Российская Академия Наук
Научный совет по комплексной проблеме «Кибернетика»

На правах рукописи

Песков Николай Владимирович

ПОИСК ИНФОРМАТИВНЫХ ФРАГМЕНТОВ ОПИСАНИЙ ОБЪЕКТОВ В ЗАДАЧАХ РАСПОЗНАВАНИЯ

05.13.17 — теоретические основы информатики
диссертация на соискание учёной степени

кандидата физико-математических наук

Научный руководитель:
д.ф.-м.н. Е.В. Дюкова

Москва – 2004

Оглавление

Введение

Рассматриваются задачи, в которых практически невозможно построение математических моделей в общепринятом смысле. Этими задачами являются задачи распознавания на основе прецедентов.

Стандартная постановка задачи распознавания заключается в следующем [65]. Исследуется некоторое множество объектов M. Объекты этого множества описываются системой признаков $\{x_1, \ldots, x_n\}$. Известно, что множество M представимо в виде объединения непересекающихся подмножеств (классов) $K_1, \ldots, K_l$. Имеется конечный набор объектов $\{S_1, \ldots, S_m\}$ из M, о которых известно, каким классам они принадлежат (это прецеденты или обучающие объекты). Требуется по предъявленному набору значений признаков, т.е. описанию некоторого объекта S из M , о котором, вообще говоря, неизвестно, какому классу он принадлежит, определить этот класс.

Для решения прикладных задач распознавания успешно применяются методы, основанные на комбинаторном анализе признаковых описаний объектов, которые особенно эффективны в случае, когда информация целочисленная и число допустимых значений каждого признака невелико. При конструировании этих методов используется аппарат дискретной математики, в частности, булевой алгебры, теории дизъюнктивных нормальных форм и теории покрытий булевых и целочисленных матриц. Основополагающими работами являются работы Ю.И. Журавлева, С.В. Яблонского и М.Н. Вайнцвайга [8, 13, 31, 85].

Главной особенностью рассматриваемых процедур распознавания, называемых в дальнейшем дискретными или логическими процедурами, является возможность получения результата при отсутствии информации о функциях распределения значений признаков и при наличии малых обучающих выборок. Не требуется также задание метрики в пространстве описаний объектов. В данном случае для каждого признака определяется бинарная функция близости между его значениями, позволяющая различать объекты и их подописания.

Основной задачей при построения дискретных процедур распознавания является поиск информативных подописаний (или фрагментов описаний) объектов. Информативными считаются такие фрагменты, которые отражают определенные закономерности в описаниях обучающих объектов, т.е. наличие или, наоборот, отсутствие этих фрагментов в классифицируемом объекте позволяет судить о его принадлежности тому или иному классу. В классических дискретных процедурах распознавания информативными считаются такие фрагменты, которые встречаются в описаниях объектов одного класса, но не встречаются в описаниях объектов остальных классов. Рассматриваемые фрагменты, как правило, имеют содержательное описание в терминах прикладной области, в которой решается задача,

и поэтому построенный алгоритм распознавания также легко интерпретируется. Однако выделение информативных подописаний во многих случаях оказывается сложным в силу чисто вычислительных трудностей переборного характера. Как правило, задача сводится к поиску тупиковых покрытий булевых и целочисленных матриц и может быть также сформулирована как задача построения нормальной формы логической функции [52, 85].

Наличие большого перебора, а также первоначально низкая производительность вычислительной техники явились причиной того, что основные усилия в течении многих лет были направлены на разработку общей теории сложности решения задач дискретного анализа информации и синтеза асимптотически оптимальных алгоритмов поиска информативных фрагментов. Полученные в данном направлении результаты позволили в определенной степени преодолеть указанные трудности и значительно усовершенствовать такие классические модели как тестовый алгоритм и алгоритм голосования по представительным наборам. Здесь следует отметить работы В.А. Слепян, В.Н. Носкова, Е.В. Дюковой и А.А. Андреева [1-3, 36-52, 74, 75, 81-83]. При этом вопросам качества распознавания не уделялось достаточного внимания. Укажем некоторые проблемы, от решения которых зависит результат распознавания.

При построении классических дискретных процедур вводится понятие элементарного классификатора. Под элементарным классификатором понимается фрагмент описания обучающего объекта. Для каждого класса строится некоторое множество элементарных классификаторов с заранее заданными свойствами и, как правило используются элементарные классификаторы, которые встречаются в описаниях объектов одного класса и не встречаются в описаниях объектов других классов, т.е характеризуют лишь некоторые из обучающих объектов данного класса. С другой стороны, наборы значений признаков, не встречающиеся в описании ни одного из обучающих объектов класса, характеризуют все объекты данного класса и с этой точки зрения являются более информативными. Поэтому актуальным является вопрос конструирования распознающих процедур, основанных на принципе «невстречаемости» наборов из допустимых значений признаков.

Одной из центральных проблем является наличие шумящих признаков, т.е. таких признаков, значения которых редко встречаются во всех классах. В частности, шумящими являются признаки, принимающие много значений. Такие признаки порождают очень большое число фрагментов, встречающихся только в одном классе, и с формальной точки зрения являющихся информативными. Однако, каждый из указанных фрагментов крайне редко встречается и в том классе, который он представляет, поэтому про него нельзя сказать, что он являются значимым.

Другая проблема - наличие в обучающей выборке объектов, лежащих на гра-

нице между классами. Каждый такой объект не является "типичным"для своего класса, поскольку его описание похоже на описания объектов из других классов. Наличие нетипичных объектов увеличивает длину фрагментов, различающих объекты из разных классов. Длинные фрагменты реже встречаются в новых объектах, тем самым увеличивается число нераспознанных объектов.

Необходимость построения эффективных реализаций для дискретных процедур распознавания напрямую связана и с вопросами изучения метрических (количественных) свойств множества информативных фрагментов. Важными задачами являются задачи оценки числа покрытий булевых и целочисленных матриц и числа допустимых и максимальных конъюнкций логических функций.

Основной целью диссертационной работы является разработка новых, эффективных в вычислительном плане, подходов к конструированию распознающих процедур дискретного характера, позволяющих повысить качество распознавания и в определенной степени решить указанные выше проблемы. Предложены методы, давшие возможность построить новые более совершенные модели и получить новые результаты, касающиеся исследования метрических свойств дискретных распознающих процедур.

В диссертационной работе введено понятие элементарного классификатора более общего вида, что позволило построить модели основанные на принципе «невстречаемости» набора из допустимых значений признака в описаниях рассматриваемых объектов. А именно предложены две новые модели алгоритмов: алгоритм голосования по антипредставительным наборам и алгоритм голосования по покрытиям классов. Практические эксперименты показали, что в определенных случаях данные алгоритмы имеют преимущество перед классическим алгоритмом голосования по представительным наборам.

Разработаны подходы к повышению эффективности алгоритмов распознавания дискретного характера, основанные на выделении для каждого класса типичных значений признаков, типичных обучающих объектов и построении информативных зон. Данные подходы позволяют снизить влияние шумящих признаков, а также повысить качество распознавания алгоритма в случае, когда в обучающей информации содержится много объектов лежащих на границе между классами.

При этом под качеством распознавания понимается качество алгоритма вне обучающей выборки (способность алгоритма к обобщению или экстраполяции), которое в данной работе оценивается по проценту правильно распознанных объектов при проведении процедуры скользящего контроля. На исследованных в работе прикладных задачах предложенные методы позволили повысить качество распознавания на 11-14%.

Получены новые результаты, касающиеся изучения метрических свойств множества покрытий целочисленной матрицы. Эти результаты использованы для

нахождения асимптотик типичных значений числа допустимых конъюнкций и типичного ранга допустимой конъюнкции двузначной логической функции заданной множеством нулей, а также оценок аналогичных характеристик множества максимальных конъюнкций.

Перейдем к более подробному изложению результатов, полученных в диссертационной работе.

При решении прикладных задач достоверная информация о структуре множества M, как правило, отсутствует, поэтому при построении алгоритма распознавания мы не можем гарантировать качество работы этого алгоритма на новых (отличных от $S_1, \ldots, S_m$) объектах. Однако, если обучающие примеры достаточно характерны для исследуемого множества объектов, то алгоритм, редко ошибающийся на обучении, будет давать неплохие результаты и на неизвестных (не входящих в обучающую выборку) объектах. В связи с этим большое внимание уделяется проблеме корректности распознающих алгоритмов. Алгоритм является корректным, если все объекты из обучающей выборки он распознает правильно.

Простейшим примером корректного алгоритма является следующий. Распознаваемый объект S сравнивается с каждым из объектов обучения $S_1, \ldots, S_m$. В случае, если описание объекта S совпадает с описанием обучающего объекта S_i, объект S относится к тому классу, которому принадлежит объект S_i, в противном случае алгоритм отказывается от распознавания. Нетрудно видеть, что описанный алгоритм является корректным, однако он не сможет распознать ни один объект, описание которого не совпадает с описанием ни одного из обучающих объектов.

Очевидно, что требование полного совпадения описаний распознаваемого объекта и одного из обучающих объектов является слишком осторожным. Анализ прикладных задач свидетельствует о том, что вопрос о близости объектов и их принадлежности одному классу можно решать на основе сравнения некоторого множества их подописаний. Поэтому возникает вопрос, как выбирать поднаборы признаков, порождающие такие подописания, по которым будут сравниваться объекты. Один из вариантов ответа на данный вопрос используется в модели алгоритмов вычисления оценок (АВО) [60, 65].

Введем следующие обозначения. Пусть H - некоторый набор из r, $r \leq n$, различных целочисленных признаков вида $\{x_{j_1}, \ldots, x_{j_r}\}$. Близость объектов $S' = (a'_1, a'_2, \ldots, a'_n)$ и $S'' = (a''_1, a''_2, ..., a''_n)$ из M по набору признаков H будем оценивать величиной

$$B(S', S'', H) = \begin{cases} 1, & \text{если } a'_{j_t} = a''_{j_t} \text{ при } t = 1, 2, ..., r; \\ 0, & \text{в противном случае.} \end{cases}$$

Принципиальная схема построения алгоритмов АВО следующая. В системе признаков $\{x_1, \ldots, x_n\}$ выделяется совокупность различных подмножеств вида

$H = \{x_{j_1}, \ldots, x_{j_r}\}$, $r \leq n$, не обязательно одинаковой мощности. В дальнейшем выделенные подмножества называются опорными множествами алгоритма, а вся их совокупность обозначается через Ω. Задаются параметры: γ_i - параметр, характерезующий представительность объекта S_i, $i = 1, 2, \ldots, m$; P_H - параметр, характерезующий представительность набора (опорного множества) H, $H \in \Omega$. Далее проводится процедура голосования или вычисления оценок. Распознаваемый объект S сравнивается с каждым обучающим объектом S_i по каждому опорному множеству. Считается, что объект S получает голос за принадлежность классу K, если $S_i \in K$ и описания объектов S и S_i совпадают по опорному множеству H (в этом случае $B(S, S_i, H) = 1$). Для каждого класса K, $K \in \{K_1, \ldots, K_l\}$, вычисляется оценка принадлежности $\Gamma(S, K)$ объекта S классу K, которая имеет вид

$$\Gamma(S, K) = \frac{1}{|W_K|} \sum_{S_i \in K} \sum_{H \in \Omega} \gamma_i \cdot P_H \cdot B(S, S_i, H),$$

где $|W_K| = |K \cap \{S_1, \ldots, S_m\}|$[1]

Объект S относится к тому классу, который имеет наибольшую оценку. Если классов с наибольшей оценкой несколько, то происходит отказ от распознавания. Очевидно, что построенный алгоритм не всегда является корректным. Для корректности этого алгоритма требуется выполнение системы линейных неравенств указанного ниже вида.

Для простоты пусть $l = 2$, $S_i \in K_1$ при $1 \leq i \leq m_1$, $S_i \in K_2$ при $m_1+1 \leq i \leq m$, $1 \leq m_1 \leq m-1$. Тогда система неравенств имеет вид

$$\begin{aligned}
&\Gamma(S_1, K_1) > \Gamma(S_1, K_2),\\
&\ldots\ldots\ldots\\
&\Gamma(S_{m_1}, K_1) > \Gamma(S_{m_1}, K_2),\\
&\Gamma(S_{m_1+1}, K_2) > \Gamma(S_{m_1+1}, K_1),\\
&\ldots\ldots\ldots\\
&\Gamma(S_m, K_2) > \Gamma(S_m, K_1).
\end{aligned}$$

Решение системы сводится к выбору параметров γ_i, $i = 1, 2, \ldots, m$, и P_H, $H \in \Omega$. В случае, если система несовместна, находится ее максимальная совместная подсистемы и из решения этой подсистемы определяются значения параметров γ_i и P_H.

Другой способ добиться корректности алгоритма - выбрать «хорошую» систему опорных множеств. В частности, выбрать ее так, чтобы для любого обучающего объекта $S' \notin K$ было выполнено $\Gamma(S', K) = 0$ и для любого обучающего

[1]здесь и далее $|Q|$- мощность множества Q.

объекта $S'' \in K$ было выполнено $\Gamma(S'', K) > 0$. Это можно сделать следующим образом.

Пусть H - некоторое опорное множество. Набор признаков H назовем тестом, если для любых обучающих объектов S' и S'', принадлежащих разным классам, выполнено $B(S', S'', H) = 0$. Другими словами, тест - это набор признаков, по которому различаются любые два объекта из разных классов.

Пусть Ω_T - некоторая совокупность тестов. Если совокупность опорных множеств алгоритма состоит из тестов, то очевидно, такой алгоритм является корректным при любых положительных значениях параметров γ_i, $i = 1, 2, \ldots, m$, и P_H, $H \in \Omega_T$.

Если набор признаков H_1 - тест, то любой набор признаков H_2 такой, что $H_1 \subset H_2$, также является тестом. При этом если объекты близки по H_2, то они будут близки и по H_1, если же два объекта близки по набору столбцов H_1, то они не всегда будут близки по H_2. В этом смысле более короткие тесты обладают большей информативностью, и разумно ограничивать длины тестов или строить тупиковые тесты.

Набор признаков H назовем тупиковым тестом, если выполнены следующие два условия 1) H - является тестом; 2) любое собственное подмножество набора H не является тестом. Другими словами тупиковым тестом является неукорачиваемый набор признаков, по которому любые два обучающих объекта из разных классов отличаются друг от друга.

Пусть каждый признак x_j имеет конечное множество допустимых значений N_j.

Пусть $H = \{x_{j_1}, \ldots, x_{j_r}\}$ - некоторый набор признаков, $S = (a_1, ..., a_n)$ - объект из обучающей выборки. Фрагмент $(a_{j_1}, \ldots, a_{j_r})$ описания объекта S обозначим через (S, H).

Каждый тест H порождает множество фрагментов описаний объектов вида (S_i, H), $i = 1, 2, \ldots, m$, где S_i - обучающий объект, причем каждый из этих фрагментов встречается в некотором классе и не встречается в остальных. При распознавании объектов производится голосование по множеству всех таких фрагментов. Впервые модель тестового алгоритма описана в [31].

Рассмотрим пример. Пусть обучающая выборка состоит из объектов $S_1 = (0, 1, 1, 0)$, $S_2 = (1, 2, 0, 1)$, $S_3 = (0, 1, 0, 1)$, $S_4 = (1, 2, 1, 0)$, $S_5 = (1, 1, 0, 1)$, $S_6 = (1, 1, 1, 2)$, при этом объекты S_1,S_2 и S_3 принадлежат первому классу, а объекты S_4,S_5 и S_6 - второму.

В данном примере тупиковыми тестами являются наборы признаков $\{x_1, x_2, x_3\}$, $\{x_1, x_2, x_4\}$ и $\{x_2, x_3, x_4\}$. Если использовать тестовый алгоритм, то объект $S = (0, 1, 2, 1)$ не будет отнесен ни к одному из классов, однако фрагмент $(0, 1)$, порождаемый парами $(S_1, \{x_1, x_2\})$ и $(S_3, \{x_1, x_2\})$, содержится в S и соответствующих

объектах из первого класса и не содержится в объектах из второго класса, что дает нам основание полагать, что распознаваемый объект более близок к первому классу. Таким образом, если при построении алгоритмов распознавания перейти от рассмотрения опорных множеств признаков к анализу фрагментов описаний объектов, можно строить менее осторожные и при этом корректные процедуры. Примерами таких процедур являются алгоритмы голосования по представительным наборам или алгоритмы типа "Кора" [8, 13].

Фрагмент описания объекта S' из класса K вида (S', H) назовем представительным набором для K, если для любого обучающего объекта S'', не принадлежащего классу K, имеет место $B(S', S'', H) = 0$. Фрагмент описания объекта S' из класса K вида (S', H) назовем тупиковым представительным набором для K, если выполнены два условия: 1) для любого обучающего объекта $S'' \notin K$ имеет место $B(S', S'', H) = 0$; 2) для любого набора H', $H' \subset H$, найдется обучающий объект $S'' \notin K$, для которого $B(S', S'', H') = 1$.

В классической модели алгоритма голосования по (тупиковым) представительным наборам для каждого класса K строится множество (тупиковых) представительных наборов, обозначаемое далее через $\mathfrak{T}(K)$. Распознавание объекта S осуществляется на основе процедуры голосования. В простейшей модификации для оценки принадлежности объекта S классу K используется величина

$$\Gamma_1(S, K) = \frac{1}{|\mathfrak{T}(K)|} \sum_{(S', H) \in \mathfrak{T}(K)} B(S, S', H),$$

Очевидно, что все фрагменты описаний обучающих объектов, порожденные некоторым тестом являются представительными наборами. Очевидно также, что не все представительные наборы, порожденные тупиковым тестом, являются тупиковыми представительными наборами, т.е. в алгоритме голосования по представительным наборам строится больше коротких фрагментов описаний объектов, следовательно, он менее осторожный и реже отказывается от распознавания.

Теперь мы можем описать общую схему конструирования дискретных процедур распознавания с использованием понятия элементарного классификатора [56, 58, 77, 87].

Пусть H - некоторый набор из r различных признаков вида $H = \{x_{j_1}, \ldots, x_{j_r}\}$, $\sigma = (\sigma_1, \ldots, \sigma_r)$, σ_i - допустимое значение признака x_{j_i}, $i = 1, 2, \ldots, r$. Набор σ назовем элементарным классификатором, порожденным признаками из H.

Близость объекта $S = (a_1, \ldots, a_n)$ из M и элементарного классификатора $\sigma = (\sigma_1, \ldots, \sigma_r)$, порожденного набором признаков H, будем оценивать величиной

$$B(\sigma, S, H) = \begin{cases} 1, & \text{если } a_{j_t} = \sigma_t \text{ при } t = 1, 2, \ldots, r; \\ 0, & \text{в противном случае.} \end{cases}$$

Множество всех элементарных классификаторов, порожденных наборами признаков из $\{x_1, \ldots, x_n\}$, обозначим через C. Таким образом, $C = \{(\sigma, H)\}$, где $H \subseteq \{x_1, \ldots, x_n\}$, $H = \{x_{j_1}, \ldots, x_{j_r}\}$, $\sigma = (\sigma_1, \ldots, \sigma_r)$, $\sigma_i \in N_j$, при $i = 1, 2, \ldots, r$. Каждый распознающий алгоритм A для каждого класса K, $K \in \{K_1, \ldots, K_l\}$, строит некоторое подмножество $C^A(K)$ множества C. Обозначим

$$C^A = \bigcup_{j=1}^{l} C^A(K_j).$$

Распознавание объекта S осуществляется на основе вычисления величины $B(\sigma, S, H)$ для каждого элемента (σ, H) множества $C^A(K)$, $K \in \{K_1, \ldots, K_l\}$, т.е. по каждому элементу множества $C^A(K)$ осуществляется процедура голосования. В результате вычисляется оценка $\Gamma(S, K)$ принадлежности объекта S классу K. Таким образом, каждый распознающий алгоритм A из рассматриваемого семейства определяется множеством элементарных классификаторов $C^A(K)$ и способом вычисления оценки $\Gamma(S, K)$, которая получается на основе голосования по элементарным классификаторам из $C^A(K)$.

