*На правах рукописи*

ГЕНРИХОВ ИГОРЬ ЕВГЕНЬЕВИЧ

# ПОСТРОЕНИЕ И ИССЛЕДОВАНИЕ ПОЛНЫХ РЕШАЮЩИХ ДЕРЕВЬЕВ ДЛЯ ЗАДАЧ КЛАССИФИКАЦИИ ПО ПРЕЦЕДЕНТАМ

05.13.17 – теоретические основы информатики

Автореферат диссертации на соискание ученой степени

кандидата физико-математических наук

Москва – 2013

Работа выполнена на кафедре теоретической информатики и дискретной математики математического факультета в Федеральном государственном бюджетном образовательном Учреждении Высшего профессионального образования Московский педагогический государственный университет

Научный руководитель: доктор физико-математических наук, профессор Дюкова Елена Всеволодовна

Официальные оппоненты: Сенько Олег Валентинович, доктор физико-математических наук, ведущий научный сотрудник, Федеральное государственное бюджетное Учреждение науки Вычислительный центр им. А. А. Дородницына Российской академии наук

Середин Олег Сергеевич, кандидат физико-математических наук, доцент, Федеральное государственное бюджетное образовательное Учреждение Высшего профессионального образования Тульский государственный университет

Ведущая организация: Федеральное государственное бюджетное Учреждение науки Институт математики и механики им. Н. Н. Красовского Уральского отделения Российской академии наук

Защита диссертации состоится «___»_____ 2013 г. в ____ часов на заседании диссертационного совета Д 002.017.02 при Федеральном государственном бюджетном Учреждении науки Вычислительный центр им. А. А. Дородницына Российской академии наук по адресу: 119333, Москва, ул. Вавилова, дом 40, конференц-зал.

С диссертацией можно ознакомиться в библиотеке Федерального государственного бюджетного Учреждения науки Вычислительный центр им. А. А. Дородницына Российской академии наук.

Автореферат разослан «___»_____ 2013 г.

Ученый секретарь

Диссертационного совета

Д 002.017.02, д.ф.-м.н., профессор В. В. Рязанов

## ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

**Актуальность темы.** Одной из центральных задач распознавания образов является задача распознавания по прецедентам. Известным инструментом решения данной задачи являются деревья решений. Синтез классического решающего дерева (РД) представляет собой итерационный процесс. Как правило, для построения очередной внутренней вершины дерева выбирается признак, который наилучшим образом удовлетворяет некоторому критерию ветвления, т.е. наилучшим образом разделяет текущее множество обучающих объектов. По каждой ветви, исходящей из построенной вершины, осуществляется спуск и строится либо лист дерева, либо новая внутренняя вершина. Каждому листу приписан один из классов, и, как правило, в листе содержится вся информация, позволяющая сделать вывод о принадлежности распознаваемого объекта классу, который приписан данному листу. Основные достоинства метода: решающее правило строится быстро, получается достаточно простым и хорошо интерпретируемым, появляется дополнительная возможность изучать влияние отдельных признаков. Однако, если при построении дерева несколько признаков удовлетворяют критерию ветвления в равной или почти равной мере, то выбор одного из них происходит случайным образом. Поэтому, в зависимости от выбранного признака, построенные деревья могут существенно отличаться как по составу используемых признаков, так и по своим распознающим качествам.

В работе Е. В. Дюковой и Н. В. Пескова[1] предложена конструкция РД, названная полным решающим деревом (ПРД). В отличие от классического РД, в ПРД на каждой итерации строится так называемая полная вершина, которой соответствует набор признаков $X$ . В набор $X$ попадают все признаки, удовлетворяющие выбранному критерию ветвления. Далее по аналогии с классическим РД проводится ветвление по каждому из признаков, входящих в $X$ . В ПРД, в отличие от классического РД, описание распознаваемого объекта $S$ может принадлежать разным листьям ПРД. Поэтому выбор класса, которому

[1] Djukova E. V., Peskov N. V. A classification algorithm based on the complete decision tree // Pattern Recognition and Image Analysis. – 2007. – Vol. 17, no. 3. – Pp. 363-367.

принадлежит объект $S$, осуществляется на основе коллективного голосования по листьям дерева (голосования по большинству). Предложенная модель РД была успешно продемонстрирована её авторами на примере усовершенствования алгоритма построения допустимых разбиений (АДР).

В работах зарубежных ученых В. Бунтина (1992) и Р. Кохави (1997) была рассмотрена конструкция РД, близкая к конструкции ПРД. Однако использование предложенных В. Бунтиным и Р. Кохави моделей РД требовало существенных затрат времени и памяти. В связи с чем решалась оптимизационная задача определения допустимого числа признаков, образующих полную вершину, и полные вершины строились не на всех ярусах РД.

Следует отметить, что конструкция ПРД изучалась в указанной выше работе Е. В. Дюковой и Н. В. Пескова исключительно на задачах с целочисленной информацией низкой значности, в частности, бинарной. В большинстве прикладных задач признаки описаны в веществennозначном пространстве, при этом в признаковых описаниях объектов могут встречаться пропуски (случай неполной информации). Кроме того, часто встречаются задачи, в которых обучающие объекты неравномерно распределены по классам (в этом случае можно указать пару классов таких, что число обучающих объектов в одном из них существенно больше числа обучающих объектов в другом).

Актуальной задачей является исследование обобщающей способности РД (надежности и качества принятия решения). С этой целью применяются различные подходы, в частности подходы, использующие такие понятия, как емкость семейства, из которого выбирается классификатор, рademахеровская сложность этого семейства и отступ (margin) обучающего объекта от границы класса, которому принадлежит этот объект. С изучением обобщающей способности связана проблема переобученности (overfitting, overtraining) распознающего алгоритма, возникающая из-за того, что алгоритм слишком «подогнан» под обучающую информацию, и, как следствие, качество распознавания на новых объектах оказывается плохим.

Представляют интерес задачи изучения комбинаторной сложности РД

(получения оценок мощности множества вершин и мощности множества ребер) и временной сложности синтеза РД.

