношений $O$.

Рассмотрим следующие величины, зависящие от обучающей выборки:

$$\delta^* = \frac{(\sqrt{m}-1)^2}{lm}\,; \quad t^* = \frac{\ln m}{\delta^*}\,.$$

**Алгоритм 1** Построение предиката с достаточно большим значением $J_t(B,K)$

**ПРОЦЕДУРА** `BuildPrettyGoodPredicate`($\mathbb{K}^{\pm}$, $T$, $t$, $\delta$, $r$)
**Параметры:** $\mathbb{K}^{\pm}$ — классы и их дополнения; $T$ — обучающая выборка; $t$ — число выполненных итераций; $\delta > 0$ — пороговый параметр; $r$ — рекомендуемая мощность $G_2^+$;
**Выход:** $B$ — предикат, добавляемый в семейство $Z_K$, $K \in \mathbb{K}^{\pm}$, такой, что $J_t(B,K) > \delta$;

инициализировать $\mathbb{K}_t(\delta) := \{K \in \mathbb{K}^{\pm} \colon \delta_t(T \cap K, T \setminus K, K) > \delta\}$;
**если** $\mathbb{K}_t(\delta) = \varnothing$ **то** $\mathbb{K}_t(\delta) := \mathbb{K}^{\pm}$;
выбрать случайный $K$ из распределения вероятностей $W_t(T \cap K, K), K \in \mathbb{K}_t(\delta)$;
упорядочить объекты $T \cap K$ и $T \setminus K$ по убыванию весов $w_t(S_i, K)$;
найти числа $r_1$ и $r_2$ такие, что
1) набор $G_1^+$, состоящий из первых $r_1$ объектов упорядоченного $T \cap K$ и набор $G^-$, состоящий из первых $r_2$ объектов упорядоченного $T \setminus K$, удовлетворяют условию $\delta_t(G_1^+, G^-, K) > \delta$,
2) значение $r_1 r_2$ минимально и
3) $r_1 \leqslant r$;
**если** найти $r_1$ и $r_2$, удовлетворяющие пункту 3), не удалось, **то**
    найти числа $r_1$ и $r_2$ такие, что выполнен пункт 1) и значения $r_1$ и $r_2$ минимальны;
    в качестве $G_1^+$ и $G_2^+$ взять первые $r_1$ объектов упорядоченного $T \cap K$;
**иначе**
    в качестве $G_1^+$ взять первые $r_1$ объектов упорядоченного $T \cap K$;
    в качестве $G_2^+$ взять первые $r$ объектов упорядоченного $T \cap K$;
в качестве $G^-$ взять первые $r_2$ объектов упорядоченного $T \setminus K$;
по матрице сравнения $L_{G^-}(G_1^+, G_2^+)$ найти предикат $B = B_{(U,O,G)}$ с наибольшей информативностью $I_t(B,K)$;

**Алгоритм 2** Построение логического корректора методом бустинга

**Вход:** $T$ — обучающая выборка;
$t_{\max}$ — максимальное число итераций;
$r$ — рекомендуемая мощность $G_2^+$;
**Выход:** $A_{t_{\max}}$ — логический корректор;

инициализировать семейств предикатов $Z_K := \varnothing$, $\forall K \in \mathbb{K}^{\pm}$;
**для** $t = 1, \ldots, t_{\max}$
    вычислить веса объектов $\tilde{w}_{t-1}(S_i, K') := \exp(-M_{t-1}(S_i, K'))$, $(S_i, K') \in D$;
    найти $K \in \mathbb{K}^{\pm}$ и предикат $B$ для добавления в $Z_K$ вызовом процедуры `BuildPrettyGoodPredicate`($\mathbb{K}^{\pm}$, $T$, $t-1$, $\ln m / t_{\max}$, $r$);
    вычислить вес $\alpha_B$ по формуле 4;
    добавить $B$ в $Z_K$;

**Утверждение 12.** *Если $\delta < \delta^*$, то процедура* `BuildPrettyGoodPredicate` *выбирает $K \in \mathbb{K}^{\pm}$ и строит предикат $B$ такие, что $J_t(B,K) > \delta$.*

**Доказательство.** Заметим, что

$$\delta_t(T \cap K, T \setminus K, K) = \frac{\tilde{W}_t(T,K)}{\hat{Q}(A_t)} \left( \sqrt{W_t(T \cap K, K)} - \sqrt{w_t^*(T \setminus K, K)} \right)^2 .$$

