**Таврический национальный университет им. В.И. Вернадского**

На правах рукописи

**ДЮЛИЧЕВА Юлия Юрьевна**

УДК 519.68

# МОДЕЛИ КОРРЕКЦИИ РЕДУЦИРОВАННЫХ БИНАРНЫХ РЕШАЮЩИХ ДЕРЕВЬЕВ

01.05.01 – Теоретические основы информатики и кибернетики

Диссертация на соискание ученой степени кандидата физико-математических наук

Научный руководитель:
доктор физико-математических наук,
профессор
ДОНСКОЙ Владимир Иосифович

Симферополь – 2004

# ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ ……. 4

Раздел 1. Методы синтеза и редукции решающих деревьев: обзор и задачи исследования ……... 12

1.1. Постановка задачи распознавания и основные понятия теории решающих деревьев …… 12

1.2. Задача синтеза решающих деревьев. Выбор критерия ветвления……. 23

1.3. Задача редукции ветвей решающих деревьев. Выбор критерия остановки ветвления …….. 37

1.4. Выводы ……. 49

Раздел 2. Методы оценивания решающих деревьев ……. 53

2.1. Оценивание бинарного решающего дерева на основе подхода к закономерности как неслучайности ……... 53

2.2. Обоснование редукции ветвей решающего дерева на основе подхода к закономерности как неслучайности …….. 60

2.3. Оценка VCD для основных классов решающих функций, представленных решающими деревьями …… 63

2.3.1. Оценка *VCD* класса решающих функций, представленных решающим деревом ограниченного ранга …….. 67

2.3.2. Оценка *VCD* класса решающих функций, представленных бинарным решающим деревом с двумя классами…… 69

2.4. Выводы……... 71

Раздел 3. Алгоритмы синтеза и принятия решений эмпирическим решающим лесом ……... 73

3.1. Эмпирический решающий лес: основные определения ……..….. 73

3.2. ***DFBSA*** – последовательный алгоритм синтеза эмпирического

решающего леса по ссылкам …….…… 75
3.3. Алгоритмы принятия решений эмпирическим решающим лесом 79
3.4. Алгебраическая алгоритмическая модель коррекции *r*-некорректного эмпирического леса ……... 84
3.5. Оценка *VCD* класса решающих функций, представленных *r*-редуцированным эмпирическим лесом ……..……. 89
3.6. Выводы …….. 95
Раздел 4. Программная реализация и апробация алгоритма ***DFBSA*** синтеза эмпирического решающего леса……. 97
4.1. Программная реализация алгоритма синтеза эмпирического решающего леса …… 97
4.2. Апробация алгоритма ***DFBSA*** синтеза эмпирического решающего леса…….. 100
4.3. Выводы …….. 110

ЗАКЛЮЧЕНИЕ ……..... 111

# ВВЕДЕНИЕ

Диссертационная работа посвящена исследованию и усовершенствованию алгоритмов обучения и распознавания, основанных на построении бинарных решающих деревьев; разработке обоснованных правил редукции бинарных решающих деревьев, основанных на оценивании конъюнктивных закономерностей; созданию последовательной процедуры синтеза совокупности решающих деревьев – алгоритма синтеза эмпирического решающего леса – и методов коррекции совокупности редуцированных решающих деревьев как набора эвристических процедур принятия решений.

**Актуальность темы.**

Задача обучения распознаванию по прецедентам является одной из центральных задач кибернетики. Специалистов в области распознавания образов решающие деревья (РД) привлекают возможностью “сжатого” описания заданного множества обучающих прецедентов, простотой программной реализации РД. На фоне других классифицирующих моделей решающие деревья выгодно выделяются возможностью представления выявленных эмпирических закономерностей в легко воспринимаемом и интерпретируемом для специалистов различных отраслей знаний виде. Важным является и тот факт, что РД позволяют выявлять информативные подсистемы признаков из их исходной совокупности, обращая внимание исследователя на скрытые связи между объектами.

Решающие деревья являются важнейшим инструментом реализации алгоритмических отображений, аппроксимирующих начальную (прецедентную) информацию, и средством синтеза структурных моделей закономерностей. Начиная от первых научных работ Ховленда (Hoveland), Ханта (Hunt) и А.Ш. Блоха в 50-60-х годах XX века, исследователи и разработчики информационных систем во всем мире по сей день активно изучают методы анализа, синтеза и редукции РД и публикуют результаты по теории и практическому использованию решающих деревьев. В связи с нежелательностью усложнения РД, являющегося результатом “перенастройки” на обучающую выборку, особенно актуальной

является проблема обоснования ограничения сложности решающих деревьев. Эта проблема с точки зрения структуры РД связана с редукцией решающих деревьев.

Однако большинство существующих алгоритмов синтеза и редукции РД основаны на интуитивных соображениях исследователей и не имеют четкого математического обоснования. В диссертации приведено теоретическое обобщение и новое решение научной проблемы разработки математически обоснованных методов редукции РД и коррекции редуцированных деревьев, как методов выявления новых структурных закономерностей в данных, путем построения совокупности решающих деревьев достаточно простой структуры.

Решающие деревья эффективно применяются на практике, что подтверждено многочисленными публикациями и экспериментальными данными, в частности, для решения задач диагностики заболеваний, геологического прогнозирования, распознавания письменных знаков, речи, изображений и во многих других практических приложениях.

Практическая полезность исследований и разработок по теме диссертации определяется в существенной мере и тем, что в последние годы интерес к решающим деревьям повысился в связи с активными разработками интеллектуализированного программного обеспечения, развитием информационных технологий Machine Learning и Data Mining.

**Связь с научными программами, планами, темами.**

Диссертационная работа выполнена согласно плану научно-исследовательской работы кафедры информатики Таврического национального университета им. В.И. Вернадского по госбюджетной тематике на 2001-2004 г.г. «Разработка гибридной универсальной программной оболочки для построения экспертных систем и логических систем поддержки принятия решений», № государственной регистрации работы 0102U001575.

**Цели диссертационной работы:**

1. Обосновать целесообразность редукции бинарных решающих деревьев, используя методы дискретной математики и вероятностного оценивания.

2. Найти и обосновать методы понижения сложности древообразных классификаторов, сохранив требование их корректности на достоверной обучающей информации.
3. Предложить и обосновать методы принятия решений, основанные на коррекции совокупности редуцированных решающих деревьев как набора эвристических (и в общем случае некорректных) процедур.

Для достижения поставленных целей в диссертационной работе решаются следующие **задачи**:

1. Разработать вероятностный критерий отсечения (редукции) ветвей бинарного решающего дерева, имеющих число внутренних вершин, превышающее заданное значение ранга *r*. Обосновать такую редукцию с точки зрения неслучайности (закономерности) обнаружения в эмпирической выборке конъюнктивной закономерности ранга *r*.
2. На основе оценок *VCD* (сложности класса решающих правил по теории Вапника-Червоненкиса) для классов алгоритмов распознавания, определяемых бинарными решающими деревьями с ограничением на число вершин, обосновать целесообразность усложнения правил распознавания и процедур коррекции решений.
3. Разработать методы построения корректной совокупности решающих деревьев (эмпирического решающего леса), обеспечивающей возможность точной настройки на обучающую выборку с одновременным соблюдением ограничения на ранг ветвей РД.
4. Получить оценку сложности эмпирического решающего леса и изучить другие его свойства как специального семейства алгоритмов распознавания.
5. Разработать алгоритмы коррекции совокупности некорректных эмпирических решающих деревьев, обеспечивающие повышение точности классификации.

6. Создать необходимое программное обеспечение и провести эксперименты на реальных данных с целью подтверждения теоретических результатов, полученных в диссертации.

Для решения поставленных задач используются методы дискретной математики, алгебры, теории множеств, теории вероятностей и математической статистики и широкий арсенал средств теоретической информатики и кибернетики, включая статистическую теорию обучения Вапника-Червоненкиса и алгебраическую теорию Ю.И.Журавлёва коррекции эвристических процедур принятия решений.

**Научная новизна полученных результатов**. В диссертационной работе получены следующие новые результаты, которые выносятся на защиту.

1. Обоснован процесс редукции ветвей РД на основе вероятностного подхода к оцениванию эмпирических закономерностей как неслучайностей. Получены оценки случайного обнаружения в стандартных обучающих таблицах конъюнктивных закономерностей как ветвей РД заданного ранга и в целом – оценки возможности “случайного” обнаружения РД-структуры заданной сложности.
2. Предложен новый алгоритм синтеза совокупности решающих деревьев с ограничением на ранг ветвей.
3. Решена проблема синтеза эмпирического решающего леса, как решающего правила, в котором соблюдается ограничение на ранг ветвей (конъюнкций) РД, и сохраняется возможность правильной классификации всех объектов обучающей выборки.
4. На основе емкостной характеристики Вапника-Червоненкиса исследована сложность и получена оценка *VCD* класса решающих правил, порождаемых эмпирическим решающим лесом.
5. На основе алгебраического подхода к распознаванию построена модель алгебраической коррекции *r*-некорректного эмпирического леса.

6. Получено экспериментальное подтверждение того, что эмпирический решающий лес в среднем на множестве решаемых задач обеспечивает более точное распознавание, чем отдельные РД.

**Практическое значение полученных результатов** диссертационной работы состоит в возможности применения разработанных в ней алгоритмов синтеза и принятия решений эмпирическим решающим лесом для построения информационных интеллектуализированных систем, выявления скрытых структурных закономерностей в данных, построения логических описаний классов объектов в виде дизъюнктивных нормальных форм. В частности, как результат решения практической задачи, в диссертации построен эмпирический решающий лес, позволяющий распознавать типы (классы) вибрионов и изучать особенности этих классов на основе найденных конъюнктивных закономерностей.

**Личный вклад**.

Все представленные в диссертации результаты получены лично автором. Научному руководителю профессору Донскому В.И., соавтору работ [20, 22, 23], принадлежат постановки задач и часть совместно проведенных экспериментов с решающими деревьями на модельных экспериментальных данных с использованием программного комплекса “Дуэль” [21].

**Апробация результатов диссертации**.

Результаты работы докладывались и обсуждались на Международной научно-практической конференции “Знание – Диалог - Решение” (Санкт-Петербург, июнь, 2001 г.); Международной конференции по индуктивному моделированию (Львов, май, 2002 г.); IV Международной научной конференции «Интеллектуализация обработки информации» (Алушта, июнь, 2002 г.); Международной научной конференции “On Problems of Decision Making and Control under Uncertainties” (Алушта, сентябрь, 2003 г.); 11-й Всероссийской конференции “Математические методы распознавания образов” (Пущино, ноябрь, 2003 г.); на научных конференциях профессорско-преподавательского состава факультета математики и информатики Таврического национального

университета им. В.И. Вернадского (2002, 2003 гг.); на научном семинаре кафедры информатики Таврического национального университета им. В.И. Вернадского.

**Публикации**. Перечисленные результаты отражены в 11 публикациях, среди которых 3 статьи в научных изданиях из списка научных квалификационных изданий Украины, утвержденных ВАК Украины; 5 статей в научных журналах и сборниках научных работ; 3 публикации в тезисах научных конференций.

**Структура и объем работы**. Содержание работы изложено в рукописи, состоящей из 108 страниц, 12 рисунков, 8 таблиц.

Диссертационная работа включает введение, четыре раздела и заключение.

В разделе 1 диссертационной работы приводится постановка задачи обучения распознаванию по прецедентам; обращается внимание на актуальность и полезность использования решающих деревьев в задаче обучения распознаванию по прецедентам; рассматриваются основные свойства бинарных решающих деревьев; приводится обзор результатов, связанных с современными подходами к синтезу, редукции и другим моделям коррекции решающих деревьев. На основе выводов по данному разделу формируются цели и задачи исследования, направленные на разрешение дилеммы между точной настройкой на начальную обучающую информацию и ограничением сложности решающих деревьев, обеспечивая возможность обобщения свойств обучающей информации с гарантированной точностью.

В разделе 2 диссертационной работы на основании вероятностного подхода к оцениванию эмпирических закономерностей обоснован процесс редукции бинарных решающих деревьев, введен функционал качества, позволяющий оценивать случай, когда решение принимается “некомпетентной” ветвью – ветвью решающего дерева максимального ранга. Поставлена задача минимизации функционала качества и доказано, что минимум функционала качества достигается на равномерных деревьях, т.е. деревьях, ранг ветвей которых отличается не более чем на единицу. В подразделах 2.3.1 и 2.3.2 получены оценки

*VCD* для основных классов решающих функций, представленных решающими деревьями, на основании которых можно утверждать, что оптимизация по числу листьев более обоснована с точки зрения теории Вапника-Червоненкиса и обеспечивает существенно более высокую скорость равномерной сходимости при обучении, чем оптимизация по рангу деревьев.

В разделе 3 диссертационной работы описана новая классифицирующая модель – эмпирический решающий лес (подраздел 3.1); приведен алгоритм синтеза эмпирического решающего леса (подраздел 3.2), а также различные алгоритмы принятия решений эмпирическим решающим лесом (подраздел 3.3) – последовательный алгоритм принятия решений с переходами по “ссылкам”, алгоритм принятия решений на основе наиболее “компетентной” ветви РД эмпирического решающего леса, алгоритм принятия решений на основе “голосования” ветвей РД эмпирического решающего леса; построены непротиворечивые логические описания классов в виде дизъюнктивных нормальных форм по эмпирическому решающему лесу; на основе алгебраического подхода построена модель коррекции *r*-некорректного эмпирического решающего леса (подраздел 3.4); получена оценка *VCD* *r*-редуцированного эмпирического леса (подраздел 3.5), позволяющая сделать вывод о том, что переход от единичного решающего дерева к *r*-редуцированному эмпирическому лесу не изменяет порядка *VCD*. В тоже время обеспечивается коррекция, позволяющая настроиться по обучающей выборке на правильную классификацию как можно большего числа объектов.

В разделе 4 диссертационной работы описывается программная реализация алгоритма синтеза эмпирического решающего леса и алгоритма принятия решений эмпирическим лесом с переходами по “ссылкам”; демонстрируются результаты практического применения алгоритма синтеза эмпирического решающего леса для классификации базы данных патогенных вибрионов и аэромонад, вызывающих желудочно-кишечные заболевания. На основе построенного программного комплекса методом статистического моделирования проводится сравнение характеристик одного решающего дерева и эмпирического

решающего леса, причем существенным является тот факт, что при синтезе эмпирического решающего леса и отдельного решающего дерева используется один и тот же критерий для выбора признаковых предикатов во внутренние вершины деревьев. Проведенные эксперименты достоверно продемонстрировали повышение точности распознавания объектов, не участвовавших ранее в обучении, эмпирическим решающим лесом по сравнению с отдельным решающим деревом.

В заключении подводятся итоги диссертационной работы.

# Раздел 1
# МЕТОДЫ СИНТЕЗА И РЕДУКЦИИ РЕШАЮЩИХ ДЕРЕВЬЕВ: ОБЗОР И ЗАДАЧИ ИССЛЕДОВАНИЯ

## 1.1. Постановка задачи распознавания и основные понятия теории решающих деревьев

*Распознавание образов* (РО) представляет собой раздел кибернетики [38], связанный с моделированием некоторых творческих аспектов мыслительной деятельности человека, таких, в частности, как способность узнавать (классифицировать) предметы и явления окружающего мира, формировать новые понятия.

Методы распознавания образов хорошо зарекомендовали себя при решении сложных прикладных задач, возникающих, прежде всего, в плохо формализованных областях, при наличии трудно выявляемых нетривиальных закономерностей между признаками. К настоящему времени разработано несколько основных направлений в теории распознавания, объединяющих сотни конкретных алгоритмов и методов. В качестве таковых следует отметить перцептронные модели, ведущие свое начало от работ Ф.Розенблатта [38, 42], метод потенциальных функций [38], статистические модели распознавания [3, 42, 70, 161], модели распознавания, основанные на построении кусочно-линейных (или более сложных) разделяющих поверхностей в признаковом пространстве [31, 32, 33, 34], алгоритмы, основанные на построении решающих деревьев [2, 7, 8, 9, 12, 22, 41, 42, 44, 45, 52, 60], структурные (лингвистические) методы [42, 49], модели частичной прецедентности [33, 42], алгебраический подход Ю.И.Журавлева [30, 31, 32, 33], нейросетевые алгоритмы [42], методы, основанные на теории нечетких множеств [42] и др. Основанные на различных идеях, гипотезах и принципах, а также их сочетаниях, эти подходы имеют свои достоинства и недостатки, различные требования к исходным данным,

ограничения на области применения. Однако все они сводятся к поиску полезной информации в пространстве возможных объектов и ее корректному обобщению.

Герберт Симон определил обучение следующим образом. “*Обучение* – это любое изменение в системе, приводящее к улучшению решения задачи при ее повторном предъявлении или к решению другой задачи на основе тех же данных” [42]. Это краткое определение затрагивает множество вопросов, связанных с разработкой обучаемых программ. Обучение обычно подразумевает обобщение на основе накопленного опыта. Поскольку область всевозможных обучающих данных обычно достаточно широка, обучающая система заведомо не может обработать все возможные объекты, и полученный ограниченный выборочный опыт она должна корректно распространить на недостающие объекты. Такая задача *эмпирической индукции* (induction) – обобщения обучающих данных (выборки) – является центральной для обучения и лежит в основе теории распознавания образов. Для большинства задач имеющихся в наличии данных недостаточно, чтобы гарантировать получение в результате обобщения точного решения. Поэтому обучающие системы неизбежно должны обобщать информацию эвристически, т.е. отбирать те аспекты, которые, скорее всего, окажутся полезными в будущем. Такой критерий отбора иногда называют *индуктивным порогом* (inductive bias) [42].

*Суть задачи обучения распознаванию по прецедентам* состоит в следующем [19].

Известно, что некоторое множество $M$ может быть представлено в виде объединения конечного числа собственных подмножеств $K_1, K_2, \ldots, K_\ell$ так что $M = \bigcup_{j=1}^{\ell} K_j$. Эти подмножества называются *классами*, а элементы множества $M$ – *допустимыми объектами*. Допустимые объекты, информация о которых получена на основе опыта, называют *прецедентами*. Точной информации о классах нет: для них неизвестны ни аналитическое описание, ни алгоритмический способ проверки принадлежности произвольного допустимого объекта некоторому классу. Заданы лишь конечные подмножества

$M_1 \subset K_1, M_2 \subset K_2, \ldots, M_\ell \subset K_\ell$, для которых известно точно: если допустимый объект $x$ принадлежит подмножеству $M_j$, то он принадлежит классу $K_j$ $\left(j = \overline{1,\ell}\right)$. Можно сказать, что предикаты "$x \in K_j$" частично заданы на множестве $\{M_1 \cup \ldots \cup M_\ell\} = M_0 \subset M$.

Требуется, используя только указанное множество $M_0$, называемое *обучающим*, найти правило распознавания, позволяющее для любого допустимого объекта $x \in M$ вычислить предикаты "$x \in K_j$", $j = \overline{1,\ell}$. Сложность данной задачи заключается в том, что, располагая ограниченной информацией, необходимо её экстраполировать, а сделать это можно различными способами.

Если $K_j \cap K_q = \varnothing$ для $1 \le j < q \le \ell$, то говорят, что поставленная задача является *задачей обучения распознаванию с непересекающимися классами*. В противном случае, если найдутся такие $j$ и $q$, что $K_j \cap K_q \neq \varnothing$, речь идет о *задаче с пересекающимися классами*. Случай пересекающихся классов может быть сведен к случаю непересекающихся классов путем выделения областей пересечения в “новые” классы. Очевидно, что для задач с непересекающимися классами, экстраполирующее правило определяется некоторым разбиением признакового пространства на $\ell$ областей.

При решении задачи распознавания образов по прецедентам используется следующая эмпирическая гипотеза [36]: считается, что при выборе объектов подмножества $M_0$ из множества $M$ не делается предпочтения одного объекта другому, т.е. объекты подмножества $M_0$ из генеральной совокупности $M$ выбираются *случайным образом*.

*В геометрической интерпретации* задача обучения распознаванию по прецедентам имеет следующий смысл. Пусть выбрано некоторое признаковое пространство $X^n$ размерности *n,* каждый реальный объект заменяется своей векторной моделью, т.е. описанием в виде последовательности значений признаков. Естественно называть модель объектов *точкой* в многомерном пространстве. Точки из множеств $M_1, \ldots, M_\ell$, образующие в пространстве

признаков $X^n$ конечные множества, должны быть разделены некоторым набором гиперповерхностей так, чтобы в любом полученном в результате разделения $X^n$ подмножестве содержались точки только одного класса из обучающего множества $M_0$.

Совокупность векторов-объектов из обучающего множества $M_0$ может быть записана в виде таблицы $T_{mn\ell}$, называемой *стандартной таблицей обучения*, где $m$ – число объектов обучающего множества; $n$ – размерность признакового пространства (число признаковых предикатов (признаков)); $\ell$ - число классов. В таблице обучения $T_{mn\ell}$ отдельно выделен столбец с номером $n$+1, называемый *целевым*, содержащий метки классов, которым принадлежат объекты обучающего множества. Обычно считается, что задана непротиворечивая таблица обучения, т.е. в ней нет ни одной пары одинаковых объектов с указанной принадлежностью к разным классам.

В настоящей диссертационной работе к обучающей информации – таблице $T_{mn\ell}$ - предъявляется дополнительное условие: значения предикатов "$x \in K_j$", $j = \overline{1,\ell}$, являются *достоверными* на множестве $T_{mn\ell}$. В дальнейшем всюду таблицы обучения $T_{mn\ell}$ полагаются *булевыми*: если $(x_1,\ldots,x_j,\ldots,x_n) = \tilde{x} \in T_{mn\ell}$, то $x_i \in \{0,1\}$, $i = \overline{1,n}$.