Перечисленные вопросы в основном исследованы для бинарного РД (БРД). Результаты, полученные для БРД, легко обобщаются на k-арные РД. Случай ПРД является более сложным, так как появляется новый тип вершин, существенно усложняющий конструкцию дерева.

**Целью диссертации** является построение и исследование распознающих процедур на основе ПРД с энтропийным критерием ветвления, решающих задачу распознавания с веществennозначной, неполной и неравномерно распределенной информацией, изучение обобщающей способности ПРД и получение оценок комбинаторной и временной сложности ПРД.

**Научная новизна.** Все основные результаты работы являются новыми. Развит подход к решению задачи распознавания по прецедентам, позволяющий более полно по сравнению с другими подходами, базирующимися на РД, использовать обучающую информацию. Впервые построены и изучены распознающие процедуры на основе ПРД с энтропийным критерием для задач классификации по прецедентам. За счет применения взвешенного голосования по листьям ПРД усовершенствовано решающее правило, применявшееся в первоначальной конструкции ПРД. Для оценки качества построенных алгоритмов использован скользящий контроль, ROC-анализ и представление классификатора в виде выпуклой корректирующей процедуры.

Для задачи с бинарной информацией получены верхние оценки емкости семейства ПРД. На прикладных задачах исследована эмпирическая радемахеровская сложность ПРД. Получена оценка обобщающей способности ПРД, учитывающая отступы обучающих объектов.

Предложен и апробирован метод снижения переобученности ПРД, основанный на вычислении средней глубины дерева и среднего числа обучающих объектов, описания которых попадают в лист дерева. Снижение переобученности ПРД достигается за счет обрезания ветвей дерева.

Для задачи с $m$ объектами и $n$ признаками (при фиксированной глубине

дерева и при условии, что из каждой вершины ПРД, не являющейся полной, выходит ровно две дуги) получены значения числовых характеристик ПРД и оценки времени синтеза ПРД. В указанном случае найдено максимальное число ребер ПРД, максимальное число полных вершин ПРД, максимальное число вершин ПРД, не являющихся полными, и максимальное число листьев ПРД. В общем случае получены оценки времени синтеза ПРД. Показано, что эти оценки могут быть улучшены в некоторых важных частных случаях.

В работе применялись **методы и понятия** распознавания образов, дискретной математики, комбинаторики, теории информации, теории сложности алгоритмов.

**Теоретическая и практическая значимость.** Результаты, полученные в диссертационной работе, могут быть использованы в теоретических исследованиях, касающихся построения эффективных реализаций для процедур распознавания на основе РД, а также при разработке спецкурсов по распознаванию образов, преподаваемых в госуниверситетах. Практическая значимость подтверждена тестированием на большом числе прикладных задач из различных областей.

**Апробация работы**. Результаты диссертационной работы докладывались и обсуждались на научных семинарах кафедры «Теоретической информатики и дискретной математики» в МПГУ; на научных семинарах ВЦ РАН; на мероприятии «Круглый стол молодых ученых по приоритетным направлениям развития науки», проводимом в МПГУ, 2007 г. и 2008 г.; на 14-ой Всероссийской конференции «Математические методы распознавания образов», гор. Суздаль, 2009 г.; на 8-ой Международной конференции «Интеллектуализация обработки информации», гор. Пафос (Кипр), 2010 г.; на 15-ой Всероссийской конференции «Математические методы распознавания образов», гор. Петрозаводск, 2011 г.

**Публикации.** По результатам диссертации опубликовано 9 печатных работ; из них 4 статьи входят в список ВАК РФ. Описания основных результатов включались в научные отчеты по проектам РФФИ 07-01-00516-а, 10-01-00770-а и в отчеты по грантам президента РФ по поддержке ведущих научных школ НШ

№5294.2008.1 и НШ №7950.2010.1.

**Структура и объем работы.** Диссертация состоит из введения, четырех глав, заключения, списка литературы и приложения. Содержание диссертации изложено на 169 страницах. Список использованной литературы содержит 132 наименования отечественных и зарубежных источников.

**СОДЕРЖАНИЕ ДИССЕРТАЦИОННОЙ РАБОТЫ**

В автореферате сохранена нумерация разделов, определений и теорем, используемая в диссертационной работе.

**Во введении** обоснована актуальность темы, сформулированы цель и задачи диссертационной работы, перечислены основные результаты, приведено описание структуры диссертации.

**В главе 1** рассмотрены методы построения РД в распознающих алгоритмах АДР и С4.5. Показана связь алгоритмов АДР и С4.5 с распознающими алгоритмами, основанными на поиске информативных фрагментов в признаковых описаниях объектов.

**В разделе 1.1** введены основные понятия, используемые в работе.

Пусть исследуется некоторое множество объектов $M$ . Объекты из $M$ описываются системой признаков $\{x_1,...,x_n\}$ . Известно, что множество $M$ представимо в виде объединения непересекающихся подмножеств, называемых классами $K_1,...,K_l$, $l \geq 2$. Дано конечное множество объектов (прецедентов или обучающих объектов) $\{S_1,...,S_m\}$ из $M$ , о которых известно, каким классам они принадлежат. Требуется по описанию в системе признаков $\{x_1,...,x_n\}$ некоторого объекта $S$ из $M$ определить класс, которому принадлежит объект $S$ .

В данном разделе описывается процедура построения РД для задачи распознавания по прецедентам с целочисленными признаками.

Обозначим через $T$ и $X(T)$ соответственно подмножество обучающих объектов и подмножество признаков, рассматриваемые на текущем шаге построения РД. На первом шаге $T = \{S_1,...,S_m\}$, $X(T) = \{x_1,...,x_n\}$.

На текущем шаге синтеза РД строится либо внутренняя вершина, либо лист

дерева. Для построения внутренней вершины выбирается некоторый признак $x$ из $X(T)$, который наилучшим образом разделяет объекты из разных классов. Пусть $V$ – множество различных значений признака $x$, встречающихся в описаниях объектов из $T$. В этом случае из вершины, соответствующей признаку $x$, строится $|V|$ дуг, помеченных разными числами из $V$. Далее по каждой дуге осуществляется спуск и строится либо лист дерева, либо следующая внутренняя вершина. При спуске по дуге с меткой $\sigma$, $\sigma \in V$, из $T$ удаляются те объекты, для которых значение признака $x$ не равно $\sigma$, из $X(T)$ удаляется признак $x$. Данный процесс синтеза РД повторяется до тех пор, пока не выполнится условие останова.