Например, если A - алгоритм вычисления оценок, то множество C^A состоит из элементарных классификаторов, порождаемых теми фрагментами обучающих объектов, которые порождаются опорными множествами алгоритма A. Если же A - алгоритм голосования по представительным наборам, то $C^A(K)$ - некоторое подмножество множества представительных наборов класса K. В обоих случаях оценка $\Gamma(S, K)$ получается на основе суммирования величин $B(\sigma, S, H)$, где $(\sigma, H) \in C^A(K)$.

В общем случае элементарный классификатор $(\sigma_1, \ldots, \sigma_r)$, порожденный признаками из H, может обладать одним из следующих трех свойств:

1) каждый фрагмент вида (S', H), где $S' \in K$, совпадает с $(\sigma_1, \ldots, \sigma_r)$;

2) не все, а лишь некоторые фрагменты вида (S', H), где $S' \in K$, совпадают с $(\sigma_1, \ldots, \sigma_r)$;

3) ни один фрагмент вида (S', H), где $S' \in K$, не совпадает с $(\sigma_1, \ldots, \sigma_r)$.

Первая ситуация встречается крайне редко, поэтому работать с наборами значений признаков, для которых выполняется свойство 1, не представляется возможным. Существенное различие в информативности следующих двух свойств заключается в том, что свойство 2 характеризует лишь некоторое подмножество обучающих объектов из K, а свойство 3 все объекты из K. Следовательно, в случае, когда важно рассматривать класс K изолированно от других классов, напрашивается вывод о большей информативности таких наборов значений признаков, для которых выполнено свойство 3. В указанном случае аргументом за отнесение распознаваемого объекта S в класс K более естественно считать ситуацию, когда набор значений признаков не присутствует у всех объектов из класса K и не присутствует у объекта S.

Классические дискретные процедуры распознавания основаны на построении элементарных классификаторов, которые встречаются в описании некоторых объектов рассматриваемого класса (обладают свойством 2). В данной работе предлагаются корректные процедуры, основанные на построении элементарных классификаторов, не встречающихся в описании ни одного объекта рассматриваемого класса (обладающих свойством 3). Этими моделями являются модель голосования по покрытиям класса и модель голосования по антипредставильным наборам [54, 56, 57, 77, 87]. В ряде случаев указанные модели позволяют повысить качество распознавания и требуют меньших вычислительных затрат.

В модели голосования по покрытиям класса множество $C^A(K)$ состоит из таких элементарных классификаторов, которые не встречаются в описаниях объектов класса K. Такие элементарные классификаторы будем назвать покрытиями класса K. Элементарный классификатор (σ, H) из $C^A(K)$ голосует за принадлежность распознаваемого объекта S классу K, если σ не встречается в описании объекта S по набору признаков H. Принадлежность объекта S классу K (в простейшей модификации) оценивается величиной

$$\Gamma_2(S,K) = \frac{1}{|C^A(K)|} \sum_{(\sigma,H)\in C^A(K)} (1 - B(\sigma,S,H)) .$$

Элементарный классификатор (σ, H) назовем антипредставительным набором класса K, если он не совпадает ни с одним из фрагментов вида (S', H) , где S' - обучающий объект из класса K, и совпадает хотя бы с одним фрагментом вида (S'', H) , где S'' - обучающий объект, не принадлежащий классу K. Процедура голосования аналогична процедуре голосования, используемой в алгоритме голосования по покрытиям класса. Принадлежность объекта S классу K (в простейшей модификации) оценивается величиной

$$\Gamma_3(S,K) = \frac{1}{|C^A(K)|} \sum_{(\sigma,H)\in C^A(K)} (1 - B(\sigma,S,H)) .$$

Нетрудно показать, что представительный набор класса K является антипредствительным для любого другого класса, в случае $l = 2$ антипредставительный набор является представительным для противоположного класса, в случае, когда $l > 2$ антипредставительный набор для K не всегда является представительным для какого-либо другого класса.

Алгоритмы голосования по покрытиям класса и антипредставительным наборам являются корректными. Здесь корректность достигается за счет того, что для любого обучающего объекта $S' \in K$ $\Gamma(S', K) = 1$ (за принадлежность S' классу K голосуют все элементарные классификаторы из $C^A(K)$) и $\Gamma(S', K') < 1$ при $K' \neq K$, $K' \in \{K_1, \ldots, K_l\}$.

Тестирование на реальных задачах из области медицинского прогнозирования показало, что в некоторых случаях алгоритм голосования по покрытиям класса показывает преимущество перед классическим алгоритмом голосования по представительным наборам.

В случае, когда $l > 2$, использование новых процедур дает выигрыш по времени.

Выше уже было сказано, что в работе предлагаются подход, позволяющий значительно повысить эффективность алгоритмов распознавания в случае, когда в обучающей выборке содержится много объектов, лежащих на границе между классами (их описания похожи на описания объектов, принадлежащих другим классам). Суть предлагаемого подхода заключается в следующем.

Пусть описание обучающего объекта S, не принадлежащего классу K, похоже на описания некоторых объектов из K. Тогда объект S «лишает» класс K некоторого множества коротких элементарных классификаторов (тестов, представительных наборов и т.д.), что существенно снижает эффективность алгоритма. Для решения указанной проблемы предлагается разбить обучающую выборку на две подвыборки, по первой (базовой) построить множество представительных наборов, по второй (контрольной) вычислить их веса. Причем разбить нужно таким образом, чтобы объекты, находящиеся на границе между классами, попали в контрольную подвыборку, а все остальные (типичные для своих классов) объекты - в базовую подвыборку. Практические эксперименты на прикладных задачах показывают, что такое разбиение увеличивает число коротких элементарных классификаторов из $C^A(K)$ и тем самым позволяет повысить качество алгоритма распознавания A [53, 55-57, 77, 78, 87].

Для выделения типичных объектов предлагаются два способа. Первый основан на оценке типичности объекта путем вычисления типичности значений признаков [53, 55, 56]. Второй способ использует процедуру скользящего контроля. В данном случае типичными считаются те объекты, которые на скользящем контроле распознаются правильно. Приводится быстрый способ вычисления оценок при голосовании по представительным и тупиковым представительным наборам для процедуры скользящего контроля [77, 78, 87].

Вопросы изучения трудоемкости и качества дискретных процедур распознавания традиционно связаны с исследованием метрических (количественных) характеристик множества элементарных классификаторов. Имеется ввиду получение асимптотических оценок для типичных значений таких характеристик как число элементарных классификаторов и длина элементарного классификатора.

Введем следующие обозначения M_{mn}^k, $k \geq 2$, - множество всех матриц размера $m \times n$ с элементами из $\{0, 1, \ldots, k-1\}$; E_k^r - множество всех наборов вида $(\sigma_1, \ldots, \sigma_r)$, где $\sigma_i \in \{0, 1, \ldots, k-1\}$.

Пусть $L \in M_{mn}^k$, $\sigma \in E_k^r$. Набор H из r различных столбцов матрицы L назовем σ-покрытием, если L^H не содержит строку σ.

Набор H из r различных столбцов матрицы L назовем тупиковым σ-покрытием, если выполнены следующие два условия:

1) L^H не содержит строку σ;

2) L^H содержит (с точностью до перестановки строк) подматрицу вида

$$\begin{bmatrix} \beta_1 & \sigma_2 & \sigma_3 & \dots & \sigma_{r-1} & \sigma_r \\ \sigma_1 & \beta_2 & \sigma_3 & \dots & \sigma_{r-1} & \sigma_r \\ \dots & \dots & \dots & \dots & \dots & \dots \\ \sigma_1 & \sigma_2 & \sigma_3 & \dots & \sigma_{r-1} & \beta_r \end{bmatrix},$$

где $\beta_p \neq \sigma_p$ при $p = 1, 2, \dots, r$. Такую подматрицу будем называть σ-подматрицей. При $\sigma = (0, \dots, 0)$ и $k = 2$ тупиковое σ-покрытие является неприводимым покрытием, а σ-подматрица - единичной подматрицей.

Покажем, что понятия (тупикового) представительного набора, (тупикового) антипредставительного набора и (тупикового) покрытия класса можно ввести используя понятие (тупикового) покрытия целочисленной матрицы.

Пусть $K \in \{K_1, \dots, K_l\}$. Таблицу обучения можно рассматривать как пару матриц L_1 и L_2, где L_1 - матрица, состоящая из описаний обучающих объектов из класса K, а L_2 - матрица, состоящая из описаний остальных обучающих объектов.

Очевидно, что элементарный классификатор вида $(\sigma_1, \dots, \sigma_r)$, задаваемый парой (S_i, H), $S_i \in K$, $H = \{x_{j_1}, \dots, x_{j_r}\}$, является (тупиковым) представительным набором для K тогда и только тогда, когда набор столбцов матрицы L_1 с номерами $j_1, \dots, j_r$ не является $(\sigma_1, \dots, \sigma_r)$-покрытием, а набор столбцов L_2 с номерами $j_1, \dots, j_r$ является (тупиковым) $(\sigma_1, \dots, \sigma_r)$-покрытием.

Элементарный классификатор вида $(\sigma_1, \dots, \sigma_r)$, задаваемый парой (S_i, H), $S_i \in \overline{K}$, является (тупиковым) антипредставительным набором для K тогда и только тогда, когда набор столбцов матрицы L_2 с номерами $j_1, \dots, j_r$ не является $(\sigma_1, \dots, \sigma_r)$-покрытием, а набор столбцов L_1 с номерами $j_1, \dots, j_r$ является (тупиковым) $(\sigma_1, \dots, \sigma_r)$-покрытием.

Элементарный классификатор $(\sigma_1, \dots, \sigma_r)$, порождаемый набором столбцов H, является (тупиковым) покрытием класса K тогда и только тогда, когда набор столбцов матрицы L_1 с номерами $j_1, \dots, j_r$ является (тупиковым) $(\sigma_1, \dots, \sigma_r)$-покрытием.

Таким образом, можно считать, что рассматриваемые модели распознающих процедур основаны на поиске покрытий целочисленных матриц, образованных описаниями обучающих объектов. Следовательно, изучение метрических свойств множества элементарных классификаторов для данных моделей алгоритмов распознавания связано с изучением метрических свойств множества покрытий цело-

численных матриц.

Введем обозначения:

Ψ_k^0 - интервал $(\log_k mn, n)$;

Ψ_k^1 - интервал

$$(\frac{1}{2}\log_k mn - \frac{1}{2}\log_k \log_k mn - \log_k \log_k \log_k n,$$
$$\frac{1}{2}\log_k mn - \frac{1}{2}\log_k \log_k mn + \log_k \log_k \log_k n);$$

$a_n \approx b_n$ означает, что $\lim(a_n/b_n) = 1$ при $n \to \infty$.

Пусть $C(L,\sigma)$ - множество всех пар вида (H,σ) , где H - σ-покрытие матрицы L, $B(L,\sigma)$ - множество всех пар вида (H,σ), где H - тупиковое σ-покрытие матрицы L, $S(L,\sigma)$ - совокупность всех σ-подматриц матрицы L.

Положим

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

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

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

Нас будут интересовать асимптотические оценки чисел $|C(L)|$, $|B(L)|$ и $|S(L)|$ для почти всех матриц L из M_{mn}^k при $n \to \infty$. Выявление типичной ситуации будет связано с высказываниями типа «Для почти всех матриц L из M_{un}^k при $n \to \infty$ выполнено свойство β», причем свойство β также может иметь предельный характер. Это означает, что доля тех матриц из M_{un}^k для которых с ε-точностью выполнено свойство β, стремиться к 1 и одновременно ε стремиться к 0 при $n \to \infty$.

Ранее в [37-40, 42, 44, 45, 47, 48, 50, 52, 86] изучался случай, когда число строк в матрице по порядку меньше числа столбцов, а именно случай, когда $m^\alpha \leq n \leq k^{m^\beta}$, $\alpha > 1$, $\beta < 1$. Показано, что в данном случае величина $|B(L)|$ почти всегда (для почти всех матриц L из M_{mn}^k) при $n \to \infty$ асимптотически совпадает с величиной $|S(L)|$ и по порядку меньше числа покрытий. На основании этого факта был построен асимптотически оптимальный алгоритм поиска покрытий из $B(L)$.

На его основе в [90,91] построен точный алгоритм с полиномиальной временной задержкой (при наличии незначительного числа повторяющихся шагов).

В данной работе рассмотрен прямо противоположный случай, а именно, когда $n^\alpha \leq m \leq k^{n^\beta}$, $\alpha > 1$, $\beta < 1$ [54, 56, 57, 87].

Получены асимптотики типичных значений числа подматриц из $S(L)$ и порядка подматрицы из $S(L)$, а именно доказана

Теорема 3.3.1. *Если $n^\alpha \le m \le k^{n^\beta}$, $\alpha > 1$, $\beta < 1$, то для почти всех матриц L из M_{mn}^k при $n \to \infty$ справедливо*

$$|S(L)| \approx \sum_{r \in \Psi_k^1} C_n^r C_m^r r!(k-1)^r k^{r-r^2},$$

и порядки почти всех подматриц из $S(L)$ принадлежат интервалу Ψ_k^1.

Пусть

$$S_1(L) = \bigcup_{r \ge \log_k mn} \bigcup_{\sigma \in E_k^r} S(L, \sigma).$$

Тогда справедлива

Теорема 3.3.2. *Для почти всех матриц $L \in M_{mn}^k$ при $n \to \infty$ имеет место*

$$|S_1(L)| = 0.$$

Из теоремы 3.3.2 следует, что почти все матрицы в M_{mn}^k не имеют σ-подматриц, порядок которых больше $\log_k mn$.

Для практически общего случая в [54, 56, 57, 87] получены асимптотики типичных значений величины $|C(L)|$ и длины покрытия из $C(L)$, а именно, доказана

Теорема 3.2.1. *Если $m \le k^{n^\beta}$, $\beta < 1$, то для почти всех матриц L из M_{mn}^k при $n \to \infty$ имеет место*

$$|C(L)| \approx \sum_{r \in \Psi_k^0} C_n^r k^r$$

и длины почти всех покрытий из $C(L)$ принадлежат интервалу Ψ_k^0.

Пусть $r_0 = \log_k m - \log_k(\log_k m \ln kn)$ и пусть

$$C_1(L) = \bigcup_{r \le r_0} \bigcup_{\sigma \in E_k^r} C(L, \sigma).$$

Тогда справедлива

Теорема 3.2.2. *Для почти всех матриц $L \in M_{(mn)}^k$ при $n \to \infty$ имеет место*

$$|C_1(L)| = 0.$$

Из теоремы 3.2.2 следует, что при $r \le \log_k m - \log_k(\log_k m \ln kn)$ почти все матрицы не имеют σ-покрытий длины r.

На основе сравнения оценок, приведенных в теоремах 3.3.1 и 3.2.1, доказана

Теорема 3.3.3. *Если $n^\alpha \le m \le k^{n^\beta}$, $\alpha > 1$, $\beta < 1/2$, то для почти всех матриц L из M_{mn}^k при $n \to \infty$ справедливо $|S(L)|/|B(L)| \to \infty$.*

Покажем, что задачу нахождения покрытий целочисленной матрицы с элементами из множества $\{0, 1, \ldots, k-1\}$ можно сформулировать как задачу построения сокращенной ДНФ двузначной логической функции, заданной на k-ичных n-мерных наборах [45, 47, 48, 52, 86].

Пусть на наборах из E_k^n задана всюду или частично определенная логическая функция f, принимающая значения из множества $\{0, 1\}$, A_f - множество единиц этой функции, B_f - множество ее нулей. Введем ряд определений. Пусть переменная x принимает значения из множества E_k^n, $\sigma \in E_k^n$. Положим

$$x^\sigma = \begin{cases} 1, & \text{если } x = \sigma; \\ 0, & \text{если } x \neq \sigma. \end{cases}$$

Элементарной конъюнкцией (э.к.) называется выражение вида $x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$, где все x_{j_i} различны. Число r называется рангом конъюнкции. Интервал истинности э.к. $\mathfrak{B}$ будем обозначать через $N_{\mathfrak{B}}$.

Введем понятия допустимой, неприводимой и максимальной конъюнкции для функции f.

Э.к. $\mathfrak{B}$ называется допустимой для f, если $N_{\mathfrak{B}} \cap A_f \neq \emptyset$ и $N_{\mathfrak{B}} \cap B_f = \emptyset$.

Э.к. $\mathfrak{B}$ называется неприводимой для f, если не существует элементарной конъюнкции $\mathfrak{B}'$ такой, что $N_{\mathfrak{B}'} \supset N_{\mathfrak{B}}$ и $N_{\mathfrak{B}'} \cap B_f = N_{\mathfrak{B}} \cap B_f$.

Э.к. $\mathfrak{B}$ называется максимальной для F, если $\mathfrak{B}$ допустимая конъюнкция и не существует допустимой конъюнкции $\mathfrak{B}'$ такой, что $N_{\mathfrak{B}'} \supset N_{\mathfrak{B}}$.

Пусть A_f состоит из наборов $(\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1n})$, …, $(\alpha_{u1}, \alpha_{u2}, \ldots, \alpha_{un})$, B_f - из наборов $(\beta_{11}, \beta_{12}, \ldots, \beta_{1n})$, …, $(\beta_{u1}, \beta_{u2}, \ldots, \beta_{un})$.

Из наборов A_f составим матрицу L_1 вида

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \ldots & & & \\ \alpha_{u1} & \alpha_{u2} & \ldots & \alpha_{un} \end{bmatrix}$$

Из наборов B_f составим матрицу L_2 вида

$$\begin{bmatrix} \beta_{11} & \beta_{12} & \ldots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \ldots & \beta_{2n} \\ \ldots & & & \\ \beta_{u1} & \beta_{u2} & \ldots & \beta_{un} \end{bmatrix}$$

Утверждение 4.1.1. *Э.к. $x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$ является допустимой для F тогда и только тогда, когда набор столбцов с номерами $j_1, \ldots, j_r$ является $(\sigma_1, \ldots, \sigma_r)$-покрытием матрицы L_2.*

Утверждение 4.1.2. *Э.к.* $x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r}$ *является неприводимой для* F *тогда и только тогда, когда в наборе столбцов с номерами* $j_1, \dots, j_r$ *содержится* $(\sigma_1, \dots, \sigma_r)$*-подматрица матрицы* L_2*.*

Утверждение 4.1.3. *Э.к.* $x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r}$ *является максимальной для* F *тогда и только тогда, когда набор столбцов с номерами* $j_1, \dots, j_r$ *является тупиковым* $(\sigma_1, \dots, \sigma_r)$*-покрытием матрицы* L_2*.*

Утверждение 4.1.4. *Э.к.* $x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r}$ *является максимальной для* f *тогда и только тогда, когда набор столбцов с номерами* $j_1, \dots, j_r$ *является тупиковым* $(\sigma_1, \dots, \sigma_r)$*-покрытием матрицы* L_2 *и подматрица матрицы* L_1*, образованная столбцами с номерами* $j_1, \dots, j_r$*, содержит хотя бы одну из строк вида* $(\sigma_1, \dots, \sigma_r)$*.*

В силу утверждений 4.1.1 - 4.1.3 для числа и ранга допустимых конъюнкций двузначной логической функции, заданной на k-ичных n-мерных наборах можно сформулировать утверждения, аналогичные теоремам 3.2.1 - 3.2.2, а также получить верхнюю ассимптотическую оценку числа максимальных конъюнкций.

Настоящая работа состоит из введения пяти глав и заключения.

В главе 1 описан классический алгоритм голосования по представительным наборам и предлагаются модели алгоритмов голосования по антипредставительным наборам и покрытиям классов.

В главе 2 предложены методы предварительного анализа обучающей информации, направленные на снижение влияния шумящих признаков и объектов, лежащих на границе между классами. Предлагаются методы оценки информативности отдельных значений признаков, фрагментов описаний объектов, а также типичности обучающих объектов. Приведен быстрый способ вычисления оценок при голосовании по представительным и тупиковым представительным наборам для процедуры скользящего контроля.

Глава 3 посвящена получению асимптотических оценок типичных значений числа покрытий и длины покрытия, а также числа σ-подматриц и размера σ-подматриц.

В главе 4 сформулированы основные принципы конструирования дискретных процедур распознавания с использованием аппарата логических функций. Рассмотрена связь между задачей нахождения покрытий целочисленной матрицы и задачей построения совершенной ДНФ двузначной логической функции заданной на k-ичных наборах. Получены асимптотики типичных значений для числа допустимых конъюнкций для указанных функций и ранга допустимой конъюнкции, а также верхние асимптотические оценки числа максимальных конъюнкций.