Задача обучения распознаванию образов по прецедентам состоит из двух этапов: обучение по эмпирическим данным и собственно распознавание. Традиционно обучение проводят не на всех объектах непротиворечивой таблицы обучения, а лишь на части; другая часть объектов составляет *контрольное множество* (проверочное множество). Результатом процесса обучения является *эмпирическая закономерность*, представленная в виде решающего правила (РП) - алгоритма, позволяющего “вычислять” метку класса для любого допустимого объекта. Подчеркнем, что задача обучения распознаванию образов по прецедентам, входящая в круг важнейших задач теоретической и прикладной информатики, предполагает алгоритмическую реализуемость, вычислимость

решающих правил. На основе объектов контрольного множества осуществляют контроль качества построенного РП посредством вычисления коэффициента ошибок (частоты ошибок) как отношения числа правильно “распознанных” объектов контрольного множества к их общему числу.

В диссертационной работе рассматриваются алгоритмы построения решающих правил распознавания с одновременным выбором информативных признаков, представимые как в виде отдельных решающих деревьев, так и в виде совокупности решающих деревьев – эмпирического решающего леса.

*Привлекательность решающих деревьев* определяется возможностями:

- построения иерархических классификаторов, определяющих структурные закономерности в данных;
- синтеза наборов эмпирических закономерностей в виде конъюнкций и продукций;
- простой организации процедуры принятия решений;
- параметрической оптимизации, основанной на теоретико-множественной вложенности классов решающих правил, определяемых деревьями с меньшим допустимым числом листьев, в классы решающих правил, определяемых деревьями с большим допустимым числом листьев.

*Привлекательность эмпирического решающего леса*, подробно изученного в диссертационной работе, обосновывается возможностью выявления бóльшего, чем в случае построения одного РД, числа закономерностей в данных; максимального использования всей начальной обучающей информации; незначительным усложнением процедуры принятия решений; настройки на правильную классификацию как можно большего числа обучающих объектов, а также незначительным усложнением класса решающих правил, порождаемых эмпирическим решающим лесом по сравнению с классом РП, порождаемых одним решающим деревом.

Для дальнейшего изложения понадобятся следующие основные понятия и факты.

**Определение 1.1** [19]**.** *Деревом* называется конечный связный ациклический граф с выделенной вершиной, называемой *корневой* вершиной (*корнем*) дерева.

Конструктивно дерево можно определить рекуррентно следующим образом [1].

1. Единственная вершина является деревом; эта же вершина является корнем этого дерева.
2. Пусть $n$ – это вершина, а $T_1, T_2, \ldots, T_k$ - деревья с корнями $n_1, n_2, \ldots, n_k$ соответственно. Можно построить новое дерево, сделав $n$ “родителем” вершин $n_1, n_2, \ldots, n_k$. В этом дереве $n$ будет корнем, а $T_1, T_2, \ldots, T_k$ - поддеревьями этого корня. Вершины $n_1, n_2, \ldots, n_k$ называются “*сыновьями” вершины n*.

*Путем из вершины* $n_1$ *в вершину* $n_k$ называется последовательность вершин $n_1, n_2, \ldots, n_k$, где для всех $i$, $1 \le i \le k$, вершина $n_i$ является “родителем” вершины $n_{i+1}$. *Длиной пути* называется число, на единицу меньшее числа вершин, составляющих этот путь. *Глубина вершины* определяется как длина пути (он единственный) от корня до этой вершины.

**Определение 1.2** [19]. *Бинарным* (ориентированным) *деревом* (БД) называется дерево, имеющее следующие свойства:

- в корневую вершину не входит ни одна дуга;
- любая другая вершина имеет в точности одну входящую дугу и только либо две, либо ни одной выходящей дуги.

Вершины бинарного решающего дерева, имеющие две выходящие дуги, называются *внутренними* или *нетерминальными вершинами*, а остальные вершины – *терминальными или листьями*.

**Определение 1.3** [19, 61]. *Решающим* называется бинарное дерево (БРД), обладающее следующими свойствами:

- каждая внутренняя вершина помечена одним из признаковых предикатов, определяемых заданной таблицей обучения $T_{mn\ell}$ (вид признакового предиката определяется его областью допустимых значений);

- две дуги, выходящие из внутренней вершины помечены значениями, принимаемыми предикатом в вершине, из которой они выходят;

- концевые вершины (листья) помечены метками классов $\omega_1, \omega_2, \ldots, \omega_\ell$;

- ни в одной ветви дерева нет двух одинаковых вершин.

Легко видеть, что БРД представляет собой алгоритмическое отображение $A_{BDT} : B^n \to \{\omega_1, \omega_2, \ldots, \omega_\ell\}$, где $B^n = \{0,1\}^n$ – множество вершин единичного $n$-мерного куба, $\{\omega_1, \omega_2, \ldots, \omega_\ell\}$ - множество меток классов, образующих разбиение множества $M$.

На рисунке 1.1 для примера представлено бинарное решающее дерево с указанием решающих правил для каждого из классов $\omega_1$ и $\omega_2$.

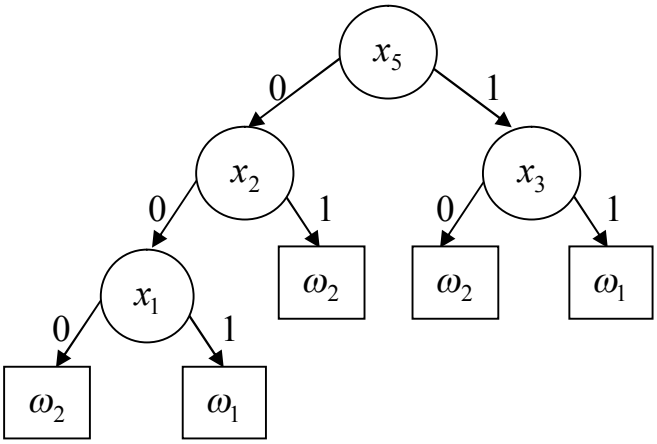


Рис 1.1. Бинарное решающее дерево, задающее решающие правила для классов $\omega_1 : x_1\bar{x}_2\bar{x}_5 \vee x_3x_5$ и $\omega_2 : \bar{x}_1\bar{x}_2\bar{x}_5 \vee x_2\bar{x}_5 \vee \bar{x}_3x_5$

**Определение 1.4** [19]. Решающее дерево, не совершающее ни одной ошибки на $T_{mn\ell}$, называется *корректным решающим деревом относительно таблицы обучения*, в противном случае – *некорректным решающим деревом*.

Если $T_{mn\ell}$ - *корректная обучающая информация* (не содержит одинаковых объектов с разными метками классов), то по ней всегда может быть построено, и вообще говоря, не единственное, корректное БРД. Таким образом, при построении БРД по стандартной таблице обучения следует различать два типа

задач: построение БРД по корректной таблице обучения и построение БРД по некорректной таблице обучения. Для корректных достоверных таблиц $T_{mn\ell}$ любое некорректное на обучающей информации БРД заведомо не способно к безошибочной классификации произвольных объектов – булевых наборов.

В рамках концепции корректных и достоверных обучающих таблиц (*КД-таблиц*) или, будем говорить, КД-подхода, с теоретической точки зрения ошибки в $T_{mn\ell}$ исключены. Но на практике может оказаться, что начальная информация не является КД-таблицей. Возможен следующий *эвристический прием изучения начальной информации* $T_{mn\ell}$.

Для любого булевого набора $\tilde{\alpha} \in T_{mn\ell}$, принадлежащего классу $\omega(\tilde{\alpha})$, вычисляются числа $k_1(\tilde{\alpha})$ наборов $\tilde{\beta} \in T_{mn\ell}$ таких, что $\rho(\tilde{\alpha}, \tilde{\beta}) \le \theta$, $1 \le \theta \le n$, и $\omega(\tilde{\beta}) = \omega(\tilde{\alpha})$, и $k_2(\tilde{\alpha})$ наборов $\tilde{\beta} \in T_{mn\ell}$ таких, что $\rho(\tilde{\alpha}, \tilde{\beta}) \le \theta$ и $\omega(\tilde{\beta}) \ne \omega(\tilde{\alpha})$. Здесь $\rho(\cdot,\cdot)$ - расстояние Хэмминга, $\theta$ - параметр. Обозначим $k_\Delta(\tilde{\alpha}) = k_2(\tilde{\alpha}) - k_1(\tilde{\alpha})$. Пусть $\tilde{\alpha}^* = \underset{(\tilde{\alpha} \in T_{mn\ell}) \wedge (k_\Delta(\tilde{\alpha}) > 0)}{\arg\max} k_\Delta(\tilde{\alpha})$. Тогда шару с центром $\tilde{\alpha}^*$ радиуса $\theta$ (окрестности точки $\tilde{\alpha}^*$) соответствует наибольшее значение $k_\Delta(\tilde{\alpha}^*)$ по сравнению с другими шарами радиуса $\theta$ с центрами $\tilde{\alpha} \in T_{mn\ell}$. В таком случае $\tilde{\alpha}^*$ объявляется "*подозрительной*" *точкой* или *кандидатом на исключение из* $T_{mn\ell}$. Осуществляется синтез экстремального БРД по информации $T_{mn\ell}$ (полагаем, построено экстремальное $\mu_1$-БРД $T_{\mu_1}$) и по информации $T_{mn\ell} \setminus \{\tilde{\alpha}^*\}$ (полагаем, построено экстремальное $\mu_2$-БРД $T_{\mu_2}$). Чем больше величина $\mu_1 - \mu_2$, тем больше оснований исключить $\tilde{\alpha}^*$ из $T_{mn\ell}$, получая лучшие статистические оценки надежности БРД $T_{\mu_2}$.

Для возможности построения корректного на обучающей информации БРД, использующего только признаки с номерами $i_1, i_2, \ldots, i_s$, необходимо и достаточно, чтобы множество $\{i_1, i_2, \ldots, i_s\}$ было тестом таблицы $T_{mn\ell}$ [45]. Известно, что для почти всех таблиц при $n \to \infty$, $m \to \infty$ и условии $\lim\limits_{n\to\infty} m / 2^{n/2} = 0$, для любого

положительного $\varepsilon = o(1)$ любые $2(1+\varepsilon)\log_2 m$ столбцов таблицы образуют тест [12, 46]. Следовательно, *для широкого класса произвольных булевых таблиц при синтезе БРД можно получить набор корректных деревьев, использующих полностью или частично разные переменные*.

Вышеуказанные доводы выступают в пользу синтеза совокупности эмпирических БРД (эмпирического решающего леса) для принятия решений. Это обеспечивает, как будет показано ниже, повышение точности распознавания и расширяет возможности применения набора БРД в случае наличия большого числа пропусков в информации, поступающей для принятия решения по синтезированному набору деревьев. Каждое отдельное дерево леса можно рассматривать в качестве эксперта, способного "судить" (делать вывод) *по разным признакам*, т.е. в своей *области* "*компетентности*".

Деревья, отличающиеся от бинарных, могут иметь более двух ребер, выходящих из некоторой внутренней вершины. Рассмотрим вкратце класс деревьев, называемый *k*-решающими.

Обозначим через $E_k$ множество $\{0,1,\ldots,k-1\}$, $k>2$. Функция $f = f(x_1, x_2, \ldots, x_n)$ называется *функцией k-значной логики*, если на всяком наборе $\tilde{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ значений переменных $x_1, x_2, \ldots, x_n$, где $\alpha_i \in E_k$, значение $f(\tilde{\alpha})$ также принадлежит множеству $E_k$.

**Определение 1.5.** Решающие деревья, у которых из каждой вершины выходят не более *k* ребер, реализующие алгоритмические отображения вида $A_{k,\ell} = \left\{ f : \underbrace{E_k \times \ldots \times E_k}_{n} \to \{0,1,\ldots,(\ell-1)\} \right\}$ называются *k-решающими деревьями* (*k*-РД).

В [19] была обоснована полнота класса функций алгебры логики, представимых в виде бинарного решающего дерева (БРД). Обобщим этот результат и покажем, что имеет место полнота класса функций *k*-значной логики, представимых в виде *k*-РД.

**Теорема 1.1.** Любое алгоритмическое отображение из класса $A_{k,\ell}$ может быть построено в виде *k*-РД.

Доказательство. Пусть $f \in A_{k,\ell}$. Прямой проверкой легко убедиться в справедливости разложения $f$ по одной (любой) переменной:

$$f(x_1,\ldots,x_{i-1},x_i,x_{i+1},\ldots,x_n) = \bigvee_{\{\sigma\in E_k\}} I_\sigma^*(x_i)\,\&\, f(x_1,\ldots,x_{i-1},\sigma,x_{i+1},\ldots,x_n) \quad (1.1)$$

Здесь $\alpha\vee\beta=\max\{\alpha,\beta\}, \alpha\,\&\,\beta=\min\{\alpha,\beta\}$, $I_\sigma^*(x)=\begin{cases}0, & x\neq\sigma,\\ \max\{(k-1),(\ell-1)\}, & x=\sigma.\end{cases}$

На первом шаге построения алгоритмического отображения $f$ реализуется корневая вершина дерева, представленная на рисунке 1.2.

Если хотя бы одна из функций $f_\sigma, \sigma\in E_k$, полученных после построения корневой вершины и не зависящих от $x_i$, где $f(x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n)=f(x_1,\ldots,x_{i-1},\sigma,x_{i+1},\ldots,x_n)$, не является константой, то к ней, как к функции $(n-1)$-ой переменной, снова применяется разложение вида (1.1), определяющее следующий шаг построения *k*-РД.

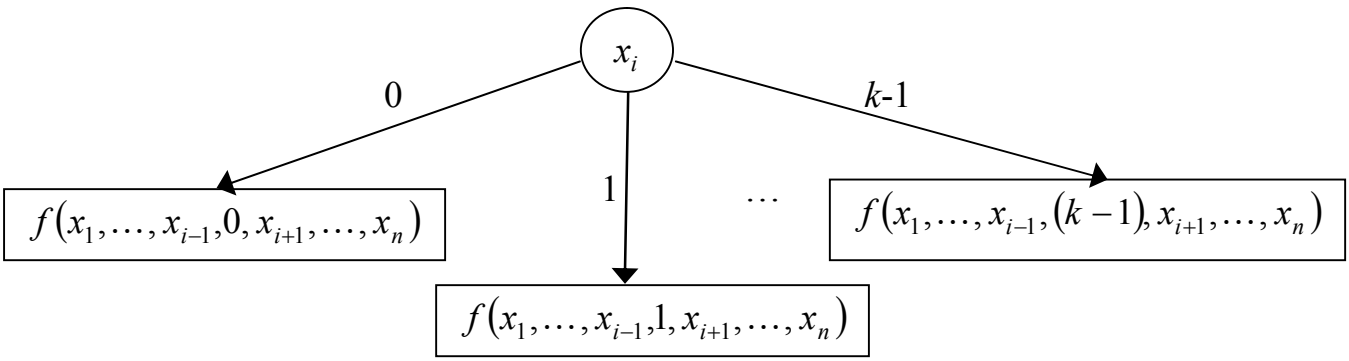


Рис.1.2. Первый шаг ветвления. Из корневой вершины выходят *k* ребер.

Если же для некоторого $\sigma\in E_k$ выполняется $f(x_1,\ldots,x_{i-1},\sigma,x_{i+1},\ldots,x_n)=\gamma=const$, т.е. оставшиеся незафиксированными переменные не являются существенными, то лист дерева, соответствующий ребру $x_i=\sigma$, становится терминальным и помечается константой $\gamma$ □.

Пусть "$x_i=\sigma$", где $\sigma\in E_k$ - набор из $nk$ предикатов, $i\in\{1,2,\ldots,n\}$. Используя такие предикаты, можно любое *k*-РД преобразовать в эквивалентное

(реализующее то же самое алгоритмическое отображение) БРД. Для этого следует каждую внутреннюю вершину $k$-РД заменить бинарным поддеревом, показанным на рисунке 1.3.

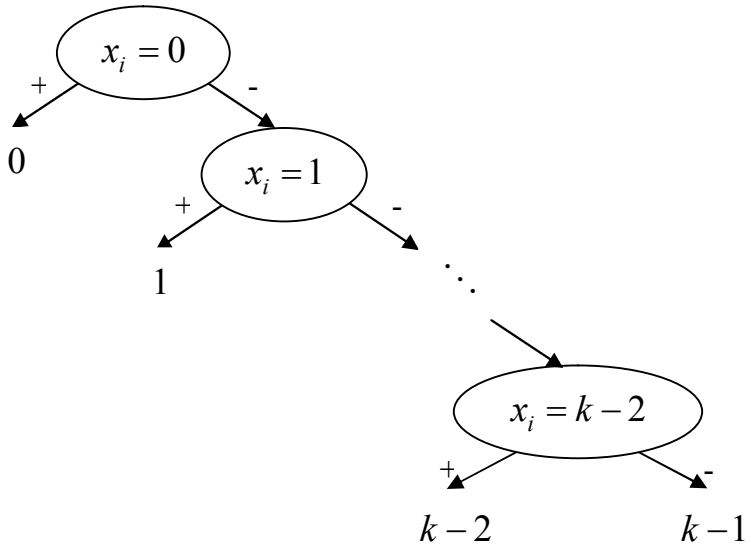


Рис.1.3. Бинарное поддерево, заменяющее внутренние вершины $k$-РД.

Следует заметить, что подсчет числа различных бинарных решающих деревьев представляет интерес в связи с получением оценок статистической надежности решающих правил, основанных на построении деревьев, и изучением алгоритмов синтеза БРД с наименьшим числом листьев [13]. Однако, точная формула для нахождения числа БРД $d(n,k,\mu)$ с $n$ внутренними вершинами, $\mu$ листьями и $k$-значными метками листьев неизвестна. Известна лишь асимптотическая оценка

**Теорема 1.2** [13]. $d(n,k,\mu) \sim (\mu-1)!\,(k(k-1))^{\mu-1} n(n-1)^{\mu-2}$ при $n \to \infty$.

Для дальнейшего изложения понадобится следующее неравенство [13]. Число булевых функций от $n$ переменных $b(n,2,\mu)$, представимых в виде БРД с ровно $\mu$ листьями, удовлетворяет неравенству

$$b(n,2,\mu) < (\mu-1)!\,2^{\mu-1} n^{\mu-1} \qquad (1.2)$$

Отметим некоторые важные *свойства бинарных решающих деревьев*:

1. БРД с ограниченным (и небольшим) числом листьев определяют для случая двухэлементных (0,1) решений чрезвычайно узкий класс булевых функций,

асимптотически (при числе аргументов $n \to \infty$) сколь угодно узкий по сравнению даже с классом линейных булевых функций $L(n) \subset P_2(n)$ [13];

2. Как отмечалось выше, любая булева функция из $P_2(n)$ может быть представлена в виде БРД [19];
3. Если $\mu$ - число листьев, то для класса $D(n, \mu)$ булевых функций, представимых в виде БРД не более чем с $\mu$ листьями, при условии $2 \le \mu \le 2^n$ ($n$-число признаков) справедливо включение $D(n, \mu) \subset D(n, \mu + 1)$ [15].

Свойства 1,2,3 и положения статистической теории обучения [3] обосновывают *возможность оптимизационного синтеза БРД*, корректного на непротиворечивой таблице обучения $T_{mn\ell}$, путем минимизации параметра $\mu$.

4. Синтез БРД с минимальным числом листьев $\mu$ по непротиворечивой таблице обучения $T_{mn\ell}$ является сложной экстремальной задачей из класса *NPC* [19, 96];
5. Построенное БРД с $\mu$ листьями далее позволяет со сложностью $O(n)$ получить логическое описание синтезированных правил для распознавания классов объектов в виде дизъюнктивных нормальных форм (ДНФ) [19]. Конъюнкции, входящие в эти ДНФ, являются эмпирическими закономерностями и могут быть использованы, кроме прочего, для синтеза эмпирических продукций, пополняющих базы знаний.

Свойства 3,4 обосновывают целесообразность построения эвристических алгоритмов синтеза БРД, близких к оптимальным, и оправдывают усилия разработчиков интеллектуализированного программного обеспечения, настойчиво проявляемые в этом направлении.

### 1.2. Задача синтеза решающих деревьев. Выбор критерия ветвления

Алгоритмы синтеза решающих деревьев по непротиворечивой таблице обучения $T_{mn\ell}$ обычно состоят из следующих основных этапов: выбор

признакового предиката во внутреннюю вершину РД согласно некоторой эвристике, т.е. выбор критерия ветвления, выбор условия для прекращения ветвления – получение терминальных вершин, т.е. выбор критерия остановки, и определение метки класса, приписываемой терминальной вершине.

Различные эвристики, используемые в алгоритмах синтеза РД, зависят от особенностей информации, представленной в непротиворечивой таблице обучения, прежде всего, от шкалы [36, 41], в которой измеряются значения признаков; от наличия пропусков в данных; от числа анализируемых объектов по отношению к числу признаков таблицы обучения, от наличия "шума" в начальной информации и.т.п. В случае бинарных признаков, задача синтеза РД в случае непересекающихся классов – это задача построения разбиения куба $B^n$ на интервалы наименьшего ранга или наибольшей размерности, покрывающие как можно больше точек одного и того же класса. Напомним, что подмножество $N_K \subseteq B^n$ называется *интервалом r-го ранга*, если оно соответствует элементарной конъюнкции $K$ $r$-го ранга, $N_k = \{\tilde{x} \in B^n \mid K(\tilde{x}) = 1\}$. В случае вещественных (непрерывных) признаков, задача синтеза РД обычно сводится к построению разбиения признакового пространства $R^n$ поверхностями достаточно простого вида [34].