Пусть ветвь РД порождена внутренними вершинами $x_{j_1},...,x_{j_r}$, $\sigma_i$ – метка дуги, выходящей из вершины $x_{j_i}$, $i=1,...,r$. Тогда рассматриваемой ветви можно поставить в соответствие пару $(B,K)$, где $K \in \{K_1,...,K_l\}$, а $B$ – элементарная конъюнкция (э.к.) вида $x_{j_1}^{\sigma_1}...x_{j_r}^{\sigma_r}$, в которой $x^{\sigma}=1$, если $x=\sigma$, и $x^{\sigma}=0$ иначе.

Пусть $N_B$ – интервал истинности э.к. $B$. По построению $\{S_1,...,S_m\} \bigcap N_B \neq \varnothing$.

**Определение 1.11.** Ветвь РД называется *корректной*, если наборы из $\{S_1,...,S_m\} \bigcap N_B$ являются описаниями объектов только одного класса. В этом случае интервал $N_B$ называется *допустимым*.

**Определение 1.12.** РД называется *корректным*, если каждая ветвь РД является корректной.

Наиболее часто применяется следующее решающее правило. Пусть построено РД с $\mu$ листьями $(B_1,K_{j_1}),...,(B_\mu,K_{j_\mu})$. Из построения РД следует, что если найдется $i$, $i \in \{1,...,\mu\}$, такое, что $S \in N_{B_i}$, то $S$ относится к классу $K_{j_i}$. В противном случае происходит отказ от распознавания.

**В разделе 1.2** описан алгоритм АДР, строящий корректное БРД. Алгоритм АДР разбивает множество описаний обучающих объектов на непересекающиеся

интервалы, которые являются допустимыми интервалами. Такое разбиение называется оптимальным допустимым разбиением, если оно состоит из минимального числа допустимых интервалов. Синтез оптимального допустимого разбиения является NP-полной задачей. Для построения допустимого разбиения, близкого к оптимальному, используются эвристические критерии ветвления.

**В разделах 1.3 и 1.4** описан алгоритм C4.5 для следующих двух случаев: 1) для задачи с целочисленной информацией; 2) для задачи с вещественнозначной информацией и с наличием пропусков в признаковых описаниях объектов.

В разделе 1.3 рассмотрен случай целочисленной информации.

Выбор признака для построения внутренней вершины РД в алгоритме C4.5 осуществляется на основе энтропийного критерия. Опишем данный критерий.

Пусть $V=\{k_1,...,k_j\}$ – множество текущих значений признака $x$, $x\in X(T)$. Текущее множество обучающих объектов $T$ разбивается на подмножества $T^{(k_1)},...,T^{(k_j)}$, где $T^{(u)}$, $u\in\{k_1,...,k_j\}$, состоит из объектов, для которых значение признака $x$ равно $u$. Обозначим через $f(K_i,T)$, $i\in\{1,...,l\}$, – число объектов из множества $T$, относящихся к классу $K_i$. Вероятность $P_i(T)$, $i\in\{1,...,l\}$, того, что случайно выбранный объект из множества $T$ будет принадлежать классу $K_i$, равна $f(K_i,T)/|T|$.

Количество информации (энтропия), необходимое для определения класса, которому принадлежит объект из множества $T$, вычисляется по формуле $\mathrm{Info(T)}=-\sum_{i=1}^{l}P_i(T)\log P_i(T)$.

Количество информации, необходимое для определения класса, которому принадлежит объект из множества $T$ при разбиении $T$ по значениям признака $x$, оценивается величиной $\mathrm{Info}(x)=\sum_{u\in V}(|T^{(u)}|/|T|)\mathrm{Info}(T^{(u)})$.

Информационный выигрыш (information gain), получаемый при выборе признака $x$, вычисляется по формуле $\mathrm{Gain}(x)=\mathrm{Info(T)}-\mathrm{Info}(x)$.

Потенциальная информация, получаемая при разбиении множества $T$ по значениям признака $x$, оценивается величиной

$$\text{SplitInfo}(x) = -\sum_{u \in V} (|T^{(u)}|/|T|)\log(|T^{(u)}|/|T|).$$

В качестве признака для ветвления выбирается признак $x$ из $X(T)$, для которого величина $\text{GainRatio}(x) = \text{Gain}(x)/\text{SplitInfo}(x)$ (нормированный информационный выигрыш) принимает наибольшее значение.

В разделе 1.4 дано описание алгоритма C4.5 в случае вещественнозначной информации и наличия пропусков в признаковых описаниях объектов.

В рассматриваемом случае для ветвления из внутренней вершины РД, соответствующей признаку $x$, $x \in X(T)$, осуществляется бинарная перекодировка текущих значений признака $x$ с помощью выбора «оптимального» порога $d(x)$.

Порог $d(x)$ выбирается из множества текущих порогов для признака $x$. При этом под текущим *порогом* для признака $x$ понимается полусумма двух соседних значений из упорядоченного множества текущих различных значений признака $x$. Для каждого из текущих порогов осуществляется бинарная перекодировка и определяется информативность признака $x$ по энтропийному критерию $\text{GainRatio}(x)$. В качестве *оптимального* текущего порога $d(x)$ берется тот порог, для которого $\text{GainRatio}(x)$ имеет максимальное значение.

Данная процедура повторяется для каждого признака из $X(T)$. После этого из $X(T)$ выбирается наиболее информативный по критерию $\text{GainRatio}(x)$ признак $x_{opt}$ с оптимальным порогом $d(x_{opt})$. Далее строится внутренняя вершина с меткой $(x_{opt}, d(x_{opt}))$. Следует отметить, что при ветвлении из этой вершины признак $x_{opt}$ не удаляется и рассматривается на следующем шаге наравне с другими признаками.