В главе 5 приведены результаты тестирования предложенных в работе подходов на реальных задачах. Исследованы задачи прогнозирования результатов лечения онкобольных и анализа результатов социологического опроса. Проведено

сравнение новых моделей с классическим алгоритмом голосования по представительным наборам.

Глава 1

Модели дискретных (логических) процедур распознавания, основанные на построении покрытий классов

Рассматривается стандартная постановка задачи распознавания для случая, когда объекты описаны набором из целочисленных признаков $\{x_1, \ldots, x_n\}$. Пусть N_j - множество допустимых значений признака x_j, $j \in \{1, 2, \ldots, n\}$, и пусть исследуемое множество объектов M представимо в виде объединения подмножеств (классов) $K_1, \ldots, K_l$. Имеется конечный набор объектов $\{S_1, \ldots, S_m\}$ из M, о которых известно, каким классам они принадлежат (обучающая выборка). Обучающие объекты представлены своими описаниями. Требуется по предъявленному набору значений признаков, т.е. описанию некоторого объекта S из M, о котором, вообще говоря, неизвестно, какому классу он принадлежит, определить этот класс.

Дискретные (логические) процедуры распознавания хорошо себя зарекомендовали при решении задач, в которых объекты описаны целочисленными признаками невысокой значности. В подобных задачах довольно сложно ввести адекватную меру близости между объектами, поскольку такая мера должна учитывать неравнозначность признаков (при этом заранее не известно какие признаки важнее). При дискретном подходе к задачам распознавания не требуется наличия указанной меры близости между объектами и в отличии от статистических не требуют наличия большого объема обучающей информации и принятия дополнительных гипотез вероятностного характера.

При решении прикладных задач очень важным этапом является содержательная интерпретация решения, принятого распознающим алгоритмом. Дискретный подход к задачам распознавания основан на комбинаторном анализе описаний обучающих объектов с целью выделения наиболее информативных подописаний объектов, составленных из допустимых значений признаков. Можно сказать, что суть дискретного подхода состоит в поиске логических закономерностей в описании объектов обучающей выборки. Поэтому не составляет особого труда содержательно обосновать полученный результат.

Обычно информативными считаются те наборы из допустимых значений признаков, которые встречаются в описаниях объектов одного класса, но не встречаются в описаниях объектов остальных классов. При этом исходные описания объектов задаются в виде наборов значений целочисленных признаков. Дискретные методы привели к появлению целого ряда эвристик, называемых дискретными или логическими процедурами распознавания (например, тестовые алгоритмы, алгоритмы типа «Кора» или алгоритмы голосования по представительным наборам, логические решающие деревья и т.д.). В данной главе предлагаются новые модели дискретного характера, которые основанны на принципе «невстречаемости» набора из допустимых значений признаков в описаниях рассматриваемых объектов. А именно предложены две новые модели алгоритмов: алгоритм голосования по антипредставительным наборам и алгоритм голосования по покрытиям классов. В определенных случаях предложенные модели позволяют повысить качество распознавания и требуют меньших вычислительных затрат при большом числе классов.

1.1 Основные определения

Одно из основных понятий дискретного подхода - понятие элементарного классификатора. В классическом варианте элементарными классификаторами являются фрагменты описаний обучающих объектов [8, 13].

Пусть $H = \{x_{j_1}, \ldots, x_{j_r}\}$ - набор из r различных признаков, $S' = (a_1, ..., a_n)$, - объект из обучающей выборки. Набор $(a_{j_1}, \ldots, a_{j_r})$ будем называть фрагментом описания объекта S', порожденным признаками из H, и обозначать через (S', H).

Расширим понятие элементарного классификатора.

Пусть H - набор из r различных признаков вида $\{x_{j_1}, \ldots, x_{j_r}\}$, $\sigma = (\sigma_1, \ldots, \sigma_r)$, σ_i - допустимое значение признака x_{j_i}, $i = 1, 2, \ldots, r$. Набор σ назовем элементарным классификатором, порожденным признаками из H.

Множество всех элементарных классификаторов, порожденных наборами признаков из $\{x_1, \ldots, x_n\}$, обозначим через C. Очевидно, что при построении алгоритма распознавания нет смысла рассматривать все возможные элементарные классификаторы. Например, не представляют интереса подописания, которые по данному набору признаков встречаются в описаниях объектов всех классов, поскольку указанные элементарные классификаторы не характеризуют ни один из классов (можно сказать, что такие наборы значений признаков выражают закономерность, не имеющую отношения к рассматриваемой задаче распознавания). Кроме того, разные процедуры распознавания основаны на разных принципах, например, в основе тестового алгоритма лежит отделение всех объектов одного класса от всех объектов остальных классов, а в алгоритме голосования по

представительным наборам - отделения каждого объекта класса от всех объектов остальных классов. Поэтому элементарный классификатор может являться информативным при построении одного распознающего алгоритма и не информативным при построении другого. Следовательно, каждый распознающий алгоритм A определяется некоторым подмножеством C^A множества C. Собственно говоря, для каждого класса K, $K \in \{K_1, \ldots, K_l\}$, строится подмножество $C^A(K)$ множества C и

$$C^A = \bigcup_{j=1}^{l} C^A(K_j).$$

Элементарный классификатор $(\sigma_1, \ldots, \sigma_r)$, порожденный признаками из H, по отношению к классу K может обладать одним из следующих трех свойств:

1) каждый фрагмент вида (S', H), где $S' \in K$, совпадает с $(\sigma_1, \ldots, \sigma_r)$;

2) не все, а лишь некоторые фрагменты вида (S', H), где $S' \in K$, совпадают с $(\sigma_1, \ldots, \sigma_r)$;

3) ни один фрагмент вида (S', H), где $S' \in K$, совпадает с $(\sigma_1, \ldots, \sigma_r)$.

Первая ситуация встречается крайне редко, поэтому работать с наборами значений признаков, для которых выполняется свойство 1, не представляется возможным. Существенное различие в информативности следующих двух свойств заключается в том, что свойство 2 характеризует лишь некоторое подмножество обучающих объектов из K , а свойство 3 все объекты из K . Следовательно, в случае, когда важно рассматривать класс K изолированно от других классов, напрашивается вывод о большей информативности таких наборов значений признаков, для которых выполнено свойство 3. В указанном случае аргументом за отнесение объекта S в класс K более естественно считать ситуацию, когда набор значений признаков не присутствует у всех объектов из класса K и не присутствует у распознаваемого объекта S. Собственно говоря, классические дискретные процедуры распознавания основаны на поиске элементарных классификаторов, которые встречаются в описании некоторых объектов рассматриваемого класса. В данной главе предлагаются процедуры, основанные на построении элементарных классификаторов, не встречающихся в описании ни одного объекта рассматриваемого класса.

При описании дискретных процедур используется понятие покрытия целочисленной матрицы. Введем следующие обозначения: M_{mn}^k, $k \geq 2$, - множество всех матриц размера $m \times n$ с элементами из $\{0, 1, \ldots, k-1\}$; E_k^r - множество всех k-ичных наборов длины r.

Пусть $L \in M_{mn}^k$, $\sigma \in E_k^r$. Набор H из r различных столбцов матрицы L назовем σ-покрытием, если подматрица L^H матрицы L, образованная столбцами из H, не содержит строки σ. Набор из различных столбцов матрицы L назовем тупиковым σ-покрытием, если выполнены два условия: 1) подматрица L^H не со-

держит строку $\sigma = (\sigma_1, \ldots, \sigma_r)$; 2) если $p \in \{1, 2, \ldots, r\}$, то L^H содержит хотя бы одну из строк вида $(\sigma_1, \ldots, \sigma_{p-1}, \beta_p, \sigma_{p+1}, \sigma_n)$, где $\beta_p \neq \sigma_p$.

Таблицу обучения T_{mn} можно рассматривать как пару матриц L_1 и L_2 , где L_1 - матрица, состоящая из описаний обучающих объектов из класса K, а L_2 - матрица, состоящая из описаний остальных обучающих объектов. Тогда очевидно, что элементарный классификатор вида $(\sigma_1, \ldots, \sigma_r)$, порожденный фрагментом парой (S_i, H), $S_i \in K$, $H = \{x_{j_1}, \ldots, x_{j_r}\}$, является (тупиковым) представительным набором для K тогда и только тогда, когда набор столбцов матрицы L_1 с номерами $j_1, \ldots, j_r$ не является $(\sigma_1, \ldots, \sigma_r)$-покрытием, а набор столбцов L_2 с номерами $j_1, \ldots, j_r$ является (тупиковым) $(\sigma_1, \ldots, \sigma_r)$-покрытием. При построении тестового алгоритма используется еще более общее понятие покрытия (Е-покрытие [45,47]), а также $(0, \ldots, 0)$-покрытия булевой матрицы (матрицы сравнения обучающей таблицы).

Пусть даны два объекта $S' = (a'_1, a'_2, \ldots, a'_n)$ и $S' = (a''_1, a''_2, ..., a''_n)$. Близость объекта S' и S'' по набору признаков H, $H = \{x_{j_1}, \ldots, x_{j_r}\}$, будем оценивать величиной

$$B(S', S'', H) = \begin{cases} 1, & \text{если } a'_{j_t} = a''_{j_t}, \text{ при } t = 1, 2, ..., r; \\ 0, & \text{в противном случае.} \end{cases}$$

Близость объекта S и элементарного классификатора $\sigma = (\sigma_1, \ldots, \sigma_r)$, порожденного набором признаков H, $H = \{x_{j_1}, \ldots, x_{j_r}\}$, будем оценивать величиной

$$B(\sigma, S', H) = \begin{cases} 1, & \text{если } a'_{j_t} = \sigma_t, \text{ при } t = 1, 2, \ldots, r; \\ 0, & \text{в противном случае.} \end{cases}$$

Пусть $K \in \{K_1, \ldots, K_l\}$, $\overline{K} = \{K_1, \ldots, K_l\} \setminus K$.

Конкретная модель алгоритма распознавания A определяется принципом построения множества C^A и оценкой $\Gamma(S, K)$ принадлежности объекта S классу K, которая вычисляется на основе голосования по элементарным классификаторам из $C^A(K)$. В тестовых алгоритмах и алгоритмах голосования по представительным наборам, считается, что элементарный классификатор σ из $C^A(K)$, порожденный набором признаков H, дает голос за принадлежность объекта S классу K, если $B(\sigma, S, H) = 0$. Далее объект S относится к тому классу, для которого оценка принадлежности наибольшая (если таких классов несколько, то происходит отказ от распознавания).

1.2 Классическая модель голосования по представительным наборам

Модель голосования по представительным наборам основана на поиске для каждого класса таких фрагментов описаний объектов, которые встречаются в этом классе и не встречаются в остальных. Если такой фрагмент описания встречается и в распознаваемом объекте, то считается, что объект близок к рассматриваемому классу. Таким образом происходит отделение каждого объекта одного класса от всех объектов остальных классов, при этом все объекты данного класса рассматриваются независимо друг от друга.

Фрагмент описания объекта S' из класса K вида (S', H) назовем представительным набором для K, если для любого обучающего объекта S'', не принадлежащего классу K, имеет место $B(S', S'', H) = 0$. Фрагмент описания объекта S' из класса K вида (S', K) назовем тупиковым представительным набором для K, если: 1) для любого обучающего объекта S'' из $\overline{K}$ имеет место $B(S', S'', H) = 0$; 2) для любого набора H', $H' \subset H$, найдется обучающий объект S'' из $\overline{K}$, для которого $B(S', S'', H') = 1$.

В классической модели алгоритма голосования по (тупиковым) представительным наборам множество $C^A(K)$ состоит из множества (тупиковым) представительных наборов для K. Распознавание объекта S осуществляется на основе процедуры голосования: если представительный набор для класса K встречается в описании распознаваемого объекта S, то данный представительный набор голосует за принадлежность объекта S к классу K. Далее объект относится к тому классу, который наберет больше голосов. В простейшей модификации для оценки принадлежности объекта S классу K используется величина

$$\Gamma_1(S, K) = \frac{1}{|C^A(K)|} \sum_{(\sigma,H)\in C^A(K)} B(\sigma, S, H)$$

(здесь и далее $|N|$- мощность множества N).

Построение множества тупиковых представительных наборов для K сводится к построению тупиковых σ-покрытий матрицы, составленной из описаний объектов из $\overline{K}$. , и последующей проверкой условия, что построенный элементарный классификатор встречается в K.

Короткие представительные наборы чаще встречаются в распознаваемых объектах, следовательно, они обладают большей информативностью, поэтому при решении прикладных задач для повышения качества распознавания и сокращения вычислительных затрат как правило рассматривают только короткие представительные наборы. Это могут быть представительные наборы ограниченной длины или же тупиковые представительные наборы.

1.3 Модели голосования по антипредставительным наборам и по покрытиям класса

Предлагаемые новые модели основаны на построении для каждого класса таких элементарных классификаторов, которые по данному набору признаков не встречаются в описаниях ни одного объекта класса . Если такой элементарный классификатор не встречается и в описании распознаваемого объекта , то считается, что этот объект близок к рассматриваемому классу. Таким образом, ищутся закономерности, присущие всем обучающим объектам рассматриваемого класса, т.е. каждый элементарный классификатор характеризует весь класс целиком.

Новые модели основаны на построении σ-покрытий матриц, образованных описаниями обучающих объектов каждого класса: модель голосования по покрытиям класса и модель голосования по антипредставительным наборам класса. Использование этих моделей позволяет несколько снизить вычислительные затраты в случае, когда $|K| < |\overline{K}|$, например, при большом числе классов. Ниже приводится описание этих моделей.

В модели голосования по (тупиковым) σ-покрытиям класса множество $C^A(K)$ состоит из таких элементарных классификаторов, которые порождаются (тупиковыми) σ-покрытиями матрицы, образованной описаниями обучающих объектов класса K. В отличии от классического алгоритма здесь элементарный классификатор из $C^A(K)$ голосует за принадлежность распознаваемого объекта классу K, если этот элементарный классификатор не встречается в описании рассматриваемого объекта. Принадлежность объекта S классу K (в простейшей модификации) оценивается величиной

$$\Gamma_2(S,K) = \frac{1}{|C^A(K)|} \sum_{(\sigma,H)\in C^A(K)} (1 - B(\sigma,S,H)).$$

Рассмотрим теперь модель с антипредставительными наборами. Элементарный классификатор σ, порожденный (тупиковым) σ-покрытием класса K и набором признаков H, является (тупиковым) антипредставительным набором, если он совпадает хотя бы с одним фрагментом вида (S', H) , где S' - обучающий объект из $\overline{K}$. Процедура голосования такая же, как в алгоритме голосования по покрытиям класса. Принадлежность объекта S классу K (в простейшей модификации) оценивается величиной

$$\Gamma_3(S,K) = \frac{1}{|C^A(K)|} \sum_{(\sigma,H)\in C^A(K)} (1 - B(\sigma,S,H)).$$

Заметим, что представительный набор для класса K является антипредставительным для $\overline{K}$.

Нетрудно показать, в случае $l = 2$ при использовании обеих моделей объект S будет отнесен к одному и тому же классу. В самом деле, пусть A_1 - алгоритм голосования по представительным наборам, A_2 - алгоритм голосования по антипредставительным наборам. Как уже было сказано, $C^{A_1}(K_1) = C^{A_2}(K_2) = C_1$, $C^{A_1}(K_2) = C^{A_2}(K_1) = C_2$. Пусть в распознаваемом объекте S с представительными наборами класса K_1 совпадают q_1 фрагментов, а с представительными наборами класса K_2 - q_2. Тогда

$$\Gamma_1(S, K_1) = \frac{q_1}{C^{A_1}(K_1)} = \frac{q_1}{C_1}, \qquad \Gamma_1(S, K_2) = \frac{q_2}{C^{A_1}(K_2)} = \frac{q_2}{C_2},$$

$$\Gamma_3(S, K_1) = \frac{C^{A_2}(K_1) - q_2}{C^{A_2}(K_1)} = \frac{C_2 - q_2}{C_2} = 1 - \frac{q_2}{C_2},$$

$$\Gamma_3(S, K_2) = \frac{C^{A_2}(K_2) - q_1}{C^{A_2}(K_2)} = 1 - \frac{q_1}{C_1}.$$

Таким образом, $\Gamma_1(S, K_1) > \Gamma_1(S, K_2)$ тогда и только тогда, когда $\Gamma_3(S, K_1) > \Gamma_3(S, K_2)$.

Рассмотрим теперь случай, когда $l > 2$. Очевидно, что представительный набор для класса K_i является антипредставительным для каждого из классов $K_1, \ldots, K_{i-1}, K_{i+1}, \ldots, K_l$. Однако антипредставительный набор для K_i может не являться представительным набором ни для одного из классов $K_1, \ldots, K_{i-1}, K_{i+1}, \ldots, K_l$. В самом деле, если некий фрагмент описания встречается в описаниях объектов из классов K_{i_1} и K_{i_2}, но не встречается в K_i, тогда он является антипредставительным набором для K_i, но не является представительным ни для K_{i_1}, ни для K_{i_2}. Таким образом, если A_1 - алгоритм голосования по представительным наборам, A_2 - алгоритм голосования по антипредставительным наборам, то справедливо:

$$C^{A_1}(K_1) \bigcup \ldots \bigcup C^{A_1}(K_l) \subseteq C^{A_2}(K_1) \bigcup \ldots \bigcup C^{A_2}(K_l)$$

Кроме того, каждый антипредставительный набор является покрытием, но не каждое покрытие является антипредставительным набором, поэтому, если A_3 - алгоритм голосования по покрытиям класса, то справедливо следующее включение:

$$C^{A_1} \subseteq C^{A_2} \subseteq C^{A_3}$$

Наиболее трудоемкий этапам при построении как классического алгоритма, так и новых моделей является нахождение множества покрытий матрицы составленной из описаний обучающих объектов. В новых моделях в отличие от классических (где покрытия строятся для $\overline{K}$) покрытия строятся для класса K. Таким образом, при большом числе классов новые модели требуют меньших вычислительных затрат.

Модели апробированы на реальной информации. Результаты тестирования приводятся в п.5.1.

Глава 2

Методы повышения эффективности дискретных процедур распознавания

Одним из подходов повышения качества распознавания и снижения вычислительных затрат является проведение предварительного анализа обучающей информации. Целью такого анализа является оценка основных информативных характеристик обучающей выборки, в частности, оценка информативности признаков, значений признаков, выделение наиболее представительных (типичных) объектов.

В данной главе описана методика предварительного анализа обучающей информации, основанная на нахождении в обучающем материале «информативной зоны» и типичных для своих классов объектов. Информативная зона выделяется на основе оценки типичности значений каждого признака.

Предложено два способа выделения типичных объектов: 1) типичными считаются объекты, описания которых состоят из типичных значений признаков; 2) типичными являются объекты, которые правильно распознаются на скользящем контроле. При использовании первого способа вычислительные затраты незначительны. Второй способ довольно трудоемкий. Для уменьшения вычислительных затрат приводится быстрый способ вычисления оценок при голосовании по представительным и тупиковым представительным наборам для процедуры скользящего контроля.

Выделение типичных объектов позволяет значительно повысить эффективность алгоритмов распознавания в случае, когда в обучающей выборке содержится много объектов, лежащих на границе между классами. В этом случае обучающая выборка разбивается на две подвыборки: на базовую (содержащую «типичные» для своих классов объекты) и контрольную (содержащую "нетипичные"для своих классов объекты) подвыборки. По первой строится множество элементарных классификаторов, а по второй вычисляются их веса.

Разработанная методика позволяет снижать влияние шумящих признаков. Шумящими, в частности, являются признаки принимающие слишком много значений, их значения редко встречаются во всех классах. Поэтому такие признаки порождают большое число уникальных элементарных классификаторов, редко

участвующих в голосовании.

Предложенные методы апробированы на реальной информации, результаты тестирования приводятся в главе 5.