Синтез разбиения признакового пространства интервалами максимальной размерности приводит к задаче построения оптимального решающего дерева с минимальным числом листьев. Исследованию этой задачи посвящены работы [15, 19, 96], в которых обосновывается NP полнота задачи поиска корректного БРД с минимальным числом листьев по непротиворечивой таблице обучения. Доказательство Л.Хьяфила и Р. Ривеста основано на полиномиальной сводимости задачи о точном покрытии, для которой установлена NP-полнота, к изучаемой задаче [96]. Доказательство этого же факта Донским В.И. основано на полиномиальной сводимости задачи целочисленного 0-1-линейного программирования общего вида к задаче поиска РД с минимальным числом листьев [15]. В работе Zantema H., Bodlaender H. [160] обоснована NP полнота

задачи поиска РД с минимальным числом листьев, дающего эквивалентные заданному РД решения. Заметим, что два решающих дерева дают эквивалентные решения, если они реализуют одну и ту же функцию, т.е. “вычисляют” одну и ту же метку класса для каждого допустимого объекта.

Задача синтеза РД непосредственно связана с задачей выбора информативной подсистемы признаков, состоящей в указании части признаков из непротиворечивой таблицы обучения, в пространстве которых заданные множества объектов, представляющие разные классы, разделяются достаточно просто [36, 38]. Синтезированные по обучающей информации РД определяют набор *эмпирических закономерностей*. В общем случае сформулировать понятие закономерности представляется довольно сложным. Загоруйко Н.Г., Ёлкина В.Н., Лбов Г.С. в работе [36] под *закономерностью* понимают “…устойчивое формализованное правило, фиксирующее зависимость между различными частями таблицы”.

В связи с важностью критериев отбора признаков при ветвлении представляет интерес работа Ю.И. Журавлева, в которой предлагается оценивать информативность признаков с помощью тупиковых тестов непротиворечивой таблицы обучения. При этом наиболее информативным считается тот признак, который вошел в большее число минимальных достаточных наборов признаков (тупиковых тестов). В [12, 15] было доказано, что используя тупиковый тест бинарной таблицы обучения, можно построить РД с числом листьев $\mu$: $k \le \mu \le 2^r$, где $k$ – число классов, $r$ – число признаков, образующих тупиковый тест. Тогда можно предложить тривиальный переборный алгоритм синтеза оптимального РД: сначала надо построить все тупиковые тесты таблицы обучения, затем на этих тупиковых тестах построить РД и среди построенных деревьев выбрать оптимальное (с наименьшим числом листьев). Однако, реализация такого алгоритма весьма затруднительна, т.к. эффективных алгоритмов построения всех тупиковых тестов нет, а построение всех тупиковых тестов – сложная переборная задача [46]. Известно, что среди РД с одинаковым числом листьев предпочтительнее то, в котором наборы из обучающей выборки

распределены равномерно по интервалам, соответствующим конъюнкциям решающей булевой функции [19].

Остается открытым вопрос о том, насколько один признак, выбранный в вершину РД, обеспечивает “лучшее” разбиение признакового пространства, чем другой. Для дальнейшего изложения понадобятся следующие определения.

Пусть для разбиения интервала $N_t$ на некотором шаге выбирается признак $x_k$, $N_t = N_t^{Right} \cup N_t^{Left}$, $N_t^{Right} \cap N_t^{Left} = \varnothing$.

Обозначим $A(k) = \{\tilde{a} \in T_{mn\ell} : \tilde{a} \in N_t^{Right}\}$; $B(k) = \{\tilde{a} \in T_{mn\ell} : \tilde{a} \in N_t^{Left}\}$.

**Определение 1.6** [19]. Будем говорить, что признак $x_k$ обладает *свойством полной отделимости*, если множество $A(k)$ содержит объекты только одного класса; множество $B(k)$ содержит объекты только одного класса и классы объектов из $A(k)$ и $B(k)$ различны.

**Определение 1.7** [19]. Будем говорить, что признак $x_k$ *обладает свойством частичной отделимости*, если множество $A(k)$ или множество $B(k)$ содержит объекты только одного класса.

Значительная часть эвристических алгоритмов синтеза РД отдают предпочтение в первую очередь признакам, обладающим свойством полной или частичной отделимости. Предполагается, что такие признаки определяют решающее правило, которое будет допускать минимальное число ошибок на объектах контрольной выборки [61, 68].

В зависимости от выбора алгоритма разбиения, по имеющемуся набору обучающих объектов, может быть построено некоторое множество решающих деревьев, позволяющих корректно классифицировать эти объекты. В этом случае следует выбрать РД, которое с наибольшей вероятностью позволит корректно классифицировать объекты, не участвовавшие в обучении. Обычно, таким деревом считают в некотором смысле “простейшее” РД, безошибочное на всех обучающих объектах. В основу такого предположения положена проверенная временем эвристика, согласно которой *предпочтение отдается простоте без дополнительных ограничений*. Этот принцип впервые был сформулирован в 1324

году философом-схоластом Вильямом из Оккама (William of Occam) и получил название “бритвы Оккама” (Occam’s Razor). “Глупо прилагать больше усилий, чем нужно для достижения цели… Не стоит приумножать сущности сверх необходимого”. Более современная версия этого принципа сводится к выбору простейшего ответа, соответствующего исходным данным [42, 56, 120].

Если объекты описаны произвольными наборами значений переменных (булевыми, вещественными, качественными), иначе говоря, являются точками некоторого произвольного допустимого множества $X$, то для построения БРД используется система признаковых предикатов $\{P_j : X \to \{0,1\},\ j = \overline{1,n}\}$. Некоторые признаковые предикаты могут быть заданы при постановке задачи, другие – найдены путём использования статистических или метрических методов анализа начальных данных [44, 45].

Различные эвристические методы построения решающих деревьев отличаются друг от друга по типу признаковых предикатов, рассматриваемых во внутренних вершинах решающего дерева. Известны следующие основные типы признаковых предикатов, используемых в алгоритмах синтеза РД [45]:

1) простейшие признаковые предикаты по порогу(-ам) одного признака, представленные в виде неравенств (обычно используются для вещественных признаков);
2) признаковые предикаты типа проверки принадлежности значения признака к некоторому множеству его значений (обычно используются для дискретных признаков);
3) признаковые предикаты типа проверки принадлежности объекта локальной области, полученной в результате разбиения признакового пространства;
4) признаковые предикаты, использующие линейную дискриминантную функцию;
5) признаковые предикаты, использующие квадратическую дискриминантную функцию.

Приведем обзор наиболее известных алгоритмов синтеза решающих деревьев.

Один из таких широко известных алгоритмов синтеза РД – *ID3 алгоритм* – предложен Р. Куинланом. Он включает в себя метод отбора множества объектов, по которым строится РД, и тест на статистическую независимость выбираемых признаков. Куинлан [8, 9, 10, 42, 45, 63, 68, 71] предложил использовать теоретико-информационную меру (Е-меру), основанную на энтропии для оценивания меры неопределенности в классификации, возникающей при использовании рассматриваемого признака во внутренней вершине РД. Полагается, что наибольшую классифицирующую силу имеет признак с наименьшей Е-мерой.

*Алгоритм ID3* построения решающего дерева можно описать следующим образом:

1) если все объекты обучающего множества точно из одного класса, то решающее дерево есть лист, содержащий в качестве ответа метку этого класса;
2) в противном случае:
   a) определить признак $x_{best}$, обладающий минимальной Е-оценкой по сравнению с другими признаками;
   b) для каждого значение $v_{best,j}$ признака $x_{best}$ построить ветвь от $x_{best}$ к решающему поддереву, построенному рекурсивно на всех обучающих объектах со значением $v_{best,j}$ признака $x_{best}$.

Пусть далее $X = \{x_1, x_2, \ldots, x_n\}$ – множество признаков, используемых для описания прецедентов, для каждого признака $x_i$ множество его возможных значений обозначим $V_i$; $v_{ij}$ - индивидуальные значения признака $x_i$, $i = 1,2,\ldots,n$, $j = 1,2,\ldots,|V_i|$. *E*-оценка – это результат определения информационной функции *E* признака в вершине.

Для данной вершины в случае двух классов имеем: $p$ - число объектов первого класса; $q$ - число объектов второго класса; $p_{ij}$ - число объектов первого класса со значением $v_{ij}$ признака $x_i$; $q_{ij}$ - число объектов второго класса со значением $v_{ij}$ признака $x_i$. Тогда $E(x_i)=\sum_{j=1}^{|V_i|}\frac{p_{ij}+q_{ij}}{p+q}I(p_{ij},q_{ij})$, где энтропия $I(\cdot,\cdot)$, как обычно в теории информации [39, 50, 153], определяется выражением:

$$I(a,b)=\begin{cases}0, & a=0;\\ 0, & b=0;\\ -\frac{a}{a+b}\log_2\frac{a}{a+b}-\frac{b}{a+b}\log_2\frac{b}{a+b}, & \textit{иначе.}\end{cases}$$

Выбор признака с минимальной E-оценкой равносилен выбору признака с максимальным приростом информации, определяемым как $I(p,q)-E(x_i)$, $i=1,2,\ldots,n$.

Алгоритм ID3 позволяет построить корректное относительно начальной информации решающее дерево, на основе которого каждый объект из исходного непротиворечивого обучающего множества будет классифицироваться правильно [9, 10, 47].

Среди недостатков ID3 алгоритма можно выделить следующие:

- E-мера предоставляет при построении дерева преимущество признакам с бóльшим числом возможных значений. Отметим, что в работах Гупала А.М., Цветкова А.М. [9,10] приведен улучшенный вариант E-меры – мера отношения, устраняющая этот недостаток;
- наличие “горизонтального” эффекта (horizon effect) – алгоритм выбирает признаки, основываясь на “локальной” мере, т.е. осуществляет на каждом шаге синтеза “локальное” разбиение подмножества объектов, которые “попали” во внутреннюю вершину РД;

- относится к числу GREEDY алгоритмов, т.е. в случае выбора “неоптимального” признака не способен осуществить возврат на уровень вверх с целью замены “неоптимального” признака.

Для справедливости нужно заметить, что указанные недостатки характеризуют бóльшую часть эвристических алгоритмов синтеза РД.

Известно семейство алгоритмов ID4, ID5, ID5R, также использующих энтропийный критерий, создание которых было направлено на устранение недостатков алгоритма ID3 и повышение его эффективности. Так, алгоритм ID4 строит решающее дерево и модифицирует его по мере того, как становится доступным новый объект. ID5 и ID5R устраняют недостатки алгоритма ID4. Существенное отличие между ними заключается в правиле перестройки решающего дерева. В ID5 и ID5R признак с наименьшей энтропийной мерой поднимается в вершину, а структура РД ниже этой вершины сохраняется, в отличие от алгоритма ID4, в котором все поддеревья ниже этой вершины отбрасываются, а дальнейшее ветвление осуществляется на основании значений признака, помещенного в эту вершину. ID5R в отличие от ID5 модифицирует поддеревья рекурсивно.

В автореферате диссертации Цветкова А.М. [48] приведены оценки сложности этих алгоритмов, представленные в таблице 1.1, где $n$ - число признаков, $b$-максимальное количество значений признаков, $m$ - число объектов из обучающего множества.

Таблица 1.1. Оценки сложности алгоритмов семейства ID

| Алгоритм | Операции | Вычисление E-меры |
|---|---|---|
| ID3 | $O\left(mn^2\right)$ | $O\left(b^n\right)$ |
| ID5R | $O\left(m \cdot n \cdot b^n\right)$ | $O\left(m \cdot b^n\right)$ |
| ID5 | $O\left(m \cdot b^n\right)$ | $O\left(m \cdot n^2\right)$ |

ID3 – алгоритм не позволял обрабатывать большие массивы информации, характеризующиеся противоречивыми данными, т.е. при наличии в таблице двух

и более одинаковых объектов с указанной принадлежностью к разным классам; пропусками в данных (отсутствие отдельных значений признаков, возможно, из-за очень высокой стоимости их получения); признаками, принимающими значения из непрерывного интервала (такие данные подлежат дискретизации и последующей группировке). Решение этих проблем привело к созданию нового поколения алгоритмов обучения, основанных на построении решающих деревьев. Наиболее известным из них является алгоритм C4.5 [93, 106, 129, 131]. Он использует в качестве критерия ветвления *оценку прироста* – теоретико-информационную меру, которая оценивает пригодность признака по относительной величине прироста информации. Во внутреннюю вершину РД выбирается признак с максимальной оценкой прироста.

Куинланом [131] предложен также подход, позволяющий использовать обучающие множества большого объема, называемый далее методом “окна” (windowing method). “Окном” называется произвольное подмножество объектов всего обучающего множества. Суть этого метода состоит в произвольном выборе “окна” из обучающего множества и построении по нему решающего дерева. В случае если “окно” правильно расклассифицирует оставшиеся объекты обучающего множества, работа метода завершается. Иначе, в “окно” добавляются неправильно расклассифицированные объекты обучающего множества, и по ним снова строится РД. Процесс продолжается до полного исчерпания неправильно расклассифицированных объектов.

В работах [45, 61, 109, 110, 138] описан эвристический критерий разбиения, используемый в методе CART (Classification and Regression Trees). В случае двух классов используется критерий $F^{(CART1)} = \max\left\{\hat{p}(t)\cdot i(t) - \left[\hat{p}\left(t_{Left}\right)\cdot i\left(t_{Left}\right) + \hat{p}\left(t_{Right}\right)\cdot i\left(t_{Right}\right)\right]\right\}$, а для многоклассовой задачи – $F^{(CART2)} = \max\left\{\hat{p}(t)\cdot i_G(t) - \left[\hat{p}\left(t_{Left}\right)\cdot i_G\left(t_{Left}\right) + \hat{p}\left(t_{Right}\right)\cdot i_G\left(t_{Right}\right)\right]\right\}$, где $i(t)$ - так называемая функция засоренности: $i(t) = \hat{P}(\omega_1 \mid t)\cdot \hat{P}(\omega_2 \mid t)$; $i_G(t)$ - функция засоренности – критерий Gini, определяемая $i_G(t) = \sum_{i \neq k} \hat{P}(\omega_i \mid t)\cdot \hat{P}(\omega_k \mid t)$,

$i,k=1,2,\ldots,\ell$; $\hat{p}(t)$ - оценка вероятности попадания объектов обучающей выборки объема $m$ в вершину $t$ в случае известных априорных вероятностей $q_i$ классов $\omega_i$, $i=1,2,\ldots,\ell$, равных $\hat{p}(t)=\sum_{i=1}^{\ell}\frac{q_i n_t^{(i)}}{m_i}$, а в случае неизвестных априорных вероятностей $\hat{p}(t)=\frac{n_t}{m}$; $\hat{P}(\omega_i \mid t)$ - оценка условной вероятности класса $\omega_i$ в вершине $t$, равная $\hat{P}(\omega_i \mid t)=q_i\frac{n_t^{(i)}}{m_i}\Big/\sum_{i=1}^{\ell}q_i\frac{n_t^{(i)}}{m_i}$ (в случае известных априорных вероятностей классов) или $\hat{P}(\omega_i \mid t)=\frac{n_t^{(i)}}{n_t}$ (в случае неизвестных априорных вероятностей классов); $m_i$ - число объектов обучающей выборки класса $\omega_i$; $n_t^{(i)}$ - число объектов класса $\omega_i$ в вершине t; $n_t$ - количество объектов всех классов $\omega_i$ в вершине t, т.е. $n_t=\sum_{i=1}^{\ell}n_t^{(i)}$; $\ell$ - число классов.

В работе [137, 151] предложен алгоритм, использующий расстояние Колмогорова-Смирнова для выбора признаковых предикатов во внутренние вершины РД. В случае двух классов требуется найти пороговое значение $\alpha$, которое разобьет объекты обучающего множества согласно значениям вещественного признака $x_k$ на два подмножества объектов: те, для которых $x_k<\alpha$, и те, для которых $x_k\geq\alpha$. Предполагается, что заданы плотности распределения вероятностей $f_A(x_k)$ и $f_B(x_k)$ для классов $A$ и $B$ соответственно; коэффициенты ошибок классификации для классов $A$ и $B$ равны, и равны априорные вероятности появления объектов каждого класса. Тогда оптимальный по Бейесу порог $\alpha$ определяется как число, минимизирующее вероятность того, что признаковый предикат, определяющий решающее правило, произведет неправильную классификацию. Считается, что заданы также кумулятивные (интегральные) функции распределения $F_A(x_k)$ и $F_B(x_k)$, соответствующие плотностям распределения вероятностей $f_A(x_k)$ и $f_B(x_k)$. Оптимальное пороговое значение $\alpha$ максимизирует величину $|F_A(\alpha)-F_B(\alpha)|$, и это максимальное

значение называется *расстоянием Колмогорова-Смирнова между двумя распределениями для признака* $x_k$. Расстояние Колмогорова-Смирнова вычисляется для каждого признака, и во внутреннюю вершину выбирается признак, имеющий максимальное расстояние Колмогорова-Смирнова между распределениями.

Сложность использования данного алгоритма состоит в том, что обычно не заданы ни плотности распределения вероятностей, ни функции распределения вероятностей, и непосредственное их вычисление затруднительно. Тогда функции распределения вероятности "аппроксимируются" посредством замены априорных вероятностей частотами – отношением числа объектов каждого класса, которые "попали" в каждый из интервалов возможного разбиения к общему числу объектов этого класса. В качестве недостатка следует отметить то, что ошибка аппроксимации распределений авторами не учитывается.

Лбовым Г.С. и др. в работе [41] были предложены эвристические алгоритмы формирования логических решающих функций с выделением признаковых предикатов, позволяющих строить понятия при разнотипных признаках. Алгоритм CORAL ориентирован на распознавание $K$ классов ($K \geq 2$). При распознавании объекта в каждой вершине дерева проверяется истинность высказывания, представляющего собой конъюнкцию простых высказываний. При обучении и распознавании допускаются пропуски значений признаков. Алгоритм DW строит решающее правило в более простом виде, чем алгоритм CORAL: в каждой вершине дерева проверяется истинность лишь простого высказывания. Алгоритм DW дает локально-оптимальное решение и предназначен для распознавания двух классов, пропуски значений признаков не допускаются. К сожалению, алгоритмы CORAL и DW носят исключительно эвристический характер и не имеют строгого обоснования.

Работы [58, 82, 90, 97, 102, 114, 129, 141, 142, 143, 144] посвящены методам добавления повторяющихся элементов (bagging) и усиления (boosting) часто используемым в комбинации с методами синтеза и редукции решающих деревьев, что позволяет повысить их эффективность. В методе *баггинга* или *добавления*

*повторяющихся элементов* производится взвешенное голосование классификаторов, обученных на различных подвыборках данных, либо на различных частях признакового описания объектов. Выделение подмножеств объектов и/или признаков производится, как правило, случайным образом. Метод *бустинга* или *усиления* также является разновидностью взвешенного голосования, однако классификаторы строятся последовательно, и процесс увеличения различий между ними управляется детерминированным образом, т.е. для каждого классификатора, начиная со второго, веса обучающих объектов пересчитываются так, чтобы классификатор как можно точнее настраивался на тех объектах, на которых чаще ошибались все предыдущие классификаторы. Веса классификаторов вычисляются исходя из числа допущенных ими ошибок.

Заметим, что использование достаточно "изощренных" эвристик оправдано сложностью задачи синтеза экстремальных по статистическим свойствам РД.

Донским В.И. предложен ряд критериев выбора признаков – $Z_1$, D, Ω-критерии, а также алгоритм LISTBB, комбинирующий и определяющий порядок использования перечисленных одношаговых алгоритмов.

Опишем подробно *D-критерий* [15, 19] *максимальной отделимости пар разных классов*. Пусть $T_{m_t n} \subseteq T_{mn\ell}$ такое подмножество наборов, что $T_{m_t n} \subset N_t$; $K_t(i)$ число пар объектов различных классов из $T_{m_t n}$, которые имеют различное значение по переменной $x_i$. Если $D(i^*) = \max\limits_{\{i\}} K_t(i)$ и для разбиения выбирается признак $x_{i^*}$, то будем говорить, что используется D-критерий ветвления.

Обобщим D-критерий ветвления, предложенный в [19] для бинарных РД, на случай синтеза $k$-РД, при $k$>2. Будем полагать, что обучающая информация состоит из $m$ векторов, случайно и независимо выбранных из $E_k^n$ (декартовой степени $n$ множества $E_k$), для каждого из которых достоверно известно, какому из $\ell$ классов он принадлежит, причем среди них нет одинаковых векторов с указанной принадлежностью к разным классам, т.е. задана стандартная таблица

$T_{mn\ell}$, $n+1$ столбец которой служит для указания меток классов и не используется при выполнении теоретико-множественных операций над таблицей.

**Определение 1.7.** *k-значным интервалом ранга r в* $E_k^n$ называется множество $N_r = \{(x_1, x_2, \ldots, x_n) \in E_k^n \mid x_{i_1} = \sigma_1, x_{i_2} = \sigma_2, \ldots, x_{i_r} = \sigma_r\}$, где $\sigma_1, \sigma_2, \ldots, \sigma_r \in E_k$; $0 \le r \le n$. Набор номеров переменных $I_r = \{i_1, i_2, \ldots, i_r\}$ называется *направлением интервала*, а набор значений $(\sigma_1, \sigma_2, \ldots, \sigma_r)$ - *кодом интервала*.

Если $r = 0$, то $N_r = E_k^n$; если $r = n$, то $N_r$ состоит из единственной точки, принадлежащей $E_k^n$.

