На правах рукописи

Песков Николай Владимирович

Поиск информативных фрагментов описаний объектов в задачах распознавания

Специальность 05.13.17 — теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание учёной степени
кандидата физико-математических наук

Москва – 2004

Работа выполнена в Научном совете по комплексной проблеме "Кибернетика"

Научный руководитель:	доктор физико-математических наук Е. В. Дюкова
Официальные оппоненты:	доктор технических наук Л. М. Местецкий кандидат физико-математических наук М. Н. Вялый
Ведущая организация:	Московский педагогический государственный университет

Защита состоится « ____ » _____ 2004 г. в _____ часов на заседании диссертационного совета Д002.017.02 Вычислительного центра им. А. А. Дородницына РАН по адресу: 119991, Москва, ул. Вавилова, 40.

С диссертацией можно ознакомиться в библиотеке Вычислительного Центра им. А. А. Дородницына РАН.

Автореферат разослан « ____ » _____ 2004 г.

Учёный секретарь
диссертационного совета
д.ф.-м.н. В. В. Рязанов

Общая характеристика работы

Актуальность темы.

Для решения прикладных задач распознавания успешно применяются методы, основанные на комбинаторном анализе признаковых описаний объектов. Эти методы особенно эффективны в случае, когда информация целочисленная и число допустимых значений каждого признака невелико. Основополагающими работами являются работы Ю.И. Журавлева, С.В. Яблонского и М.Н. Вайнцвайга.

Главной особенностью рассматриваемых процедур распознавания, называемых в дальнейшем дискретными или логическими процедурами, является возможность получения результата при отсутствии информации о функциях распределения 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. Например, информативными считаются такие фрагменты, которые встречаются в описаниях обучающих объектов одного класса, но не встречаются в описаниях обучающих объектов остальных классов. Рассматриваемые фрагменты, как правило, имеют содержательное описание в терминах той 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 покрытий целочисленной матрицы и может быть также сформулирована как задача построения дизъюнктивной нормальной формы логической функции. При этом особую сложность представляет поиск тупиковых покрытий.

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. При этом вопросам качества распознавания не уделялось достаточное внимание. 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 рассматриваемого класса и не встречаются в описаниях объектов других классов, т.е характеризуют лишь некоторые из обучающих объектов 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 слишком много значений. Такие признаки порождают очень большое число фрагментов, позволяющих различать объекты разных классов, 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 границе 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 с вопросами изучения метрических (количественных) свойств множества информативных фрагментов. Важными и технически очень сложными являются задачи получения асимптотических оценок для типичных значений числа (тупиковых) покрытий и длины (тупикового) покрытия целочисленной матрицы, а также задачи получения аналогичных оценок для допустимых и максимальных конъюнкций логической функции.

Целью 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.

Научная новизна. В диссертационной работе введено более общее по сравнению с ранее используемым понятие элементарного классификатора, что позволило построить модели, основанные на поиске наборов из допустимых значений признаков, не встречающихся в описаниях обучающих объектов класса.

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.

Получены новые результаты, касающиеся изучения метрических свойств множеств покрытий и тупиковых покрытий целочисленной матрицы.

Методы исследования. В работе использовался аппарат дискретной математики, в частности алгебры логики, теории дизъюнктивных нормальных форм логических функций. Применялись методы построения покрытий булевых и целочисленных матриц, а также методы получения асимптотических оценок для типичных значений количественных характеристик множеств покрытий и тупиковых покрытий целочисленной матрицы.

Теоретическая и практическая ценность. Результаты, полученные в диссертационной работе, могут быть использованы в теоретических исследованиях, касающихся построения эффективных реализаций для моделей дискретных (логических) процедур распознавания. Эффективность предложенных подходов подтверждена решением практических задач из области медицинского прогнозирования и анализа результатов социологических опросов.

Апробация работы. Результаты, изложенные в диссертации, докладывались на Всероссийских конференциях «Математические методы распознавания образов IX, X и XI»(Москва, ВЦ РАН, 1999, 2001, 2003 гг.); Международной конференции «Интеллектуализация обработки информации - 2002» (Симферополь, ТНУ); Международной конференции «Распознавание образов и анализ изображений: новые информационные технологии - VI» (Великий Новгород, 2002г.); Международной конференции студентов и аспирантов по фундаментальным наукам «Ломоносов 2003» (Москва, МГУ); семинарах отдела «Вычислительных методов прогнозирования» ВЦ РАН и лабораторий «Кибернетических методов информатики» и «Распознавания образов и прогнозирования» НСК РАН.