2.1 Методы оценки информативных характеристик обучающей выборки

Для повышения эффективности дискретных процедур распознавания полезно оценить информативность признаков, их отдельных значений, а также фрагментов описаний объектов. Традиционно в качестве меры важности признака x_j, $j \in \{1, 2, \ldots, n\}$ рассматривалась величина

$$I_j = |C_j^A|/|C^A|,$$

где C_j^A - подмножество таких элементарных классификаторов из C^A, в которых содержится признак x_j [46, 65]. При этом в качестве A использовался тестовый алгоритм или алгоритм голосования по представительным наборам. Во многих случаях такой способ оценки информативности признаков дает не очень хорошие результаты. Например, если признак является тестом, тогда он входит лишь в один тупиковый тест и согласно формуле для I_j имеет очень маленький вес. Однако, очевидно, что указанный признак является важным и обладает большой информативностью. С другой стороны малоинформативные шумящие признаки входят в большое число тестов (представительных наборов) и, следовательно, имеют большой вес I_j.

Пусть A - алгоритм голосования по представительным наборам. Частично проблему шумящих признаков можно решить, введя порог p_{min} минимальной встречаемости представительного набора в базовой подвыборке. То есть рассматривать только те представительные наборы, которые в базовой подвыборке встречаются не менее p_{min} раз. Используя это дополнительное условие шумящие многозначные признаки не получают самой большой оценки по информативности. Однако возможно, что их информативный вес будет всё-таки довольно большим, поскольку не всегда можно взять большой порог p_{min}, из-за того, что структура класса такова, что каждый представительный набор встречается не очень большое число раз. В данном случае помогает следующий прием.

Для снижения вычислительных трудностей разделим случайным образом обучающую выборку на две подвыборки: базовую и контрольную. По базовой будем строить множество представительных наборов, а по контрольной вычислять их веса. Для вычисления весов представительных наборов можно воспользоваться различными функциями. Например, пусть ω - представительный набор класса K , $K \in \{1, \ldots, l\}$, порождаемый парой (S', H), где S' - объект из базовой выборки,

и пусть $\delta(K,\omega)$ - число объектов в контрольной выборке, за которых представительный набор голосует "правильно", $\delta(\overline{K},\omega)$ - число объектов в контрольной выборке, за которых он голосует "неправильно". Тогда в качестве функции $\nu_{(S',H)}$ (вес представительного набора) возьмем функцию:

$$\nu_{(S',H)} = \frac{1+\delta(K,\omega)}{1-\delta(\overline{K},\omega)}.$$

В общем случае функция, по которой вычисляется вес представительного набора должна, обладать следующими свойствами: она должна монотонно возрастать по числу объектов из контрольной выборки, за которые представительный набор голосует правильно и монотонно убывать по числу объектов, за которые он голосует не правильно.

Для решения проблемы шумящих признаков не обязательно при помощи варьирования p_{min} добиваться того, чтобы оценка их информативности была низкой при любом разбиении на базовую и контрольную выборку. Достаточно, чтобы оценка информативности шумящих признаков при различных разбиениях (имеется ввиду, что мощности контрольной и обучающей выборки постоянны, меняется только их состав) была неустойчива по разбиению информации на две подвыборки, при том что по остальным признакам наблюдается некоторая устойчивость. При решении практических задач, чтобы выявить шумящие признаки достаточно было вычислить веса признаков для трех-четырех разных разбиений.

Как обобщение описанного выше метода выявления шумящих признаков рассмотрим метод оценки информативности признаков, основанный на вычислении весов так называемых (p,q)-представительных наборов.

Зададим целые числа p и q, такие, что

$$1 \le p \le \min_{1\le i\le l}|K_i|, \qquad 0 \le q \le \min_{1\le i\le l}|\overline{K}_i|, \qquad p > q.$$

Элементарный классификатор σ, порожденный признаками из H, назовем (p,q)-представительным набором для класса K_i, если не менее чем для p объектов S' из класса K_i справедливо $B(\sigma,S',H)=1$ и не более чем для q объектов S'' из $\overline{K}_i$ справедливо $B(\sigma,S'',H)=1$.

В частности представительный набор является (1,0)-представительным набором.

Очевидны следующие свойства (p,q)-представительных наборов.

1) Если набор значений признаков σ не удовлетворяет порогу p, то любой набор содержащий σ, также не удовлетворяет порогу p.

2) Если σ не удовлетворяет порогу q, то любой поднабор σ не удовлетворяет порогу q.

Пусть σ является (p,q)-представительным набором класса K, p_σ - его встречаемость в обучающей выборке в классе K, q_σ - встречаемость в в остальных

классах ($p_\sigma \geq p$, $q_\sigma \leq q$). Тогда информативный вес набора будем вычислять по следующей формуле:

$$\nu(\sigma) = (p_\sigma - p) + (q - q_\sigma).$$

Использование свойств 1) и 2) дает возможность существенно снизить вычислительные трудности при построении (p, q)-представительных наборов.

За счет варьирования p и q можно существенно снизить влияние шумящих признаков. Дело в том, что, разрешив представительным наборам встречаться в других классах (увеличив q), можно увеличить минимальную встречаемость в своем классе (увеличить p). В практической реализации для снижения вычислительных затрат используется следующий метод. Множество (p, q)-представительных наборов строится по случайной подвыборке (60-70 %) обучающей выборки. После этого проверяется, являются ли построенные представительные наборы (p, q)-представительными наборами для всей обучающей выборки, и в случае положительного результата вычисляются их информативные веса. Далее вычисляется информативный вес каждого признака как сумма весов (p, q)-представительных наборов, в которые он входит.

Приведенные методы оценки информативности признаков показывают хорошие результаты в прикладных задачах(см. п.5.2). Однако их недостатком является большая вычислительная сложность указанных методов. Указанные методы трудно применимы для предварительного анализа обучающей информации.

Достаточно эффективным и не требующим больших вычислительных затрат являются предлагаемый ниже метод оценки информативности признаков и отдельных значений признаков, основанный на вычислении близости между объектами по отдельным признакам.

Пусть $S' \in K_i$, $i \in \{1, 2, \ldots, l\}$, $j \in \{1, 2, \ldots, n\}$. Положим

$$\overline{K} = \{K_1, \ldots, K_l\} \setminus K, \qquad \mu_{ij}^{(1)}(S') = \frac{1}{K_i} \sum_{S'' \in K_i} B(S', S'', \{x_j\}),$$

$$\mu_{ij}^{(2)}(S') = \frac{1}{\overline{K}_i} \sum_{S'' \in \overline{K_i}} B(S', S'', \{x_j\})$$

Величины $\mu_{ij}^{(1)}(S')$ и $\mu_{ij}^{(2)}(S')$ характеризуют близость объекта S' соответственно к своему классу и к другим классам. Величину

$$\mu_{ij}(S') = \mu_{ij}^{(1)}(S') - \mu_{ij}^{(2)}(S')$$

назовем весом значения признака x_j для объекта S'. Будем говорить, что значение признака x_j для S' является типичным, если $\mu_{ij}(S') > \mu$, где μ - порог информативности значения признака, $-1 < \mu < 1$. В качестве μ, например, можно взять 0. Тогда значение признака будет являться типичным для класса, если в этом классе оно встречается чаще, чем в остальных.

Множество типичных значений признаков в таблице обучения образует так называемую информативную зону. Далее при построении множества элементарных классификаторов имеет смысл анализировать только те значения признаков, которые попадают в информативную зону, тем самым уменьшается перебор при построении распознающего алгоритма. Кроме того в информативную зону не попадают значения признаков, которые очень редко в встречаются во всех классах (шумящие). Поэтому использование информативной зоны позволяет также снизить влияние шумящих признаков.

Исследование типичности значений признаков позволяет также оценить сложность (в смысле возможности построения качественного алгоритма распознавания) решаемой задачи.

2.2 Выделение типичных объектов в классе для задач распознавания. Разбиение обучающей выборки на базовую и контрольную

При решении прикладной задачи распознавания интересно попытаться оценить эффективность построенного алгоритма при распознавании объектов, не входящих в обучающую выборку. Например, воспользоваться хорошо известным методом скользящего контроля. К сожалению, в ряде прикладных задач алгоритмы, описанные в главе 1, не всегда показывают достаточную высокую эффективность. Такая ситуация возникает, когда классы плохо отделяются друг от друга (в каждом классе есть много объектов, описания которых похожи на объекты, не принадлежащие данному классу). В этом случае построенный алгоритм зачастую, хотя и хорошо распознает "известные"ему объекты (объекты, которые участвовали в построении алгоритма), но плохо распознает "новые"объекты. В данном разделе предлагается подход, позволяющий повысить качество распознающих алгоритмов за счет выделения "типичных"для своих классов объектов. Этот подход продемонстрирован ниже на примере модели голосования по представительным наборам.

Объекты лежащие на границе между классами плохо распознаются и, повидимому, они не позволяют строить короткие представительные наборы. Пусть описание обучающего объекта, не принадлежащего классу K , похоже на описания некоторых объектов из K. Тогда данный объект "лишает"класс K некоторого множества коротких представительных наборов, и это существенно снижает эффективность алгоритма. Для решения указанной проблемы предлагается разбить обучающую выборку на две подвыборки, по первой (базовой) построить множество представительных наборов, по второй (контрольной) вычислить их веса. Причем разбить нужно таким образом, чтобы объекты, находящиеся на границе

между классами, попали в контрольную подвыборку, а все остальные (типичные) объекты - в базовую подвыборку. Практические эксперименты на прикладных задачах подтверждают гипотезу о том, что такое разбиение увеличивает число коротких представительных наборов и тем самым позволяет повысить качество алгоритма распознавания.

Для выделения типичных объектов предлагаются два способа. Первый основан на оценке типичности объекта путем вычисления типичности значений признаков из его описания. Проводится следующая процедура. В таблице обучения для каждого значения признака вычисляется величина $\mu_{ij}(S')$, $i \in \{1, 2, \ldots, l\}$, $j \in \{1, 2, \ldots, n\}$, $S' \in K_i$. (см. разд. 2.1) Задается число μ, где μ - порог информативности значения признака, $-1 < \mu < 1$.

Пусть дано целое число p, $1 \leq p \leq n$. Объект S' будем считать типичным для класса K_i по порогу p , если неравенство $\mu_{ij}(S') > \mu$ выполняется не менее, чем для p признаков.

Заметим, что пороги μ и p можно выбирать из эвристических соображений, например положить $\mu = 0$, $p = [n/2]$. Тогда значение признака x_j для S' будет типичным для класса K_i , если в K_i оно встречается чаще, чем в $\overline{K_i}$. Объект S' будет типичным для K_i, если не менее половины значений признаков в его описании типичны для K_i.

Описанный метод позволяет очень быстро оценить типичность обучающих объектов по отношению к своим классам. Его недостатком является то, что информативность (типичность) значений признака вычисляется независимо от других признаков, т.е. не учитывается тот факт, что, вообще говоря, фрагмент описания некоторого объекта может быть типичен для одного из классов (в этом классе данный фрагмент встречается значительно чаще, чем в остальных класса), но при этом значения признаков которые составляют этот фрагмент встречаются в разных классах примерно одинаковое число раз и не являются типичными ни для одного из классов. Кроме того, необходимо настраивать параметры μ и p, что не очень удобно.

Ниже предлагается метод выделения типичных объектов на основе проведении процедуры скользящего контроля, который заключается в следующем. Из обучающей выборки исключается одни объект S_i, $i \in \{1, 2, \ldots, m\}$. По оставшейся подвыборке $\{S_1, \ldots, S_m\} \backslash S_i$ строится распознающий алгоритм (например, используется модель алгоритма голосования по представительным наборам в классическом варианте). Далее этот алгоритм применяется для распознавания объекта S_i. Объект S_i считается типичным для своего класса, если построенный алгоритм распознал его правильно, и нетипичным для своего класса, если алгоритм отнес его к другому классу или отказался от распознавания. Описанная процедура повторяется для всех объектов обучающей выборке.

Пусть обучающая выборка одним из описанных выше способов разбита на базовую и контрольную подвыборки. По базовой построим множество представительных наборов. Сопоставим каждому построенному представительному набору некий вес, который вычисляется по контрольной подвыборке.

Пусть ω - представительный набор класса K , $K \in \{1, \ldots, l\}$, порождаемый парой (S', H), где S' - объект из базовой выборки, и пусть $\delta(K, \omega)$ - число объектов в контрольной выборке, за которых представительный набор голосует "правильно", $\delta(\overline{K}, \omega)$ - число объектов в контрольной выборке, за которых он голосует "неправильно". Тогда в качестве функции $\nu_{(S',H)}$ (вес элементарного классификатора) можно рассматривать следующие функции:

$$\nu_{(S',H)} = \delta(K, \omega);$$

$$\nu_{(S',H)} = \frac{1 + \delta(K, \omega)}{1 - \delta(\overline{K}, \omega)}.$$

Принадлежность объекта S классу K будем оценивать величиной

$$\Gamma_1(S, K) = \frac{1}{|C^A(K)|} \sum_{(S',H) \in C^A(K)} \nu_{(S',H)}(1 - B(S, S', H)),$$

В качестве информативного веса признака x_j будем рассматривать величину

$$I_j = \frac{\sum\limits_{(S',H) \in C^A(K), x_j \in H} \nu_{(S',H)}}{\sum\limits_{(S',H) \in C^A(K)} \nu_{(S',H)}}$$

Данный подход был апробирован на реальных задачах. Результаты тестирования приведены в главе 5.

2.3 Быстрый метод вычисления оценок при голосовании по представительным наборам для процедуры скользящего контроля

При использовании скользящего контроля в алгоритме голосования по представительным наборам для вычисления оценок обычно используется следующая процедура. Для каждого $i \in \{1, \ldots, m\}$ по выборке $\{S_1, \ldots, S_m\} \setminus S_i$ строится множество представительных наборов, по которым вычисляются величины $\Gamma(S_i, K)$ и $\Gamma(S_i, \overline{K})$. Очевидно, что для больших задач эта процедура требует существенных вычислительных затрат. Ниже предлагается метод, который сокращает время счета примерно в m раз.

Пусть $K \in \{K_1, \ldots, K_l\}$, $S \in K$, $S \in \{S_1, \ldots, S_m\}$. Для простоты рассмотрим случай $l = 2$.

Введем следующие обозначения:

$Q_1(S,K)$ - множество всех таких наборов признаков H, $H=\{x_{j_1},\ldots,x_{j_r}\}$, для которых ни один фрагмент вида (S',H) ,$S'\in\{S_1,\ldots,S_m\}$, $S'\notin K$, не совпадает с фрагментом (S,H) ($Q_1(S,K)$ - множество всех таких наборов признаков H, для которых фрагмент (S,H) является представительным набором для класса K);

$N_1(S,H)$ - число фрагментов вида (S',H), $H\in Q_1(S,K)$, $S'\in\{S_1,\ldots,S_m\}$, $S'\in K$, совпадающих с фрагментом (S,H), считая сам фрагмент (S,H).

$Q_2(S,K)$ - множество всех таких наборов признаков H, $H=\{x_{j_1},\ldots,x_{j_r}\}$, для которых выполнены два следующих условия: 1) ни один фрагмент вида (S',H), $S'\in\{S_1,\ldots,S_m\}$, $S'\in K$, $S'\neq S$, не совпадает с (S,H) ; 2) по крайней мере, один фрагмент вида (S'',H), $S''\in\{S_1,\ldots,S_m\}$, $S''\notin K$, совпадает с (S,H);

$N_2(S,H)$, $H\in Q_2(S,K)$, - число фрагментов вида (S'',H), $S''\in\{S_1,\ldots,S_m\}$, $S''\notin K$, совпадающих с (S,H) ;

$$\lambda_{11}(S,K)=\sum_{H\in Q_1(S,K)}[N_1(S,H)-1];$$

$$\lambda_{12}(S,K)=\sum_{H\in Q_1(S,K)}[N_1(S,H)-1]+\sum_{S'\in KS'\neq S}\sum_{H\in Q_1(S,K)}N_1(S',H);$$

$$\lambda_{21}(S,K)=\sum_{H\in Q_2(S,K)}N_2(S,H);$$

$$\lambda_{22}(S,K)=\sum_{H\in Q_2(S,K)}N_2(S,H)+\sum_{S'\in\overline{K}}\sum_{H\in Q_1(S,\overline{K})}N_1(S',H);$$

Нетрудно видеть, что

$$\Gamma(S,K)=\lambda_{11}(S,K)/\lambda_{12}(S,K) \tag{2.1}$$

$$\Gamma(S,\overline{K})=\lambda_{21}(S,K)/\lambda_{22}(S,K) \tag{2.2}$$

Таким образом, для вычисления требуемых оценок необходимо найти, во-первых, представительные наборы для класса K (эти наборы участвуют в построении множеств $Q_1(S,K)$, $S\in K$) и, во-вторых, найти так называемые $(1,q)$-представительные наборы для K, т.е. фрагменты, которые по данному набору признаков встречаются в K в точности один раз, а в другом классе в точности q раз.

Вычисление величин $\lambda_{11}(S,K)$, $\lambda_{12}(S,K)$, $\lambda_{21}(S,K)$ и $\lambda_{22}(S,K)$, $S\in\{S_1,\ldots,S_m\}$, $K\in\{K_1,\ldots,K_l\}$, может быть организовано следующим образом. Первоначально полагаем эти величины равными нулю. Пусть на очередном шаге найден представительный набор для K, порожденный объектами $S_{i_1},\ldots,S_{i_p}$. Тогда величины $\lambda_{11}(S_{i_t},K)$ и $\lambda_{12}(S_{i_t},K)$ при $t=1,2,\ldots,p$ увеличиваем на $p-1$, а величину

$\lambda_{12}(S_j, K)$ при $j \notin \{i_1, \ldots, i_p\}$ увеличиваем на p. Если же на очередном шаге найден $(1, q)$-представительный набор для K , порожденный объектом S, то величины $\lambda_{21}(S, K)$ и $\lambda_{22}(S, K)$ увеличиваем на q.

Формулы (2.1) и (2.2) легко обобщаются на случай $l > 2$ и на случай тупиковых представительных наборов. Очевидно, что при голосовании по тупиковым представительным наборам в качестве $Q_1(S, K)$ нужно взять множество всех таких наборов признаков H, $H \subseteq \{x_1, \ldots, x_n\}$, для которых фрагмент (S, H) является тупиковым представительным набором для класса K , и в качестве $Q_2(S, K)$ - множество всех таких наборов признаков H, $H \subseteq \{x_1, \ldots, x_n\}$, для которых выполнены три следующих условия: 1) ни один фрагмент вида (S', H), $S' \in \{S_1, \ldots, S_m\}$, $S' \in K$, $S' \neq S$, не совпадает с (S, H); 2) по крайней мере, один фрагмент вида (S'', H), $S'' \in \{S_1, \ldots, S_m\}$, $S' \notin K$, совпадает с (S, H) ; и 3) для каждого t , $t \in \{1, 2, \ldots, r\}$, в K можно указать строку S'_t , $S'_t \neq S$, такую, что $(S'_t, H^{(t)}) = (S, H^{(t)})$, где $H^{(t)} = H \setminus \{x_{j_t}\}$.

Глава 3

Метрические свойства множества σ-покрытий целочисленной матрицы

Традиционно вопросы изучения трудоемкости и качества дискретных процедур распознавания связаны с исследованием метрических (количественных) характеристик множества элементарных классификаторов [1-3, 36-52, 74, 75, 81-83]. В частности, такая информация как типичное число элементарных классификаторов и типичная длина элементарного классификатора позволяет оценить требуемые вычислительные ресурсы, лучше организовать память компьютера и тем самым понизить необходимые требования к вычислительной технике при программной реализации дискретных алгоритмов. Изучение метрических свойств множества элементарных классификаторов напрямую связано с получением асимптотических оценок типичных значений числа (тупиковых) σ-покрытий и длины (тупикового) σ-покрытия целочисленных матриц. Изучаются также метрические свойства подматриц специального вида.

Ранее в [37-40, 42, 44, 45, 47, 48, 50, 52, 86] изучался случай, когда число строк в матрице по порядку меньше числа столбцов, а именно случай, когда $m^\alpha \leq n \leq k^{m^\beta}$, $\alpha > 1$, $\beta < 1$. Показано, что в данном случае величина $|B(L)|$ почти всегда (для почти всех матриц L из M_{mn}^k) при $n \to \infty$ асимптотически совпадает с величиной $|S(L)|$ и по порядку меньше числа покрытий. На основании этого факта был построен асимптотически оптимальный алгоритм поиска покрытий из $B(L)$.