В случае наличия пропусков по признаку $x$, $x \in X(T)$, предполагается, что пропущенные значения признака $x$ вероятностно распределены пропорционально частоте появления встречающихся значений. Поэтому при вычислении $\text{GainRatio}(x)$ учитываются все объекты из $T$ (при спуске из внутренней вершины, соответствующей признаку $x$, объекты из $T$, имеющие пропуски по признаку $x$, не удаляются).

**В разделе 1.5** показано, что классификаторы, основанные на построении РД, по конструкции близки к моделям логических процедур распознавания, например к алгоритмам голосования по представительным наборам и тестовым алгоритмам распознавания.

Пусть $H=\{x_{j_1},...,x_{j_r}\}$, $r\le n$, – набор из $r$ различных признаков, где $x_{j_i}\in\{x_1,...,x_n\}$, $i=1,...,r$, и пусть $S\in M$, $S=(a_1,...,a_n)$. Набор $H$ выделяет в описании объекта $S$ фрагмент $(S,H)=(a_{j_1},...,a_{j_r})$.

**Определение 1.13.** Фрагмент $(S,H)$, $S\in K$, называется представительным набором для класса $K$, если для любого обучающего объекта $S'\notin K$ имеет место $(S,H)\neq(S',H)$.

В алгоритмах голосования по представительным наборам на этапе обучения для каждого класса $K$, $K\in\{K_1,...,K_l\}$, строится некоторое множество представительных наборов.

Пусть листу РД соответствует пара $(B,K)$, $B=x_{j_1}^{\sigma_1}...x_{j_r}^{\sigma_r}$. Тогда, очевидно, $(\sigma_1,...,\sigma_r)$ – фрагмент, содержащийся в описаниях обучающих объектов $S'$ из класса $K$ таких, что $S'\in N_B$. Нетрудно также видеть, что в случае корректной ветви набор $(\sigma_1,...,\sigma_r)$ – представительный набор для класса $K$.

Из сказанного следует, что алгоритм АДР строит некоторое множество представительных наборов, так как этот алгоритм строит корректное РД. В дереве, построенном алгоритмом С4.5, ветви не всегда являются корректными. Алгоритм С4.5 нацелен на построение «почти» представительных наборов для каждого класса $K$, т.е. таких фрагментов, которые наиболее часто встречаются в описаниях обучающих объектов класса $K$ и достаточно редко в описаниях обучающих объектов из других классов.

**В главе 2** предложены алгоритмы синтеза ПРД для случаев целочисленной информации, веществennозначной информации, наличия пропусков и неравномерного распределения обучающих объектов по классам.

**В разделе 2.1** приведены основные понятия, используемые при построении

ПРД, и для указанных выше случаев описаны основные принципы синтеза ПРД.

Опишем структуру ПРД.

На каждом шаге построения ПРД строится либо лист дерева, либо формируется набор из различных признаков $X = \{x_{t_1},...,x_{t_q}\}$ , $X \subseteq X(T)$ , образующий полную вершину. Далее из этой полной вершины строится ровно $|X|$ дуг с метками $t_1,...,t_q$ , каждая из которых входит в обычную вершину, соответствующую признаку $x_{t_i}$ , $x_{t_i} \in X$ , $i = 1,...,q$ . При ветвлении из вершины, соответствующей признаку $x$ , происходит удаление признака $x$ из $X(T)$ и удаление некоторых обучающих объектов из $T$ .

**Определение 2.1.** Ветвью в ПРД называется путь, начинающийся в корне дерева и заканчивающийся в вершине ПРД.

Пусть $v$ – лист дерева, порожденная ветвью с обычными вершинами $x_{j_1},...,x_{j_r}$ , и $\sigma_i$ , $i \in \{1,...,r\}$, – метка дуги, выходящей из вершины $x_{j_i}$ .

**Определение 2.2**. Набор $N_v = (\alpha_1,...,\alpha_n)$ называется порождающим набором для листа $v$, если $\alpha_{j_i} = \sigma_i$ при $i = 1,...,r$ , и $\alpha_j = "*"$ при $j \notin \{j_1,...,j_r\}$.

Каждой висячей вершине ПРД ставится в соответствие пара $(N_v, \{\omega_v^1,...,\omega_v^l\})$ , где $N_v$ – порождающий набор для вершины $v$ , $\{\omega_v^1,...,\omega_v^l\}$ – вектор оценок за классы, способ вычисления этих оценок будет указан ниже.

**Определение 2.6.** Глубиной ветви в ПРД называется число обычных вершин, которые содержит эта ветвь, исключая концевую вершину ветви.

**Определение 2.7.** Ребром в ПРД называется дуга, выходящая из обычной или из полной вершины дерева.

**Определение 2.8.** Глубиной ПРД называется максимальная глубина среди всех построенных ветвей дерева.

**Определение 2.9.** Ярусом $i$ -ого уровня ($i$ -ым ярусом) в ПРД называется совокупность полных и обычных вершин, порожденных ветвями с глубиной $i-1$, а также совокупность листьев дерева, порожденных ветвями с глубиной $i$ .

Как правило, используются два способа ветвления из обычной вершины,

соответствующей признаку $x$, $x \in X(T)$. Первый способ предпочтителен в случае бинарной или целочисленной информации низкой значности, а второй в случае веществennозначной или целочисленной информации высокой значности.

Способ 1. Пусть $V$, $|V| \geq 2$, – множество различных значений признака $x$, встречающихся в описаниях объектов из $T$. В этом случае из обычной вершины $x$ выходит $|V|$ дуг, помеченных разными числами. Пусть $\sigma$, $\sigma \in V$, – метка одной из дуг, выходящих из вершины $x$. При спуске из вершины $x$ по дуге с меткой $\sigma$ удаляются те объекты из $T$, для которых значение признака $x$ не равно $\sigma$.