Публикации. По теме диссертации опубликовано 10 работ [1-10], в том числе 1 в ЖВМиМФ и 3 в журнале Pattern Recognition and Image Analysis. Описания

основных результатов, полученных в диссертации, включались в научные отчеты по проектам РФФИ 98-01-00596, 01-01-00575, 00-15-96064.

Структура и объём работы. Диссертация состоит из введения, пяти глав, заключения и списка литературы (91 наименование). Общий объем работы - 102 страницы.

Содержание работы

Во введении обосновывается актуальность темы и обсуждается круг проблем, возникающих при практическом применении дискретных процедур распознавания. Приводится краткое изложение содержания работы.

В главе 1 формулируется постановка задачи распознавания, вводится ряд основных определений, рассматривается общая схема конструирования дискретных процедур распознавания. Описывается классический алгоритм голосования по представительным наборам и новые модели: алгоритм голосования по антипредставительным наборам и алгоритм голосования по покрытиям классов.

Задача распознавания формулируется следующим образом. Исследуется некоторое множество объектов M, которые могут быть описаны в системе целочисленных признаков $\{x_1, \ldots, x_n\}$. Пусть N_j - конечное множество допустимых значений признака x_j, $j \in \{1, 2, \ldots, n\}$, и пусть 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 описанию некоторого объекта S из M , о котором, вообще говоря, неизвестно, какому классу он принадлежит, определить этот класс. Требуется также оценить важность при распознавании каждого признака и важность отдельных значений признаков.

В разделе 1.1. приводятся основные определения и обозначения. Одно из основных понятий, используемых при конструировании дискретных процедур распознавания является понятие элементарного классификатора.

Пусть $H = \{x_{j_1}, \ldots, x_{j_r}\}$ - некоторый набор признаков, $S = (a_1, ..., a_n)$ - объект из обучающей выборки. Набор $(a_{j_1}, \ldots, a_{j_r})$ будем называть фрагментом описания объекта S и обозначать через (S, H). В классическом варианте элементарными классификаторами является фрагменты описаний обучающих объектов.

В диссертационной работе введено понятие элементарного классификатора более общего вида.

Пусть H - некоторый 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.

Близость объекта $S = (a_1, \ldots, a_n)$ из M и элементарного классификатора σ

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 даны два объекта $S' = (a'_1, a'_2, \ldots, a'_n)$ и $S' = (a''_1, a''_2, ..., a''_n)$ из M. Близость объектов S' и S'' по набору признаков H будем оценивать величиной

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

Каждый распознающий алгоритм дискретного характера A для каждого класса K, $K \in \{K_1, \ldots, K_l\}$, строит некоторое подмножество $C^A(K)$ из множества всех элементарных классификаторов. Элементарные классификаторы из $C^A(K)$ и только они считаются информативными при использовании алгоритма A. Распознавание объекта S осуществляется на основе вычисления величины $B(\sigma, S, H)$ для каждого элемента σ из $C^A(K)$, т.е. для каждого класса K 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 определяется множеством построенных элементарных классификаторов и способом вычисления оценок $\Gamma(S, K_1), \ldots, \Gamma(S, K_l)$. Объект S относится к тому классу, для которого оценка максимальна. Если таких оценок несколько, то происходит отказ от распознавания.

При построении элементарных классификаторов используются понятия покрытия и тупикового покрытия целочисленной матрицы. Введем следующие обозначения: L - матрица с элементами из $\{0, 1, \ldots, k-1\}$, $k \geq 2$; $\sigma = (\sigma_1, \ldots, \sigma_r)$, $\sigma_i \in \{0, 1, \ldots, k-1\}$, при $i = 1, 2, \ldots, r$.

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 H, являющийся σ-покрытием, назовем тупиковым σ-покрытием, если L^H содержит подматрицу, составленную из строк $(\beta_1, \sigma_2, \sigma_3, \ldots, \sigma_{r-1}, \sigma_r)$, $(\sigma_1, \beta_2, \sigma_3, \ldots, \sigma_{r-1}, \sigma_r)$, ..., $(\sigma_1, \sigma_2, \sigma_3, \ldots, \sigma_{r-1}, \beta_r)$, где $\beta_p \neq \sigma_p$ при $p = 1, 2, \ldots, r$. Указанную подматрицу будем называть σ-подматрицей.

Пусть $\overline{K} = \{K_1, \ldots, K_l\} \setminus K$. Элементарный классификатор σ, 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 σ; 2) не все, а лишь часть фрагментов вида (S', H), где $S' \in K$, совпадают с σ; 3) ни один фрагмент вида (S', H), где $S' \in K$, не совпадает с σ. 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 поиске элементарных классификаторов, обладающих свойством 2.