В данной главе рассмотрен прямо противоположный случай, а именно, когда $n^\alpha \leq m \leq k^{n^\beta}$, $\alpha > 1$, $\beta < 1$ [54, 56, 57, 87]. Получены асимптотики типичных значений числа σ-подматриц и порядка σ-подматрицы. Для практически общего случая получены асимптотики типичных значений числа σ-покрытий и длины σ-покрытия.

3.1 Основные определения

Введем следующие обозначения: M_{mn}^k, $k \geq 2$, - множество всех матриц размера $m \times n$ с элементами из $\{0, 1, \ldots, k-1\}$; E_k^r, $k \geq 2$, $r \leq n$, - множество всех

k-ичных наборов длины r, $Q_p(\sigma)$, $\sigma \in E_k^r$, $\sigma = (\sigma_1, \ldots, \sigma_r)$, $p \in \{1, 2, \ldots, r\}$, - множество всех таких наборов $\beta_1, \ldots, \beta_r$ в E_k^r , для которых $\beta_p \neq \sigma_p$ и $\beta_j = \sigma_j$ при $j \in \{1, 2, \ldots, r\} \setminus \{p\}$.

Пусть

W_r^n, $r \leq n$, - множество всех наборов вида $j_1, \ldots, j_r$, где $j_l \in \{1, 2, \ldots, n\}$ при $l = 1, 2, \ldots, r$ и $j_1 < \ldots < j_r$;

Ψ_0 - интервал $(log_k mn, n)$;

Ψ_1 - интервал

$$(\frac{1}{2}\log_k mn - \frac{1}{2}\log_k \log_k mn - \log_k \log_k \log n,$$
$$\frac{1}{2}\log_k mn - \frac{1}{2}\log_k \log_k mn + \log_k \log_k \log n);$$

$a_n \approx b_n$ означает, что $\lim(a_n/b_n) = 1$ при $n \to \infty$;

$a_n \leq_n b_n$ означает, что $a_n \leq b_n$ при всех достаточно больших n.

Пусть $L \in M_{mn}^k$, $\sigma \in E_k^r$.

Набор H из r различных столбцов матрицы L назовем σ-покрытием L, если подматрица L^H, образованная столбцами из H, не содержит строку σ. Набор H из r различных столбцов матрицы L назовем тупиковым σ-покрытием L, если, во-первых, подматрица L^H, образованная столбцами из H, не содержит строку σ, и, во-вторых, если $p \in \{1, 2, \ldots, r\}$, то L^H содержит хотя бы одну из строк $Q_p(\sigma)$.

Заметим, что если $\sigma = (\sigma_1, \ldots, \sigma_r)$, то набор столбцов H матрицы L является тупиковым σ-покрытием в том и только том случае, если выполнены следующие два условия:

1) L^H не содержит строку σ;

2) L^H содержит (с точностью до перестановки строк) подматрицу вида

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

где $\beta_p \neq \sigma_p$ при $p = 1, 2, \ldots, r$. Такую подматрицу будем называть σ-подматрицей.

В частности, тупиковое σ-покрытие булевой матрицы является неприводимым покрытием [44, 45, 47, 48, 50, 52, 83].

Пусть $K \in \{K_1, \ldots, K_l\}$. Таблицу обучения T_{mn} можно рассматривать как пару матриц L_1 и L_2 , где L_1 - матрица, состоящая из описаний обучающих объектов из класса K, а L_2 - матрица, состоящая из описаний остальных обучающих объектов. Тогда очевидно, что элементарный классификатор вида $(\sigma_1, \ldots, \sigma_r)$, задаваемый парой (S_i, H), $S_i \in K$, $H = \{x_{j_1}, \ldots, x_{j_r}\}$, будет (тупиковым) представительным набором для K тогда и только тогда, когда набор столбцов матрицы

L_1 с номерами $j_1, \ldots, j_r$ не является $(\sigma_1, \ldots, \sigma_r)$-покрытием, а набор столбцов L_2 с номерами $j_1, \ldots, j_r$ является (тупиковым) $(\sigma_1, \ldots, \sigma_r)$-покрытием.

Пусть $C(L, \sigma)$ - множество всех пар вида (H, σ) , где H - σ-покрытие матрицы L, $B(L, \sigma)$ - множество всех пар вида (H, σ), где H - тупиковое σ-покрытие матрицы L, $S(L, \sigma)$ - совокупность всех σ-подматриц матрицы L.

Положим

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

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

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

Далее нас будут интересовать асимптотики типичных значений чисел $|C(L)|$ и $|S(L)|$, а также оценка типичного значения отношения $|S(L)|/|B(L)|$. Выявление типичной ситуации будет связано с высказываниями типа «Для почти всех матриц L из M_{un}^k при $n \to \infty$ выполнено свойство β», причем свойство β также может иметь предельный характер. Это означает, что доля тех матриц из M_{un}^k для которых с ε-точностью выполнено свойство β, стремиться к 1 и одновременно ε стремиться к 0 при $n \to \infty$.

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

Лемма 3.1.1. *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 $n \to \infty$ верно $\boldsymbol{M}X_1(L) \approx \boldsymbol{M}X_2(L)$, $\boldsymbol{D}X_2(L)/(\boldsymbol{M}X_2(L))^2 \to 0$. Тогда для почти всех матриц L из M_{mn}^k имеет место $X_2(L) \approx X_1(L) \approx \boldsymbol{M}X_2(L)$, $n \to \infty$.*

Лемма 3.1.2. *Пусть $X(L)$ - случайная величина, определенная на M_{mn}^k, и пусть $\triangle(m, n)$ - доля тех матриц L из M_{mn}^k, для которых $X(L) = 0$. Тогда если $\boldsymbol{M}X(L) \to 0$, $n \to \infty$, то $\triangle(m, n) \to 1$, $n \to \infty$.*

3.2 Асимптотика типичных значений числа σ-покрытий и типичной длины σ-покрытия

В настоящем разделе получены асимптотики типичного числа σ-покрытий и типичной длины σ-покрытия для матрицы из M_{mn}^k в случае, когда $m \leq k^{n^\beta}$, $\beta < 1$.

Отметим, что указанное ограничение на m не является существенным, так как при $m > k^n$ матрица обязательно содержит одинаковые строки). Справедлива

Теорема 3.2.1. *Если $m \le k^{n^\beta}$, $\beta < 1$, то для почти всех матриц L из M_{mn}^k при $n \to \infty$ имеет место*

$$|C(L)| \approx \sum_{r \in \Psi_0} C_n^r k^r$$

и длины почти всех покрытий из $C(L)$ принадлежат интервалу Ψ_0.

Доказательство теоремы 3.2.1 опирается на лемму 3.1.1 и приведенные ниже леммы 3.2.1-3.2.3.

Лемма 3.2.1. *Если $m \le k^{n^\beta}$, $\beta < 1$, то при $n \to \infty$ имеет место*

$$\sum_{r=1}^{n} C_n^r k^r (1 - k^{-r})^m \approx \sum_{r \in \Psi_0} C_n^r k^r (1 - k^{-r})^m.$$

Доказательство. Положим

$$M_r = C_n^r k^r (1 - k^{-r})^m, \quad r \in \{1, 2, \ldots, n\}$$

Пусть $r \le \log_k mn + 1$. Тогда имеем

$$\frac{M_{r-1}}{M_r} = \frac{C_n^{r-1} k^{r-1} (1 - k^{-r+1})^m}{C_n^r k^r (1 - k^{-r})^m} \le \frac{r}{n - r + 1} \le \frac{\log_k mn + 1}{n - \log_k mn} \le o(1)$$

Следовательно, $\sum\limits_{r \notin \Psi_0} M_r = o(\sum\limits_{r \in \Psi_0} M_r)$, при $n \to \infty$. Лемма доказана.

Пусть $w \in W_r^n$ и $\sigma \in E_k^r$. На множестве элементарных событий $M_{mn}^k = \{L\}$ рассмотрим случайную величину $\xi_{(w,\sigma)}(L)$, равную 1, если набор столбцов с номерами из w образует σ-покрытие матрицы L, и равную 0 в противном случае. Положим

$$\xi_1(L) = \sum_{r=1}^{n} \sum_{w \in W_r^n} \sum_{\sigma \in E_k^r} \xi_{(w,\sigma)}(L), \quad \xi_1(L) = \sum_{r \in \Psi_0} \sum_{w \in W_r^n} \sum_{\sigma \in E_k^r} \xi_{(w,\sigma)}(L).$$

Заметим, $|C(L)| = \xi_1(L)$.

Обозначим через $M_{(w,\sigma)}$, $\sigma = (\sigma_1, \ldots, \sigma_r)$, множество всех матриц L в M_{mn}^k таких, что в подматрице матрицы L, образованной столбцами с номерами из w, не содержится строка $(\sigma_1, \ldots, \sigma_r)$.

Лемма 3.2.2. *Если $m \le k^{n^\beta}$, $\beta < 1$, то для почти всех матриц L из M_{mn}^k при $n \to \infty$ имеет место*

$$\boldsymbol{M}\xi_1(L) \approx \boldsymbol{M}\xi_2(L) \approx \sum_{r \in \Psi_0} C_n^r k^r.$$

Доказательство. Очевидно,

$$\mathbf{M}\xi_1(L) = \sum_{r=1}^{n} \sum_{w\in W_r^n} \sum_{\sigma\in E_k^r} \mathbf{P}(\xi_{(w,\sigma)}(L) = 1),$$

где $\mathbf{P}(\xi_{(w,\sigma)}(L) = 1)$ - вероятность того, что $\xi_{(w,\sigma)}(L) = 1$. Очевидно,

$$\mathbf{P}(\xi_{(w,\sigma)}(L) = 1) = |M_{(w,\sigma)}|/|M_{mn}^k| = (1-k^{-r})^m k^{mn}/|M_{mn}^k| = (1-k^{-r})^m.$$

Тогда, в силу леммы 3.2.1,

$$\mathbf{M}\xi_1(L) = \sum_{r=1}^{n} C_n^r k^r (1-k^{-r})^m \approx \sum_{r\in\Psi_0} C_n^r k^r (1-k^{-r})^m.$$

С другой стороны,

$$\mathbf{M}\xi_2(L) = \sum_{r\in\Psi_0} \sum_{w\in W_r^n} \sum_{\sigma\in E_k^r} \mathbf{P}(\xi_{(w,\sigma)}(L) = 1) = \sum_{r\in\Psi_0} C_n^r k^r (1-k^{-r})^m.$$

Требуемая оценка следует из того, что при $r \in \Psi_0$ в условиях леммы имеет место $(1-k^{-r})^m \geq \exp(-2mk^{-r}) \geq \exp(-2/n)$. Лемма доказана.

Лемма 3.2.3. *Если $m \leq k^{n^\beta}$, $\beta < 1$, то имеет место*

$$\frac{\boldsymbol{D}\xi_2(L)}{\boldsymbol{M}(\xi_2(L))^2}, n \to \infty.$$

Доказательство. Имеем

$$\mathbf{D}\xi_2(L) = \mathbf{M}(\xi_2(L))^2 - (\mathbf{M}\xi_2(L))^2.$$

Нетрудно видеть, что

$$\mathbf{M}(\xi_2(L))^2 = \sum_{r,l\in\Psi_0} \sum_{\substack{w_1\in W_r^n \\ w_2\in W_l^n}} \sum_{\substack{\sigma_1\in E_k^r \\ \sigma_2\in E_k^l}} |M|/k^{mn},$$

где $|M| = |M_{(w_1,\sigma_1)} \cap M_{(w_2,\sigma_2)}| \leq k^{mn}$. Отсюда получаем, что

$$\mathbf{M}(\xi_2(L))^2 \leq \sum_{r,l\in\Psi_0} C_n^r k^{r+l} \sum_{a=0}^{\min(r,l)} C_r^a C_{n-r}^{l-a} \leq \sum_{r,l\in\Psi_0} C_n^r C_n^l k^{r+l}.$$

С другой стороны, в силу леммы 3.2.2 имеем

$$(\mathbf{M}\xi_2(L))^2 \approx \sum_{r,l\in\Psi_0} C_n^r C_n^l k^{r+l}.$$

Лемма доказана.

Пусть $r_0 = \log_k m - \log_k(\log_k m \ln kn)$ и пусть

$$C_1(L) = \bigcup_{r\leq r_0} \bigcup_{\sigma\in E_k^r} C(L,\sigma), \quad \xi_3(L) = \sum_{r\leq r_0} \sum_{w\in W_r^n} \sum_{\sigma\in E_k^r} \xi_{(w,\sigma)}(L).$$

Заметим, что $|C_1(L)| = \xi_3(L)$.

Теорема 3.2.2. *Для почти всех матриц $L \in M^k_{(mn)}$ при $n \to \infty$ справедливо*

$$|C_1(L)| = 0.$$

Доказательство. Имеем

$$\mathbf{M}\xi_3(L) = \sum_{r \le r_0} \sum_{w \in W^n_r} \mathbf{P}(\xi_{(w,\sigma)}(L) = 1) = \sum_{r \le r_0} C^r_n k^r (1 - k^{-r})^m$$

(здесь $\mathbf{P}(\xi_{(w,\sigma)}(L) = 1)$ - вероятность того, что $\xi_{(w,\sigma)}(L) = 1$).

Так как при $r \le r_0$

$$a_r = C^r_n k^r (1 - k^{-r})^m \le C^r_n k^r \exp(-mk^{-r}) \le (kn)^{r - \log_k m},$$

то

$$\sum_{r \le r_0} a_r \le (kn)^{r_0 + 1 - \log_k m} \le (kn)^{-\log_k(\log_k m \cdot \ln kn) + 1},$$

а значит, $\mathbf{M}\xi_3(L) \to 0$ при $n \to \infty$. Отсюда, пользуясь леммой 3.1.2, получаем утверждение теоремы.

3.3 Асимптотика типичных значений числа σ-подматриц и порядка σ-подматрицы в случае большого числа строк

В данном разделе получены асимптотики типичного числа σ-подматриц и типичного порядка σ-подматрицы для матрицы из M^k_{mn}.

Теорема 3.3.1. *Если $n^\alpha \le m \le k^{n^\beta}$, $\alpha > 1$, $\beta < 1$, то для почти всех матриц L из M^k_{mn} при $n \to \infty$ справедливо*

$$|S(L)| \approx \sum_{r \in \Psi_1} C^r_n C^r_m r!(k-1)^r k^{r - r^2},$$

и порядки почти всех подматриц из $S(L)$ принадлежат интервалу Ψ_1 .

Доказательство теоремы опирается на лемму 3.1.1 и приводимые ниже леммы 3.3.1-3.3.5.

Пусть V^m_r, $r \le m$, - множество всех упорядоченных наборов вида $(i_1, \dots, i_r)$, где $i_l \in \{1, \dots, m\}$ при $l = 1, 2, \dots, r$ и $i_{l_1} \ne i_{l_2}$ при $l_1, l_2 = 1, 2, \dots, r$, и пусть $v \in V^m_r$, $v = (i_1, \dots, i_r)$, $w \in W^n_r$, $v = (j_1, \dots, j_r)$, $\sigma \in E^r_k$.

Матрицу L из M^k_{mn} назовем (v, w, σ)-матрицей, если в подматрице матрицы L, образованной столбцами с номерами $(j_1, \dots, j_r)$, строка с номером i_p, $p = 1, 2, \dots, r$, принадлежит $Q_p(\sigma)$. Обозначим через $M_{(v,w,\sigma)}$ множество всех (v, w, σ)-матриц в M^k_{mn}. Очевидно,

$$|M_{(v,w,\sigma)}| = (k-1)^r k^{mn - r^2} \tag{3.1}$$

Лемма 3.3.1. *Если* $v_1 \in V_r^m$, $v_2 \in V_l^m$, $w_1 \in W_r^n$, $w_2 \in W_l^n$, $\sigma_1 \in E_k^r$, $\sigma_2 \in E_r^l$ *и наборы* v_1 *и* v_2 *пересекаются по* a $(a \geq 0)$ *элементам, а наборы* w_1 *и* w_2 *пересекаются по* b $(b \geq 0)$ *элементам, то*

$$|M_{(v_1,w_1,\sigma_1)} \cap M_{(v_2,w_2,\sigma_2)}| \leq (k-1)^{r+l-a} k^{mn-r^2-l^2+ab}.$$

Доказательство. Оценим, сколькими способами можно построить матрицу из $M = M_{(v_1,w_1,\sigma_1)} \cup M_{(v_2,w_2,\sigma_2)}$. Сначала выберем те элементы, которые расположены на пересечении строк с номерами из v_1 и столбцов с номерами из w_1. Это можно сделать $(k-1)^r$ способами. Затем выбираем элементы, расположенные на пересечении строк с номерами из v_2 и столбцов с номерами из w_2, учитывая, что ab из них расположены одновременно на пересечении строк с номерами из v_1 и столбцов с номерами из w_2 ($(k-1)^{l-a}$ способов). Произвольным образом доопределяем строки матрицы с номерами из $v_1 \cup v_2$ ($k^{(r+l-a)n+ab-r^2-l^2}$ способов). Выбираем остальные строки произвольно ($k^{mn-(r+l-a)n}$ способов). Из сказанного следуют требуемые оценки для $|M|$.

Лемма 3.3.2. *[47] Если* $n^\alpha \leq m \leq k^{n^\beta}$, $\alpha > 1$, $\beta < 1$, *то при* $n \to \infty$ *имеет место*

$$\sum_{r=1}^{\min(m,n)} C_m^r C_n^r r!(k-1)^r k^{r-r^2} \approx \sum_{r \in \Psi_1} C_m^r C_n^r r!(k-1)^r k^{r-r^2}.$$

Доказательство. Пусть $p = \frac{1}{2}\log_k mn - \frac{1}{2}\log_k \log_k mn$, $q = log_k log_k log_k n$,

$$a_r = C_m^r C_n^r r!(k-1)^r k^{r-r^2}.$$

а) Пусть $r \geq p - q + 1$. Тогда имеем

$$\frac{a_{r+1}}{a_r} = \frac{(n-r)(m-r)}{r+1} k^{-2r+1} \leq \frac{mn}{p} k^{-2p-2q+3} \leq_n 4k^{-2q+3}.$$

Следовательно, $\sum_{r \in [p+q,n]} a_r = o\left(\sum_{r \in \Psi_1} a_r\right)$.

б) Пусть $r \leq p - q + 1$. Тогда имеем

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

$$\frac{2(p-q+1)}{(n-q)(m-q)} k^{2p-2q+1} \leq_n 8k^{-2q+1}$$

Следовательно, $\sum_{r \in [1,p-q]} a_r = o\left(\sum_{r \in \Psi_1} a_r\right)$ Отсюда получаем требуемое утверждение.

Лемма 3.3.3. *[47] Если* $n^\alpha \leq m$, $\alpha > 1$, *и* $r, l \leq \frac{1}{2}\log_k mn$, *то имеет место*

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

где $\delta \to 0$ *при* $n \to \infty$.

Доказательство.

Обозначим $\lambda_a = (k-1)^a k^{la} C_n^r C_r^a C_{n-r}^{l-a} / C_n^r C_{n-r}^l$. Так как

$$\frac{C_r^a C_{n-r}^{l-a}}{C_{n-r}^l} \le \left(\frac{rl}{n-r-l}\right)^a,$$

и по условию $r, l \le \frac{1}{2}\log_k mn$, то

$$\lambda_a \le \left(\frac{(k-1)\log_k^2 m}{m^{\frac{1}{2}}\,(1-(2\log_k m)/m)}\right)^a.$$

При достаточно большом n в силу того, что $n^\alpha \le m$, $\alpha > 1$,

$$\frac{(k-1)\log_k^2 m}{m^{\frac{1}{2}}\,(1-(2\log_k^m)/m)} \le 1$$

и оцениваемая сумма не превосходит

$$C_n^r C_{n-r}^l \left(1 + \frac{(k-1)\log_k^3 m}{m^{\frac{1}{2}}\,(1-(2\log_k m)/m)}\right).$$

Отсюда, пользуясь неравенством $C_{n-r}^l \le C_n^l$, получаем требуемое утверждение.