Непосредственной проверкой можно установить, что $\tilde{W}_t(T,K_1)+\cdots+\tilde{W}_t(T,K_l)=2\hat{Q}(A_t)$ и $W_t(T\cap K,K)+W_t(T\cap\overline{K},\overline{K})=1$. Поэтому всегда найдется $K\in\mathbb{K}^{\pm}$, для которого верны неравенства:

$$\frac{\tilde{W}_t(T,K)}{\hat{Q}(A_t)}\geqslant\frac{2}{l};\quad W_t(T\cap K,K)\geqslant\frac{1}{2};\quad w_t^*(T\setminus K,K)\leqslant\frac{1}{2m},$$

из которых выводится неравенство $\delta_t(T\cap K,T\setminus K,K)\geqslant\delta^*$. Это означает, что если выполнены условия доказываемого утверждения, то процедура `BuildPrettyGoodPredicate` выбирает $K$ из $\mathbb{K}^{\pm}$ такой, что $\delta_t(T\cap K,T\setminus K,K)>\delta$ и строит предикат $B$, для которого по утверждению 11 выполняется неравенство $J_t(B,K)>\delta$. Утверждение доказано. ■

Алгоритм 2 демонстрирует, как можно использовать бустинг для построения логического корректора общего вида из предикатов, необязательно являющихся корректными. Однако выход алгоритма 2 не всегда является корректной распознающей процедурой. Сформулируем достаточное условие корректности.

**Теорема 1.** *Если число итераций $t_{\max}>t^*$, то алгоритм 2 строит корректную процедуру распознавания.*

**Доказательство.** Из утверждения 12 следует, что для предиката $B$, строящегося на шаге 4 алгоритма 2, верно неравенство $J_{t-1}(B,K)>\ln m/t_{\max}$, $t\in\{1,\ldots,t_{\max}\}$. Таким образом, справедливы предпосылки следствия из утверждения 10, из которого заключаем, что распознающая процедура $A_{t_{\max}}$ корректна. Теорема доказана. ■

### 4.2 Локальные базисы классов

В данном подразделе рассматривается вопрос сокращения временных затрат поиска предикатов с высокой информативностью за счет отбрасывания части столбцов матрицы сравнения.

Пусть $K\in\mathbb{K}^{\pm}$. Обозначим матрицу сравнения $L_{T\setminus K}(T\cap K,T\cap K)$ через $L_K$. Каждый столбец матрицы сравнения $L_K$ соответствует тройке $(H,\sigma,o)$, где $(H,\sigma)$ — эл.кл. и $o$ — отношение из $\mathcal{O}^*$. Множество всех таких троек обозначим через $\mathcal{V}^*$. Мощность $\mathcal{V}^*$ даже в задаче с небольшим числом признаков может оказаться существенной. Большое число столбцов матрицы сравнения затрудняет поиск корректных предикатов. Предлагается использовать не всю матрицу сравнения, а лишь подматрицу, состоящую из части ее столбцов.

Набор $\mathcal{V}_K=\{(H_1,\sigma_1,o_1),\ldots,(H_d,\sigma_d,o_d)\}$ троек из $\mathcal{V}^*$ будем называть *локальным базисом класса* $K$, если набор эл.кл. $((H_1,\sigma_1),\ldots,(H_d,\sigma_d))$ отделяет прецеденты из $T\cap K$ от прецедентов из $T\setminus K$ с помощью набора отношений $(o_1,\ldots,o_d)$.

Ясно, что $\mathcal{V}_K$ является локальным базисом класса $K$ тогда и только тогда, когда подматрица, составленная из столбцов $\mathcal{V}_K$ матрицы сравнения $L_K$, не имеет нулевых строк, т. е. для этой подматрицы существует покрытие.

Пусть $G_1\subseteq G_2\subseteq T\cap K$. Нетрудно заметить, что если $\mathcal{V}_K$ является локальным базисом класса $K$, то подматрица, составленная из столбцов $\mathcal{V}\cup(G_2\setminus G_1)$ матрицы сравнения $L_{T\setminus K}(G_1,G_2)$, не имеет нулевых строк. Таким образом, «в рамках» локального базиса класса $K$ всегда можно найти корректный для $K$ предикат $B_{(U,O,G)}$ такой, что $G_1\subseteq G\subseteq\subseteq G_2$.