Пусть на шаге ветвления $t$ при синтезе *k*-РД разбиению подлежит интервал $N_r^{(t)}$. Для ветвления, вообще говоря, может быть выбрана любая переменная, номер которой $j$ не принадлежит направлению интервала $N_r^{(t)}$. Обозначим $K_t(j)$ число пар наборов различных классов в непустой таблице $T_{mn\ell} \cap N_r^{(t)}$, различающихся по переменной $x_j$. Если $K_t(j^*) = \max\limits_j K_t(j)$ и для ветвления выбирается переменная $x_{j^*}$, будем говорить, что используется D-критерий ветвления.

**Определение 1.8.** $(\Delta, m)$-*сужением k-значного интервала* $N_r$ ранга $r$ по переменной $x_m, m \in I_r$ называется множество точек $N_r^{(\Delta,m)} = \{(x_1, x_2, \ldots, x_m, \ldots, x_n) \in N_r \mid x_m \in \Delta, \Delta \subset E_k\}$.

$D_\Delta$ - критерий ветвления определим так, что при вычислении чисел $K_t(j)$ используется подтаблица $T_{mn\ell} \setminus N_r^{(\Delta,m)(t)}$, где $N_r^{(\Delta,m)(t)}$ - сужение интервала $N_r^{(t)}$, подлежащего разбиению. Если переменная $x_{j^*}$ выбирается по $D_\Delta$-критерию, то множество ребер, выходящих из внутренней вершины, соответствующей переменной $x_{j^*}$, состоит из группы ребер, соответствующих значениям $\sigma \in \{E_k \setminus \Delta\}$ и еще одного ребра, соответствующего множеству значений $\Delta$.

Отметим, что $D_\Delta$- критерий особенно полезен при синтезе *k*-РД по начальной информации $T_{mn\ell}$, имеющей пропуски – неизмеренные или неизвестные значения некоторых признаков. В этом случае символу $\Delta$ сопоставляется пропуск значения в таблице $T_{mn\ell}$. Совокупность строк из $T_{mn\ell}$, имеющих пропуск значения переменной $x_{j^*}$, образует подтаблицу, которая далее используется для синтеза дерева с условием, что при последующих ветвлениях переменная $x_{j^*}$ использоваться не будет. Шаг ветвления, допускающий пропуски, поясняется рисунком 1.4.

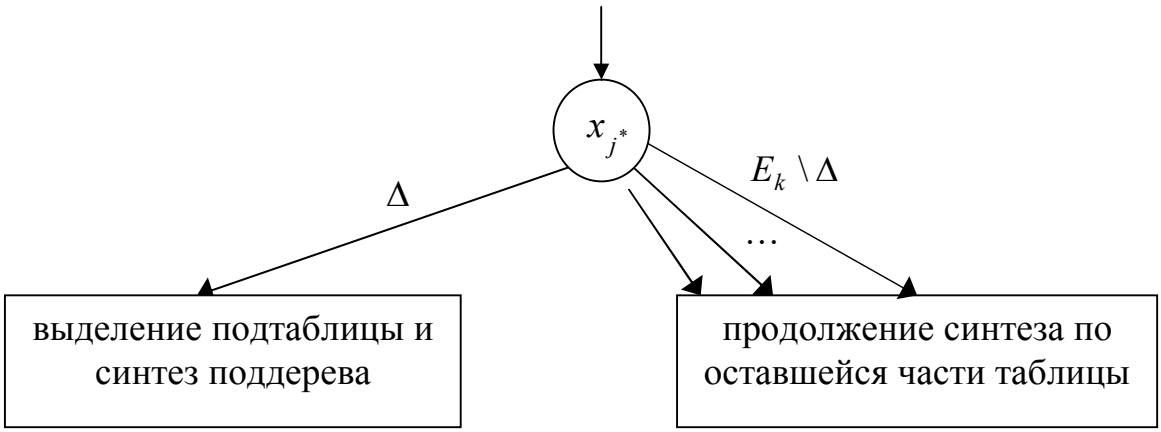


Рис.1.4. Шаг ветвления с выделением подтаблицы по множеству $\Delta$

Аналогично, могут быть обобщены и другие критерии ветвления $S_1$, $S_2$, $Z_1$, $\Omega$. На основе такого обобщения осуществляется синтез *k*-РД, близких к оптимальным, алгоритмом, который можно отнести к типу GREEDY. Класс *k*-РД при этом расширяется до класса (*k*+1)-РД, допускающих принятие решений при наличии пропусков в информации. Таким образом, алгоритмы принятия решений, основанные на решающих деревьях и допускающие работу с пропусками в начальной информации, дают возможность полнее использовать обучающую информацию.

Рассмотренные *методы синтеза решающих деревьев имеют следующие недостатки* [45]:

- являются очень чувствительными к “разбросу” объектов обучающей выборки;

- обычно требуют наличия обучающих выборок большого объема;
- построение оптимального решающего дерева часто требует больших ресурсов ЭВМ (процессорного времени и памяти);
- количество возможных альтернативных структур РД для конкретной задачи, как правило, очень велико;
- выбор признаков во внутренних вершинах РД обычно глобально неоптимальный;
- относятся к алгоритмам типа GREEDY, т.е. не способны к переоценке и, при необходимости, к замене признакового предиката, ранее выбранного во внутреннюю вершину РД;
- трудно решается задача оптимизации структуры классификатора.

Появление и разработка различных эвристических алгоритмов синтеза РД, не имеющих серьезного математического обоснования, является неизбежным эволюционным этапом развития теории и практики распознавания образов. Полезность таких эвристических алгоритмов подтверждена многочисленными экспериментами. Вопрос, какой метод синтеза решающих деревьев из большого числа известных дает лучшие результаты, требует сравнения по затрачиваемым вычислительным ресурсам при обучении, статистической надежности и точности классификации.

## 1.3. Задача редукции ветвей решающих деревьев. Выбор критерия остановки ветвления

Процесс синтеза РД чаще всего направлен на создание “идеального” классификатора, который правильно (точно) классифицирует любой объект обучающего множества, т.е. полностью “настраивается” на объекты обучающего множества в том числе, возможно, и на “зашумленные” объекты. В результате проявляется так называемый *эффект “переподгонки” или переобучения*, способствующий потере точности классификации при распознавании объектов из

множества $M \setminus M_0$ (не участвовавших в обучении) и излишнему усложнению структуры РД.

На протяжении десятилетий усилия исследователей в области обучения распознаванию на основе решающих деревьев [55, 57, 60, 65, 72, 73, 74, 76, 78, 80, 81, 84, 87, 91, 92, 104, 107, 109, 110, 111, 112, 115, 119, 124, 126, 127, 136, 139, 140, 147, 161] *были направлены на разработку стратегий избежания переподгонки на обучающем множестве*, основными из которых являются стратегии редукции или отсечения ветвей РД.

Процесс чрезмерного усложнения структуры РД (увеличение числа листьев, рост ветвей в глубину), вызванный излишним следованием “зашумленным” данным, называется *переподгонкой* (overfitting). Уточним понятие “зашумленных” объектов. Процесс извлечения объектов из генеральной совокупности для формирования обучающей выборки предполагается *случайным и независимым выбором*. Из генеральной совокупности должны быть выбраны любые произвольные, но *неискаженные*, не подверженные в процессе выбора никаким изменениям объекты. Только при таком условии обучающая таблица $T_{mn\ell}$ будет “отражать” свойства генеральной совокупности.

Будем называть таблицу $T^*_{mn\ell}$ *зашумленной*, если в процессе ее передачи в качестве начальной информации для обучения распознаванию будет изменен хотя бы один ее элемент. Иначе говоря, если $T_{mn\ell}$ - выборка, извлеченная из генеральной совокупности согласно описанной выше модели выбора, а $T^*_{mn\ell}$ - выборка, представленная в виде начальной информации для алгоритма обучения, то $T^*_{mn\ell} \neq T_{mn\ell}$.

Результатом переподгонки на обучающем множестве является РД, структура которого отражает не только основные закономерности в данных, но и влияние случайных искажений, что приводит к снижению точности распознавания объектов, не участвовавших в обучении. Процесс упрощения РД за счет отсечения избыточных ветвей с целью избежания переподгонки называется *редукцией или подрезанием* (pruning). Суть редукции состоит в удалении таких

поддеревьев РД, которые характеризуются недостаточной статистической достоверностью. Однако процесс редукции, вообще говоря, не гарантирует улучшения средней (обобщенной) точности классификатора типа РД, хотя и влечет уменьшение числа листьев. Процесс редукции обычно направлен на сокращение зависимости РД от “зашумленных” данных.

В настоящей диссертации применение редукции направлено в первую очередь на отбор и использование эмпирических закономерностей, имеющих гарантированную статистическую оценку неслучайности их извлечения, что обеспечивается редукцией ветвей деревьев.

*Предредукция* (prepruning) и *постредукция* (postpruning) – две распространенные эвристические стратегии редукции решающих деревьев. *Предредукция* или *критерий “ранней остановки”* досрочно прекращает дальнейшее ветвление в вершине РД, основываясь на некоторой эвристической мере (например, используя информационный прирост, критерий $\chi^2$ или критерий Фишера). Предредукция не является эффективным методом избежания переподгонки, поскольку принятие решения в каждой вершине РД осуществляется на основе “локальной” информации в этой вершине, т.е. на объектах, непосредственно “попавших” в эту вершину (не учитывая информацию о том, что произойдет на нижних уровнях РД). Как правило, более эффективной считается стратегия постредукции.

*Стратегия постредукции* осуществляет отсечение ветвей согласно эвристической мере (например, на основе коэффициента ошибки в каждой вершине) после того, как дерево полностью “настроится” на имеющуюся обучающую выборку. Стратегии редукции отличаются по используемым объектам для синтеза РД и редукции его ветвей: одни стратегии разделяют стандартную таблицу обучения на две подтаблицы, на объектах одной из которых осуществляются синтез РД, а на другой - редукция РД; другие не предполагают разделения исходной таблицы обучения, осуществляя и синтез и редукцию на одних и тех же прецедентах. Стратегии, требующие разделения таблицы обучения приводят к плохим результатам в случае таблиц обучения, содержащих

небольшое число объектов, тогда в качестве альтернативы предлагается использовать метод перекрестной проверки (кросс-проверки), что приводит к большим вычислительным затратам и построению множества различных деревьев.

Приведем обзор наиболее известных методов редукции РД.

1. *Редукция на основе сокращенной ошибки (Reduced Error Pruning(REP))* [60, 61, 75, 76, 78, 106]. Согласно стратегии REP обучающее множество разбивается на два подмножества: множество, предназначенное для синтеза РД, и множество, предназначенное для редукции РД. Решающее дерево $T$ полностью "настраивается" на объекты множества синтеза РД. Синтез редуцированного дерева производится на объектах множества редукции. Для каждой внутренней вершины $t$ осуществляется сравнение числа ошибок классификации на множестве редукции $I(T(t))$, допускаемых поддеревом с корневой вершиной $t$ с числом ошибок классификации на множестве редукции $I(t)$, которое возникнет в результате преобразования вершины $t$ в лист согласно мажоритарному правилу. Если $I(T(t)) \geq I(t)$, то поддерево с корневой вершиной $t$ редуцируется. Далее процесс редукции применяется к полученному редуцированному дереву до тех пор, пока для всех внутренних вершин не будет выполнено условие $I(T(t)) < I(t)$. За счет того, что объекты множества редукции не участвуют в синтезе РД, возможно получение несмещенной оценки ошибки классификации на объектах контрольного множества.

*Недостатки стратегии REP* [75, 76] заключаются в том, что

- нет четкого указания, как выбирать метки классов для листьев, получаемых в процессе редукции. Здесь существуют две возможности: использовать мажоритарное правило класса на объектах обучающего множества или на объектах множества редукции;

- в процессе редукции часто возникает ситуация, когда некоторые поддеревья РД "не получают" объекты из множества редукции, т.к. используются разные множества для синтеза и редукции РД. Такие поддеревья ("пустые" поддеревья) принято считать формируемыми за счет случайных объектов,

попавших во множество синтеза и, следовательно, они всегда редуцируются стратегией REP. Пустые поддеревья связаны с исследованием интервалов наименьшей размерности, соответствующих ветви дерева и покрывающих сравнительно небольшое число объектов обучающего множества;

- стратегия приводит к чрезмерной редукции ветвей РД, особенно в случае, когда число объектов множества редукции существенно меньше числа объектов множества синтеза РД.

2. *Редукция на основе пессимистической ошибки* (Pessimistic Error Pruning (PEP)) [60, 61, 76, 78, 106]. Стратегия использует одно и то же обучающее множество и для синтеза и для редукции РД. Пусть коэффициент ошибки в вершине $t$ на объектах обучающего множества определяется как

$$r(t)=\frac{e(t)}{n(t)},$$

где $e(t)$ - число объектов обучающего множества, не принадлежащих мажоритарному классу; $n(t)$ - число объектов обучающего множества, попавших в вершину $t$. Оценка коэффициента ошибки в вершине $t$ на объектах обучающего множества имеет вид:

$$r'(t)=\frac{e(t)+1/2}{n(t)}.$$

Аналогично, для поддерева $T(t)$ с вершиной $t$ коэффициент ошибки на обучающем множестве имеет вид:

$$r'(T(t))=\frac{\sum\limits_{j\in L(T(t))} e(j)+\frac{|L(T(t))|}{2}}{\sum\limits_{j\in L(T(t))} n(j)},$$

где $L(T(t))$ - множество листьев поддерева $T(t)$, редукция поддерева осуществляется, если $e'(t)\le e'(T(t))+SE[e'(T(t))]$, где

$SE[r'(T(t))]=\left[e'(T(t))\cdot\frac{n(t)-e'(T(t))}{n(t)}\right]^{1/2}$ - стандартная ошибка поддерева $T(t)$, вычисленная в предположении, что распределение ошибок биномиально.

Существенным недостатком PEP является *отсутствие теоретического обоснования* введенной оценки ошибок на обучающем множестве. В работе [76] отмечается, что в зависимости от типа обучающего множества, стратегия PEP может привести как к чрезмерной так и недостаточной редукции в зависимости от заданной начальной информации.

3. *Редукция на основе минимальной ошибки* (Minimum Error-Pruning (MEP)) [60, 61, 76, 78, 106]. Стратегия осуществляет редукцию, просматривая вершины дерева снизу вверх, направленную на поиск одного дерева, минимизирующего "ожидаемый коэффициент ошибки". И синтез и редукция РД осуществляются на одном и том же обучающем множестве.

В случае $k$ классов, ожидаемая вероятность того, что объект, попавший в вершину $t$ принадлежит $i$-му классу определяется по формуле $p_i(t) = \frac{n_i(t) + p_{ai} \cdot m}{n(t) + m}$, где $p_{ai}$-априорная вероятность $i$-го класса; $m$ - параметр, определяющий воздействие априорной вероятности на апостериорную вероятность $p_i(t)$. Для простоты будем полагать, что значение $m$ одинаково для всех классов. $p_i(t)$ называется $m$-вероятностной оценкой. При распознавании "нового" объекта, попавшего в вершину $t$, коэффициент ожидаемой ошибки задается как

$$Err(t) = \min_i \{1 - p_i(t)\} = \min_i \left\{ \frac{n(t) - n_i(t) + (1 - p_{ai}) \cdot m}{n(t) + m} \right\}.$$

Согласно стратегии MEP для каждой внутренней вершины $t$ сначала вычисляется ожидаемый коэффициент ошибки, называемый *статической ошибкой* $STE(t)$, затем ожидаемый коэффициент ошибки поддерева $T(t)$, называемый *динамической ошибкой* $DYE(t)$. $DYE(t)$ вычисляется как взвешенная сумма ожидаемых коэффициентов ошибки для "сыновей" вершины $t$, где вес $p_s$ - вероятность того, что объект из $t$ достигнет "сына" $s$ вершины $t$. В ранних версиях этой стратегии в качестве веса $p_s$ принималась часть обучающих объектов, достигших $s$-го "сына". Параметр $m$ выбирается произвольно.

Обычно чем больше $m$, тем “жестче” редукция. Недостатком данной стратегии является произвол в выборе значения параметра $m$. Предлагается два подхода: либо значение $m$ должен определять эксперт согласно уровню шума в данных, либо, в случае отсутствия эксперта, предлагается оценивать точность классификации редуцированных деревьев, построенных для различных значений параметра $m$, а затем среди построенной совокупности деревьев выбирать дерево с наименьшим числом листьев и минимальным эмпирическим коэффициентом ошибки.

4. *Редукция на основе критического значения* (Critical Value Pruning, CVP) [60, 61, 76, 78, 106]. Согласно этой стратегии внутренняя вершина дерева редуцируется, если значение, полученное в процессе выбора признака во внутреннюю вершину РД, не превышает фиксированного критического значения, т.е. CVP учитывает информацию, “накопленную” на этапе синтеза РД. Однако может случиться так, что условие редукции выполняется в вершине $t$ и не выполняется для ее “сыновей”. В этом случае ветвь $T(t)$ не подвергается редукции. Степень редукции изменяется в зависимости от критического значения: чем больше критическое значение, тем “жестче” редукция. CVP состоит из двух этапов:

- Редукция $T_{\max}$ с увеличением критического значения (строится совокупность РД для разных значений критического значения);
- Выбирается дерево с наименьшим числом листьев и максимальной точностью классификации среди последовательности редуцированных деревьев.

Для выбора “лучшего” РД из последовательности редуцированных РД предлагалось использовать отдельное множество редукции. Однако в этом случае не гарантируется, что в построенной последовательности существует РД “оптимальное” на множестве редукции.

5. *Редукция на основе оценки цены-сложности* (Cost-Complexity Pruning, (CCP)) [60, 61, 76]. Этот метод используется для редукции РД в упоминаемой ранее (пункт 1.2) системе CART и состоит из двух этапов:

- Выбор параметрического семейства $T_{\max}(\alpha)$ поддеревьев $\{T_0, T_1, \ldots, T_L\}$, построенных на основе РД $T_{\max}$, согласно некоторой эвристике;
- Выбор "оптимального" РД $T_i$ из параметрического семейства согласно оценке "истинных" коэффициентов ошибок деревьев (коэффициентов ошибок на объектах контрольного множества).

На первом этапе основная идея построения параметрического семейства состоит в том, что РД $T_{i+1}$ получается из $T_i$ редуцированием тех ветвей, которые определяют минимальное увеличение коэффициента ошибки на обучающем множестве по сравнению с ветвью, преобразованной в лист в результате редукции. Действительно, когда РД $T$ редуцируется в вершине $t$, его коэффициент ошибки увеличивается на величину $r(t) - r(T(t))$, в то время как число его листьев уменьшается на единицу. Таким образом, следующее отношение $\alpha = \dfrac{r(t) - r(T(t))}{|L(T(t))| - 1}$ измеряет увеличение коэффициента ошибки на обучающем множестве в результате преобразования ветви в лист, т.е. в результате редукции. Затем, $T_{i+1}$ из параметрического семейства получается путем редукции всех вершин РД $T_i$ с минимальным значением $\alpha$. Первое дерево семейства $T_0$ получается в результате редукции исходного (ранее нередуцированного) РД $T_{\max}$, последнее дерево семейства $T_L$ - корневое дерево.

На втором этапе выбирается "оптимальное" РД из параметрического семейства $T_{\max}(\alpha)$ на основе оценки точности классификации. Здесь предлагается два подхода оценивания "истинного" коэффициента ошибки для каждого дерева:

- использование множеств по методу кросс-проверки;
- использование множества редукции.

6. *Редукция, основанная на ошибке* (Error-Based Pruning (EBP)) [60, 61, 76, 78, 106]. Эта стратегия редукции РД реализована в алгоритме C4.5, упоминавшемся ранее. Отличие этой стратегии от ранее перечисленных стратегий состоит в том, что, кроме редукции, РД упрощается за счет перестройки

его структуры. EBP упрощает РД $T$ за счет перемещения ветви $T(t)$ на место "родительской" для $t$ вершины.

Подобно редукции на основе пессимистической ошибки, в EBP вычисляются оценки ошибок на объектах обучающего множества, в предположении, что ошибки биномиально распределены. Однако вместо стандартной ошибки в стратегии PEP, вычисляется доверительный интервал для расчета ошибки, основываясь на том факте, что биномиальное распределение может быть аппроксимировано нормальным распределением на выборках большого объема. Верхняя граница доверительного интервала используется для оценки ошибки в листе на объектах контрольной выборки. Верхняя граница используется для оценки вероятности ошибки, которая происходит в листе. В C4.5 доверительный интервал устанавливается заранее. Стратегия EBP осуществляет просмотр вершин РД снизу вверх. Поддерево будет замещено листом, т.е. редуцировано, если все его ветви должны быть редуцированы. Замещение выполняется, если коэффициент ошибки для "будущего" листа не больше, чем сумма коэффициентов ошибок для листьев данного поддерева.

Использование доверительного интервала, к сожалению, не имеет теоретического обоснования. Следует отметить, что и предположение о нормальном распределении требует оценивания и ограничивает область применения эвристики. В работе [76], основываясь на результатах эмпирического сравнения, отмечается склонность стратегии EBP к недостаточной редукции.

7. *Стратегии редукции РД на основе использования принципа минимальной длины описания* (Minimum Description Length Principle (MDLP)) [54, 117, 118, 132]. Стратегии редукции, основанные на принципе минимальной длины описания, отличаются друг от друга по используемой схеме кодирования и имеют два преимущества: не зависят от произвольно выбираемого параметра, определяющего "жесткость" редукции и не требуют использования вспомогательного множества – множества редукции.