В разделе 1.2 описана классическая модель алгоритма A_1 голосования по представительным наборам

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 $B(S', S'', H) = 0$. Представительный набор для класса K вида (S', H) назовем тупиковым если для любого набора H', $H' \subset H$, найдется обучающий объект S'' из $\overline{K}$, для которого $B(S', S'', H') = 1$. В качестве $C^{A_1}(K)$ берется множество (тупиковым) представительных наборов для K. В простейшей модификации для оценки принадлежности объекта S классу K используется величина

$$\Gamma_1(S, K) = \frac{1}{|C^{A_1}(K)|} \sum_{(\sigma,H)\in C^{A_1}(K)} B(\sigma, S, H)$$

(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 σ встречается в K.

В разделе 1.3. описываются алгоритм A_2 голосования по покрытиям классов и алгоритм A_3 голосования по антипредставительным наборам.

Множество $C^{A_2}(K)$ состоит из элементарных классификаторов, обладающих свойством 3. Такие элементарные классификаторы будем назвать покрытиями класса K. Элементарный классификатор, являющийся покрытием класса K и встречающийся в описаниях объектов из $\overline{K}$, будем называть антипредставительным набором. Множество $C^{A_3}(K)$ состоит из антипредставительных наборов для K.

Элементарный классификатор σ из $C^{A_2}(K)$ или $C^{A_3}(K)$, порожденный набором признаков H, голосует за принадлежность распознаваемого объекта S классу K, если $\sigma \neq (S, H)$. Принадлежность объекта S классу K (в простейшей модифи-

кации) оценивается величиной

$$\Gamma_t(S,K) = \frac{1}{|C^{A_t}(K)|} \sum_{(\sigma,H)\in C^{A_t}(K)} (1 - B(\sigma,S,H)), t = 2,3.$$

Построение множества покрытий и множества антипредставительных наборов класса K сводится к построению σ-покрытий матрицы, составленной из описаний объектов из K. При построении антипредставительных наборов дополнительно проверяется условие, что построенный элементарный классификатор σ встречается в $\overline{K}$.

В главе 2 предложены методы предварительного анализа обучающей информации, направленные на снижение влияния шумящих признаков и объектов, лежащих на границе между классами.

В разделе 2.1. предлагается методика оценки информативности (важности для распознавания) признаков, а также отдельных значений признаков.

Традиционно в качестве меры важности признака x_j, $j \in \{1,2,\ldots,n\}$, рассматривалась величина (впервые предложенна Ю.И. Журавлевым), равная отношению числа построенных элементарных классификаторов, в которых содержится признак x_j, к общему числу построенных элементарных классификаторов. Такой способ оценки информативности признаков не всегда дает хорошие результаты, например, если признак является шумящим.

В диссертационной работе предлагается метод оценки информативности признаков и отдельных значений признаков, основанный на вычислении близости между объектами по отдельным признакам.

Пусть $S' = (a'_1, a'_2, \ldots, a'_n)$ - обучающий объект, $S' \in K_i$, $i \in \{1,2,\ldots,l\}$, и $j \in \{1,2,\ldots,n\}$. Положим: $\mu_{ij}^{(1)}(S')$ - частота встречаемости a'_j в описаниях обучающих объектов из K_i, $\mu_{ij}^{(2)}(S')$ - частота встречаемости a'_j в описаниях обучающих объектов из $\overline{K_i}$. 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 признака x_j для объекта S'. Будем говорить, что значение признака x_j для S' является типичным для K_i, если $\mu_{ij}(S') > \mu$, где μ - порог информативности, $-1 < \mu < 1$. Например, в случае $\mu = 0$, 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 информативную зону. В информативную зону не попадают значения шумящих признаков. 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.

В разделе 2.2. предлагается подход, позволяющий значительно повысить эффективность алгоритмов распознавания в случае, когда в обучающей выборке содержится много объектов, лежащих на границе между классами. Суть предлагаемого подхода заключается в следующем.

Пусть описание обучающего объекта S из $\overline{K}$ 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: базовую, состоящую из типичных для своих классов обучающих объектов, и контрольную, состоящую из обучающих объектов, находящиеся на границе между классами. По базовой будем строить множество элементарных классификаторов, а по контрольной - вычислять их веса. Практические эксперименты на прикладных задачах показывают, что такое разбиение увеличивает число коротких элементарных классификаторов и тем самым позволяет повысить качество алгоритма распознавания A.