Набор $\mathcal{V}\subseteq\mathcal{V}^*$, являющийся локальным базисом для каждого из классов $K_1,\ldots,K_l$, будем называть *локальным базисом задачи*. Например, набор $\mathcal{V}_1$, состоящий из троек $(H,\sigma,o)\in\mathcal{V}^*$ таких, что эл.кл. $(H,\sigma)$ имеет ранг 1, является локальным базисом задачи.

Опишем универсальный метод построения локального базиса класса, состоящего из эл.кл. произвольного ранга.

Пусть $K \in \mathbb{K}^{\pm}$. Рассмотрим задачу распознавания с двумя классами $K$ и $\overline{K}$. Построим семейство эл.кл. $C_K$ и каждому эл.кл. $(H, \sigma) \in C_K$ присвоим ненулевой вес $\alpha_{(H,\sigma)}$. Рассмотрим распознающую процедуру

$$A_T^K(S) = \operatorname{sign}\Big(\sum_{(H,\sigma)\in C_K} \alpha_{(H,\sigma)}[H(S) = \sigma]\Big), \tag{6}$$

где $\operatorname{sign}(x)$ — функция «знак», возвращающая 1 при $x < 0$, $-1$ при $x < 0$ и 0 при $x = 0$. Будем считать процедуру $A_T^K$ корректной в случае, когда $A_T^K(S_i) = 1$, $\forall S_i \in T \cap K$, и $A_T^K(S_i) = -1$, $\forall S_i \in T \setminus K$. Построим по взвешенному семейству $C_K$ набор $\mathcal{V}_K$ такой, что каждому эл.кл. $(H, \sigma)$ из $C_K$ однозначно соответствует тройка $(H, \sigma, o) \in \mathcal{V}_K$, в которой $o = [x \leqslant y]$ при $\alpha_{(H,\sigma)} > 0$ и $o = [x \geqslant y]$ при $\alpha_{(H,\sigma)} < 0$. Очевидно, что справедливо

**Утверждение 13.** *Если распознающая процедура $A_T^K$ корректна, то набор $\mathcal{V}_K$, построенный по взвешенному семейству эл.кл. $C_K$, является локальным базисом класса $K$, причем упорядоченный набор, составленный из эл.кл. семейства $C_K$, является корректным для $K$ и имеет поляризуемую корректирующую функцию.*

Существует ряд методов построения корректных распознающих процедур вида (6), например бустинг или построение деревьев решений. В [10] лучшим алгоритмом построения локального базиса оказался бустинг-алгоритм BOOSTLO. В настоящей работе используется два метода: голосование по представительным наборам и бустинг-алгоритм, аналогичный BOOSTLO.

Практика показывает, что для прикладной задачи с большой значностью признаков редко удается построить небольшой локальный базис. Заметим, что при использовании бустинга для формирования семейств голосующих предикатов на каждой итерации ищется набор эл.кл. $U$, отделяющий некоторое подмножество прецедентов $G^+$ от подмножества прецедентов $G^-$. При этом совсем необязательно осуществлять поиск набора $U$ в локальном базисе задачи. Целесообразно на каждой итерации формировать локальный базис, ориентированный на отделение $G^+$ от $G^-$ и учитывающий текущие веса остальных прецедентов.

Были реализованы и протестированы 4 модификации логического корректора общего вида (каждая из модификаций для формирования семейств голосующих предикатов использует бустинг):

1. ОЛК1 — логический корректор, в котором предикаты строятся в рамках локального базиса задачи, состоящего из троек $(H, \sigma, o)$ таких, что ранг эл.кл. $(H, \sigma)$ равен 1 и $o \in \in \{[x \leqslant y], [x \geqslant y]\}$;
2. ОЛК2 — логический корректор, в котором предикаты строятся в рамках локального базиса задачи, построенного бустинг-алгоритмом;
3. ОЛК3 — логический корректор, в котором предикаты строятся в рамках локального базиса, формируемого на каждой итерации голосованием по представительным наборам;
4. ОЛК4 — логический корректор, в котором предикаты строятся в рамках локального базиса, формируемого на каждой итерации бустинг-алгоритмом.

## 5 Эксперименты

Новые логические корректоры были протестированы на прикладных задачах из репозитория UCI. В табл. 1 даны характеристики задач. В столбцах $l$, $m$, $n$ и $z$ приведены соответственно число классов, число строк, число столбцов и число всех представительных наборов ранга 1, характеризующее значность признаков.