Пусть $v \in V_r^m$, $w \in W_r^n$, $\sigma \in E_k^r$. На множестве элементарных событий $M_{mn}^k = \{L\}$ рассмотрим случайную величину $\eta_{(v,w)}(L,\sigma)$, равную 1, если $L \in M_{(v,w,\sigma)}$, и равную 0 в противном случае.

Положим

$$\eta_1(L) = \sum_{r=1}^{\min(m,n)} \sum_{\substack{v\in V_r^m \\ w\in W_r^n}} \sum_{\sigma\in E_k^r} \eta_{(v,w)}(L,\sigma),$$

$$\eta_2(L) = \sum_{r\in\Psi_1} \sum_{\substack{v\in V_r^m \\ w\in W_r^n}} \sum_{\sigma\in E_k^r} \eta_{(v,w)}(L,\sigma),$$

Заметим, что $|S(L)| = \eta_1(L)$.

Лемма 3.3.4. *Если $m \le k^{n^\beta}$, $\beta < 1$, то имеет место*

$$\boldsymbol{M}\eta_1(L) \approx \boldsymbol{M}\eta_2(L) \approx \sum_{r=\in\Psi_1} C_m^r C_n^r r!(k-1)^r k^{r-r^2}, n \to \infty.$$

Доказательство. Имеем

$$\mathbf{M}\eta_1(L) = \sum_{r=1}^{\min(m,n)} \sum_{\substack{v\in V_r^m \\ w\in W_r^n}} \sum_{\sigma\in E_k^r} \mathbf{P}(\eta_{(v,w)}(L,\sigma) = 1),$$

где $\mathbf{P}(\eta_{(v,w)}(L,\sigma)=1)$ - вероятность того, что $\eta_{(v,w)}(L,\sigma)=1$. Из (6.1) получаем

$$\mathbf{P}(\eta_{(v,w)}(L,\sigma)=1)=|M_{(v,w,\sigma)}|/|M_{mn}^k|=(k-1)^r k^{-r^2}.$$

Следовательно,

$$\mathbf{M}\eta_1(L)=\sum_{r=1}^{\min(m,n)} C_m^r C_n^r r!(k-1)^r k^{r-r^2},$$

$$\mathbf{M}\eta_2(L)=\sum_{r\in\Psi_1} C_m^r C_n^r r!(k-1)^r k^{r-r^2},$$

Отсюда и из леммы 3.3.2 следует утверждение леммы 3.3.3.

Лемма 3.3.5. *Если $n^\alpha \le m \le k^{n^\beta}$, $\alpha>1$, $\beta<1$, то имеет место*

$$\frac{\boldsymbol{D}\eta_2(L)}{\boldsymbol{M}(\eta_2(L))^2}\to 0, n\to\infty.$$

Доказательство. Имеем

$$\mathbf{D}\eta_2(L)=\mathbf{M}(\eta_2(L))^2-(\mathbf{M}\eta_2(L))^2. \tag{3.2}$$

Нетрудно видеть, что

$$\mathbf{M}(\eta_2(L))^2=\sum_{r,l\in\Psi_1}^{n}\sum_{\substack{v_1\in V_r^m, w_1\in W_r^n\\ v_2\in V_l^m, w_2\in W_l^n}}\sum_{\substack{\sigma_1\in E_k^r\\ \sigma_2\in E_k^l}}|M|/k^{mn},$$

где $M=M_{v_1,w_1,\sigma_1}\cap M_{v_2,w_2,\sigma_2}$. Пользуясь леммой 3.3.1, получаем

$$\mathbf{M}(\eta_2(L))^2\le\sum_{r,l\in\Psi_1}\sum_{b=0}^{\min(r,l)} k^{r+l}(k-1)^{r+l}k^{-r^2-l^2+lb}C_n^r C_r^b C_{n-r}^{l-b} C_m^r r! C_m^l l!.$$

Отсюда в силу леммы 3.3.3 имеем

$$\mathbf{M}(\eta_2(L))^2\le\sum_{r,l\in\Psi_1} C_n^r C_n^l C_m^r r! C_m^l l! k^{r+l}(k-1)^{r+l}k^{-r^2-l^2}[1+\delta(n)] \tag{3.3}$$

где $\delta(n)\to 0$, $n\to\infty$.

С другой стороны, в силу леммы 3.3.4,

$$(\mathbf{M}\eta_2(L))^2\approx\sum_{r,l\in\Psi_1} C_n^r C_n^l C_m^r r! C_m^l l! k^{r+l}(k-1)^{r+l}k^{-r^2-l^2}. \tag{3.4}$$

Из (3.2), (3.3) и (3.4) следует утверждение доказываемой леммы.

Утверждение теоремы 3.3.1 следует из лемм 3.3.4, 3.3.5 и леммы 3.1.1.

Пусть $r_1=\log_k mn$, и пусть

$$S_1(L)=\bigcup_{r\ge r_1}\bigcup_{\sigma\in E_k^r} S(L,\sigma), \eta_3=\sum_{r\ge r_1}\sum_{\substack{v\in V_r^m\\ w\in W_r^n}}\sum_{\sigma\in E_r^r}\eta_{(v,w)}(L,\sigma).$$

Заметим, что $|S_1(L)|=\eta_3(L)$.

Теорема 3.3.2. *Для почти всех матриц $L \in M_{mn}^k$ при $n \to \infty$ справедливо*

$$|S_1(L)| = 0.$$

Доказательство. Имеем

$$\mathbf{M}\eta_3(L) = \sum_{r \geq r_1} \sum_{\substack{v \in V_r^m \\ w \in W_r^n}} \sum_{\sigma \in E_r^r} \mathbf{P}(\eta_(v, w)(L, \sigma) = 1),$$

где $\mathbf{P}(\eta_(v, w)(L, \sigma) = 1)$ - вероятность того, что $\eta_(v, w)(L, \sigma) = 1$. Следовательно, в силу (3.1),

$$\mathbf{M}\eta_3(L) = \sum_{r \geq r_1} C_n^r C_m^r r!(k-1)^r k^{r-r^2}.$$

Так как при $r \geq r_1$

$$C_n^r C_m^r r!(k-1)^r k^{r-r^2} \leq \frac{(mn)^r}{r!} r^{2r-r^2} \leq \left(\frac{k^2 e}{r}\right)^r,$$

то при достаточно большом n

$$\sum_{r \geq r_1}^{n} C_n^r C_m^r r!(k-1)^r k^{r-r^2} \leq n \left(\frac{k^2 e}{\log_k mn}\right)^{\log_k mn} \to 0, \text{ при } n \to \infty.$$

Теперь утверждение теоремы следует из леммы 3.1.2.

Замечание. Из теорем 3.2.1 и 3.3.2 следует, что при $n \to \infty$ для почти всех матриц L из M_{mn}^k число тупиковых покрытий не превосходит величины

$$\sum_{r \in \Psi_2} C_n^k k^r,$$

где Ψ_2 - интервал

$$(\log_k m - \log_k(log_k m \cdot \ln kn), \log_k mn).$$

Теорема 3.3.3. *Если $n^\alpha \leq m \leq k^{n^\beta}$, $\alpha > 1$, $\beta < 1/2$, то для почти всех матриц L из M_{mn}^k при $n \to \infty$ справедливо $|S(L)|/|B(L)| \to \infty$.*

Доказательство. Согласно теореме 3.3.1 для почти всех матриц L из M_{mn}^k при $n \to \infty$ справедливо

$$|S(L)| \gtrsim \sum_{r \in \Psi} (mn)^{r-1}(1 - r/n)^r (1 - r/m)^r / r!.$$

Так как

$$(1 - r/n)^r \geq_n \exp\left(-2r^2/n\right) \geq \exp\left(\frac{-2\log_k^2 m}{n}\right) \geq \exp\left(-2n^{2\beta-1}\right)$$

и, аналогично,

$$(1 - r/m)^r \geq_n \exp\left(\frac{-2\log_k^2 m}{m}\right),$$

то имеем

$$|S(L)| \gtrsim \sum_{r\in\Psi_1} (mn)^{r-1}/[\log_k mn]! \geq (mn)^{r_1-1}/[\log_k mn]!, \tag{3.5}$$

где

$$r_1 = \frac{1}{2}\log_k mn - \frac{1}{2}\log_k \log_k mn - \log_k \log_k \log_k n.$$

В силу замечания, для почти всех матриц L из $M_m n^k$ справедливо

$$|B(L)| \leq \sum_{r\leq \log_k mn} C_n^k k^r.$$

Пусть $a_r = C_n^k k^r$. Тогда

$$\frac{a_{r-1}}{a_r} \leq \frac{r}{kn\,(1-(r-1)/n)} \to 0, n \to \infty.$$

Следовательно,

$$|B(L)| \leq_n C_n^{[\log_k mn]} mn \leq \frac{n^{\log_k mn} mn}{[\log_k mn]!}.$$

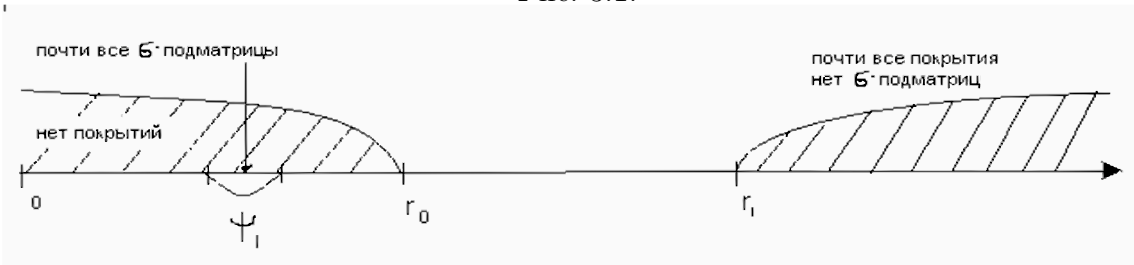
Отсюда и из (3.5) следует, что

$$|S(L)|/|B(L)| \gtrsim (mn)^{r_1-2}/n^{\log_k mn} \geq$$
$$\geq n^{(\alpha+1)(r_1-2)-\log_k mn} \geq$$
$$\geq n^{\frac{1}{2}(\alpha-1)\log_k mn-(\alpha+1)\log_k \log_k mn} \to \infty, n \to \infty.$$

Теорема доказана. Из теоремы 3.3.3 следует, что в случае, когда число строк матрицы по порядку больше числа столбцов, почти всегда величина $|S(L)|$ по порядку больше величины $|B(L)|$. Ранее (см. [39, 47, 52]) было установлено, что в противоположном случае, когда число строк в матрице по порядку меньше числа столбцов, почти всегда $|S(L)|$ асимптотически совпадает с $|B(L)|$.

Для случая $n^{\alpha} \leq m \leq k^{n^{\beta}}$, $\alpha > 1$, $\beta < 1$, результаты, полученные в теоремах 3.2.1 - 3.3.2 проиллюстрированы на рисунке 3.1.

Рис. 3.1:



Ψ_0 - интервал $(log_k mn, n)$;

Ψ_1 - интервал

$$(\frac{1}{2}\log_k mn - \frac{1}{2}\log_k \log_k mn - \log_k \log_k \log n,$$
$$\frac{1}{2}\log_k mn - \frac{1}{2}\log_k \log_k mn + \log_k \log_k \log n);$$

$r_0 = \log_k m - \log_k(\log_k m \ln kn)$, $r_1 = \log_k mn$

Глава 4

Конструирование дискретных процедур распознавания с использованием аппарата логических функций

В данной главе основные принципы конструирования дискретных процедур распознавания изложены с использованием аппарата логических функций. Рассмотрена связь между задачей нахождения покрытий целочисленной матрицы с элементами из множества $\{0, 1, \ldots, k-1\}$ и задачей построения сокращенной дизъюнктивной нормальной формы (ДНФ) двузначной логической функции, заданной на k-ичных наборах [52, 58, 86]. На основе теорем, доказанных в главе 3, получены новые результаты касающиеся изучения метрических свойства множества допустимых, неприводимых и максимальных конъюнкций логических функций, заданных множеством нулей. Полученные результаты имеют также значение для классической дискретной математики.

4.1 Связь задач построения множества элементарных классификаторов, построения нормальных форм логических функций и поиска покрытий целочисленных матриц

Все неопределяемые ниже понятия можно найти в [29].

Пусть на наборах из E_k^n задана частично определенная логическая функция f, принимающая значения из множества $\{0, 1\}$, A_f - множество единиц этой функции, B_f - множество ее нулей. Введем ряд определений.

Пусть переменная x принимает значения из множества E_k^n, $\sigma \in E_k^n$. Введем обозначение

$$x^\sigma = \begin{cases} 1, & \text{если } x = \sigma; \\ 0, & \text{если } x \neq \sigma. \end{cases}$$

Элементарной конъюнкцией (э.к.) называется выражение вида $x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$, где все x_{j_i} различны. Число r называется рангом конъюнкции. Интервал истинности

э.к. $\mathfrak{B}$ обозначается через $N_{\mathfrak{B}}$.

Э.к. $\mathfrak{B}$ называется допустимой для f, если $N_{\mathfrak{B}} \cap A_f \neq \emptyset$ и $N_{\mathfrak{B}} \cap B_f = \emptyset$.

Э.к. $\mathfrak{B}$ называется неприводимой для f, если не существует элементарной конъюнкции $\mathfrak{B}'$ такой, что $N_{\mathfrak{B}'} \supset N_{\mathfrak{B}}$ и $N_{\mathfrak{B}'} \cap B_f = N_{\mathfrak{B}} \cap B_f$ [58, 86].

Э.к. $\mathfrak{B}$ называется максимальной для F, если $\mathfrak{B}$ допустимая и не существует допустимой конъюнкции $\mathfrak{B}'$ такой, что $N_{\mathfrak{B}'} \supset N_{\mathfrak{B}}$.

Заметим, что определения допустимой, неприводимой и максимальной конъюнкций полностью переносятся и на случай всюду определенной логической функции, т.е. на случай, когда $A_f = E_k^n \setminus B_f$.

Задачу построения сокращенной ДНФ функции f обычно сводят к задаче построения сокращенной ДНФ функции F, принимающей значение 0 на наборах из B_f и значение 1 на остальных наборах из E_k^n. После построения ДНФ функции F из нее удаляют конъюнкции $\mathfrak{B}$, не обладающие свойством $N_{\mathfrak{B}} \cap A_f \neq \emptyset$ [29].

Пусть A_f состоит из наборов $(\alpha_{11}, \alpha_{12}, \ldots, \alpha_{1n})$, …, $(\alpha_{u1}, \alpha_{u2}, \ldots, \alpha_{un})$, B_f - из наборов $(\beta_{11}, \beta_{12}, \ldots, \beta_{1n})$, …, $(\beta_{u1}, \beta_{u2}, \ldots, \beta_{un})$.

Из наборов A_f составим матрицу L_1 вида

$$\begin{bmatrix} \alpha_{11} & \alpha_{12} & \ldots & \alpha_{1n} \\ \alpha_{21} & \alpha_{22} & \ldots & \alpha_{2n} \\ \ldots & & & \\ \alpha_{u1} & \alpha_{u2} & \ldots & \alpha_{un} \end{bmatrix}$$

Из наборов B_f составим матрицу L_2 вида

$$\begin{bmatrix} \beta_{11} & \beta_{12} & \ldots & \beta_{1n} \\ \beta_{21} & \beta_{22} & \ldots & \beta_{2n} \\ \ldots & & & \\ \beta_{u1} & \beta_{u2} & \ldots & \beta_{un} \end{bmatrix}$$

Очевидными являются приводимые ниже утверждения 4.1.1-4.1.4

Утверждение 4.1.1. *Э.к. $x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$ является допустимой для F тогда и только тогда, когда набор столбцов с номерами $j_1, \ldots, j_r$ является $(\sigma_1, \ldots, \sigma_r)$-покрытием матрицы L_2.*

Утверждение 4.1.2. *Э.к. $x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$ является неприводимой для F тогда и только тогда, когда в наборе столбцов с номерами $j_1, \ldots, j_r$ матрицы L_2 содержится $(\sigma_1, \ldots, \sigma_r)$-подматрица.*

Утверждение 4.1.3. *Э.к. $x_{j_1}^{\sigma_1} \ldots x_{j_r}^{\sigma_r}$ является максимальной для F тогда и только тогда, когда набор столбцов с номерами $j_1, \ldots, j_r$ является тупиковым $(\sigma_1, \ldots, \sigma_r)$-покрытием матрицы L_2.*

Утверждение 4.1.4. *Э.к.* $x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r}$ *является максимальной для* f *тогда и только тогда, когда набор столбцов с номерами* $j_1, \dots, j_r$ *является тупиковым* $(\sigma_1, \dots, \sigma_r)$*-покрытием матрицы* L_2 *и подматрица матрицы* L_1*, образованная столбцами с номерами* $j_1, \dots, j_r$*, содержит хотя бы одну из строк вида* $(\sigma_1, \dots, \sigma_r)$*.*

Построить сокращенную ДНФ логической функции можно также путем преобразования конъюнктивной формы, которая строится по наборам из B_F.

$$D_1 \& D_2 \& \dots \& D_u, \tag{4.1}$$

где $D_i = \overline{x_1^{\beta_{i1}}} \vee \overline{x_2^{\beta_{i2}}} \vee \dots \vee \overline{x_n^{\beta_{in}}}$, $i = 1, 2, \dots, u$.

Очевидно, конъюнктивная форма (4.1) реализует функцию F.

Воспользуемся равенством

$$\overline{x^\alpha} = \bigvee_{\beta \neq \alpha} x^\beta,$$

Тогда конъюнктивная форма (4.1) примет вид

$$D_1^* \& D_2^* \& \dots \& D_u^*,$$

$$D_i^* = \bigvee_{\gamma \neq \beta_{i1}} x_1^\gamma \vee \bigvee_{\gamma \neq \beta_{i2}} x_2^\gamma \vee \dots \vee \bigvee_{\gamma \neq \beta_{in}} x_n^\gamma, i = 1, 2, \dots, u.$$

Далее производится процедура логического перемножения скобок аналогичная бинарному случаю.

Рассмотрим ситуацию, когда объекты из исследуемого множества M описаны признаками, каждый из которых принимает значения из множества $\{0, 1, \dots, k-1\}$. Тогда

Элементарному классификатору (σ, H), где где $\sigma = (\sigma_1, \dots, \sigma_r)$, H - набор признаков с номерами $j_1, \dots, j_r$ поставим в соответствие э.к. $\mathfrak{B} = x_{j_1}^{\sigma_1} \dots x_{j_r}^{\sigma_r}$. Если $S = (a_1, \dots, a_n)$ - объект из множества M, то, очевидно, $B(\sigma, S, H) = 1$ тогда и только тогда, когда $(a_1, \dots, a_n) \in N_{\mathfrak{B}}$.

Покажем, что построение множества элементарных классификаторов класса K для рассмотренных в главе 1 моделей сводится к нахождению допустимых и максимальных конъюнкций для характеристической функции класса K, т.е. такой двузначной логической функции, которая на обучающих объектах из K и $\overline{K}$ принимает разные значения. Распознавание объекта $S = (a_1, \dots, a_n)$ осуществляется на основе голосования по построенным элементарным конъюнкциям. Приведем примеры характеристических функций для рассмотренных в главе 1 моделей процедур распознавания.

1. Алгоритм голосования по представительным наборам.

В данном случае характеристическая функция класса K - частичная логическая функция f_K от n переменных, принимающая значение 1 на описаниях объектов из K, значение 0 на объектах из $\overline{K}$ и не определенная на остальных наборах из E_k^n. Представительному набору класса K соответствует допустимая конъюнкция для f_K, тупиковому представительному набору соответствует максимальная конъюнкция для f_K. Допустимая (максимальная) конъюнкция $\mathfrak{B}$ голосует за принадлежность объекта S классу K, если $(a_1, \ldots, a_n) \in N_{\mathfrak{B}}$.

2. Алгоритм голосования по антипредставительным наборам.

Характеристическая функция класса K - частичная логическая функция $f_{\overline{K}}$ от n переменных, принимающая значение 0 на описаниях объектов из K, значение 1 на объектах из $\overline{K}$ и не определенная на остальных наборах из E_k^n. Антипредставительному набору класса K соответствует допустимая конъюнкция для функции $f_{\overline{K}}$, а тупиковому антипредставительному набору - максимальная конъюнкция для $f_{\overline{K}}$. Допустимая (максимальная) конъюнкция $\mathfrak{B}$ голосует за принадлежность объекта S классу K, если $(a_1, \ldots, a_n) \notin N_{\mathfrak{B}}$.