Способ 2. Для ветвления из обычной вершины, соответствующей признаку $x$, осуществляется бинарная перекодировка текущих значений признака $x$ с помощью «оптимального» порога $d(x)$ (разд. 2.4). Рассматриваемая вершина помечается парой $(x, d(x))$. Спуск из вершины $(x, d(x))$ происходит по двум ветвям, при этом левая ветвь помечается 0, а правая – 1. При спуске из вершины $(x, d(x))$ по левой (правой) ветви удаляются те объекты из $T$, для которых значение признака $x$ больше (не больше) $d(x)$.

Рассмотрим решающие правила, используемые при классификации распознаваемого объекта $S = (b_1, ..., b_n)$ с помощью ПРД в способах 1 и 2.

Способ 1. Под описанием объекта $S$ в вершине $v$ понимается вектор $S(v) = (\beta_1, ..., \beta_n)$, в котором $\beta_{j_i} = b_{j_i}$ при $i = 1, ..., r$, иначе $\beta_j = "*"$.

Способ 2. Под описанием объекта $S$ в вершине $v$ понимается вектор $S(v) = (\beta_1, ..., \beta_n)$, в котором $\beta_{j_i} = 1$, если $b_{j_i} > d(x_{j_i})$, иначе $\beta_{j_i} = 0$ при $i = 1, ..., r$, и $\beta_j = "*"$ при $j \notin \{j_1, ..., j_r\}$.

**Определение 2.10.** Лист $v$ называется *голосующим* за $S$, если $S(v) = N_v$.

Пусть $Q(S)$ – множество всех голосующих листьев ПРД за объект $S$. Для каждого $i \in \{1, ..., l\}$ вычисляется оценка принадлежности объекта $S$ классу $K_i$, имеющая вид $\Gamma(S, K_i) = \sum_{v \in Q(S)} \omega_v^i$, $i = 1, ..., l$.

Объект $S$ зачисляется в класс $K_i$, если:

$$\Gamma(S,K_i) = \max_{j=1,...,l} \Gamma(S,K_j),\ i=1,...,l,\ \ \Gamma(S,K_i) \neq \Gamma(S,K_j) \text{ при } i \neq j,\ \ j=1,...,l.$$

Если классов с максимальной оценкой несколько и среди них имеется класс $K$, который имеет наибольшее число объектов в обучающей выборке, то $S$ приписывается классу $K$. Если классов с максимальной оценкой несколько и среди них нет класса с наибольшим числом объектов в обучающей выборке, то происходит отказ от классификации объекта $S$.

Рассмотрим теперь случай, когда информация вещественнозначная и в признаковых описаниях объектов имеются пропуски.

Пусть значение признака $x$ для некоторого обучающего объекта не определено, т.е. равно "$*$". Методика построения ПРД в случае наличия пропусков направлена на сохранение исходной информации в полном объеме. Поэтому при ветвлении из вершины, соответствующей признаку $x$, рассматриваемый обучающий объект удаляется.

Стоит отметить, что при синтезе ПРД происходит лавинообразный рост вершин и ветвей. В связи, с чем увеличивается время классификации объекта $S$. Для сокращения времени классификации предложено строить только голосующие за $S$ листья дерева. Например, при ветвлении из обычной вершины, соответствующей признаку $x_{j_i}$, строится только ветвь с меткой $\sigma_i$, где $\sigma_i = \beta_{j_i}$. В случае, когда описание объекта $S$ содержит пропуски, то при классификации этого объекта признаки с пропущенными значениями исключаются из $\{x_1,...,x_n\}$. Данная методика позволяет сократить время классификации на скользящем контроле («leave-one-out», LOO) более чем в 2 раза. Если же имеется контрольная выборка и синтез ПРД для каждого объекта из контрольной выборки требует значительного времени, то целесообразнее один раз полностью построить ПРД.

**В разделе 2.2** описан алгоритм Полный C4.5, являющийся модификацией алгоритма C4.5. Разработанный алгоритм строит корректное ПРД и предназначен для обработки целочисленной информации. Для построения полной вершины предложена специальная процедура выделения набора признаков. Данная процедура заключается в выборе признаков, информативность которых равна

максимальному значению величины GainRatio($x$), $x \in X(T)$, на текущем шаге синтеза ПРД, а также признаки, информативность которых близка к максимальной информативности. Близость признаков по информативности определяется с применением текущего среднего нормированного информационного выигрыша.

Вектор оценок для висячей вершины $v$ с порождающим набором $N_v$ в алгоритме Полный C4.5 определяется следующим образом. Пусть $m_v^i$ – число обучающих объектов класса $K_i$, $i \in \{1,...,l\}$, описания которых равны $N_v$. Оценка $\omega_v^i = 1$, если $m_v^i \neq 0$, иначе $\omega_v^i = 0$.

**В разделе 2.3** описаны алгоритмы AGI.Voice, AGI.La.max, AGI.La.sum и AGI.Bias, использующие схемы синтеза ПРД из разд. 2.1 для задач классификации с веществennозначной информацией и с наличием пропусков в признаковых описаниях объектов. В этих алгоритмах процедура выделения набора признаков аналогична процедуре, используемой в алгоритме Полный C4.5. Для вычисления информативности признака $x$, $x \in X(T)$, используется модифицированный энтропийный критерий, описание которого приведено в разд. 2.4.

Вектор оценок $\{\omega_v^1,...,\omega_v^l\}$ для висячей вершины $v$ с порождающим набором $N_v$ в указанных алгоритмах определяется следующим образом. Пусть $m_v^i$, $i \in \{1,...,l\}$, – число обучающих объектов класса $K_i$, описания которых равны $N_v$, $m^i$ – число всех обучающих объектов класса $K_i$, $m_v^* = \max\limits_{i=1,...,l} m_v^i$ и пусть $m_v = \sum_{i=1}^{l} m_v^i$. Тогда

1) в алгоритме AGI.Voice оценка $\omega_v^i = 1$, если $m_v^i = m_v^*$, иначе $\omega_v^i = 0$;

2) в алгоритме AGI.La.max оценка $\omega_v^i = (m_v^i + 1)/(m_v + l)$, если $m_v^i = m_v^*$, иначе $\omega_v^i = 0$;

3) в алгоритме AGI.La.sum оценка $\omega_v^i = (m_v^i + 1)/(m_v + l)$;

4) в алгоритме AGI.Bias оценка $\omega_v^i = (m_v^i + 1)/(m^i + l)$.