Задачи по трудоемкости можно разбить на 3 группы. Задачи с номерами 1–11 имеют средний объем обучения, и поэтому для тестирования на этих задачах применяется методика 10-кратного скользящего контроля. В задачах 12 и 13 много объектов, поэтому для сокращения времени счета выборка делится только 1 раз на обучающую и тестовую. В задачах 14 и 15, наоборот, очень мало объектов, поэтому используется методика скользящего контроля по одному объекту (leave-one-out).

В тестировании помимо логических корректоров ОЛК1–ОЛК4 участвовали следующие алгоритмы распознавания:

1) Т — голосование по тестам (для каждого класса строится не более 200 тестов);
2) ПН — голосование по представительным наборам (для каждого объекта строится не более 5 представительных наборов);
3) МОН — голосование по монотонным корректным наборам эл.кл. (для каждого класса строится не более 200 наборов и эл.кл. имеют ранг 1);
4) Т* — голосование по тестам (голосующие семейства формируются бустингом);
5) ПН* — голосование по представительным наборам (голосующие семейства формируются бустингом);
6) МОН* — голосование по монотонным корректным наборам эл.кл. (голосующие семейства формируются бустингом и эл.кл. в наборах имеют ранг 1).

В табл. 2 приведены результаты счета. Показателем качества является средняя доля ошибок на тестовых выборках. Прочерки соответствуют случаям, когда алгоритм не справился с задачей за 1 ч. Время счета представлено в табл. 3.

**Таблица 1** Задачи

| № | Название | $l$ | $m$ | $n$ | $z$ |
|---|---|---|---|---|---|
| 1. | audiology | 24 | 226 | 69 | 161 |
| 2. | balance scale | 3 | 625 | 4 | 20 |
| 3. | breast cancer | 2 | 699 | 9 | 90 |
| 4. | car | 4 | 1728 | 6 | 21 |
| 5. | dermatology | 4 | 366 | 34 | 192 |
| 6. | house votes | 2 | 435 | 16 | 48 |
| 7. | kr vs kp | 2 | 3196 | 36 | 73 |
| 8. | monks-2 | 2 | 601 | 6 | 17 |
| 9. | nursery | 5 | 12960 | 8 | 27 |
| 10. | soybean large | 19 | 307 | 35 | 132 |
| 11. | tic tac toe | 2 | 958 | 9 | 27 |
| 12. | optdigits | 10 | 5620 | 64 | 914 |
| 13. | letter recognition | 26 | 20000 | 16 | 256 |
| 14. | lenses | 3 | 24 | 4 | 9 |
| 15. | soybean small | 4 | 47 | 35 | 72 |

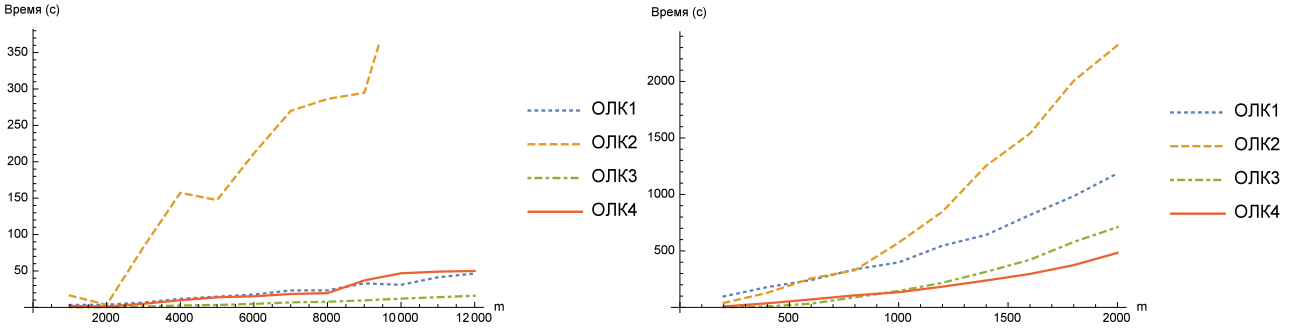
**Таблица 2** Средняя частота ошибок на тестовой выборке