3. Алгоритм голосования по покрытиям класса.

Характеристическая функция класса K - всюду определенная логическая функция $F_{\overline{K}}$, принимающая значение 0 на описаниях объектов из K и значение 1 на остальных наборах из E_k^n. Покрытию класса K соответствует допустимая для $F_{\overline{K}}$ конъюнкция, тупиковому покрытию - максимальная для $F_{\overline{K}}$ конъюнкция. Допустимая (максимальная) конъюнкция $\mathfrak{B}$ голосует за принадлежность объекта S классу K, если $(a_1, \ldots, a_n) \notin N_{\mathfrak{B}}$.

Таким образом, построение множества элементарных классификаторов для класса K сводится к следующему. Задается характеристическая функция. Далее строится ДНФ, реализующая эту функцию. Наибольшую сложность представляет построение ДНФ из максимальных конъюнкций (сокращенной ДНФ) функции f_K.

4.2 Метрические свойства дизъюнктивных нормальных форм двузначных логических функций, определенных на k-ичных n-мерных наборах

Пусть в пространстве E_k^n задана логическая функция F, принимающая значения из множества $\{0, 1\}$, B_F множество наборов, на которых функция F принимает значение 0, $|B_F| = u$. Множество всех таких функций обозначим через $\mathfrak{F}_{un}^k$.

Пусть Ψ_k^0 - интервал $(log_k un, n)$;

Ψ_k^1 - интервал

$$(\frac{1}{2}\log_k un - \frac{1}{2}\log_k\log_k un - \log_k\log_k\log n,$$
$$\frac{1}{2}\log_k un - \frac{1}{2}\log_k\log_k un + \log_k\log_k\log n).$$

$\mathfrak{C}(F,r)$ - множество всех допустимых конъюнкций ранга r для функции F ; $\mathfrak{S}(F,r)$ - множество всех максимальных конъюнкций ранга r для функции F;

$$\mathfrak{C}(F) = \bigcup_{r=1}^{n} \mathfrak{C}(F,r),$$

$$\mathfrak{S}(F) = \bigcup_{r=1}^{n} \mathfrak{S}(F,r).$$

Теорема 4.2.1. *Если $u \le k^{n^\beta}$, $\beta < 1$, то для почти всех функций F из $\mathfrak{F}_{un}^k$ при $n \to \infty$ имеет место*

$$|\mathfrak{C}(F)| \approx \sum_{r\in\Psi_k^0} C_n^r k^r$$

и ранги почти всех конъюнкций из $\mathfrak{C}(F)$ принадлежат интервалу Ψ_k^0.

Доказательство теоремы следует из утверждения 4.1.1 и теоремы 3.2.1.

Следствие. *Если $u \le 2^{n^\beta}$, $\beta < 1$, то для почти всех функций F из $\mathfrak{F}_{un}^2$ при $n \to \infty$ имеет место*

$$|\mathfrak{C}(F)| \approx \sum_{r\in\Psi_2^0} C_n^r 2^r$$

и ранги почти всех конъюнкций из $\mathfrak{C}(F)$ принадлежат интервалу Ψ_2^0.

Пусть $r_0 = \log_k u - \log_k(\log_k u \ln kn)$ и пусть

$$\mathfrak{C}_1(F) = \bigcup_{r\le r_0} \mathfrak{C}(F,r).$$

Справедлива

Теорема 4.2.2. *Для почти всех функций $F \in \mathfrak{F}_{(un)}^k$ при $n \to \infty$ справедливо*

$$|\mathfrak{C}_1(F)| = 0.$$

Доказательство теоремы следует из утверждения 4.1.1 и теоремы 3.2.2.

Теорема 4.2.3. *Если $n^\alpha \le u \le k^{n^\beta}$, $\alpha > 1$, $\beta < 1$, то для почти всех функций F из $\mathfrak{F}_{un}^k$ при $n \to \infty$ справедливо*

$$|\mathfrak{S}(F)| = o(\sum_{r\in\Psi_k^1} C_n^r C_u^r r!(k-1)^r k^{r-r^2}),$$

Доказательство теоремы следует из утверждения 4.1.3 и теорем 3.3.1 и 3.3.3.

Следствие. *Если* $n^\alpha \le u \le 2^{n^\beta}$, $\alpha > 1$, $\beta < 1$, *то для почти всех функций* F *из* $\mathfrak{F}^2_{un}$ *при* $n \to \infty$ *справедливо*

$$|\mathfrak{S}(F)| = o(\sum_{r \in \Psi^1_2} C^r_n C^r_u r! 2^{r-r^2}),$$

Пусть $r_1 = \log_k un$ и пусть

$$\mathfrak{S}_1(F) = \bigcup_{r \ge r_1} \mathfrak{S}(F, r).$$

Справедлива

Теорема 4.2.4. *Для почти всех функций* $F \in \mathfrak{F}^k_{un}$ *при* $n \to \infty$ *справедливо*

$$|\mathfrak{S}_1(F)| = 0.$$

Доказательство теоремы следует из утверждения 4.1.2 и теорем 3.3.2 и 3.3.3.

Глава 5

Апробация предложенных методов на реальных задачах

5.1 Решение задач прогнозирования результатов лечения онкозаболеваний

На задачах из области медицинского прогнозирования было проведено сравнение классических конструкций и новых моделей, описанных в данной работе. Рассматривались задачи прогнозирования результатов лечения онкозаболеваний. Остеогенная саркома - раковое заболевание костей, от которого страдают в основном молодые люди (у людей старшего возраста эта болезнь практически не встречается). К сожалению, вероятность летального исхода в случае заболевания остеогенной саркомой очень велика. Для лечения саркомы используются в основном химические методы, заключающиеся в том, что больной принимает в течение некоторого времени небольшие дозы ядовитых веществ. Так как раковые клетки растут гораздо быстрее здоровых клеток организма, то яд они потребляют быстрее. В результате раковая опухоль начинает разрушаться еще до того, как отравляется организм.

В этой области существенными для рассмотрения представляются следующие две задачи: задача выживаемости (прожил ли человек год после лечения) и задача прогнозирования патаморфоза, т.е. степени деструкции опухоли после химиотерапии (высокая или низкая степень деструкции опухоли). Как было показано предыдущими исследованиями, прогнозирование выживаемости - очень трудная задача, так как помимо состояния клеток опухоли есть масса других объективных факторов, таких как иммунитет человека, его психическое состояние, окружающая среда и т.д., воздействие которых играет немаловажную роль. Задача прогнозирования степени патаморфоза решается достаточно хорошо, поскольку в этой задаче состояние клеток опухоли играет решающую роль, а влияние остальных факторов значительно меньше.

В каждой из задач обучающая выборка состояла из 77 объектов (больных), разбитых на два класса. В задаче прогнозирования выживаемости мощности

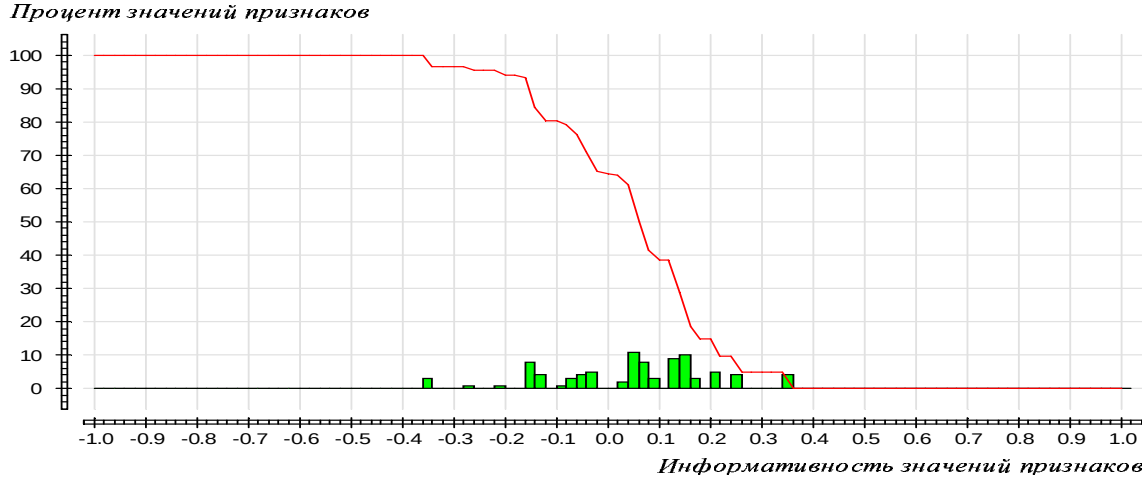
классов были 52 и 25 объектов, в задаче прогнозирования степени патаморфоза - 47 и 30 объектов. Объекты были описаны в системе из семи признаков. Содержательно признаки - это некоторые характеристики раковой опухоли. Значность каждого признака была равна трем.

Для оценки эффективности процедур распознавания использовался метод скользящего контроля.

Тестирование показало, что в модели голосования по представительным наборам при решении обеих задач достаточно ограничиться построением представительных наборов длины 3. В случае, когда при построении алгоритма распознавания добавлялись представительные наборы большей длины, эффективность алгоритма оказывалась такой же. При добавлении представительных наборов меньшей длины эффективность алгоритма понижалась. Классическая модель голосования по представительным наборам (при длине представительных наборов равной 3) показала эффективность равную 61%, на задаче выживаемости и 83% - на задаче патаморфоза. В то время как эффективность алгоритма голосования по покрытиям класса составила соответственно 75% и 62%.

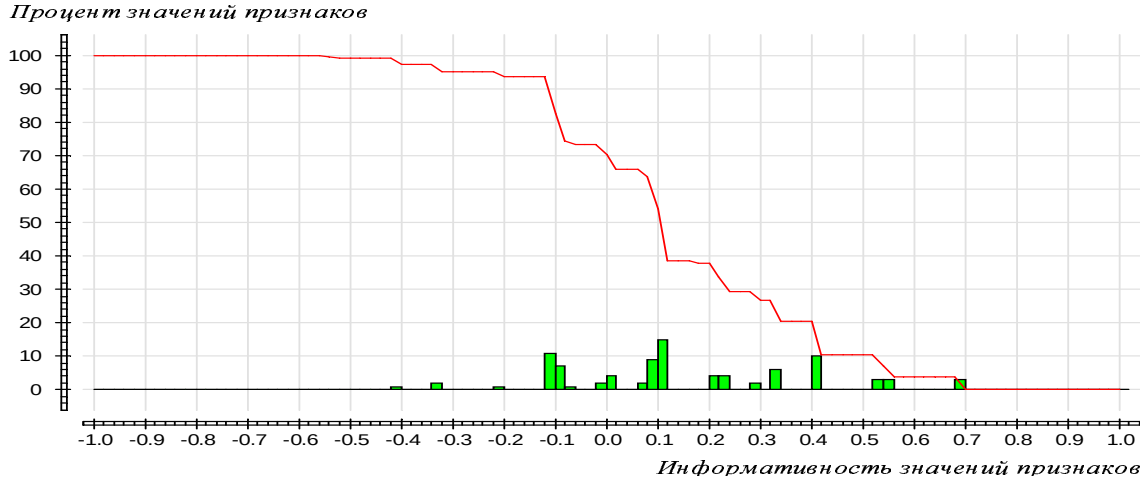
Для того чтобы разобраться в причинах низкой эффективности классического алгоритма при решении задачи выживаемости, был проведен анализ информативности значений признаков согласно методике, описанной в разд. 2. Результаты анализа для задачи выживаемости и задачи патаморфоза приведены соответственно на рис. 5.1 и 5.2. На этих рисунках отражена зависимость числа (в %) значений признаков, которые считаются типичными в зависимости, от выбора порога μ минимальной информативности. Диаграммы показывают распределение весов значений признаков (высота столбика - число (в %) значений признаков в описаниях объектов из обучающей выборки, которые имеют вес, содержащийся в данном интервале).

Рис. 5.1: Распределение типичности значений признаков в задаче выживаемости



Из рассмотрения рис.5.1 видно, что большинство значений признаков имеют вес, близкий к нулю. Это означает, что в обоих классах эти значения встречаются одинаково часто. Другими словами, объекты из разных классов трудно отделимы друг от друга, что и является причиной низкой эффективности классического алгоритма распознавания.

Рис. 5.2: Распределение типичности значений признаков в задаче патаморфоз



Из рассмотрения рис.5.2 видно, что в задаче патаморфоза часть значений признаков имеют вес близкий к нулю, но при этом много таких значений, которые обладают довольно большим весом, то есть являются очень типичными для одного из классов.

Для повышения эффективности распознающих алгоритмов был использован следующий подход. Исходная выборка была разбита на базовую и контрольную двумя способами, а именно методом скользящего контроля и методом, основанным на оценке типичности значений признаков по отношению к классам (см. разд. 2). Эффективность построенных алгоритмов оценивалась числом правильно распознанных объектов при использовании метода скользящего контроля. Более точно проводилась следующая процедура. Из обучающей выборки удалялся один объект, оставшиеся объекты разбивались на базовую и контрольную подвыборки, далее по базовой подвыборке строились представительные наборы, а по контрольной - вычислялись их веса. Проводилось голосование по представительным наборам с учетом весов и принималось решение об отнесении удаленного объекта к тому или иному классу. Эта процедура повторялась для каждого объекта из обучающей выборки. Результаты счета представлены в таблице.

	Выживаемость	Патаморфоз
Классическая модель	61	83
Разбиение методом скользящего контроля	75	94
Разбиение на основе оценки типичности значений признаков	75	92

Таким образом, использование предложенных в работе методов позволяет значительно повысить качество распознающего алгоритма. Причина повышения эффективности алгоритма заключается в следующем. Проанализируем построенные множества представительных наборов, точнее, мощности этих множеств. В задаче выживаемости при использовании классической модели число построенных представительных наборов для первого и второго классов равно соответственно 472 и 154. Если же строить представительные наборы только по типичным для классов объектам и для выделения типичных объектов использовать, например, метод скользящего контроля, то указанные величины составляют соответственно 730 и 252. В задаче прогнозирования патаморфоза число представительных наборов в классической модели - 657 и 468, а в модели с разбиением на базовую выборку и контрольную соответственно 849 и 668.

Эти данные подтверждают предположение о том, что появление нетипичных для класса объектов уменьшает число коротких представительных наборов и тем самым снижает качество построенного алгоритма.

5.2 Оценка важности признаков в задаче анализа результатов социологического опроса

Описанные в главе 2 методы были протестированы на результатах социологического опроса предоставленного Информационно-социологическим центром Российской академии государственной службы при Президенте Российской Федерации. Целью опроса было изучение отношения людей к политической жизни страны. Анкетирование проводилось в разных регионах страны. Существовало две анкеты одна для обычных людей, другая для государственных служащих. Анкеты отличались только несколькими пунктами, связанными с родом занятий. Каждая из анкет состояла из 80 вопросов, ответы на которые кодировались целыми числами. В результате кодирования число вопросов возросло до 100 (это связано с тем, что вопросы, предполагающие выбор сразу нескольких вариантов ответов, разбивались на несколько подвопросов) В опросе приняло участие 1629 людей и 806 государственных служащих. Таким образом информация представляет собой 2 таблицы размерностью 1629x100 и 806x100, где столбцы таблиц

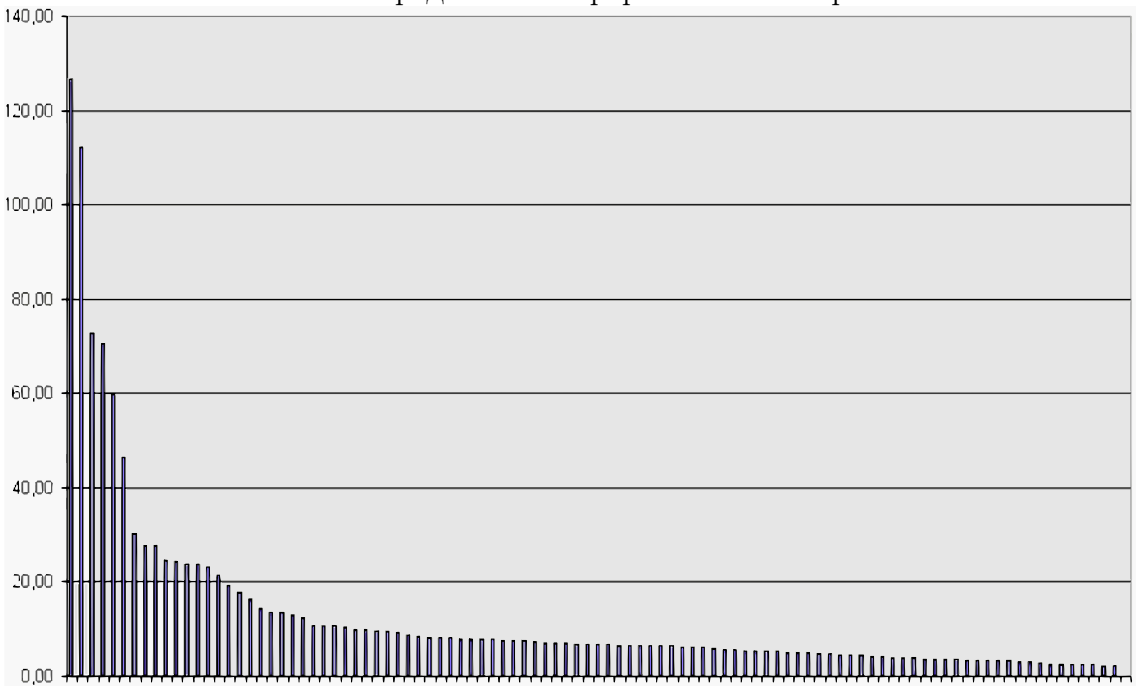
(признаки) - вопросы, а строки (объекты) - респонденты. На этой информации была поставлена следующая задача. Задавался целевой признак, например вопрос об отношении респондента к политическим партиям и движениям. Один из таких вопросов звучит примерно так: «Что преобладает в Вашем отношении к компартии Г. Зюганова и ее сторонникам?» Возможные варианты ответа: симпатия, равнодушие, неприязнь, трудно сказать. Ответами на целевой вопрос исходная таблица разбивается на 4 класса. Требовалось определить информативность признаков (значений признаков) при таком разбиении. Под информативностью признака подразумевалось то, насколько по его значениям можно судить о принадлежности объекта к тому или иному классу. При решении поставленной задачи задачи с использовании представительных наборов, пришлось отказаться от требования тупиковости представительного набора, так как проверка тупиковости значительно снижает скорость работы алгоритма. Использовались представительные наборы ограниченной длинны. Максимальная длина набора бралась равной 3. При меньшей максимальной длине большая часть объектов не содержала ни одного представительного набора. А увеличение максимальной длинны до 4, резко увеличивало время работы алгоритма. Был получен следующий результат (см. рис.5.3). Если расположить признаки в порядке убывания информативности, то, как правило, в каждом классе есть выделенная группа признаков с большой информативностью, далее идет некоторый разрыв и потом оставшиеся признаки выстраиваются в ряд с плавно уменьшающейся информативностью. Причем, как при использовании метода, основанного на построении (p, q)-представительных наборов (см. п.2.1), так и метода основанного на выделении типичных объектов (см. п.2.2) результат получался примерно одинаковым, т.е. оба метода в группу наиболее информативных признаков выделяли одни и те же признаки, а порядок признаков различался не значительно.

Еще одним интересным результатом является то, что информативность набора значений признаков может значительно (иногда на порядок) превышать вес признаков, которые его составляют. Другими словами фрагмент порожденный двумя признаками, может значительно сильнее характеризовать один из классов, чем значения каждого из указанных признаков в отдельности.

Например, доля респондентов первого класса (симпатизирующих компартии Г. Зюганова и ее сторонникам), ответивших на вопрос «Хотели бы вы поддержать какую-либо политическую партию при выводе России из кризиса?» «безусловно», среди государственных служащих составила 0,38, доля указанных респондентов, принадлежащих остальным классам - 0,15. Доля респондентов, ответивших на вопрос «Хотели бы вы поддержать Президента и правительство партию при выводе России из кризиса?» в первом классе составила 0,42, а в остальных классах 0,20.