*Procedure PruneTree* (вершина *N*)

$1^0$. *if* $N$ – лист *return* $C(M_0)+1$

$2^0$. $min\_cost_1$:=$PruneTree(N_1)$;

$3^0$. $min\_cost_2$:=$PruneTree(N_2)$;

$4^0$. $min\_cost_N := \min\{C(M_0)+1, C_{split}(N)+1+min\_cost_1+min\_cost_2\}$;

$5^0$. *if* $min\_cost_N = C(M_0)+1$

$6^0$. редуцировать "сыновей" $N_1$ и $N_2$ вершины $N$;

$7^0$. *return* $min\_cost_N$

$$C(M_0)=\sum_i m_i \log_2 \frac{m}{m_i}+\frac{\ell-1}{2}\log_2\frac{m}{2}+\log_2\frac{\pi^{\ell/2}}{\Gamma(\ell/2)},$$

где $M_0$ - обучающее множество; $\ell$-число классов; $m$ – число объектов в обучающем множестве; $m_i$ - число объектов класса $i$, $i=1,2,\ldots,\ell$. Величина $C(M_0)$ называется *стохастической сложностью*. Эвристический подход MDLP также представляется недостаточно обоснованным.

8. *Редукция на основе DI индекса* [84, 85]. Большая часть методов редукции (REP, PEP, MEP, CVP, CCP) направлена на минимизацию коэффициента ошибки классификации. Рассмотрим метод редукции, учитывающий структуру поддеревьев, т.е. и "засоренность" листьев и глубину ветвей, называемый далее DI-индекс (Depth-Impurity Index). Индекс качества РД $T$ имеет максимальное значение тогда и только тогда когда выполнены два условия:

- все листья дерева $T$ не содержат "примесей";
- глубина РД $T$ равна 1.

DI-индекс для дерева $T$ с корнем $\Omega$ определяется: $DI(T)=\sum_{i=1}^{m}\alpha_i\left(1-\varphi(\Omega_i')\right)f\left(depth(\Omega_i')\right)$, где $\Omega_1',\Omega_2',\ldots,\Omega_m'$ - листья РД $T$, $\alpha_i=\frac{|\Omega_i|}{|\Omega|}$; $\varphi$ - мера "примесей" $D$.

DI-индекс учитывает глубину РД, чем длиннее ветвь, тем меньше "качество" РД. Согласно DI-редукции, редуцируются все поддеревья дерева $T$ с корнем $\Omega$, удовлетворяющие условию $DI(T)\le 1-\varphi(\Omega)$.

В работах [139, 140] отмечается, что в случаях, когда при синтезе РД используется стратегия избежания переподгонки, возможно получение

индукторов с большей ошибкой, чем при использовании более тщательно “настроенных” на обучающую выборку деревьев. К сожалению, в этой работе отсутствуют строгие обоснования выводов, касающихся переподгонки. Иллюстративная кривая в координатах “ошибка” $(p)$ – “сложность РД” ($\mu$ - число листьев) [140], приведенная на рисунке 1.5, с нашей точки зрения, верно характеризует процесс синтеза РД.

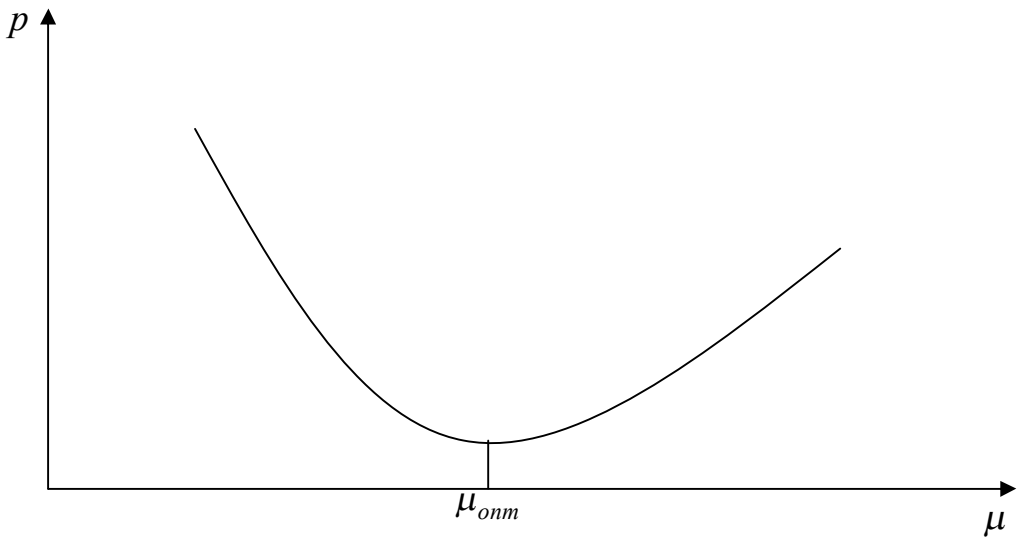


Рис 1.5. Кривая в координатах “ошибка” ($p$) - “сложность” ($\mu$ )

Действительно, при малом $\mu << \mu_{опт}$ появляется все больше ошибок на обучающей выборке (и, как следствие, при распознавании произвольных объектов). При $\mu >> \mu_{опт}, \mu = m$, получается “вырожденный” классификатор, настроенный “поточечно”.

Используя свойства единичных интервалов для почти всех булевых функций можно обосновать положения работы [140]. Нужно отметить, что принципиальное значение имеет модель начальной информации: являются ли данные таблицы обучения достоверными и непротиворечивыми (“шум” отсутствует) или в обучающем множестве возможны ошибки (наличие “шума”). Если в таблице обучения имеется “зашумленный” объект, то стратегия избежания переподгонки должна быть направлена на “отбрасывание” такого объекта: в

идеальном случае именно на нем при ошибке обучения должна остановиться подгонка.

Таким образом, любая стратегия, направленная на избежание переподгонки, может ухудшить качество распознавания [139, 140]. Однако в защиту стратегий редукции выступают многочисленные эксперименты. Эффективность стратегий редукции можно объяснить не столько природой эвристических методов, сколько удачным подбором областей для проведения экспериментов, а также подстройкой стратегий редукции под эмпирические данные за счет варьируемого параметра – уровня значимости. Следует отметить, что нет четкого теоретического обоснования, как выбирать значения для уровня значимости, в связи с чем классификатор типа РД оказывается либо недостаточно, либо чрезмерно редуцированным.

Отсутствие достаточного теоретического обоснования эффективности стратегий редукции, а также выявления их недостатков в ходе проведения многочисленных экспериментов привели к поиску других стратегий избежания переподгонки [80]. Среди них методы синтеза совокупности решающих деревьев – решающих лесов (*C4.5 decision forests*) такие как *метод случайных подпространств* [59, 93, 94, 95], позволяющий строить, в случайно выбранных подпространствах заданного признакового пространства, корректные решающие деревья, используя в качестве критерия ветвления алгоритм C4.5 (результаты практического применения этого метода представлены в [66]); *алгоритм "дровосека"* (Lumberjack algorithm) *синтеза решающего леса с переходами по ссылкам* [152] , основанный на анализе структуры РД с установкой ссылки перехода на другое дерево леса, при появлении в исходном дереве одинаковых поддеревьев; *алгоритмы наращивания отдельных вершин РД* (grafting) [156, 157]; *алгоритмы выбора признаковых предикатов с возвратом* (look ahead algorithms или backtracking method) [125]. Метод увеличения точности прогноза РД за счет "прививки" РД или наращивания отдельных вершин представляет собой процесс обработки построенного нередуцированного РД и направлено на выявление таких областей признакового пространства, которые не заняты обучающими объектами

или заняты только неверно расклассифицированными обучающими объектами, и рассмотрение возможных способов разбиения таких областей. Рассматриваются различные “альтернативные” ветви, которые могут дать лучшую классификацию “сомнительной” области, чем исходное дерево. В этом случае “альтернативная” ветвь встраивается на место внутренней вершины предшествующей листу. В работе [157] отмечается, что подобное наращивание незначительно увеличивает сложность структуры РД и в то же время увеличивает точность классификации объектов не участвовавших в обучении.

Появление как редуцирующих РД эвристик, так и наращивающих стратегий вызвано поиском компромиссных методов, позволяющих упрощать структуру исходного РД и при этом, по возможности, не уменьшать точности классификации “новых” объектов, т.е. объектов, на участвовавших ранее в обучении.

## 1.4. Выводы

Увеличение сложности структуры решающих деревьев способствует получению алгоритма распознавания, безошибочного на заданной обучающей информации (точно настроенного на обучающую выборку). Такой точно настроенный (корректный на обучающей выборке) алгоритм, имеющий структуру дерева, всегда можно построить по непротиворечивой обучающей информации, если принципиально не ограничивать сложность структуры синтезируемого решающего дерева. Более того, практически всегда точно настроенный, имеющий структуру дерева, алгоритм не является единственным. Но тогда, в соответствии со статистической теорией обучения Вапника-Червоненкиса, решения (искомые алгоритмы распознавания) будут отыскиваться в классе неограниченной емкости, что сделает саму обучаемость с гарантированной заданной точностью в общем случае невозможной. Обоснованная теоретически нежелательность усложнения

РД в процессе их синтеза (обучения) нашла практическое подтверждение в замеченном в ряде научных работ эффекте “перенастройки”.

Сложность класса решающих деревьев естественно ограничить, осуществляя редукцию ветвей – запрещая использование ветвей, содержащих число внутренних вершин бóльшее заданной пороговой величины. Ограничение сложности можно также обеспечить, ограничив общее допустимое число внутренних вершин дерева. По этому пути продвигались разработчики алгоритмов синтеза РД, “нащупывая” компромисс между излишним усложнением РД и получением как можно более высокой оценки качества синтезируемого решающего правила.

Редукция приводит к получению РД, в общем случае не обязательно корректных на обучающей таблице, сохраняя при этом возможность улучшения точности классификации объектов, не участвовавших в обучении. Но тогда в случае достоверных непротиворечивых обучающих таблиц *редукция может сразу же внести ошибку в правило классификации*, если РД некорректно относительно начальной обучающей информации.

*Возникает постановка следующей достаточно “тонкой” новой проблемы. Возможно ли (и каким образом) добиться понижения сложности древообразных классификаторов, сохранив требование их корректности на достоверной обучающей информации?* Именно в таком контексте, прежде всего, ставятся цели диссертационной работы, направленные на разрешение дилеммы, известной специалистам-разработчикам алгоритмов синтеза решающих деревьев: точнее настроиться на обучающую выборку или ограничить сложность решающего правила, обеспечивая возможность обобщения свойств обучающей информации по методу эмпирической индукции с гарантированной точностью.

Изложенные выше теоретические проблемы синтеза решающих деревьев и разработки математических моделей и методов построения широкого класса интеллектуализированных информационных обучаемых систем, использующих решающие деревья, определяют следующие *цели диссертационной работы*.

1. Обосновать целесообразность редукции бинарных решающих деревьев, используя методы дискретной математики и вероятностного оценивания.
2. Найти и обосновать методы понижения сложности древообразных классификаторов, сохранив требование их корректности на достоверной обучающей информации.
3. Предложить и обосновать методы принятия решений, основанные на коррекции совокупности редуцированных решающих деревьев как набора эвристических (и в общем случае некорректных) процедур.

Для достижения поставленных целей в диссертационной работе ставятся следующие *задачи*.

1. Разработать вероятностный критерий отсечения (редукции) ветвей бинарного решающего дерева, имеющих число внутренних вершин, превышающее заданное значение ранга $r$. Обосновать такую редукцию с точки зрения неслучайности (закономерности) обнаружения в эмпирической выборке конъюнктивной закономерности ранга $r$.
2. На основе оценок *VCD* (сложности класса решающих правил по теории Вапника-Червоненкиса) для классов алгоритмов распознавания, определяемых бинарными решающими деревьями с ограничением на число вершин, обосновать целесообразность усложнения правил распознавания и процедур коррекции решений.
3. Разработать методы построения корректной совокупности решающих деревьев (эмпирического решающего леса), обеспечивающей возможность точной настройки на обучающую выборку с одновременным соблюдением ограничения на ранг ветвей РД.
4. Получить оценку сложности эмпирического решающего леса и изучить другие его свойства как специального семейства алгоритмов распознавания.
5. Разработать алгоритмы коррекции совокупности некорректных эмпирических решающих деревьев, обеспечивающие повышение точности классификации.

6. Создать необходимое программное обеспечение и провести эксперименты на реальных данных с целью подтверждения теоретических результатов, полученных в диссертации.

# Раздел 2
# МЕТОДЫ ОЦЕНИВАНИЯ РЕШАЮЩИХ ДЕРЕВЬЕВ

## 2.1. Оценивание бинарного решающего дерева на основе подхода к закономерности как неслучайности

При большом числе признаков и объектов размерность таблицы обучения $T_{mn}$ высока, и тогда сложность ее анализа и синтеза решающих деревьев стремительно возрастает. Сложными задачами здесь являются: выявление связей между признаками, представление накопленных в таблице данных в таком виде, который не потребовал бы хранения большого объема информации, а представлял ее, по возможности, в простой и доступной для понимания форме, отражающей структурные закономерности между признаками. В сущности, именно такой формой представления и является БРД. Следует подчеркнуть, что в силу ограниченности объема заданной начальной информации по сравнению с объемом всей генеральной совокупности, выявляемая по таблице обучения закономерность неизбежно носит характер гипотезы.

Вероятностный подход к оцениванию эмпирических закономерностей и решающих правил основывается на представлении закономерности как неслучайности. В основе такого подхода лежит намеченный А.Н.Колмогоровым путь построения математической теории, позволяющей придать точный смысл понятию “случайность” как отсутствию закономерности [40]. *В соответствии с колмогоровским представлением о природе случайности, полученные в диссертации результаты, связаны со сложностью изучаемых конструктивных объектов – деревьев и леса.*

Закревский А.Д. в монографии [37] предложил рассматривать закономерность как неслучайность в следующем смысле.

Пусть обучающая информация $T_{mn}$ в задаче классификации представлена указанием $m$ булевых векторов длины $n$; $T_{mn} \subset B^n = \{0,1\}^n$. Таблица обучения $T_{mn}$

представляет собой случайную независимую выборку: каждый объект обучающей выборки получается путем случайного выбора из генеральной совокупности объектов, и выбор каждого последующего объекта не зависит от результата выбора предыдущих объектов. Предположим, объекты генеральной совокупности равновероятны: на множестве $B^n$ задано равномерное распределение $(\forall \tilde{x})\, P(\tilde{x}) = 2^{-n}$. Тогда вероятностная мера любого интервала ранга $k$ есть $2^{n-k}/2^n = 2^{-k}$ (отношение числа векторов, у которых $k$ координат зафиксировано к общему числу булевых векторов длины $n$), а вероятность того, что при случайном извлечении из $B^n$ точки $\tilde{x}$ она не будет принадлежать зафиксированному интервалу ранга $k$, есть $1 - 2^{-k}$. Иначе говоря, случайно выбранная точка не попадает в заданный интервал ранга $k$ с вероятностью $1 - 2^{-k}$. При $m$ независимых испытаниях такое непопадание в указанный интервал произойдет с вероятностью $\left(1 - 2^{-k}\right)^m$, а в любой интервал ранга $k$ - с вероятностью меньшей, чем величина $W(m,n,k) = C_n^k 2^k \left(1 - 2^{-k}\right)^m$, поскольку $C_n^k 2^k$ - число различных интервалов ранга $k$ в $B^n$. Следовательно, при случайном выборе $m$ точек для обучения из генеральной совокупности $B^n$ с равномерным распределением (формирование эмпирической таблицы $T_{mn}$) любой интервал ранга $k$ окажется *случайно пуст* с вероятностью $P(m,n,k) < W(m,n,k)$. Противоположное событие, состоящее в том, что *указанный интервал пуст неслучайно*, т.е. *закономерно*, должно иметь вероятность $1 - P(m,n,k) > 1 - W(m,n,k)$. А.Д. Закревским в [37] была приведена таблица 2.1 значений $W(m,n,k)$ для достаточно типичных случаев $m = 200, n = 100$.

По таблице 2.1 виден четкий порог дискриминации: пустые интервалы ранга $k \le 3$ почти достоверно являются закономерно пустыми при данных $m$ и $n$; интервалы большего ранга не дают никаких оснований делать выводы о закономерности, связанной с непопаданием в них точек из $T_{mn}$.

Таблица 2.1. Порог дискриминации рангов

| $k$ | $W(200,100,k)$ |
|---|---|
| 1 | $1{,}24\cdot10^{-58}$ |
| 2 | $2{,}04\cdot10^{-21}$ |
| 3 | $3{,}26\cdot10^{-6}$ |
| 4 | $1{,}56\cdot10$ |
| 5 | $4{,}21\cdot10^{6}$ |
| 6 | $3{,}27\cdot10^{9}$ |

Основная идея оценивания вероятности случайного обнаружения закономерностей – гипотез, основанная на колмогоровском подходе, предполагает использование модели генеральной совокупности с равномерным распределением объектов.

Изучая закономерность – гипотезу, можно оценить вероятность того, что она может быть обнаружена по выборке, извлеченной из генеральной совокупности с *равномерным* распределением, т.е. *случайно*. Если эта вероятность достаточно мала, то случайное обнаружение закономерности представляется маловероятным; также маловероятным в таком случае является равномерное распределение объектов генеральной совокупности.

Уточним понятие конъюнктивной закономерности ранга $r$ относительно таблицы обучения $T_{mn}$.

**Определение 2.1**. Конъюнкция $K_r = x_{i_1}^{\sigma_{i1}} \,\&\, x_{i_2}^{\sigma_{i2}} \,\&\, \ldots \,\&\, x_{i_r}^{\sigma_{i_r}}$ ранга $r$ называется *конъюнктивной закономерностью ранга r относительно таблицы обучения* $T_{mn}$, если среди $n$ столбцов этой таблицы найдутся такие $r$ столбцов с номерами $i_1, i_2, \ldots, i_r$ и не входящий в их число столбец с номером $k$, что для всех строк $\widetilde{x} \in T_{mn}$, удовлетворяющих условию $x_{i_1} = \sigma_{i_1}, x_{i_2} = \sigma_{i_2}, \ldots, x_{i_r} = \sigma_{i_r}$

переменная $x_k$ принимает одно и то же значение $\gamma_{\tilde{\sigma}}$. При этом множества $\{\tilde{x} : x_{i_1} = \sigma_{i_1}, x_{i_2} = \sigma_{i_2}, \ldots, x_{i_r} = \sigma_{i_r}\} \cap T_{mn} = T_1$ и $T_{mn} \setminus T_1$ непусты.

В задачах обучения распознаванию или формирования понятий булева переменная $x_k$ является характеристической функцией некоторого класса объектов и называется *целевой*. Например, если $x_k = 1$, то объект принадлежит указанному классу, а если $x_k = 0$ - не принадлежит.

С точки зрения теоретико-множественного подхода, конъюнктивная закономерность $K_r$ ранга $r$ соответствует интервалу $N_{K_r}$ размерности $(n-r)$, содержащему объекты только одного класса из таблицы обучения $T_{mn}$.

Поскольку для поиска закономерностей используется случайно выбранная таблица $T_{mn}$ из генеральной совокупности $\tilde{T}_{mn}$ булевых таблиц размерности $m \times n$, то конъюнктивная закономерность может быть *обнаружена случайно* в таблице $T_{mn}$, если столбец с номером $k$ случайно окажется заполненным значениями, соответствующими некоторой функции $f$. Функция $f$ ставит в соответствие набору $\sigma_{i_1}, \sigma_{i_2}, \ldots, \sigma_{i_r}$ зафиксированное значение целевого столбца. Случайность такого обнаружения оценил А.Д. Закревский [37]. Ему принадлежит следующий результат, важный для дальнейшего исследования РД; поэтому этот результат строго доказан ниже.

**Теорема 2.1.** Пусть $T_{mn}$ - булева таблица, случайно выбранная из генеральной совокупности $\tilde{T}_{mn}$ таблиц размерности $m \times n$ с равномерным распределением на $\tilde{T}_{mn}$. Тогда вероятность $P(m,n,r)$ случайного обнаружения конъюнктивной закономерности ранга $r$ относительно таблицы $T_{mn}$ удовлетворяет неравенству

$$P(m,n,r) < (n-r) C_n^r 2^{-(m-2^r)} \quad (2.1)$$

Доказательство. В соответствии с равномерным распределением таблиц на $\tilde{T}_{mn}$, для произвольного столбца таблицы $T_{mn}$ любое из его $2^m$ возможных значений равновероятно.

Пусть $A_r$ - событие, состоящее в появлении в случайной таблице $T_{mn}$ хотя бы одной закономерности ранга $r$. Событие $A_r$ всегда происходит одновременно с не менее чем одним событием $H_{f,\tilde{r},k}$, соответствующим тому, что конъюнктивная закономерность ранга $r$ определяется зафиксированным набором столбцов $\tilde{r}=(i_1,i_2,\ldots,i_r)$, зафиксированной функцией $f\in P_2(r)$ и зафиксированной целевой переменной $x_k$ (столбцом $k$). Различные события $H_{f,\tilde{r},k}$, вообще говоря, не являются несовместными: может существовать несколько разных зафиксированных закономерностей ранга $r$ одновременно. Поэтому

$$A_r=\bigcup_{(f,\tilde{r},k)\in M} A_r\cap H_{f,\tilde{r},k}, \quad (2.2)$$

где $M=\{(f,\tilde{r},k): f\in P_2(r);\tilde{r}\subset\{1,2,\ldots,n\};k\in\{1,2,\ldots,n\}\setminus\tilde{r}\}$.

Число различных событий $H_{f,\tilde{r},k}$ конечно и равно мощности множества $M$:

$$|M|=2^{2^r}C_n^r(n-r),$$

где $2^{2^r}$ - число различных функций $f\in P_2(r)$; $C_n^r$ - число способов выбора $r$ любых столбцов из $n$; $(n-r)$ - число способов выбрать целевой столбец из оставшихся.