| № | Задача | Классические | | | Бустинг | | | Логические корректоры общего вида | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | T | ПН | МОН | T* | ПН* | МОН* | ОЛК1 | ОЛК2 | ОЛК3 | ОЛК4 |
| 1. | audiology | 0,14 | 0,07 | 0,09 | 0,03 | 0,03 | 0,03 | 0,03 | 0,03 | **0,02** | 0,03 |
| 2. | b. scale | 0,92 | 0,27 | 0,46 | 0,25 | 0,2 | 0,19 | **0,18** | 0,23 | 0,23 | 0,25 |
| 3. | b. cancer | 0,21 | 0,05 | 0,24 | 0,046 | **0,044** | 0,057 | 0,061 | 0,059 | 0,065 | 0,059 |
| 4. | car | 0,97 | 0,09 | 0,27 | 0,061 | 0,032 | 0,033 | 0,013 | 0,027 | 0,022 | **0,011** |
| 5. | dermat. | 0,84 | 0,47 | 0,79 | 0,41 | 0,4 | 0,4 | **0,39** | 0,42 | 0,44 | 0,43 |
| 6. | h. votes | 0,34 | 0,06 | 0,15 | 0,07 | **0,05** | 0,07 | **0,05** | 0,06 | 0,07 | 0,08 |
| 7. | kr-vs-kp | 0,63 | 0,017 | 0,101 | 0,008 | 0,004 | **0,003** | 0,008 | 0,007 | 0,004 | **0,003** |
| 8. | monks-2 | 0,96 | 0,52 | 0,96 | 0,37 | 0,55 | 0,42 | **0,04** | 0,44 | 0,56 | 0,36 |
| 9. | nursery | 0,66 | 0,015 | 0,36 | 0,027 | 0,003 | 0,005 | 0,002 | — | **0,0019** | 0,004 |
| 10. | soybean l. | 0,19 | 0,094 | 0,131 | 0,075 | **0,064** | 0,072 | 0,078 | 0,106 | 0,083 | 0,075 |
| 11. | tic-tac-toe | 0,97 | 0,005 | 0,52 | 0,011 | 0,002 | 0,005 | 0,028 | 0,002 | **0,001** | 0,007 |
| 12. | letter r. | 0,52 | 0,21 | 0,63 | 0,21 | **0,16** | 0,25 | — | — | 0,23 | 0,25 |
| 13. | optdigits | 0,77 | 0,19 | 0,55 | 0,25 | 0,23 | 0,17 | 0,15 | — | 0,27 | **0,14** |
| 14. | lenses | 1 | **0,21** | 0,46 | 0,42 | 0,25 | 0,29 | 0,33 | 0,29 | 0,38 | 0,25 |
| 15. | soybean s. | 0,02 | **0** | **0** | **0** | **0** | **0** | **0** | 0,02 | 0,04 | **0** |

**Таблица 3** Время счета в секундах

| № | Задача | Классические | | | Бустинг | | | Логические корректоры общего вида | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | T | ПН | МОН | T* | ПН* | МОН* | ОЛК1 | ОЛК2 | ОЛК3 | ОЛК4 |
| 1. | audiology | 1,9 | **1,4** | 4,6 | 22,7 | 8,9 | 42,2 | 224,1 | 420,2 | 4,1 | 82,4 |
| 2. | b. scale | 0,6 | **0,4** | 2,4 | 0,8 | 1,1 | 2,8 | 132,5 | 1251,1 | 3,9 | 243,7 |
| 3. | b. cancer | 1,2 | **0,2** | 7,8 | 0,7 | 0,5 | 5,2 | 108,1 | 110,1 | 1,3 | 51,3 |
| 4. | car | 3,1 | **1,3** | 10,1 | 2,3 | 3,1 | 7,1 | 78,5 | 713,7 | 7,9 | 34,9 |
| 5. | dermat. | **2,8** | 15,4 | 13,2 | 40,9 | 66,5 | 118,9 | 272,4 | 689,6 | 98,9 | 345,1 |
| 6. | h. votes | 2,4 | **1,1** | 7,6 | 4,2 | 8,6 | 12,1 | 37,1 | 87,3 | 7,9 | 167,3 |
| 7. | kr-vs-kp | 36,3 | **10,2** | 94,4 | 58,9 | 79,8 | 87,5 | 226,1 | 192,1 | 84,8 | 173,6 |
| 8. | monks-2 | **0,5** | 0,6 | 1,2 | 0,9 | 1,9 | 2,1 | 15,4 | 500,6 | 5,8 | 104,1 |
| 9. | nursery | 163,9 | **20,9** | 595,5 | 87,9 | 89,3 | 224,4 | 452,9 | — | 157,9 | 589,1 |
| 10. | soybean l. | 2,5 | **2,4** | 7,7 | 28,3 | 21,2 | 79,9 | 329,3 | 868,6 | 15,9 | 249,2 |
| 11. | tic-tac-toe | 3,2 | **0,6** | 6,5 | 4,5 | 1,9 | 10,2 | 45,6 | 9,3 | 2,2 | 16,2 |
| 12. | letter r. | 58,2 | **47,1** | 790,7 | 92,3 | 233,1 | 550,5 | — | — | 363,1 | 1191,1 |
| 13. | optdigits | **25,8** | 636,2 | 277,7 | 249,6 | 1570,5 | 840,6 | 3117,2 | — | 2160,6 | 1110,8 |
| 14. | lenses | **0,01** | **0,01** | 0,03 | **0,01** | 0,03 | 0,06 | 0,09 | 4,7 | 0,03 | 2,2 |
| 15. | soybean s. | 0,4 | **0,06** | 1,1 | 0,8 | 0,1 | 1,5 | 4,8 | 0,3 | 0,1 | 0,4 |