**В разделе 2.4** описан критерий выбора оптимального порога $d(x)$ для бинарной перекодировки текущих значений признака $x$, используемый в алгоритмах AGI.Voice, AGI.La.max, AGI.La.sum и AGI.Bias. Выбор $d(x)$ позволяет оценить информативность признака $x$ и из наиболее информативных признаков построить полную вершину ПРД.

Предложенный критерий отличается от соответствующего критерия в алгоритме C4.5. Опишем эти отличия. Строятся только такие пороги, которые позволяют отличать объекты из разных классов. При вычислении информативности признака $x$, $x \in X(T)$, учитываются только те объекты из $T$, в описании которых значение признака $x$ определено.

**В главе 3** проведено тестирование разработанных алгоритмов Полный C4.5, AGI.Voice, AGI.La.max, AGI.La.sum и AGI.Bias на прикладных задачах. Качество этих алгоритмов сравнивалось с распознающими алгоритмами из системы интеллектуального анализа данных, распознавания и прогнозирования «Распознавание 2.0» (разработано в ВЦ РАН), с распознающими алгоритмами из широко распространенной системы «WEKA», основанных на построении РД, а также с алгоритмом C5.0. Качество алгоритмов оценивалось методом LOO функционалами $\tilde{\Theta} = m^+/m$, $\Theta = \sum_{i=1}^{l} q_i / l$ и $Q = \min\limits_{i=1,\ldots,l} q_i$, где $q_i$ – процент правильно распознанных объектов класса $K_i$, $m^+$ – общее число правильно распознанных объектов.

На примере алгоритма AGI.Bias рассмотрены вопросы снижения переобученности ПРД. Предложена методика снижения эффекта переобучения ПРД с использованием числовых характеристик дерева. Проведен анализ ПРД как выпуклой корректирующей процедуры.

**В разделе 3.1** проведено сравнение алгоритма Полный C4.5 с алгоритмами Полный АДР (усовершенствованный с помощью ПРД алгоритм АДР) и C4.5. Тестирование проводилось на прикладных задачах из различных областей.

Вычислялась величина $\tilde{\Theta}$. Алгоритм Полный C4.5 показал лучшие результаты.

**В разделах 3.2 и 3.3** проведено сравнение алгоритмов AGI.La.max, AGI.Voice, AGI.La.sum и AGI.Bias с алгоритмом C5.0 и с алгоритмами из системы «Распознавание 2.0», а именно с нейронной сетью (НС), алгоритмом вычисления оценок (ABO), стохастическим алгоритмом голосования по тупиковым тестам (ГТТ), линейным дискриминантом Фишера (ЛДФ), алгоритмом построения РД с применением генетического метода (ГМ), методом синтеза БРД (МБРД) и с методом k ближайших соседей (МБС). Алгоритмы AGI.La.sum и AGI.Bias также сравнивались с алгоритмами из системы «WEKA», а именно с алгоритмом J48 (аналог алгоритма C4.5), SimpleCART (аналог алгоритма CART), RandomForest («бэггинг» над РД) и LADTree («бустинг» над РД). Тестирование алгоритмов проводилось на 25 реальных задачах, собранных в отделе Математических проблем распознавания и методов комбинаторного анализа ВЦ РАН. Вычислялись величины $\Theta$ и $Q$.

Следует отметить, что функционал $\tilde{\Theta}$ не рассматривался, так как при неравномерном распределении обучающих объектов по классам как правило $\tilde{\Theta} \geq \Theta$.

Наилучшие результаты среди разработанных алгоритмов показал алгоритм AGI.Bias, который нацелен на случай неравномерного распределения объектов по классам. Качество алгоритма AGI.Bias по показателю Q превосходит все указанные в тестировании алгоритмы на большинстве задач. По показателю $\Theta$ алгоритм AGI.Bias лучше алгоритмов C5.0, МБРД, ГМ, J48, ABO, SimpleCART, LADTree, и немного уступает алгоритму НС на большинстве задач. При сравнении с другими алгоритмами по показателю $\Theta$ алгоритм AGI.Bias на одних типах задач (с целочисленной информацией, с пропусками) немного превосходит эти алгоритмы, а на других задачах (с вещественнозначной информацией, без пропусков) немного уступает.

На шести задачах проведен ROC-анализ алгоритмов НС, ГМ, ГТТ и AGI.La.sum. Методика ROC-анализа позволяет выявить соотношение ошибок первого и второго рода. Ошибка первого рода – отнесение объекта из первого

класса во второй класс. Ошибка второго рода – отнесение объекта из второго класса в первый класс. Лучшим алгоритмом стал алгоритм ГТТ, второй результат показал алгоритм AGI.La.sum, третий результат – алгоритм НС, четвертый результат – ГМ. По результатам ROC-анализа алгоритмов AGI.La.sum и AGI.Bias показано преимущество алгоритма AGI.Bias.

**В разделе 3.4** предложена процедура снижения эффекта переобучения ПРД. Первый этап процедуры – синтез начального ПРД с использованием методики, описанной в разд. 2.1, и вычисление характеристик дерева, а именно: средней глубины дерева, обозначаемой $D$, и среднего числа обучающих объектов $J$, описания которых попадают в лист дерева. Второй этап процедуры – синтез итогового ПРД. Итоговое ПРД отличается от начального ПРД тем, что текущая ветвь обрывается и строится лист дерева, если глубина текущей ветви превышает значение $D$ или число объектов в текущем множестве обучающих объектов меньше значения $J$. Также рассматривались варианты построения итогового ПРД с использованием одной из указанных характеристик и с небольшими отступами от точных значений.

Указанная методика исследована на примере алгоритма AGI.Bias. Счет на прикладных задачах показал эффективность предложенной методики, в среднем снижение переобученности ПРД с комбинацией «$D$ и $J$» составило 5%.