По теореме умножения зависимых событий и теореме сложения совместных событий без учета вероятностей совместного появления событий $H_{f,\tilde{r},k}$ согласно (2.2), можно оценить вероятность события $A_r$:

$$P(A_r)=P\left(\bigcup_{(f,\tilde{r},k)\in M} A_r\cap H_{f,\tilde{r},k}\right)<\sum_{(f,\tilde{r},k)\in M} P(H_{f,\tilde{r},k})P(A_r/H_{f,\tilde{r},k}).$$

Событие $A_r$ при условии $H_{f,\tilde{r},k}$ наступает тогда, когда столбец, соответствующий целевой переменной $x_k$, принимает единственный набор значений из $2^m$ наборов значений длины $m$, в точности определяемый функцией $f$. Обозначим этот набор $\tilde{\alpha}_f$. Событие, состоящее в появлении набора $\tilde{\alpha}_f$ в столбце $k$ таблицы $T_{mn}$,

обозначим $G_{\tilde{\alpha}_f,k}$. Очевидно, что $A_r / H_{f,\tilde{r},k} \subset G_{\tilde{\alpha}_f,k}$, откуда следует $P\left(A_r / H_{f,\tilde{r},k}\right) < P\left(G_{\tilde{\alpha}_f,k}\right) = 2^{-m}$.

Следовательно, $P\left(A_r\right) < \sum\limits_{(f,\tilde{r},k)\in M} 2^{-m} P\left(H_{f,\tilde{r},k}\right) < 2^{-m} 2^{2^r} C_n^r \left(n-r\right) = \left(n-r\right) C_n^r 2^{-\left(m-2^r\right)}$. □

На практике этот результат приводит к поиску конъюнкций, ранг которых, как правило, *не должен превышать число семь* [37].

Если оценка вероятности $P(m,n,r)$ мала, и в таблице $T_{mn}$, извлеченной из произвольной генеральной совокупности таблиц, обнаружена закономерность ранга $r$, то правомерно считать, что эта закономерность будет появляться и на других таблицах, извлекаемых из этой совокупности, и отражать обнаруженное свойство генеральной совокупности.

С точностью до колмогоровского положения "*закономерность* ≡ *неслучайность*", можно заключить: если случайное появление закономерности - гипотезы имеет вероятность $P(m,n,r)$, то ее неслучайное появление имеет вероятность $W(m,n,r) = 1 - P(m,n,r)$. Иначе говоря, если в эмпирической таблице $T_{mn}$ обнаружена закономерность ранга $r$, то она имеет место на всей генеральной совокупности с вероятностью $W(m,n,r) > 1 - \left(n-r\right) C_n^r 2^{-\left(m-2^r\right)}$, отражая ее свойство.

Конъюнктивная закономерность определяет импликативное правило принятия решения:

$$\left(x_{i_1}^{\sigma_{i_1}} \;\&\; x_{i_2}^{\sigma_{i_2}} \;\&\; \ldots \;\&\; x_{i_r}^{\sigma_{i_r}}\right) \to x_k^{\gamma_{\tilde{\sigma}}} \qquad (2.3),$$

т.е. определяет значение целевой переменной $x_k$ при условии, когда вектор $\tilde{x} = \left(x_1, x_2, \ldots, x_n\right)$ попадает в интервал $N_{K_r}$, соответствующий конъюнкции $K_r$, содержащейся в левой части импликации (2.3). Поэтому можно говорить о "*неслучайном выводе решения*" при помощи конъюнктивной закономерности с вероятностью $W(m,n,r)$.

**Определение 2.2**. Если конъюнктивная закономерность выполняется для любой таблицы $T_{mn}$, извлеченной из некоторой генеральной совокупности $\widetilde{T}_{mn}$, то эта конъюнктивная закономерность называется *абсолютной*.

Если конъюнктивная закономерность обнаружена случайно, то она появилась на некоторой данной таблице $T_{mn}$, но может являться абсолютной лишь с некоторой вероятностью. Теорема 2.1 дает оценку этой вероятности случайного вывода решения при помощи эмпирической конъюнктивной закономерности.

Если бинарное решающее дерево $BDT_{\mu,m,n}$ с $\mu$ листьями корректно на некоторой таблице $T_{mn}$, то оно определяет набор конъюнктивных закономерностей $K_{r_1}, K_{r_2}, \ldots, K_{r_\mu}$, обеспечивающих вывод решения. Эти конъюнкции попарно ортогональны, поскольку соответствуют различным ветвям дерева [15]. Указанные выше свойства $BDT_{\mu,m,n}$ и неравенство 2.1 позволяют получить следующую оценку.

**Теорема 2.2**. Вероятность $P_{BDT}(\mu,m,n)$ того, что при использовании бинарного решающего дерева $BDT_{\mu,m,n}$ с $\mu$ листьями, корректного на эмпирической таблице обучения $T_{mn}$, будет получен случайный вывод о свойстве $x_k$ для произвольного объекта $\widetilde{x} = (x_1, x_2, \ldots, x_{k-1}, x_{k+1}, \ldots, x_n)$, выбранного равновероятно из множества булевых векторов длины $n-1$, удовлетворяет неравенству

$$P_{BDT}(\mu,m,n) < \sum_{j=1}^{\mu} (n - r_j) C_n^{r_j} 2^{-\left(m + r_j - 2^{r_j}\right)} \qquad (2.4),$$

где $r_1, r_2, \ldots, r_\mu$ - ранги конъюнкций, соответствующих ветвям дерева.

Доказательство. Конъюнкции, соответствующие ветвям дерева $BDT_{\mu,m,n}$, попарно ортогональны, и для любого допустимого объекта $\widetilde{x}$ решение всегда принимается ровно одной из них. По теореме умножения зависимых событий и теореме сложения несовместных событий, получим:

$$P_{BDT}(\mu,m,n) = \sum_{j=1}^{\mu} P\left(N_{K_j}\right) P\left(Сл / N_{K_j}\right),$$

где события $Сл/N_{K_j}$ соответствуют случайному выводу решения при использовании конъюнкции $K_j$, а $P\left(N_{K_j}\right)=2^{-r_j}$ - вероятность того, что при предъявлении объекта $\tilde{x}$ он будет классифицироваться ветвью бинарного решающего дерева $BDT_{\mu,m,n}$, соответствующей конъюнкции $K_j$, т.е. попадет в интервал $N_{K_j}$. $P\left(Сл/N_{K_j}\right)=P\left(m,n,r_j\right)$ и оценено неравенством (2.1). Поэтому

$$P_{BDT}\left(\mu,m,n\right)<\sum_{j=1}^{\mu}2^{-r_j}\left(n-r_j\right)C_n^{r_j}2^{-\left(m-2^{r_j}\right)}=\sum_{j=1}^{\mu}\left(n-r_j\right)C_n^{r_j}2^{-\left(m+r_j-2^{r_j}\right)}.\ \square$$

**Замечание 2.1**. В случае, когда целевая переменная (целевой столбец) задана отдельно от эмпирической таблицы обучения $T_{mn}$, т.е. не входит в число ее столбцов, неравенство аналогичное (2.4) имеет вид

$$P^{*}_{BDT}\left(\mu,m,n\right)<\sum_{j=1}^{\mu}C_n^{r_j}2^{-\left(m+r_j-2^{r_j}\right)} \qquad (2.5),$$

где $P^{*}_{BDT}\left(\mu,m,n\right)$ - вероятность случайного вывода целевого свойства деревом, построенным по стандартной начальной информации с $n$ булевыми признаками.

**Замечание 2.2**. Оценки (2.4), (2.5) характеризуют среднее значение *(математическое ожидание) оценки* вероятности случайного вывода на множестве всех допустимых объектов $\tilde{x}$ при заданных рангах $r_1,r_2,\ldots,r_\mu$ конъюнкций, соответствующих ветвям БРД $BDT_{\mu,m,n}$.

## 2.2. Обоснование редукции ветвей решающего дерева на основе подхода к закономерности как неслучайности

*Редукция решающего дерева* – процесс упрощения структуры РД, направленный на избежание “переподгонки” на объектах обучающего множества. Интуитивно ясно: если структура РД усложняется за счет роста отдельных ветвей дерева, то такие ветви должны редуцироваться, как статистически ненадежные. В противном случае, будет построено решающее правило, “настроенное” на

отдельные объекты (небольшие группы объектов) выборки и даже, возможно, на зашумленные объекты.

Обозначим $D_{\mu,m,n}$ - класс всевозможных решающих деревьев с ровно $\mu$ листьями, которые могут быть построены на объектах стандартной таблицы обучения $T_{mn}$. Исходя из оценки (2.5) , полученной в пункте 2.2, введем на множестве $D_{\mu,m,n}$ функционал качества $\varphi : D_{\mu,m,n} \to R$ следующего вида:

$$\varphi(d) = \max_{1 \le j \le \mu} \left( C_n^{r_j} 2^{-\left(m + r_j - 2^{r_j}\right)} \right) \qquad (2.6).$$

Функционал $\varphi(d)$ является оценкой “худшего” случая при использовании для принятия решения самой “неблагоприятной” ветви дерева – ветви дерева наибольшего ранга. Это обосновывается следующей леммой 2.1.

**Лемма 2.1**. Величина $h(n,m,r) = C_n^r 2^{-\left(m + r - 2^r\right)}$ монотонно возрастает с ростом ранга ветви $r$, при $r \ge 2$ и $n \ge r+1$.

Доказательство. Пусть $a(r,r+1) = h(n,m,r+1)/h(n,m,r)$, тогда

$$a(r,r+1) = \frac{C_n^{r+1} 2^{-m} 2^{-(r+1)} 2^{2^{r+1}}}{C_n^r 2^{-m} 2^{-r} 2^{2^r}} = \frac{(n-r)}{2(r+1)} 2^{2^r}.$$

Убедимся, что $a(r,r+1) > 1$, при $r \ge 2$ и $n \ge r+1$. Действительно, т.к. по условию $n \ge r+1$, то $n - r \ge 1$ и тогда $\frac{(n-r)}{2(r+1)} 2^{2^r} \ge \frac{1}{2(r+1)} 2^{2^r} = \frac{2^{2^r-1}}{r+1}$. При $r = 1$, получим равенство $2^{2^r-1} = r+1$, но при $r \ge 2$ выполняется $\frac{2^{2^r-1}}{r+1} > 1$, поскольку экспонента $2^{2^r-1}$ растет быстрее линейной функции $r+1$ с ростом ранга ветви $r$. □

**Лемма 2.2**. Минимум функционала $\varphi(d)$ на множестве РД $D_{\mu,m,n}$

$$\varphi(d^*) = \min_{d \in D_{\mu,m,n}} \max_{1 \le j \le \mu} \left( C_n^{r_j} 2^{-\left(m + r_j - 2^{r_j}\right)} \right)$$

достигается при $\mu = 2^k, k \in N$, в случае, если $d^*$ - полное дерево; а при $\mu \neq 2^k$ - в случае, когда дерево $d^*$ обладает следующим свойством: $\max\limits_{1 \le j < q \le \mu} \left| r_j - r_q \right| \le 1$, где $r_1, r_2, \ldots, r_\mu$ - ранги конъюнкций, соответствующих ветвям РД $d^*$.

Доказательство. Рассмотрим два взаимно исключающих случая согласно условию леммы.

1) Пусть $\mu = 2^k, k \in N, r_j = k, j = \overline{1, \mu}$, т.е. случай полного дерева $d^*$ с $\mu$ листьями, тогда максимальный ранг ветви равен $r_{\max} = k$. Пусть $\varphi(d^*) = \max\limits_j \left( C_n^{r_j} 2^{-\left(m + r_j - 2^{r_j}\right)} \right)$. Переход от полного дерева $d^*$ к любому другому дереву $d$ повлечет удаление вершины из хотя бы одной ветви и добавление вершины в хотя бы одну другую ветвь, поскольку число листьев $\mu$ должно оставаться неизменным. Тогда $r_{\max} \ge k + 1$ и по лемме 2.1 о монотонности функционала $\varphi(d) > \varphi(d^*)$.

2) Пусть $\mu \neq 2^k, k \in N$. Рассмотрим любое дерево $d \in D_{\mu,m,n}$. Оно не может быть полным в силу ограничения на $\mu$ и, следовательно, среди рангов $r_1, r_2, \ldots, r_\mu$ найдутся различные. Выберем $r_{\min} = \min\{r_1, r_2, \ldots, r_\mu\}$ и $r_{\max} = \max\{r_1, r_2, \ldots, r_\mu\}$. Удалив одну внутреннюю вершину, предшествующую листу из ветви с рангом $r_{\max}$, и добавив ее в ветвь, соответствующую $r_{\min}$, получим уменьшение величины $\hat{r}_{\max} = r_{\max}(\hat{d})$, где $\hat{d}$ - новое дерево, полученное из $d$ путем указанной перестройки, но тогда и только тогда, когда $r_{\max} - r_{\min} \ge 2$. По лемме 2.1 это повлечет уменьшение значения функционала $\varphi(\hat{d}) < \varphi(d)$. □

*Рангом ветви* будем называть ранг соответствующей этой ветви конъюнкции.

**Определение 2.3**. Бинарное решающее дерево называется *равномерным*, если ранги его ветвей отличаются не более чем на единицу.

**Теорема 2.3**. В классе эмпирических решающих деревьев $D_{\mu,m,n}$ наименьшую равномерную оценку (2.6) вероятности получения случайного заключения имеют равномерные деревья.

Действительно, по лемме 2.2, с учетом определения 2.3 минимум функционала достигается на равномерных деревьях.

Теорема 2.3 является теоретическим обоснованием процесса редукции ветвей РД, доставляющих функционалу (2.6) максимум. Редукция направлена на упрощение структуры РД с целью “приблизить” полученное в результате синтеза решающее дерево к равномерному РД. Это предварительное соображение будет развито ниже и послужит обоснованием алгоритма синтеза совокупности эмпирических решающих деревьев.

## 2.3. Оценка VCD для основных классов решающих функций, представленных решающими деревьями

Изложим ряд положений теории Вапника-Червоненкиса, которые необходимы для понимания полученных ниже результатов, а также результатов пункта 3.4.

Пусть $X$ - множество, $S$ - некоторая система его подмножеств; $X^\ell = x_1, x_2, \ldots, x_\ell$ - последовательность элементов из $X$ длины $\ell$. Каждое множество $A \in S$ определяет подпоследовательность $X_A$ этой последовательности, состоящую из тех элементов, которые принадлежат $A$. Говорят, что $A$ *индуцирует* подпоследовательность $X_A$ на последовательности $X^\ell$ [3].

Обозначим $\Delta^S(x_1, x_2, \ldots, x_\ell)$ - число различных подпоследовательностей индуцированных всеми множествами $A \in S$. Очевидно, что $\Delta^S(x_1, x_2, \ldots, x_\ell) \le 2^\ell$.

Функция $m^S(\ell) = \max\limits_{x_1,x_2,\ldots,x_\ell} \Delta^S(x_1, x_2, \ldots, x_\ell)$, где максимум берется по всем последовательностям длины $\ell$, называется *функцией роста системы множеств* $S$. Для случая задач обучения распознаванию, когда используется эмпирическая выборка и некоторое решающее правило, функция роста имеет большое значение и поэтому уточняется ниже.

Если $S$ - множество решающих правил, $S = \{F(x,\alpha), \alpha \in A\}$, где $\alpha$ - параметр, и $x_1, x_2, \ldots, x_\ell$ - выборка из множества допустимых объектов $X$, то эта выборка может быть разделена на два класса не более чем $2^\ell$ способами. Такое разделение всякий раз фиксируется выбранным правилом $F(x,\alpha)$, подпоследовательностью $X_{F(x,\alpha)}$ последовательности $X^\ell$ и ее дополнением в $x_1, x_2, \ldots, x_\ell$. Иначе говоря, с помощью правила $F(x,\alpha)$ множество $X^\ell$ делится на два подмножества: состоящее из таких $x$, что $F(x,\alpha) = 1$ и (второе) множество таких, что $F(x,\alpha) = 0$.

Число способов разделения выборки зависит и от класса решающих правил $\{F(x,\alpha), \alpha \in A\}$, и от состава выборки $x_1, x_2, \ldots, x_\ell$. Это число в рассматриваемом случае и есть $\Delta^S(x_1, x_2, \ldots, x_\ell)$.

В задачах обучения распознаванию каждый объект $x$ выборки снабжается меткой $\omega$ класса, которому он принадлежит. Положим, $\omega \in \{0,1\}$ и рассмотрим систему событий $S(\alpha) = \{(x,\omega) \mid (\omega - F(x,\alpha))^2 = 1\}$, соответствующую ошибкам решающего правила $F(x,\alpha)$ на объектах $x \in X$. Обучающая выборка представляется в виде $(x_1,\omega_1), (x_2,\omega_2), \ldots, (x_\ell,\omega_\ell)$, и система событий $S(\alpha)$ индуцирует на ней $\Delta(S(\alpha); (x_1,\omega_1), (x_2,\omega_2), \ldots, (x_\ell,\omega_\ell)) = \Delta^S(x_1, x_2, \ldots, x_\ell)$ различных подвыборок.

**Определение 2.4** [3]. Функция $m^S(\ell) = \max\limits_{x_1,x_2,\ldots,x_\ell} \Delta^S(x_1, x_2, \ldots, x_\ell)$, где максимум берется по всевозможным выборкам длины $\ell$, называется *функцией роста системы событий, образованной решающими правилами* $\{F(x,\alpha), \alpha \in A\}$.

**Теорема 2.4** [3]. Функция роста либо тождественно равна $2^{\ell}$, либо при $\ell > h$ мажорируется функцией $1{,}5\,\ell^h/h!$, то есть

$$m^S(\ell)\begin{cases} \equiv 2^{\ell}, & \text{если } h=\ell, \\ < 1{,}5\dfrac{\ell^h}{h!}, & \text{если } \ell > h, \end{cases}$$

где $h+1$ - минимальный объем выборки, при котором нарушается условие $m^S(\ell)\equiv 2^{\ell}$.

Число $h$ может служить *мерой разнообразия* класса решающих правил.

**Определение 2.5** [3]. *Класс решающих правил имеет ёмкость* $h$, если справедливо неравенство $m^S(\ell) < 1{,}5\,\ell^h/h!$, $\ell > h$. Если же $m^S(\ell) = 2^{\ell}$, то говорят, что *ёмкость класса решающих правил бесконечна*.

Например, для класса линейных решающих правил в $R^n$ максимальное число точек $h$, которое можно с помощью *гиперплоскости, проходящей через начало координат*, разбить на два класса всеми возможными способами, равно *n*. Поэтому при $\ell > h$ функция роста оценивается неравенством $m^S(\ell) < 1{,}5\,\ell^n/n!$, и класс линейных отделителей имеет емкость *n*.

В зарубежной литературе емкость *h* класса $\{F(x,\alpha),\alpha \in A\}$ решающих правил называют *VCD* (Vapnik-Chervonenkis Dimension) класса $\{F(x,\alpha),\alpha \in A\}$. Таким образом, *ёмкостная характеристика решающих правил VCD оценивает отношение сложности решающего правила к его разделяющей способности*.

Если $P(\alpha)$- вероятность ошибки при использовании правила $F(\alpha)$, а $\upsilon(\alpha)$ - частота этой ошибки на эмпирической выборке, то имеет место неравенство [3]

$$P\left\{\sup_{\alpha\in A}|P(\alpha)-\upsilon(\alpha)|>\varepsilon\right\} < 6m^S(2\ell)e^{-\frac{\varepsilon^2\ell}{4}},$$

где $\varepsilon > 0$. Эта оценка может быть нетривиальной, когда VCD класса $\{F(x,\alpha),\alpha \in A\}$ конечна и равна *h*. Тогда $m^S(\ell) < 1{,}5\,\ell^h/h!$ и

$$P\left\{\sup_{\alpha\in A}|P(\alpha)-\upsilon(\alpha)|>\varepsilon\right\} < 9\frac{(2\ell)^h}{h!}e^{-\frac{\varepsilon^2\ell}{4}} \to 0, \text{ при } \ell\to\infty \qquad (2.7).$$

С ростом $\ell$ правая часть неравенства (2.7) стремится к нулю и притом тем быстрее, чем меньше ёмкость класса $h$.

Таким образом, *конечная величина VCD обеспечивает равномерную сходимость частот ошибок к вероятностям при росте длины обучающей выборки и позволяет оценить скорость этой сходимости, поэтому получение VCD для различных классов решающих правил представляет значительный интерес.*

Если класс решающих правил конечен, $\left|\{F(x,\alpha),\alpha\in A\}\right| = M < \infty$, то очевидно, $VCD(\{F(x,\alpha),\alpha\in A\}) \le \log_2 M$ (2.8). Поэтому оценивание мощности конечных классов решающих правил позволяет, используя последнее неравенство, получить для них оценку VCD, а при детерминистской постановке задачи обучения распознаванию для случая нулевой эмпирической ошибки использовать неравенство [3]

$$P\left\{\sup_{\alpha\in A} P(\alpha) > \varepsilon\right\} < M(1-\varepsilon)^{\ell},$$

или, с учетом соотношения $(1-\varepsilon) < e^{-\varepsilon}$, $\varepsilon > 0$,

$$P\left\{\sup_{\alpha\in A} P(\alpha) > \varepsilon\right\} < Me^{-\varepsilon\ell}.$$

Заметим, что никакая нижняя оценка $L < M$ числа решающих правил $M$ конечного класса $\{F(x,\alpha),\alpha\in A\}$, вообще говоря, не дает возможности оценки его *VCD*.

Далее рассматриваются оценки *VCD* для основных классов решающих деревьев, а именно: оценки *VCD* решающих деревьев ограниченного ранга и оценки *VCD* бинарного решающего дерева с двумя классами.

### 2.3.1. Оценка *VCD* класса решающих функций, представленных решающим деревом ограниченного ранга

В этом пункте обобщаются результаты работы [145] с целью последующего сравнения различных вариантов оценки сложности решающих деревьев.