Доля респондентов «безусловно» желающих поддержать какую-либо поли-

Рис. 5.3: Распределение информативности признаков



тическую партию и не желающих поддержать Президента и правительство при выводе России из кризиса в первом классе составила 0,22, а в остальных классах 0,01. Таким образом, совокупность указанных ответов гораздо сильнее характеризует первый класс, чем каждый из ответов в отдельности. Следовательно, рассмотрение совокупностей признаков в некоторых задачах дает более интересные результаты.

ЗАКЛЮЧЕНИЕ

Работа посвящена исследованию дискретных процедур распознавания. При построении этих процедур важнейшим этапом является поиск информативных фрагментов признаковых описаний объектов. В работе предложены новые подходы к поиску таких фрагментов.

(1) Обобщено понятие элементарного классификатора и построены новые модели процедур дискретного характера, основанные на выделении таких наборов допустимых значений признаков, которые не встречаются в признаковых описаниях обучающих объектов класса. В определенных случаях предложенные модели позволяют повысить качество распознавания и требуют меньших вычислительных затрат в случае большого числа классов.

(2) Разработаны подходы к повышению эффективности алгоритмов распознавания, основанные на выделении для каждого класса типичных значений признаков и типичных обучающих объектов. Данные подходы позволяют существенно снизить влияние шумящих признаков, а также повысить качество распознавания в случае, когда в обучающей выборке содержится много объектов, лежащих на границе между классами.

(3) Предложен быстрый способ вычисления оценок при голосовании по представительным наборам для процедуры скользящего контроля, который позволяет значительно сократить время счета по сравнению с традиционно применявшимся методом.

(4) Получены асимптотики типичных значений числа покрытий и длины покрытия целочисленной матрицы для практически общего случая.

(5) Получены асимптотики типичных значений числа σ-подматриц и ранга σ-подматрицы для случая, когда число строк в матрице значительно превосходит число столбцов. Показано, что в этом случае число σ-подматриц по порядку больше числа тупиковых σ-покрытий.

(6) Получены новые оценки, касающиеся метрических свойств допустимых и максимальных конъюнкций двузначной логической функции, заданной множеством нулей.

Литература

[1] Андреев А. Е. Некоторые вопросы тестового распознавания образов // ДАН СССР. 1981, Т. 255, № 4. С. 781-734.

[2] Андреев А. Е. О тупиковых и минимальных тестах // ДАН СССР. 1981, Т. 256, № 3. С. 521-524.

[3] Андреев А. Е. Об асимптотическом поведении числа тупиковых тестов и минимальной длины теста для почти всех таблиц // Проблемы кибернетики. Вып. 41. М.: Наука, 1984. С. 117-141.

[4] Айзенберг Н.Н., Журавлев Ю.И., Пилюгин С.В. Применение сверточных алгебр для построения корректных распознающих алгоритмов //Ж. вычж-сл. матем. и матем. физ. 1987. Т. 27, № 6. С. 912-923.

[5] Асланян А., Журавлев Ю. И. Об одном подходе к построению эффективных алгоритмов распознавания // Ж. вычисл. матем. и матем. физ. 1985. Т. 25, № 2. С. 283-291.

[6] Асланян Л. А. Об одном методе распознаваний, основанном на разделении классов дизъюнктивными нормальными формами // Кибернетика. 1975. № 5. С. 103-110.

[7] Асланян Л. А. Алгоритмы распознавания с логическими отделителями // Сб. работ по матем. кибернетике. Вып. 1. М.: ВЦ АН СССР, 1976. С. 116-131.

[8] Баскакова Л. В., Журавлев Ю. И. Модель распознающих алгоритмов с представительными наборами и системами опорных множеств // Ж. вычисл. матем. и матем. физ. 1981. Т. 21, № 5. С. 1264-1275.

[9] Бушманов О. Н. Класс алгоритмов распознавания, основанный на поиске эмпирических закономерностей // М: ВЦ РАН. 1992. 21 с.

[10] Бушманов О. Н., Дюкова Е. В., Журавлев Ю. И., Кочетков Д. В., Рязанов В. В. Система анализа и распознавания образов // Распознавание, классификация, прогноз (математические методы и ИХ применение). М.: Наука, 1989. Вып. 2. С. 250-273.

[11] Бушманов О. Н., Дюкова Е. В., Рязанов В. В. Система распознавания, таксономии и анализа стандартных данных // Тез. III Всесоюзной конф. "Мат. методы распознавания образов".

[12] Валев В., Беликов М., Дюкова Е. В, Программный комплекс для решения задач распознавания на основе построения эмпирических закономерностей . М.: ВЦ АН СССР, 1988. 20 с.

[13] Вайнцвайг М. Н. Алгоритм обучения распознаванию образов «Кора» // Алгоритмы обучения распознаванию образов. М.: Сов. радио, 1973. С. 82-91.

[14] Вапник В. Н., Червоненкис А. Я. Теория распознавания образов. М.: Наука, 1974. 418 с,

[15] Вапник В. Н. Восстановление зависимостей по эмпирическим данным. М.: Наука, 1979. 448 с.

[16] Вапник В. Н. Индуктивные принципы поиска эмпирических закономерностей // Распознавание, классификация, прогноз (математические методы и их применение). М.: Наука, 1988. Вып. 1.С. 17-81 .

[17] Васильев В, И. Распознающие системы. .Киев: Наукова думка,1983. 466 С.

[18] Вентцель Н. С. Теория вероятностей // М.: Наука, 1964.

[19] Вешторт А. М., Зуев Ю. А., Краснопрошин В. В. Двухуровневая схема распознавания с логическим корректором // Распознавание, классификация, прогноз (математические методы и их применение). М.: Наука, 1989. Вып. 2. С. 73-98.

[20] Волжков Ю, К., Дюкова Е. В., Левашов Е. А., Рязанов В. В. Применение методов распознавания образов для прогнозирования свойств твердых сплавов группы СТИМ, М: МИСИС, 1988. С. 40-43.

[21] Гельфанд И. М., Губерман Ш. А., Шифрин М, А. Прогнозирование и распознавание в медицинских задачах // Распознавание, классификация, прогноз, М.: Наука, 1989, С, 201-228,

[22] Глаголев В, В. Некоторые оценки дизъюнктивных нормальных форм функции алгебры логики // Проблемы кибернетики. М. Наука,1967. ВЫП. 19, С. 75-94. Глаз А. Б. Параметрическая и структурная адаптация правил в задачах распознавания, Рига: Зинатне, 1988. 172 с.

[23] Гольдберг С. И. Об одном методе распознавания образов "Совокупный антисиндром"// Вычисл. системы, Новосибирск. 1978. Вып. 76. С. 83-90.

[24] Горелик А. Л., Гуревич И. Б., Скрипник В. А. Современное состояние проблемы распознавания. М.: Радио и связь, 1984. 160 с.

[25] Горелик А. Л., Скрипник В. А, Методы распознавания. М.: Высшая школа, 1984. 208 с.

[26] Гренандер У. Лекции по теории образов. М.: Мир, Т. 1. 1979. 384 с.; Т, 2. 1981. 448 с.; Т. 3. 1983, 432 с.

[27] Гуревич И. Б., Журавлев Ю. И. Минимизация булевых функций и эффективные алгоритмы распознавания // Кибернетика. 1974,, № 3. С. 16-20.

[28] Денисова Р.А. Метод синтеза тупиковых представительных наборов для k-значных таблиц, М.: ВЦ АН СССР, 1984. 29 с.

[29] Дискретная математика и математические вопросы кибернетики / Под ред. С.Б. Яблонского, О.Б. Лупанова, М.: Наука"1974. 312 с.

[30] Дмитриев А. И., Журавлев Ю. И., Кренделев Ф. П. О математических принципах классификации предметов или явлений // Дискретный анализ. Новосибирск: ИМ СО АН СССР, Вып. 7. 1966. С. 1-17

[31] 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,

[32] Долгоруков А. Ю., Дюкова Е. В. Об одном способе вычисления информационных характеристик обучающей // Тезисы Всероссийской конференции "Математические методы распознавания образов (ММРО-6)". г. Москва, 1993. С. 22-23.

[33] Донской В. И. Алгоритмы обучения, основанные на построении решающих деревьев // Ж. вычисл. матем. и матем. физ. 1982. Т. 22, № 4. С. 963-974.

[34] Донской В. И. Слабоопределенные задачи линейного булева программирования с частично заданным множеством допустимых решений // Ж. вычисл. матем. и матем. физ. 1988. Т. 28, № 9, С. 1379-1385.

[35] Дуда Р., Харт П. Распознавание образов и анализ сцен. М.: Мир. 1976. 511 с.

[36] Дюкова Е. В. Об одном алгоритме построения тупиковых тестов для бинарных таблиц // Сборник работ по дискретной математике. М.: ВЦ АН СССРб 1976. Вып. 1. С. 167-185

[37] Дюкова Е. В. Об асимптотически оптимальном алгоритме построения тупиковых тестов // ДАН СССР. 1977. Т. 233, № 4. С. 527-530

[38] Дюкова Е. В. Об асимптотически оптимальном алгоритме построения тупиковых тестов для одного класса k-значных таблиц // Тез. докл. IV Всесоюз. конф. по проблемам теоретический кибернетики. Новосибирск: Изд-во СО АК СССР, 1977. С. 197-199.

[39] Дюкова Е. В. Об асимптотически оптимальном алгоритме построения тупиковых, тестов для бинарных таблиц // Проблемы кибернетики. М.: Наука, 1978. Вып. 34. С. 169-186.

[40] Дюкова Е. В. Построение тупиковых тестов для k-значных таблиц // ДАН СССР. 1978. Т.238. № 6. С. 1279-1282

[41] Дюкова Е.В., Журавлев Ю.И., Зенкин А.А. и др. Пакет прикладных программ для решения задач распознавания и классификации (ПАРК). М.: ВЦ АН СССР, 1981. 23 с.

[42] Дюкова Е. В. Асимптотически оптимальные тестовые алгоритмы в задачах распознавания // Проблемы кибернетики. Вып. 39, М.: Наука, 1982, С. 165-199.

[43] Дюкова Е. В., Рязанов В. В. 0 решении прикладных задач алгоритмами распознавания, основанными на принципе голосования. М.: ВЦ АН СССР, 1986. 26 с.

[44] Дюкова Е. В. 0 метрических свойствах алгоритмов типа «Кора» для одного класса бинарных таблиц. М.: ВЦ АН СССР, 1986, 17 с.

[45] Дюкова Е. В. О сложности реализации некоторых процедур распознавания // Ж. вычисл. матем. и матем. физ. 1987. Т. 27, № 1. С. 114–127.

[46] Дюкова Е. В. Об одной параметрической модели алгоритмов распознавания типа "Кора". М.: ВЦ АН СССР, 1988, 23 с.

[47] Дюкова Е. В. Алгоритмы распознавания типа "Кора": сложность реализации и метрические свойства // Распознавание, классификация, прогноз (математические методы и их применение). М.: Наука"1989. Вып. 2. С. 99-125.

[48] Дюкова Е. В. О решении систем булевых квазинельсоновского типа // Вопросы кибернетики / Дискретная математика. Методы и применение / М.: АН СССР, Научный Совет по комплексной проблеме "Кибернетика", 1983. с. 5-19.

[49] Дюкова Е. В., Карнеева М. Л. Модели алгоритмов, основанные на различных способах перекодировки исходной информации // Матем. методы в распознавании образов и дискретной оптимизации. М.: ВЦ АН СССР, 1990. С. 43-56.

[50] Дюкова Е. В. Асимптотические оценки некоторых характеристик множества представительных наборов и задача об устойчивости // Ж. вычисл. матем. и матем. физ. 1995. Т. 35, № 1. С. 122-184.

[51] Дюкова Е.В., Журавлев Ю.М., Рудаков К.В. Об алгебро-логическом синтезе корректных процедур распознавания на базе элементарных алгоритмов // Ж. вычисл. матем. и матем. физ. 1996. Т. 36, 1 8. С. 217-225.

[52] Дюкова Е. В., Журавлев Ю. И. Дискретный анализ признаковых описаний в задачах распознавания большой размерности // Ж. вычисл. матем. и матем. физ. 2000. Т. 40. №8. С.1264-1278.

[53] Дюкова Е. В., Песков Н. В. О некоторых подходах к вычислению информативных характеристик обучающей выборки // Докл. Всеросс. Конф. "Матем. методы распознавания образов - 9". М.: АЛЕВ-В, 1999, С. 181-183.

[54] Дюкова Е. В., Песков Н. В. О дискретных процедурах распознавания, основанных на построении покрытий классов // Докл. Всеросс. Конф. "Матем. методы распознавания образов - 10", М.: АЛЕВ-В, 2001, С. 48-51.

[55] Дюкова Е.В., Песков Н.В. Информативность признаков, отдельных значений признаков и фрагментов описаний объектов // Докл. Всеросс. конф. "Математические методы распознавания образов 10", М.: АЛЕВ-В, 2001. С. 44 47.

[56] Дюкова Е.В., Песков Н.В. Поиск информативных фрагментов описаний объектов в дискретных процедурах распознавания // Ж. вычисл. матем. и матем. физ. 2002. Том 42, № 5, С. 741-753.

[57] Дюкова Е.В., Инякин А.С., Песков Н.В. О некоторых направлениях современных исследований в области дискретного анализа информации в проблеме распознавания // Труды межд. Конф. "РОАИ-6-2002", Великий Новгород, 2002. Т. 1. С. 203-208.

[58] Дюкова Е.В. Дискрентные (логические) процедуры распознавания: принципы конструирования, сложность реализации и основные модели //Учебное пособие для студентов Математических факультетов педвузов. М: МПГУ 2003 г. 30 с.

[59] Журавлев Ю. И. 0б одном классе не всюду определенных функций алгебры логики // Дискретный анализ. Новосибирск: ИМ СО АН СССР, Вып. 2, 1964. С. 23-27.

[60] Журавлев Ю. И., Никифоров В. В. Алгоритмы распознавания, основанные на вычислении оценок // Кибернетика. 1971. JI 3. С. 1-11.

[61] Журавлев Ю. И., Мирошник С. Н., Швартин С. М. Об одном подходе к оптимизации в классе параметрических алгоритмов распознавания // Ж. вычисл. матем. и матем. физ.. 1975. Т. 16, № 1, С, 209-218.

[62] Журавлев Ю. И. Экстремальные алгоритмы в математических моделях для задач распознавания и классификации // ДАН СССР.

[63] Журавлев Ю. И. Непараметрические задачи распознавания образов // Кибернетика. 1976. № 6. С. 93-103.

[64] Журавлев Ю. И., Зенкин А. А., Зенкин А. И., Исаев И. В., Кольцов П. П., Кочетков Д, В., Рязанов В. В. Задачи распознавания или классификации со стандартной обучающей информацией // Ж. вычисл. матем. и матем. физ. 1980, Т. 20, № 5. С. 1294-1309.

[65] Журавлёв Ю. И. Об алгебраическом подходе к решению задач распознавания или классификации // Пробл. кибернетики. М.: Наука, 1978. Вып. 33. С. 5-68.

[66] Журавлев Ю. И., Платоненко И. М. Об экономном умножении булевых уравнений: // Ж. вычиел. матем. и матем. физ. 1984. Т. 20, № 5. С. 1294-1309.

[67] Журавлев Ю. И., Коган А. Ю. Реализация булевых функций с малым числом нулей дизъюнктивными нормальными формами и смежные задачи // ДАН СССР. 1985. Т. 285, № 4. С. 795-799.

[68] Журавлев Ю.И. Об алгоритмах распознавания с представительными наборами (о логических алгоритмах). Ж. вычисл. матем. и матем. физ. 2002. Том 42, № 5, С. 1425-1435.

[69] Катериночкина Н. Н. Поиск максимального верхнего нуля монотонной функции алгебры логики // ДАН СССР. 1975, Т. 224, № 3. С. 557-560,

[70] Коган Ю. А. О дизъюнктивных нормальных формах булевых функций, с малым числом нулей // Ж. вычисл. матем. и матем. физ. 1987. Т, 27. 16, С. 924-931.

[71] Константинов Р. М., Королева З. Е., Кудрявцев В. Б. О комбинаторно-логическом подходе к задачам прогноза рудоносности // Проблемы кибернетики. Вып.30. М.: Наука, 1975. С. 5-33.

[72] Кузнецов В. Е. Об одном стохастическом алгоритме вычисления информативных характеристик таблиц по методу тестов // Дискретный анализ. Новосибирск: ИМ СО АН СССР, 1973. Вып. 23. С. 8-23.

[73] Мадатян Х. А. Оценка числа представительных наборов для одного класса бинарных таблиц // Math. Problems In Compiit, Theory. Banach Center Pitbls, Warsaw, 1988. V. 21. P. 513-522.

[74] Носков В. Н. О тупиковых и минимальных тестах для одного класса таблиц // Дискретный анализ. Новосибирск: ИМ СО АН СССР, 1968. вып. 12. С. 27-49.

[75] Носков В. Н., Слепян В. А. О числе тупиковых тестов для одного класса таблиц // Кибернетика. 1972. № 1. С. 60-65.

[76] Платоненко И. М. О реализации алгоритмов типа "Кора"с помощью решения систем булевых уравнений специального вида. М.: ВЦ АН. СССР, 1983. 21 с.

[77] Песков Н.В. О некоторых подходах к конструированию дискретных процедур распознавания // Сообщения по прикладной математике. М.: ВЦ РАН, 2002. 28с.

[78] Песков Н.В. Об одном подходе к повышению эффективности алгоритмов распознавания // Интеллектуализация обработки информации: тезисы докладов Международной конференции, Симферополь, 2002. С. 73-74.

[79] Рудаков К. В. О числе гиперплоскостей, разделяющих конечные множества в эвклидовом пространстве // ДАН СССР, 1976. Т. 231, № 6. С. 1296-1299.

[80] Сапоженко А. А. Оценка длины и числа тупиковых д.н.ф. для почти всех не всюду определенных булевых функций // Матем. заметки. 1980. Т. 28. № 2. С. 279-300.

[81] Слепян В.А. Параметры распределения, тупиковых тестов и веса столбцов в бинарных таблицах // Дискретный анализ. Новосибирск: ИМ СО АН' СССР, 1969. Вып. 14. С. 28-43.

[82] Слепян В. А. О числе тупиковых тестов и о мерах информативности столбца для почти всех бинарных таблиц // ДАН СССР. 1987. Т. 297, № 1, С. 43-46.

[83] Слепян В. А. Длина минимального теста для некоторого класса таблиц // Дискретный анализ. Новосибирск: ММ СО АН СССР, 1973. Вып. 23. С. 59-71.

[84] Слуцкая Т.Л. Алгоритмы вычисления информационных весов // Дискретный анализ. Новосибирск: ИМ СО АН СССР, 1966, Вып. 12. С. 75-90.

[85] Чегис И.А., Яблонский С.В. Логические способы контроля электрических схем // Труды Матем. ин-та им. В.А.Стеклова АН СССР. 1958, Т. 51. С. 270-360.

[86] Djukova E. V., Zhuravlev Yu. l. Diskrete methods of information analysis and algorithm synthesis // J. Pattern Recognition and Image Analys., 1997. V. 7. № 2. P. 192-207.

[87] Djukova E.V., Peskov N.V. Selection of Typical Objects in Classes for Recognition Problems // J. Pattern Recognition and Image Analysis. 2002. V. 12. No. 3. P. 243 249.

[88] Djukova E.V., Inyakin A.S., Peskov N.V. Recent Trends in Discrete Analysis of Information in Recognition Problems // J. Pattern Recognition and Image Analysis. 2003. V. 13. No. 3. P. 11-13.

[89] Djukova E.V., Inyakin A.S., Peskov N.V. Methods of Combinatorial Analysis in Synthesis of Efficient Recognition Algorithms // J. Pattern Recognition and Image Analysis. 2003. V. 13. No. 3. P. 426-432.

[90] Djukova E.V. Discrete Recognition Procedures: The Complexity of Realization // J. Pattern Recognition and Image Analysis. 2003. V. 13. No. 1. P. 8-10.

[91] Djukova E.V. Discrete (Logical) Recognition Procedures: Principles of Construction, Complexity of Realization and Basic Modeles // J. Pattern Recognition and Image Analysis. 2003. V. 13. No. 3. P. 417-425.