**В разделе 3.5** на прикладных задачах проведено исследование ПРД как выпуклой корректирующей процедуры, так как решающее правило на основе ПРД представимо в виде выпуклой комбинации базовых классификаторов. Показано, что ошибка выпуклой корректирующей процедуры для ПРД меньше по сравнению с классическим РД.

**В главе 4** получены значения числовых характеристик ПРД (точные оценки мощности множества вершин и мощности множества ребер) при фиксированной глубине дерева, когда из каждой обычной вершины в ПРД выходят ровно две дуги. Для бинарной информации получены верхние оценки емкости семейства ПРД. Найдены оценки времени синтеза ПРД. На основе понятий из теории отступов получена оценка обобщающей способности ПРД. На реальных задачах

проведено исследование рademахеровской сложности ПРД.

**В разделе 4.1** получены значения числовых характеристик ПРД в зависимости от глубины дерева $k$. Найдено максимальное число полных вершин $FV(k)$, максимальное число обычных вершин $SV(k)$, максимальное число листьев $L(k)$ и максимальное число ребер $R(k)$. Показано, что $k \le \min\{n, m-1\}$.

Обозначим через $A_n^k$ – число размещений из $n$ по $k$.

**Теорема 4.1.1.** Имеет место

$$FV(k) = \sum_{i=1}^{k} 2^{i-1} A_n^{i-1},\ SV(k) = \sum_{i=1}^{k} 2^{i-1} A_n^i,\ L(k) = 2^k A_n^k,\ R(k) = 3SV(k).$$

Найденные оценки также являются оценками аналогичных характеристик для бинарного ПРД (БПРД).

Известно, что для классического БРД справедливо

$$R(k) = 2SV(k),\ L(k) = 2^k \text{ и } SV(k) = L(k) - 1.$$

Полученные результаты свидетельствуют о значительной комбинаторной сложности ПРД по сравнению с классическим РД.

**В разделе 4.2** с помощью метода pVCD, предложенного В. И. Донским, получены верхние оценки емкости семейства БПРД. Метод pVCD основан на понятии Колмогоровской сложности конечных объектов и заключается в получении сжатого двоичного описания элемента рассматриваемого семейства.

Пусть $FD_{\mu,l,n,k}$ – семейство БПРД с не более $\mu$ листьями и глубиной не более $k$, в котором БПРД строится для задачи распознавания с $n$ признаками и $l$ классами. Обозначим через $\mathrm{VCD}(FD_{\mu,l,n,k})$ – емкость семейства $FD_{\mu,l,n,k}$.

Получена оценка $\mathrm{VCD}(FD_{\mu,l,n,k}) \le \mu(k+32l) + ((k+2)\mu - 1)\lceil \log(n+2) \rceil$.

Для сравнения оценка емкости семейства БРД с не более $\mu$ листьями, в котором БРД строится для задачи с $n$ признаками и двумя классами, не превышает величины $(\mu-1)\left(\lceil \log n \rceil + \lceil \log(\mu+3) \rceil\right)$.

Таким образом, полученная оценка емкости семейства БПРД превосходит оценку емкости семейства БРД примерно в $k+2$ раза. Также показано, что

основными величинами, влияющими на оценку емкости семейства ПРД, являются число висячих вершин и глубина дерева. Тем самым за счет уменьшения числа висячих вершин и глубины дерева происходит уменьшение сложности семейства ПРД, и, следовательно, снижается переобученность ПРД. Этот вывод согласуется с результатами разд. 3.4.

**В разделе 4.3** на примере алгоритма AGI.Bias получены оценки времени $T(n,m,l)$ синтеза ПРД.

Пусть $k$ – глубина ПРД.

**Теорема 4.3.1.** Имеет место следующая оценка времени построения ПРД:

$$T(n,m,l) = O(A_n^k (m-k)(m-k+1)l) + O(A_n^k (m-k)(n-k+1)).$$

**Теорема 4.3.2.** Если на каждом шаге построения ПРД в полную вершину попадает не более $\lceil n'/2 \rceil$ признаков, где $n' = |X(T)|$, то

$$T(n,m,l) = O((k+1)(m-k)(m-k+1)l\, A_n^k / 2^k) + O((k+1)(m-k)(n-k+1)\, A_n^k / 2^k).$$

**Теорема 4.3.3.** Если на каждом шаге построения ПРД в полную вершину попадает не больше трех признаков из $X(T)$, то

$$T(n,m,l) = O(3^{k-1}(m-k)(m-k+1)l) + O(3^k (m-k)(n-k+1)).$$

Полученные оценки времени синтеза ПРД существенно превосходят оценки времени синтеза классического РД.

**В разделе 4.4** с использованием понятий из теории отступов получена оценка обобщающей способности ПРД. На реальных задачах проанализированы распределения отступов обучающих объектов.

**Определение 4.6.** Отступом объекта $S$ называется величина $\text{margin}(S) = \bar{\omega}^k(S) - \max_{j \neq k}\{\bar{\omega}^j(S)\}$, где $k$ – номер правильного класса для $S$, $\bar{\omega}^j(S)$ – оценка классификатора за принадлежность $S$ классу $K_j$, $\sum_{j=1}^{l} \bar{\omega}^j(S) = 1$.

Далее рассматривается стандартная задача распознавания с двумя классами, при этом метки классов принадлежат множеству $Y = \{-1,1\}$. Пусть $D$ – распределение на $M \times Y$, $T = \{S_1,...,S_m\}$ – множество обучающих объектов выбранных независимо и случайно из $D$.

Решающая функция $FDT(S)$ на основе ПРД с $\mu$ листьями для объекта $S$ может быть представлена в следующем виде:

$$FDT(S) = \text{sgn}(f(S)) = \text{sgn}\left[\sum_{i=1}^{\mu} \alpha_i \sigma_i B_i(S)\right],$$

где $\sigma_i \in Y$ – метка класса $i$-ого листа, $\sum_{i=1}^{\mu} \alpha_i = 1$. Если описание объекта $S$ для $i$-ого листа принадлежит $N_{B_i}$, то $B_i(S) = 1$, иначе $B_i(S) = 0$, $i = 1,..., \mu$.