Ранг БРД $T$ определяется индуктивно.

**Определение 2.6** [145]. Если БРД $T$ содержит единственную вершину, то его ранг $r$=0. Если $T$ содержит корень, левое поддерево $T_0$ ранга $r_0$ и правое поддерево $T_1$ ранга $r_1$, то

$$rank(T)=\begin{cases} 1+r_0, & \text{если } r_0 = r_1, \\ \max\{r_0, r_1\}, & \text{если } r_0 \neq r_1. \end{cases}$$

БРД $T$ имеет ранг больше чем $r$ тогда и только тогда, когда полное бинарное дерево глубины $r$+1 может быть “встроено” в $T$ [145].

Обозначим $rDT_n$, следуя работе [145], класс булевых функций от $n$ переменных, представимых в виде БРД ранга не более $r$, и $VCD(rDT_n)$ - ёмкость этого класса.

**Пример 2.1.** На рисунке 2.1 представлены два бинарных решающих дерева одинакового ранга, имеющие различную сложность, традиционно оцениваемую как число внутренних вершин или листьев дерева, учитывая, что число листьев на единицу больше числа внутренних вершин.

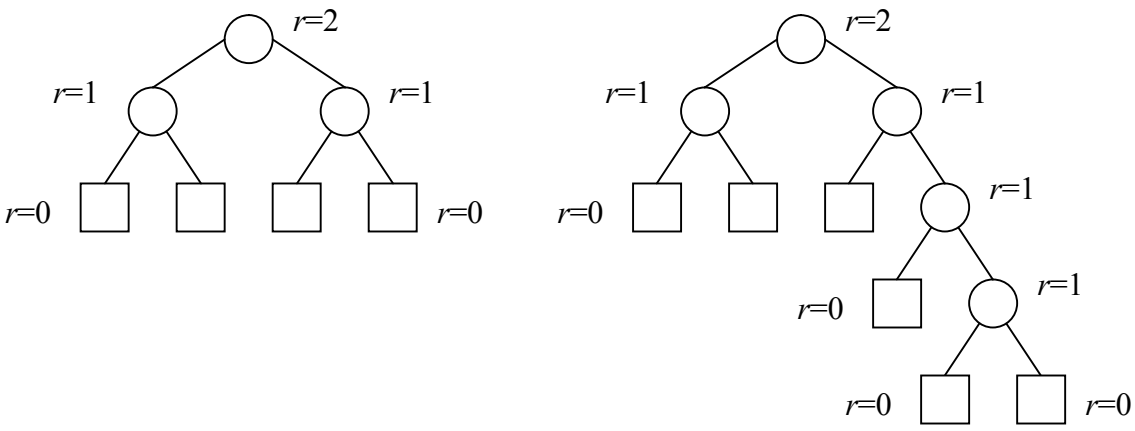


Рис. 2.1. БРД одинакового ранга 2, имеющие различную сложность $\mu$.

Пример показывает, что такая характеристика, как ранг дерева, предложенная в работе [145], согласно определению 2.6, не представляется обоснованной как показатель качества БРД. При достаточно большой размерности признакового пространства $n$ может оказаться, что дерево небольшого ранга $r << n$ будет иметь глубину близкую к $n$, что должно приводить к редукции некоторых ветвей. Поэтому результат работы [145] не дает возможности оценивать БРД, построенные при помощи алгоритмов обучения распознаванию, адекватно их сложности и нечувствителен к переподгонке.

**Теорема 2.5** [145]. $VCD\left(rDT_n\right)=\sum_{i=0}^{r} C_n^i$.

Для дальнейшего изучения VCD решающих деревьев понадобиться следующее

**Следствие 2.1.** $VCD\left(rDT_n\right)=\Theta\left(n^r\right)$ при любой заданной константе $r$ и $n \to \infty$.

Доказательство. С учетом теоремы 2.5, получим оценку сверху и снизу $VCD\left(rDT_n\right)$. Оценка сверху для $VCD\left(rDT_n\right)$:

$$\sum_{i=0}^{r} C_n^i = C_n^0 + C_n^1 + \ldots + C_n^r < 1 + n + \frac{n\cdot(n-1)}{2!} + \ldots + \frac{n\cdot(n-1)\cdot\ldots\cdot(n-r+1)}{r!} < 1 + n +$$

$$+ \frac{n^2}{2!} + \ldots + \frac{n^r}{r!} < 1 + n + n^2 + \ldots + n^r = \frac{n^{r+1}-1}{n-1} < \frac{n^{r+1}}{n-1} = \frac{n}{n-1}\cdot n^r < 2n^r,$$ при $n > 2$. Поэтому $\sum_{i=0}^{r} C_n^i = O\left(n^r\right)$. Оценка снизу для $VCD\left(rDT_n\right)$:

$$\sum_{i=0}^{r} C_n^i = C_n^0 + C_n^1 + \ldots + C_n^r > C_n^r = \frac{n\cdot(n-1)\cdot\ldots\cdot(n-r+1)}{r!} > \frac{(n-r)^r}{r!}.$$

Таким образом, $\frac{(n-r)^r}{r!} < \sum_{i=0}^{r} C_n^i$ при любой заданной константе $r$, откуда несложно получить соотношение $n^r = O\left(\sum_{i=0}^{r} C_n^i\right)$ при $n \to \infty$. □

### 2.3.2. Оценка *VCD* класса решающих функций, представленных бинарным решающим деревом с двумя классами

БРД, использующее булевы переменные и метки только двух классов, обеспечивают вычисление соответствующих им булевых функций (б.ф.). Любая б.ф. может быть представлена некоторым БРД [19], поэтому класс БРД как решающих правил очень широк и имеет неограниченную емкость по определению 2.5. Конструктивные ограничения, накладываемые на БРД, позволяют выделять конечные подклассы БРД конечной емкости. Таким являлось рассмотренное в пункте 2.3.1 ограничение на величину ранга деревьев.

Рассмотрим *конечное множество БРД с не более чем* $\mu$ *листьями, обозначим его и соответствующий этому множеству класс булевых функций* $BDT(\mu, n)$.

**Теорема 2.6.** $\max(\mu, \log_2 n) \le VCD(BDT(\mu, n)) < \mu \log_2(n\mu)$.

Доказательство. Покажем, что $VCD(BDT(\mu, n)) \ge \mu$. Действительно, корректная обучающая выборка состоит из попарно различных булевых векторов, поэтому при зафиксированном числе листьев $\mu$ для любых $\mu$ векторов можно получить любую классификацию, если каждому вектору соответствует единственный лист дерева, и тогда можно любыми $2^{\mu}$ способами расставить метки классов.

Теперь покажем, что $VCD(BDT(\mu, n))$ может быть больше, чем $\mu$. Пусть $\mu = 2$. Укажем три булевых вектора, представленные в таблице 2.2, которые могут быть расклассифицированы любым способом при $n = 8$. Каждый из вариантов классификации получается, если в единственной внутренней вершине поместить одну из переменных $x_1, \ldots, x_8$.

С учетом примера 2.1, БРД при $n = 2$ и $\mu = 4$ (см. рис.2.1), очевидно, что достижимая нижняя граница VCD определяется неравенством $\max(\mu, \log_2 n) \le VCD(BDT(\mu, n))$.

Таблица 2.2. Три булевых объекта, допускающих классификацию любым способом при $n=8$.

| | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ | $x_7$ | $x_8$ |
|---|---|---|---|---|---|---|---|---|
| $\widetilde{\alpha}_1$ | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| $\widetilde{\alpha}_2$ | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 0 |
| $\widetilde{\alpha}_3$ | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |

Покажем, что $VCD(BDT(\mu,n))$ может быть больше, чем $\max\{\log_2 n, \mu\}$. Пусть $\mu=3$. Укажем четыре объекта, которые могут быть расклассифицированы любым способом при $n=4$ $(\log_2 4=2)$.

| $x_1$ | $x_2$ | $x_3$ | $x_4$ |
|---|---|---|---|
| 0 | 0 | 0 | 1 |
| 0 | 0 | 1 | 0 |
| 0 | 1 | 0 | 0 |
| 1 | 0 | 0 | 0 |

Учитывая, что $BDT(\mu,n)<(\mu-1)!\,2^{\mu-1}n^{\mu-1}$ [13], получаем $VCD(BDT(\mu,n))<\log_2(\mu-1)!+(\mu-1)+(\mu-1)\log_2 n$.

Известна оценка $\ln n!<\left(n+\frac{1}{2}\right)\ln(n+1)-n$ [51], используя которую, получаем:

$$\log_2(\mu-1)!<\left(\mu-\frac{1}{2}\right)\log_2\mu-(\mu-1)\log_2 e<\mu\log_2\mu-(\mu-1),$$

$$VCD(BDT(\mu,n))<\log_2(\mu-1)!+(\mu-1)+(\mu-1)\log_2 n<\mu\log_2\mu+\mu\log_2 n=\mu\log_2(n\mu).$$

Таким образом, с учетом полученных неравенств, имеем: $VCD(BDT(\mu,n))<\mu\log_2(n\mu)$ (2.7) □.

**Следствие 2.2.** При любом $\mu=const$ и $n\to\infty$ имеет место оценка $VCD(BDT(\mu,n))=\Theta(\log_2 n)$.

Доказательство. Очевидно, что $\log_2 n\le\max(\mu,\log_2 n)$. Следовательно, $\log_2 n\le VCD(BDT(\mu,n))$ и $\log_2 n=O(VCD(BDT(\mu,n)))$. Воспользуемся

неравенством (2.7) получим $VCD\left(BDT\left(\mu,n\right)\right)<\mu\log_2\left(n\mu\right)<\mu\log_2\left(n^{\mu}\right)=\mu^2\log_2 n$. Следовательно, $VCD\left(BDT\left(\mu,n\right)\right)=O\left(\log_2 n\right)$. Поэтому $VCD\left(BDT\left(\mu,n\right)\right)=\Theta\left(\log_2 n\right)$ □.

Сравнивая порядок $VCD\left(BDT\left(\mu,n\right)\right)=\Theta\left(\log_2 n\right)$ с полученным в пункте 2.3.1 результатом $VCD\left(rDT_n\right)=\Theta\left(n^r\right)$, убеждаемся, что *оптимизация БРД по параметру $\mu$ - числу листьев более обоснована с точки зрения теории Вапника-Червоненкиса, чем оптимизация по рангу дерева, определенному в работе [144], и обеспечивает существенно более высокую скорость сходимости*.

## 2.4. Выводы

1. Получена оценка вероятности того, что при использовании бинарного решающего дерева, корректного на эмпирической таблице обучения, *будет сделан случайный вывод о свойстве целевой переменной $x_k$* для произвольного объекта $\tilde{x}=\left(x_1,x_2,\ldots,x_{k-1},x_{k+1},\ldots,x_n\right)$, выбранного равновероятно из множества булевых векторов длины $n-1$. Эта оценка имеет вид:

$$P_{BDT}\left(\mu,m,n\right)<\sum_{j=1}^{\mu}\left(n-r_j\right)C_n^{r_j}2^{-\left(m+r_j-2^{r_j}\right)}$$

Поэтому *неслучайность обнаружения по эмпирической выборке длины $m$ обобщенной закономерности, имеющей структуру БРД с $\mu$ листьями рангов $r_1,r_2,\ldots,r_{\mu}$, оценивается вероятностью большей*, *чем* $1-\sum_{j=1}^{\mu}\left(n-r_j\right)C_n^{r_j}2^{-\left(m+r_j-2^{r_j}\right)}$.

2. Показано, что оценка $C_n^r2^{-\left(m+r-2^r\right)}$ монотонно возрастает с ростом $r$. Поэтому *неслучайность обнаружения конъюнктивной закономерности, соответствующей ветви БРД ранга $r$, в случае выделенного целевого столбца оценивается тем большей вероятностью, чем меньше ранг $r$*. Это позволяет применять *критерий редукции ветвей в виде неравенства* $C_n^r2^{-\left(m+r-2^r\right)}<\varepsilon$ для заданной величины $\varepsilon\in\left(0;1\right)$.

3. Предложен функционал качества

$$\varphi(d)=\max_{1\le j\le \mu}\left( C_n^{r_j}\, 2^{-\left(m+r_j-2^{r_j}\right)} \right),$$

позволяющий оценивать случай, когда решение принимается ветвью РД максимального ранга. Установлено, что минимум функционала качества достигается на равномерных деревьях, что обосновывает необходимость редукции ветвей дерева, имеющих максимальный ранг.

4. На основе сравнения оценок сложности РД ограниченного ранга $VCD(rDT_n)=\Theta(n^r)$ и РД с ограниченным числом листьев $VCD(BDT(\mu,n))=\Theta(\log_2 n)$ установлено следующее. Оптимизация БРД по числу листьев более обоснована с точки зрения теории Вапника-Червоненкиса и обеспечивает существенно более высокую скорость равномерной сходимости при обучении, чем оптимизация по рангу дерева [145].

# Раздел 3
# АЛГОРИТМЫ СИНТЕЗА И ПРИНЯТИЯ РЕШЕНИЙ ЭМПИРИЧЕСКИМ РЕШАЮЩИМ ЛЕСОМ

## 3.1. Эмпирический решающий лес: основные определения

Основой многих современных алгоритмов синтеза решающих деревьев является “жадный” метод разделения и захвата (divide-and-conquer method). Этот метод не позволяет переоценить полученное разбиение после выбора признакового предиката во внутреннюю вершину РД, т.е. в случае неудачи – вернуться и выбрать для разбиения другой признаковый предикат (признак). И в целом, большинство существующих методов обучения при построении решающих деревьев учитывают лишь часть признаков, не используя всю заданную начальную информацию.

Предлагаемый далее новый алгоритм построения индуктивной модели эмпирического решающего леса (ЭРЛ) реализует стратегию, близкую по логике к стратегии *“предредукции” (prepruning) или “ранней остановки” с “возвратом”*. Построение эмпирического решающего леса направлено на поиск системы ветвей леса, правильно классифицирующей все объекты непротиворечивой обучающей выборки на основе конъюнктивной закономерности заданного порогового ранга. Под *критерием “ранней остановки”* здесь можно понимать невыполнение ограничения на глубину *хотя бы одной ветви* РД; в результате создается “ссылка” на следующее решающее дерево. Но понимать в качестве “возврата” построение решающего дерева по “ссылке” с пересмотром признаков, вообще говоря, неверно. При переходе по ссылке в первую очередь используются признаки, еще не участвовавшие в обучении, и осуществляется построение нового, дополнительного РД как очередной компоненты эмпирического решающего леса. Предыдущее РД сохраняется как составная часть системы закономерностей, определяющей ЭРЛ.

Таким образом, индуктивная модель ЭРЛ является совокупностью решающих деревьев вместе с некоторым правилом их использования при распознавании объектов. В качестве такого правила далее формулируются условия перехода от одного РД к следующему. При этом уточняется порядок решающих деревьев и собственно переход, как ссылка на корневую вершину следующего дерева.

**Определение 3.1.** *Областью отказа эмпирического бинарного решающего дерева* называется интервал $N_{K_j}$, соответствующий ветви дерева с рангом $r_j > r$, где $r$ - заданное значение, ограничивающее ранг.

Предполагается, что в область отказа эмпирического бинарного решающего дерева попадают объекты, вызывающие излишнее усложнение структуры БРД (рост числа листьев, рост ранга отдельных ветвей).

**Определение 3.2.** *Ссылкой $c_{12}$ дерева $d_1$ на дерево $d_2$* называется указатель на корневую вершину дерева $d_2$, размещенный в каждом листе дерева $d_1$, соответствующем некоторой области отказа.

**Определение 3.3. У**порядоченный набор эмпирических решающих деревьев $D_r = (d_1, d_2, \ldots, d_q)$ со ссылками $c_{12}, c_{23}, \ldots, c_{q-1q}$ называется *эмпирическим решающим лесом*.

Очевидно, что эмпирический решающий лес определяет решающие правила распознавания принадлежности объектов классам.

**Определение 3.4.** Эмпирический решающий лес называется *r-корректным относительно таблицы* $T_{mn}$, если входящие в него деревья $(d_1, d_2, \ldots, d_q)$ не содержат ветвей ранга, превышающего *r*, последнее по порядку дерево $d_q$ не имеет ветвей, соответствующих областям отказа и решающее правило, соответствующее эмпирическому решающему лесу, безошибочно определяет класс каждого объекта из таблицы $T_{mn}$. Если же эмпирический решающий лес ошибочно классифицирует хотя бы один объект таблицы $T_{mn}$, он называется *r-некорректным относительно этой таблицы*.

Обозначим ДНФ $\bigvee\limits_{j\in W_\omega^p} K_j^p = F^p(\omega)$, где $p=\overline{1,q}$ - индекс дерева $d_p \in D_r$; $K_j^p$ - конъюнкция, соответствующая ветви дерева $d_p$; $W_\omega^p$ - множество номеров ветвей, имеющих метку $\omega$ в дереве $d_p$.

**Определение 3.5.** ДНФ $\bigvee\limits_{p=1}^{q} F^p(\omega) = H_{D_r}(\omega)$ называется *множественным логическим описанием класса* $\omega$ *по эмпирическому лесу* $D_r$.

Множественное логическое описание может рассматриваться и как решающая функция для класса, однако она не определяет, какой именно конъюнкцией ДНФ принимается решение.

Разные деревья леса могут быть построены на различных и даже непересекающихся подмножествах признаков. Это, например, следует из результатов, полученных при изучении тупиковых тестов булевых таблиц [46]. Поэтому множественные логические описания для двух различных классов могут оказаться неортогональными, что вызовет неоднозначность решения.

Отмеченные свойства множественных логических описаний классов обосновывают необходимость разработки специальных процедур синтеза решающих правил на основе эмпирического решающего леса, обеспечивающих однозначный вывод решений.

## 3.2. *DFBSA* – последовательный алгоритм синтеза эмпирического решающего леса по ссылкам

В этом пункте описывается новый *алгоритм синтеза r-редуцированного эмпирического леса по ссылкам*, именуемый далее ***DFBSA*** (***D***ecision ***F***orest ***B***uilding ***S***equencing ***A***lgorithm). Главная идея синтеза решающих правил на основе совокупности решающих деревьев состоит в том, что при попадании классифицируемого объекта, не участвовавшего в обучении, в область отказа дерева $d_1$, классификация этого объекта должна осуществляться другим деревом

$d_2$ на основе конъюнктивной закономерности ограниченного ранга. Аналогично, решающее дерево $d_2$ “ссылается” на РД $d_3$ эмпирического решающего дерева и.т.д. Такой подход, конечно, предполагает задание некоторого упорядочения решающих деревьев ЭРЛ $d_1, d_2, \ldots, d_q$. Схема эмпирического леса представлена на рисунке 3.1.

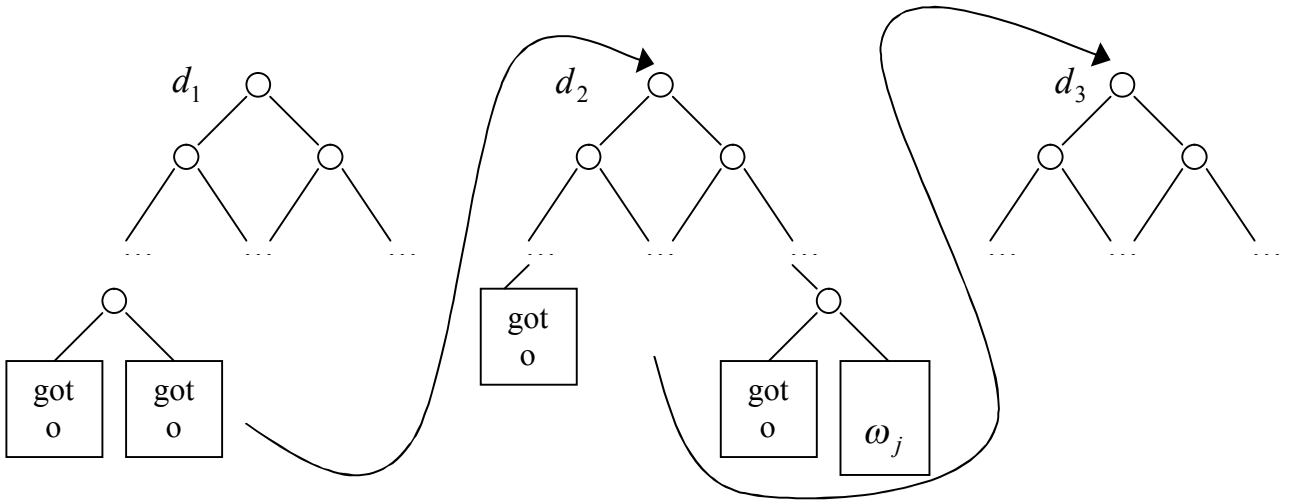


Рис. 3.1. Схема эмпирического решающего леса

Перейдем к строгой формулировке процедур, реализующих указанную идею.

*Этапы синтеза r-редуцированного эмпирического леса:*

$1^0$. Синтезируется дерево $d_1$ одним из известных методов (см. пункт 1.2). Если оно корректно на $T_{mn}$ и ранги его ветвей не превышают заданный ранг $r$, то синтез завершен, и лес состоит из одного дерева $D_r = (d_1)$. Иначе – перейти на $2^0$.