Предложено два способа выделения типичных объектов: 1) типичными считаются объекты, описания которых состоят в основном 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.

В разделе 2.3. приводится быстрый способ вычисления оценок при голосовании по представительным наборам для процедуры скользящего контроля, позволяющий сократить временные затраты примерно в m раз.

Глава 3 посвящена изучению метрических свойств множества покрытий целочисленной матрицы.

В разделе 3.1 приводятся необходимые определения и обозначения. Пусть Ψ_0 - интервал $(log_k un, n)$; Ψ_1 - интервал

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

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

Пусть M^k_{un}, $k \geq 2$, - множество всех матриц размера $u \times n$ с элементами из $\{0, 1, \ldots, k-1\}$; $L \in M^k_{un}$; E^r_k - множество всех k-ичных наборов длины r; $\sigma \in E^r_k$, $\sigma = (\sigma_1, \ldots, \sigma_r)$. Пусть далее $C(L, \sigma)$ - 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

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

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

Нас будут интересовать асимптотические оценки чисел $|C(L)|$, $|B(L)|$ и $|S(L)|$ для почти всех матриц L из M_{un}^k при $n \to \infty$. Выявление типичной ситуации будет связано с высказыванием 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 основном изучался случай, когда число строк в матрице по порядку меньше числа столбцов, а именно, когда $u^\alpha \le n \le k^{u^\beta}$, $\alpha > 1$, $\beta < 1$. Е.В. Дюковой показано, что в данном случае величина $|B(L)|$ в типичной ситуации асимптотически совпадает с величиной $|S(L)|$ и по порядку меньше числа покрытий. На основании этого факта ею был построен асимптотически оптимальный алгоритм поиска покрытий из $B(L)$.

В разделе 3.2 для практически общего случая получены асимптотики типичных значений величины $|C(L)|$ и длины покрытия из $C(L)$, а именно, доказана

Теорема 3.2.1. *Если $u \le k^{n^\beta}$, $\beta < 1$, то для почти всех матриц L из M_{un}^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.

Пусть $C_1(L)$ - множество всех покрытий из $C(L)$, длины которых не превосходят $\log_k u - \log_k(\log_k u \ln kn)$. Справедлива

Теорема 3.2.2. *Для почти всех матриц $L \in M_{(un)}^k$ при $n \to \infty$ имеет место $|C_1(L)| = 0$.*

В разделе 3.3 для случая $n^\alpha \le u \le k^{n^\beta}$, $\alpha > 1$, $\beta < 1$, получены асимптотики типичных значений числа подматриц из $S(L)$ и порядка подматрицы из $S(L)$, а именно доказана

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

$$|S(L)| \approx \sum_{r \in \Psi_1} C_n^r C_u^r r!(k-1)^r k^{r-r^2}$$

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

Пусть $S_1(L)$ - множество всех σ-подматриц из $S(L)$, ранги которых не меньше $\log_k un$. Справедлива

Теорема 3.3.2. *Для почти всех матриц $L \in M_{un}^k$ при $n \to \infty$ имеет место $|S_1(L)| = 0$.*

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

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

В главе 4 сформулированы основные принципы конструирования дискретных процедур распознавания с использованием аппарата логических функций. Рассмотрена связь между задачей построения ДНФ двузначной логической функции, заданной на k-ичных n-мерных наборах, и задачей нахождения покрытий матрицы из M_{mn}^k. На основе теорем, доказанных в главе 3, получены асимптотики типичных значений для числа допустимых конъюнкций и ранга допустимой конъюнкции указанной функций, а также верхняя асимптотическая оценка 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 алгоритмом голосования по представительным наборам. Результаты сравнения показывают, что предложенные в работе подходы в ряде случаев имеют преимущество перед классическими, если для оценки качества алгоритма использовать процедуру скользящего контроля.

В заключении приводится краткая формулировка результатов, выносимых на защиту.

Результаты, выносимые на защиту

(1) Введено более общее по сравнению с ранее используемым понятие элементарного классификатора и построены новые модели распознающих процедур дискретного характера, основанные на принципе «невстречаемости» набора из допустимых значений признаков в описаниях объектов класса.

(2) Разработаны подходы к повышению эффективности алгоритмов распознавания, основанные на выделении для каждого класса типичных значений признаков и типичных обучающих объектов.

(3) Предложен быстрый способ вычисления оценок при голосовании по представительным наборам для процедуры скользящего контроля.

(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 превосхо-

дит 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.

Публикации по теме диссертации

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

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

[8] 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.

[9] 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.

[10] 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.