На 14 задачах лидируют алгоритмы, в которых применяется бустинг для формирования голосующих семейств. На 11 задачах лидируют новые модели. Лучшими среди новых являются ОЛК3 и ОЛК4, в которых локальный базис формируется на каждой итерации, причем эти логические корректоры демонстрируют хорошие результаты на задачах с большой значностью признаков и имеют сравнительно небольшое время счета почти на всех задачах.




(а) nursery (б) optdigits

**Рис. 2** Зависимость времени обучения логических корректоров от размера выборки

Для выявления наилучшей стратегии построения локального базиса с точки зрения времени счета проведена следующая серия экспериментов. Выбраны две задачи с достаточно большим числом объектов: nursery и optdigits. Задача nursery отличается от задачи optdigits тем, что имеет сравнительно небольшую значность и существенно неравномерное распределение объектов по классам. Для задачи nursery было сформировано 60 случайных подвыборок по 5 подвыборок каждого из размеров $1000, 2000, \ldots, 12000$. Для задачи optdigits было сформировано 50 подвыборок по 5 подвыборок каждого из размеров $200, 400, \ldots, 2000$. На рис. 2, *а* и 2, *б* изображены графики зависимости усредненного времени счета логических корректоров ОЛК1–ОЛК4 от размера подвыборки.

Очевидно, самой неудачной является модификация ОЛК2. Время работы ОЛК2 быстро увеличивается с ростом объема обучения, напрямую связанного с мощностью строящегося логическим корректором локального базиса задачи.

На задаче nursery ОЛК3 является наилучшим. Строящиеся логическим корректором ОЛК3 локальные базисы имеют небольшую мощность, поскольку в задачах с малой значностью признаков, как правило, представительные наборы имеют высокую информативность.

На задаче optdigits ОЛК4 обгоняет ОЛК3, начиная с размера подвыборки 800. Бустинг-алгоритм, использующийся в ОЛК4 для формирования локального базиса, не требует корректности эл.кл. Большая значность признаков в задаче optdigits приводит к тому, что в локальный базис, формируемый ОЛК3, попадает много малоинформативных представительных наборов, что плохо сказывается на времени счета.

## 6 Заключение

В работе введены понятия корректного и представительного предиката. С помощью этих понятий сформулированы классические определения теста, представительного набора, корректного эл.кл., а также определение корректного набора эл.кл.

Предложен способ конструирования корректного предиката по корректному набору эл.кл., учитывающий характер монотонности корректирующей функции по каждой ее переменной. С каждым эл.кл. набора связывается определенное отношение, использующе-еся при сравнении прецедентов с распознаваемыми объектами. Приведены условия, при которых набор эл.кл. имеет поляризуемую или монотонную корректирующую функцию. Приведен пример модельной задачи, на котором явно демонстрируются преимущества предикатов введенной конструкции.

Построена общая модель голосования по представительным предикатам, в терминах которой могут быть описаны процедуры голосования по корректным эл.кл. и по корректным наборам эл.кл.