Пусть $R_j$ – множество предикатов для признака $x_j$, э.к. $B_i$ – конъюнкция предикатов, $i \in \{1,..., \mu\}$. Обозначим через $U = \bigcup_{i=1}^{n} R_i$.

**Теорема 4.4.1.** Пусть $VCD(U)$ – емкость семейства $U$, и пусть $\delta > 0$. Тогда с вероятностью не меньше $1-\delta$ над множеством $T$, для каждой решающей функции $FDT(S)$ и для любого $\theta > 0$ справедливо

$$P_D[\text{error}] \le 2P_T(\text{margin} \le \theta) + c/m(1/\theta^2 (v \ln m + \ln d) \ln (m\theta^2/v) + \ln(1/\delta)),$$

где $P_T(\text{margin} \le \theta)$ – вероятность того, что отступ для случайно выбранного объекта из $T$ не превысит величины $\theta$, $P_D[\text{error}]$ – вероятность ошибки ПРД на объекте $S \in M$, $v = \sum_{i=1}^{\mu} \alpha_i d_i VCD(U)$, $d = \max_i d_i$, где $d_i$ – глубина $i$-ого листа.

Таким образом, получено, что выпуклая комбинация (ПРД) сглаживает прогнозы базовых классификаторов (РД).

В данном разделе на прикладных задачах показано, что при построении ПРД отступы обучающих объектов увеличиваются, при этом увеличивается доля объектов с положительным отступом, тем самым повышается вероятность правильно распознать произвольный объект.

**В разделе 4.5** проведено исследование эмпирической радемахеровской сложности ПРД.

**Определение 4.7.** Эмпирической радемахеровской сложностью (ЭРС) семейства $H$ на обучающей выборке $T$ называется величина $R_T(H) = \mathbf{E}_{\vec{r}}\left[\sup_{h \in H} \left| m^{-1} \sum_{j=1}^{m} r_j h(S_j) \right| \right]$, где $\vec{r} = (r_1,..., r_m)$ – вектор радемахеровских переменных, независимых от обучающей выборки.

Важность данного понятия отражена в следующем неравенстве. Пусть $P(h)$ – вероятность ошибки функции $h$ на произвольном объекте, $\nu(h)$ – частота ошибок функции $h$ на обучении. Тогда для любого $\delta > 0$ и для любой функции $h \in H$ с вероятностью не меньшей $1-\delta$ справедливо

$$P(h) \le \nu(h) + R_T(H) + 3\sqrt{\ln(2/\delta)/2m}\,.$$

На прикладных задачах показано, что ЭРС полного дерева существенно ниже ЭРС классического РД за счет использования взвешенного голосования и удаления просмотренного признака при спуске из обычной вершины. При этом применение полных вершин практически не изменяет ЭРС дерева, этот факт также подтверждается тем, что выпуклая комбинация (ПРД) не увеличивает сложность (разд. 4.4). Таким образом, вероятность ошибки классификации с помощью ПРД будет ниже, чем при применении классического РД.

**В заключении** приведены основные результаты, полученные в работе.

**ОСНОВНЫЕ РЕЗУЛЬТАТЫ ДИССЕРТАЦИИ**

1. Разработаны и исследованы распознающие процедуры на основе ПРД с энтропийным критерием, использующие различные варианты взвешенного коллективного голосования по листьям ПРД (для задач с целочисленной и вещественнозначной информацией, с наличием пропусков в признаковых описаниях объектов и с неравномерным распределением обучающих объектов по классам).

2. Исследованы вопросы обобщающей способности ПРД. С помощью метода pVCD получена верхняя оценка емкости семейства БПРД. Для задачи с двумя классами на основе теории отступов получена оценка обобщающей способности семейства ПРД. Экспериментально изучена эмпирическая радемахеровская сложность ПРД. Разработана и исследована методика снижения переобученности ПРД.

3. При фиксированной глубине дерева и при условии, что из каждой обычной вершины ПРД выходят ровно две дуги, получены точные оценки мощности множества вершин и мощности множества ребер ПРД, а также оценки времени синтеза ПРД.

## СПИСОК ПУБЛИКАЦИЙ ПО ТЕМЕ ДИССЕРТАЦИИ


1. Генрихов И. Е., Дюкова Е. В. Полные решающие деревья в задачах классификации по прецедентам // Всеросс. конф. Математические методы распознавания образов-15. – М.: МАКС Пресс, 2011. – С. 84-87.
2. Генрихов И. Е., Дюкова Е. В. Исследование комбинаторных свойств и сложности построения полных решающих деревьев // Всеросс. конф. Математические методы распознавания образов-15. – М.: МАКС Пресс, 2011. – С. 88-91.
3. Генрихов И. Е. Построение полного решающего дерева на базе алгоритма C4.5 // Сообщение по прикладной математике. – М.: ВЦ РАН, 2009. – 24 с.
4. Генрихов И. Е., Дюкова Е. В. Усовершенствование алгоритма C4.5 на основе использования полных решающих деревьев // Всеросс. конф. Математические методы распознавания образов-14. – М.: МАКС Пресс, 2009. – С. 104-107.
5. Генрихов И. Е., Дюкова Е. В. Построение и исследование распознающих процедур на основе полных решающих деревьях // Междунар. конф. Интеллектуализация обработки информации-8. – М.: МАКС Пресс, 2010. – С. 117-120.
6. **Генрихов И. Е., Дюкова Е. В. Классификация на основе полных решающих деревьев // Ж. вычисл. матем. и матем. физ. – 2012. – Т. 52, № 4. – С. 750-761.**
7. **Генрихов И. Е. Исследование переобученности распознающих процедур на основе полных решающих деревьев // Программные продукты и системы. – 2011 – № 4 (96). – С. 141-147.**
8. **Генрихов И. Е. О сложности построения полных решающих деревьев // Естественные и технические науки. – 2012. – № 1 (57). – С. 327-336.**
9. **Genrikhov I. E. Synthesis and analysis of recognizing procedures on the basis of full decision trees // Pattern Recognition and Image Analysis. – 2011. – Vol. 21, no. 1. – Pp. 45-51.**