Поиск корректных предикатов с максимальной информативностью сведен к дискретным задачам, являющимся специальными случаями известной задачи о покрытии булевой матрицы. Сложность решения этих задач существенно зависит от размеров входной матрицы. Входной матрицей при обучении логических корректоров является специальная матрица сравнения, строящаяся по прецедентной информации. Предложено использовать не всю матрицу сравнения, а лишь ее небольшую подматрицу. Набор столбцов этой подматрицы формируется путем построения локального базиса. Набор строк меняется итеративно в зависимости от весов объектов, вычисляемых бустинг-алгоритмом. Предикаты, строящиеся по подматрице, вообще говоря, не являются корректными. Однако получена теоретическая оценка числа итераций бустинг-алгоритма, достаточного для формирования семейств предикатов, голосование по которым является корректной процедурой распознавания.

Эксперименты показали, что логические корректоры общего вида на большинстве тестовых задач ошибаются реже ранее построенных моделей. Преимущество особенно заметно на задачах с большой значностью.

На время счета влияет общая стратегия и алгоритм формирования локального базиса. В случае, когда локальный базис строится для всей задачи и не меняется на последующих итерациях, его мощность, а следовательно и время обучения, оказываются достаточно большими. Если же локальный базис перестраивать на каждой итерации, настраиваясь на объекты, которые вызывают у ранее построенных предикатов наибольшие затруднения, то мощность локального базиса, как правило, не велика.

Одним из дальнейших направлений исследования видится обобщение предложенных моделей на случай, когда на множестве значений признаков заданы определенные отношения порядка. Практический интерес представляют цепи, решетки, полурешетки.

## Литература

[1] *Дмитриев А. И., Журавлев Ю. И., Кренделев Ф. П.* Об одном принципе классификации и прогноза геологических объектов и явлений // Известия Сиб. отд. АН СССР, Геология и геофизика, 1968. Т. 5. С. 50–64.

[2] *Баскакова Л. В., Журавлев Ю. И.* Модель распознающих алгоритмов с представительными наборами и системами опорных множеств // ЖВМ и МФ, 1981. Т. 21. № 5. С. 1264–1275.

[3] *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.

[4] *Журавлев Ю. И* Об алгоритмах распознавания с представительными наборами (о логических алгоритмах) // ЖВМ и МФ, 2002. Т. 42. № 9. С. 1425–1435.

[5] *Дюкова Е. В., Журавлев Ю. И., Песков Н. В., Сахаров А. А.* Обработка вещественнозначной информации логическими процедурами распознавания // Искуственный интеллект, НАН Украины, 2004. № 2. С. 80–85.

[6] *Журавлев Ю. И.* Корректные алгебры над множеством некорректных (эвристических) алгоритмов. Ч. I // Кибернетика, 1977. Т. 13. №. 4. С. 5–17.

[7] *Воронцов К. В.* Оптимизационные методы линейной и монотонной коррекции в алгебраическом подходе к проблеме распознавания // ЖВМ и МФ, 2000. Т. 40. № 1. С. 166–176.

[8] *Дюкова Е. В., Журавлев Ю. И., Рудаков К. В.* Об алгебро-логическом синтезе корректных процедур распознавания на базе элементарных алгоритмов // ЖВМ и МФ, 1996. Т. 36. № 8. С. 215–223.

[9] *Dyukova E. V., Zhuravlev Yu. I., Sotnezov M. R.* Construction of an ensemble of logical correctors on the basis of elementary classifiers // Pattern Recogn. Image Anal., 2011. Vol. 21. No. 4. P. 599–605.

[10] *Dyukova E. V., Prokofjev P. A.* Models of recognition procedures with logical correctors // Pattern Recogn. Image Anal., 2013. Vol. 23. No. 2. P. 235–244.

[11] *Дюкова Е. В., Любимцева М. М., Прокофьев П. А.* Об алгебро-логической коррекции в задачах распознавания по прецедентам // Машинное обучение и анализ данных, 2013. Т. 1. № 6. С. 705–713.

[12] *Любимцева М. М.* Логические корректоры в задачах распознавания // Сб. тезисов лучших дипломных работ факультета ВМК МГУ 2014 года. — М: МАКС ПРЕСС, 2014. С. 47–49.

[13] *Воронцов К. В.* О проблемно-ориентированной оптимизации базисов задач распознавания // ЖВМ и МФ, 1998. Т. 38. № 5. С. 870–880.

[14] *Schapire R. E., Singer Y.* Improved boosting using confidence-rated predictions // Machine Learning, 1999. Vol. 37. No. 3. P. 297–336.

[15] *Carr R. D., Doddi S., Konjevod G., Marathe M. V.* On the red-blue set cover problem // 11th ACM-SIAM Symposium on Discrete Algorithms Proceedings, 2000. P. 345–353.

[16] *Peleg D.* Approximation algorithms for the label-cover max and red-blue set cover problems // J. Discrete Algorithms, 2007. Vol. 5. No. 1. P. 55–64.

[17] *Miettinen P.* On the positive–negative partial set cover problem // Inform. Proc. Lett., 2008. Vol. 108. No. 4. P. 219–221.

[18] *Дюкова Е. В., Прокофьев П. А.* Построение и исследование новых асимптотически оптимальных алгоритмов дуализации // Машинное обучение и анализ данных, 2014. Т. 1. № 8. С. 1048–1067.

## References

[1] Dmitriev, A. N., Yu. I. Zhuravlev, and F. P. Krendelev. 1966. Mathematical principles of classification of patterns and scenes. *Discrete Anal. (Inst. Mat. Sib. Otd. Akad. Nauk SSSR, Novosibirsk)* 7:3–11.

[2] Baskalova, L. V., and Yu. I. Zhuravlev. 1981. A model of recognition algorithms with representative samples and systems of supporting sets. *Comput. Math. Math. Phys.* 21(5):189–199.

[3] Dyukova, E. V., and N. V. Peskov. 2002. Search for informative fragments in descriptions of objects in discrete recognition procedures. *Comput. Math. Math. Phys.* 42(5):711–723.

[4] Zhuravlev, Yu. I. 2002. Recognition algorithms with representative sets (logic algorithms). *Comput. Math. Math. Phys.* 42(9):1372–1382.

[5] Djukova, E. V., Yu. I. Zhuravlev, N. V. Peskov, and A. A. Saharov. 2004. Processing a real-valued information with logical recognition procedures. *Artificial Intelligence* 2:80–85. (In Russian.)

[6] Zhuravlev, Yu. I. 1977. Correct algebras over sets of incorrect (heuristic) algorithms. I. *Cybernetics* 13(4):489–497.

[7] Vorontsov, K. V. 1998. Optimization methods for linear and monotone correction in the algebraic approach to the recognition problem // *Comput. Math. Math. Phys.* 40(1):159–168.

[8] Dyukova, E. V., Yu. I. Zhuravlev, and K. V. Rudakov. 1996. Algebraic-logic synthesis of correct recognition procedures based on elementary algorithms. *Comput. Math. Math. Phys.* 36(8):1161–1167.

[9] Dyukova, E. V., Yu. I. Zhuravlev, and M. R. Sotnezov. 2011. Construction of an ensemble of logical correctors on the basis of elementary classifiers. *Pattern Recogn. Image Anal.* 21(4):599–605.

[10] Dyukova, E. V., and P. A. Prokofjev. 2013. Models of recognition procedures with logical correctors. *Pattern Recogn. Image Anal.* 23(2):235–244.

[11] Djukova, E. V., M. M. Lyubimtseva, and P. A. Prokofjev. 2013. Algebraic-logical correction in recognition problems. *J. Mach. Learn. Data Anal.* 1(6):705–713.

[12] Lyubimtseva, M. M. 2014. Logical correctors in pattern recognition. *Abstracts of the best theses of the Faculty CMC MSU 2014*. 47–49. (In Russian.)

[13] Vorontsov, K. V. 1998. The task-oriented optimization of bases in recognition problems. *Comput. Math. Math. Phys.* 38(5):838–847.

[14] Schapire, R. E., and Y. Singer. 1999. Improved boosting using confidence-rated predictions. *Machine Learning* 37(3):297–336.

[15] Carr, R. D., S. Doddi, G. Konjevod, and M. V. Marathe. 2000. On the red-blue set cover problem. *11th ACM-SIAM Symposium on Discrete Algorithms Proceedings*. 345–353.

[16] Peleg, D. 2007. Approximation algorithms for the label-cover max and red-blue set cover problems. *J. Discrete Algorithms* 5(1):55–64.

[17] Miettinen, P. On the positive–negative partial set cover problem. 2008. *Inform. Proc. Lett.* 108(4):219–221.

[18] Djukova, E. V., and P. A. Prokofjev. 2014. Construction and investigation of new asymptotically optimal algorithms for dualization. *J. Mach. Learn. Data Anal.* 1(8):1048–1067.