$2^0$. Обозначим через $X = \{x_1, x_2, \ldots, x_n\}$ - исходное множество признаков. Пусть $X_{d_1}$ - множество признаков, использованных для синтеза дерева $d_1$. Если дерево $d_1$ имеет ветвь ранга, больше чем $r$, то эта ветвь *редуцируется* со ссылкой на дерево $d_2$. Редукция ветви РД состоит в замене $(r+1)$-ой вершины терминальной вершиной специального вида – ссылкой. Конъюнкция, соответствующая редуцированной ветви, определяет область отказа. Решающее дерево, имеющее ссылки (непустые области отказа), называется далее *r-*

*редуцированным*. Ссылки на РД $d_2$ устанавливаются для всех областей отказа РД $d_1$. Затем синтезируется решающее дерево $d_2$, причем при ветвлении сначала используются признаки множества $X \setminus X_{d_1}$, если оно не пусто и признаков достаточно для синтеза дерева $d_2$. Если $X \setminus X_{d_1} = \varnothing$, то изменяется порядок выбора признаков по сравнению с порядком, использованным при синтезе РД $d_1$. Пусть $N_{K_{j1}}$ - область отказа дерева $d_1$. $T_2 = T_{mn} \cap N_{K_{j1}}$ - объекты разных классов выборки $T_{mn}$, попавшие в область отказа. Дерево $d_2$ строится по таблице $T_2$, а только затем достраивается на объектах $T_{mn} \setminus T_2$. Получается, вообще говоря, *r*-некорректное дерево $d_2$ и лес $D_r = (d_1, d_2)$. Если $D_r$ корректен, то синтез завершен. Иначе – синтезируется дерево $d_3$ (повторяются действия шага $2^0$), если не выполняется одно из условий остановки синтеза, описанное ниже.

Рисунок 3.2 иллюстрирует основные этапы синтеза каждого РД эмпирического решающего леса.

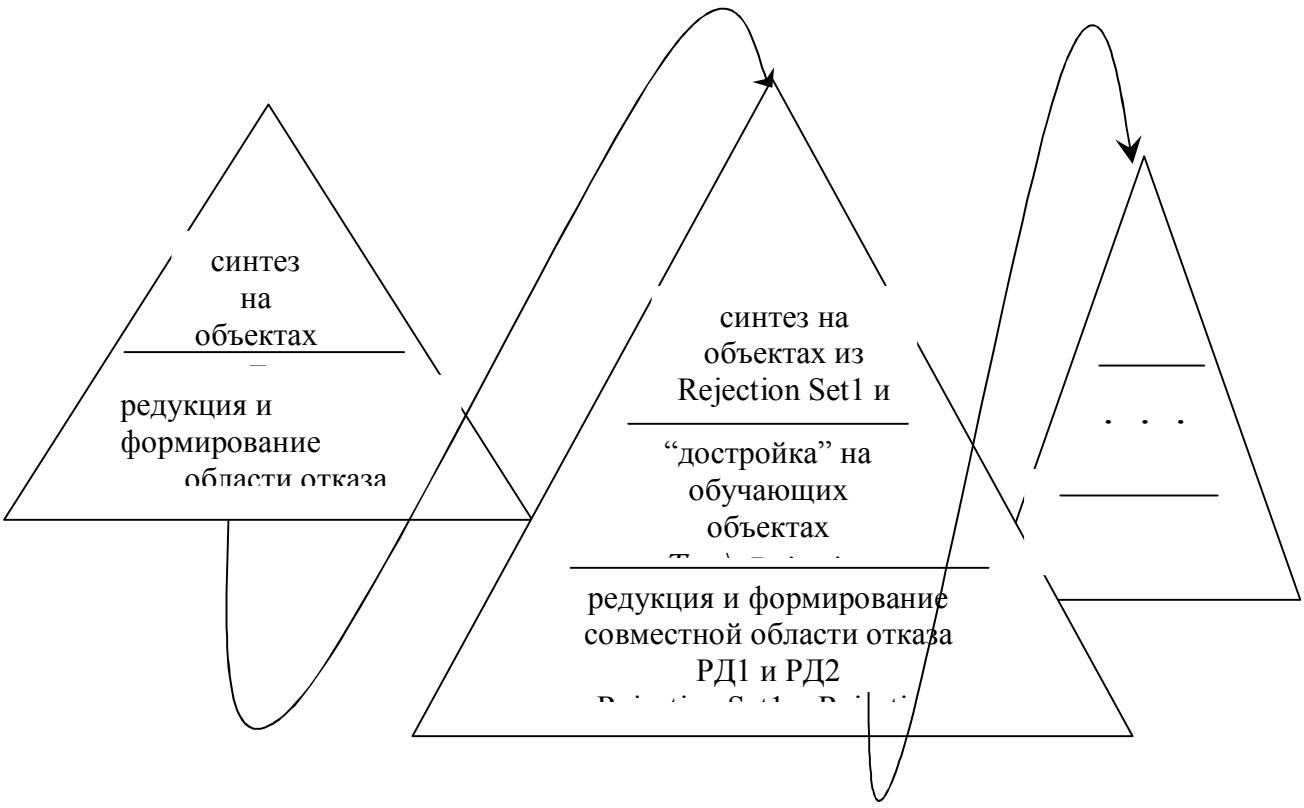


Рис 3.2. Основные этапы синтеза эмпирического решающего леса

*Критерии остановки синтеза эмпирического решающего леса*:

1. получен *r*-корректный эмпирический лес;
2. попытки синтеза новых деревьев выполнены более заданного числа раз;
3. превышен допустимый объем памяти, который могут занимать решающие деревья ЭРЛ.

Решающие деревья построенного эмпирического леса содержат концевые вершины двух видов: концевую вершину, содержащую метку класса (лист в обычном понимании) и концевую вершину, содержащую отказ от обучения – cut (редукция или отсечение).

Следующая теорема представляет *критерий существования r-корректного эмпирического решающего леса* для заданной непротиворечивой обучающей таблицы $T_{mn}$.

**Теорема 3.1.** Для существования *r*-корректного эмпирического решающего леса относительно таблицы обучения $T_{mn}$ необходимо и достаточно, чтобы для каждого объекта $\tilde{x} \in T_{mn}$ существовал интервал $N_{\tilde{x}}^r$ ранга не больше, чем *r* такой, что $\tilde{x} \in N_{\tilde{x}}^r$ и во множестве $N_{\tilde{x}}^r \cap T_{mn}$ содержались объекты только одного и того же класса.

Доказательство. Необходимость. Если эмпирический решающий лес является *r*-корректным, то каждый объект $\tilde{x} \in T_{mn}$ правильно классифицируется ветвью ранга больше чем *r* хотя бы одной ветвью некоторого дерева леса. Этой ветви соответствует интервал $N_{\tilde{x}}^r$ и ветвь имеет метку, а не ссылку (признак редукции). Поэтому все объекты подтаблицы $N_{\tilde{x}}^r \cap T_{mn}$ принадлежат одному и тому же классу, определяемому указанной меткой.

Достаточность. Полагая выполненным условие теоремы, укажем процедуру синтеза *r*-корректного эмпирического леса. Пусть $\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_m$ - совокупность всех объектов таблицы $T_{mn}$; $\omega(\tilde{x}_1), \omega(\tilde{x}_2), \ldots, \omega(\tilde{x}_m)$ - классы, которым принадлежат объекты; $N_{\tilde{x}_1}^r, N_{\tilde{x}_2}^r, \ldots, N_{\tilde{x}_m}^r$ - интервалы, определяемые условием теоремы; $K_1^r, K_2^r, \ldots, K_m^r$ - конъюнкции, соответствующие этим интервалам. Построим *m*

бинарных решающих деревьев, представленных на рисунке 3.3 так, что *j*-ое БРД правильно классифицирует объект $\tilde{x}_j$ ветвью, соответствующей конъюнкции $K_j^r$ (лист $\omega(\tilde{x}_j)$).

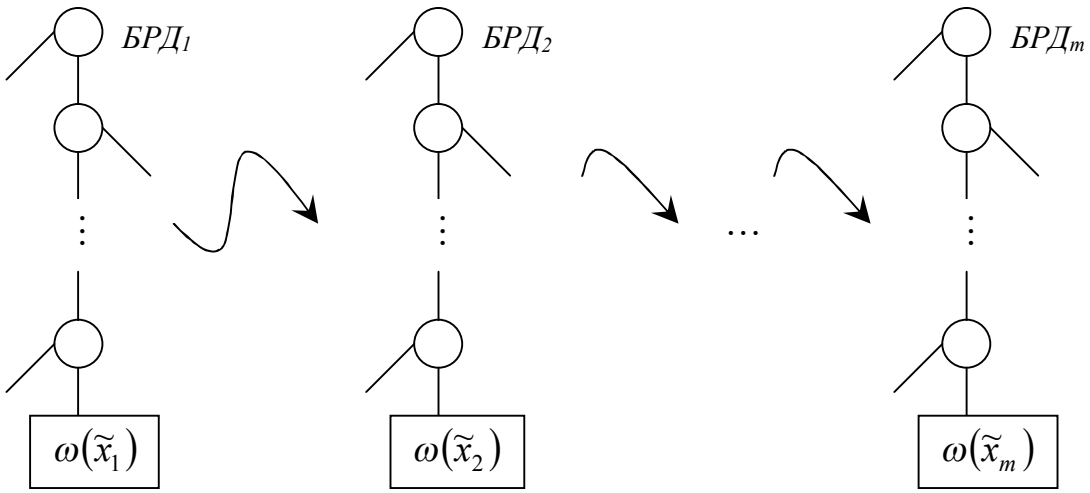


Рис.3.3. *m* бинарных решающих деревьев.

Остальные листья *j*-го БРД, $j = 1,2,\ldots,m$, помечаются либо меткой некоторого класса, если соответствующая листу ветвь определяет интервал, в который попадают объекты только одного и того же класса из $T_{mn}$, либо ссылку на следующее дерево. В *БРД*$_m$ вместо ссылок можно поставить метку любого класса. Очевидно, построенный эмпирический решающий лес правильно классифицирует все объекты из $T_{mn}$. □

### 3.3. Алгоритмы принятия решений эмпирическим решающим лесом

С учетом структуры построенного эмпирического леса можно предложить несколько различных *алгоритмов принятия решений r-корректным эмпирическим лесом*:

- *Последовательный алгоритм с переходами по "ссылкам"* реализует условный переход на первое по порядку ссылок дерево

эмпирического леса, которое для заданного объекта позволяет принять решение с допустимой оценкой вероятности неслучайности конъюнктивной закономерности (если пороговый ранг *r* заранее выбран с учетом такой допустимой оценки, последнее равносильно “выходу” на метку класса).

- *Алгоритм принятия решений на основе наиболее “компетентной” ветви* предполагает просмотр всех деревьев эмпирического леса и выбор ветви с меткой класса, соответствующей интервалу, в который попадают классифицируемые объекты, и имеющей минимальный ранг. Проигрывая в скорости получения решения, этот алгоритм позволяет выбрать ветвь с наилучшей статистической оценкой.
- *Алгоритм “голосования” ветвей* использует широко применяемое мажоритарное правило, являющееся упрощенным вариантом процедуры алгебраической коррекции.

Опишем подробно каждый из них.

*Алгоритм принятия решений на основе построенного r-корректного эмпирического леса по ссылкам* состоит в следующем. Предоставляется описание объекта – набор значений *n* булевых признаков. Указатель устанавливается на первое дерево. Согласно значениям признаков осуществляется “прохождение” ветви и определение соответствующего листа. Если этот лист помечен меткой класса – объект опознан. Иначе, если это ссылка, указатель устанавливается на следующее (определенное порядком синтеза при выполнении алгоритма ***DFBSA***) дерево. Если в итоге будет получена терминальная ссылка без метки класса – принимается решение об отказе от опознавания предъявленного объекта.

Очевидно, что результат работы этого алгоритма определяется построенным лесом и предопределенным порядком просмотра деревьев леса. Переход на новое дерево соответствует классификации заново, по измененной системе признаков, другой ветвью. Предыдущий отказ от решения *принципиально никакого значения для последующей классификации не имеют*.

*Алгоритм распознавания на основе наиболее "компетентной" ветви эмпирического решающего леса*. Указатель устанавливается на первое РД решающего леса. Согласно значениям признаков осуществляется "прохождение" ветви и определение соответствующего листа. Если лист содержит метку класса, то ранг ветви запоминается. Далее осуществляется переход на следующее дерево в поисках ветви наименьшего ранга, оканчивающейся листом с меткой класса, которому принадлежит объект. Таким образом, объект оказывается распознанным ветвью, имеющей наилучшую статистическую оценку (наименьший ранг).

*Алгоритм распознавания на основе голосования ветвей эмпирического решающего леса*. Каждое дерево ЭРЛ можно рассматривать как эксперта, принимающего решение в своей области "компетентности" (совокупности непересекающихся интервалов допустимого ранга). Каждый эксперт-дерево при поступлении описания объекта может либо проголосовать за определенный класс либо воздержаться от голосования (ветвь заканчивается меткой cut). Сначала указатель устанавливается на первое дерево ЭРЛ. Согласно значениям признаков осуществляется "прохождение" ветви и определение соответствующего листа. Если лист помечен меткой класса, эксперт-дерево голосует за этот класс. В результате алгоритма осуществляется просмотр всех деревьев эмпирического леса и подсчитывается число голосов, отданное экспертами-деревьями в пользу каждого класса. Класс, набравший наибольшее число голосов, определяет метку класса объекта, поступившего для распознавания. Данный алгоритм принятия решений эффективен в случае двух классов. В противном случае может оказаться, что установить мажоритарный класс достаточно сложно.

Все эти алгоритмы распознавания целесообразно применять, когда построен корректный эмпирический лес. В случае некорректного эмпирического леса целесообразно применение алгебраического корректора (см. пункт 3.4 данного раздела диссертации).

Далее рассмотрим *построение непротиворечивых логических описаний классов в виде ДНФ по эмпирическому решающему лесу*. Для описания классов в виде ДНФ, привычных и полезных при анализе единичных РД, в ЭРЛ фигурирует

существенно более сложная конструкция. Прежде, чем приступить к ее построению, заметим, что *процесс принятия решения и его описание - существенно разные вещи*. Корректный ЭРЛ определяет разбиение куба $B^n$ на области, соответствующие классам и, соответственно, однозначные решающие правила (алгоритмические функции классов). Действительно, из определения алгоритма распознавания с переходами по "ссылкам", очевидно, что для любого объекта $\tilde{x} \in B^n$ либо однозначно определяется его метка класса $\omega(\tilde{x})$ (в этом случае завершится выполнение алгоритма распознавания с переходами по ссылкам) либо ссылка на следующее дерево (метка листа cut), причем для корректного леса в последнем по порядку РД ссылок нет, следовательно, $\tilde{x}$ получает ровно одну метку класса.

Пусть $W_i : B^n \to \{\omega_1, \omega_2, \ldots, \omega_\ell, \Delta\}$ - алгоритмическое отображение, определяемое $i$-м РД леса $D_r = (d_1, d_2, \ldots, d_q)$, где $\ell$-число классов, $\Delta$-отказ от решения.

Областью компетентности $i$-го решающего дерева $d_i$ будем называть множество $CompTree(d_i) = \bigcup_{j=1}^{\ell} \{\tilde{x} \in B^n \mid \omega_j(\tilde{x}) \neq \Delta\}$, а областью компетентности упорядоченного множества $d_1, d_2, \ldots, d_q$ решающих деревьев - $CompSet(q) = \bigcup_{i=1}^{q} CompTree(d_i)$.

Покажем, что $CompSet(1) \subseteq CompSet(2) \subseteq \ldots \subseteq CompSet(q)$.

Действительно, $CompSet(1) = CompTree(d_1)$ - множество, попав в которое классифицируемый объект обязательно будет опознан. В противном случае, он попадает в область отказа первого решающего дерева $d_1$: $B^n \setminus CompSet(1)$ и по "ссылке" переадресовывается второму решающему дереву $d_2$. В РД $d_2$ этот объект может попасть в область $CompTree(d_2)$ и тогда, очевидно, $CompSet(1) \subset CompSet(2)$. Если же любой объект $\tilde{x} \in B^n \setminus CompSet(1)$ не попадает в

область компетентности $CompTree(d_2)$, то $CompSet(1)=CompSet(2)$. Аналогично, $CompSet(j)\subseteq CompSet(j+1)$ для $j=2,3,\ldots,q-1$.

**Теорема 3.2**. Эмпирический решающий лес $\{d_1,d_2,\ldots,d_q\}$ корректен относительно стандартной таблицы обучения $T_{mn\ell}$ тогда и только тогда, когда $CompSet(q)=B^n$.

Доказательство. Достаточность. Если $CompSet(q)=B^n$, то любой объект $\tilde{x}\in T_{mn\ell}$ принадлежит хотя бы одной области компетентности и тогда он правильно классифицируется ветвью ограниченного ранга.

Необходимость. Предположим, $CompSet(q)\neq B^n$. Тогда $B^n\setminus CompSet(q)$ - непустая область отказа. Отказ формируется только тогда, когда ветвь редуцируется. Редукция в свою очередь, происходит только тогда, когда в интервал соответствующий редуцируемой ветви, попадает неправильно классифицируемый объект. □

Очевидно, каждое последующее по порядку, определенному алгоритмом принятия решений, РД леса "вычисляет" отображение $W_i$ только на сужении $D_{i-1}=B^n\setminus CompSet(i-1)$, которое может быть описано в виде логической формулы $F_{i-1}(x_1,x_2,\ldots,x_n)=\begin{cases}1, & \tilde{x}\in D_{i-1}\\ 0, & \tilde{x}\notin D_{i-1}\end{cases}$. Эта формула может быть представлена некоторой ДНФ $D_{i-1}=K_1^{i-1}\vee K_2^{i-1}\vee\ldots\vee K_\upsilon^{i-1}$. Если в $d_i$ определено решающее правило $L^i_{1,\omega_j}\vee L^i_{2,\omega_j}\vee\ldots\vee L^i_{u_{\omega_j},\omega_j}$ для некоторого класса $\omega_j$ в виде ДНФ, то для логического описания этого класса следует использовать выражение $\left(\mathrm{K}_1^{q-1}\vee\mathrm{K}_2^{q-1}\vee\ldots\vee\mathrm{K}_{\beta_{q-1}}^{q-1}\right)\left(L^q_{1,\omega_j}\vee L^q_{2,\omega_j}\vee\ldots\vee L^q_{u_{\omega_j},\omega_j}\right)$. Но, несмотря на это, сложные формулы определяют не используемые для вывода решения конъюнкции, а лишь логическое описание решения вместе с областями компетентности.

*Построение ДНФ класса по ЭРЛ как описания решения*.

$1^0$. Взять все ветви первого решающего дерева леса, помеченные метками классов, и "расписать" конъюнкции по классам, получая ДНФ $D_1(\omega_j)$, $j=\overline{1,\ell}$, как

описание *j*-го класса по первому РД. Записать ДНФ $R_1$, соответствующую ветвям, помеченным меткой cut ($R_1$ - описание области отказа первым РД эмпирического леса).

$i^0$. Пусть построены $D_{i-1}(\omega_j)$, $R_{i-1}$ ($R_{i-1}, i \neq 1$ - описание пересечения областей отказа $i-1$ решающих деревьев эмпирического леса). По $d_i$ построим описание j-го класса в виде рекурсивной процедуры: $D_i(\omega_j) = D_{i-1}(\omega_j) \vee (R_{i-1} \wedge D_i(\omega_j))$.

Заметим, что решение, определяемое ЭРЛ согласно алгоритму распознавания с переходами по ссылкам, *всякий раз принимается одной конъюнкцией ограниченного ранга*, но, возможно, в условиях отказа предыдущих по порядку РД леса. Важным *свойством* любой такой *принимающей решение конъюнкции является её корректность на обучающей таблице* $T_{mn\ell}$: она выполняется только на объектах одного класса.

## 3.4. Алгебраическая алгоритмическая модель коррекции *r*-некорректного эмпирического леса

Сначала изложим ряд понятий алгебраического подхода Ю.И. Журавлева [19, 31, 32, 33, 34], которые необходимы для понимания полученных ниже результатов.

Пусть далее $K_1, K_2, \ldots, K_\ell$ - классы допустимых объектов $S \in \{S\}$; $(P_1, P_2, \ldots, P_\ell) = \widetilde{P}_\ell$ - основные предикаты; $(S \in K_j) \Leftrightarrow (P_j(S) = 1)$. $I \in \{I\}$ - стандартная допустимая начальная информация; $\{S_1, S_2, \ldots, S_q\} = \widetilde{S}_q$ - произвольный набор допустимых объектов; $A(I, \widetilde{S}_q, \widetilde{P}_\ell) = \|\beta_{ij}\|_{q \times \ell}$ - алгоритм распознавания; $A_c(I, \widetilde{S}_q, \widetilde{P}_\ell) = \|\alpha_{ij} : \alpha_{ij} = P_j(S)\|_{q \times \ell}$ - корректный алгоритм; $(S'_1, S'_2, \ldots, S'_m) = \widetilde{S}'_m$ - обучающая выборка; $P_j(S_k)$ известны для $j = \overline{1,\ell}; k = \overline{1,m}$; $Z(I, \widetilde{S}_q, \widetilde{P}_\ell)$ - задача распознавания.

Рассмотрим класс некорректных алгоритмов распознавания $\{T\}$, использующих БРД, причем $T\left(I,\widetilde{S}_q,\widetilde{P}_\ell\right)=\left\|\beta_{ij}\right\|_{q\times\ell}$. Пусть $\widetilde{p}=\{p_1,p_2,\ldots\}$ - счетное множество признаковых предикатов.

**Определение 3.6** [33]. Множество признаковых предикатов называется *полным относительно задачи Z*, если для любой пары различных объектов $S_1$ и $S_2$ из разных классов найдется предикат $p_i$ такой, что $p_i(S_1)\neq p_i(S_2)$.

Заметим, что корректность БРД на начальной информации, вообще говоря, не обеспечивает корректности для произвольных допустимых объектов $\widetilde{S}_q\subset\{S\}$. Однако, в линейном замыкании $L(T)$ корректный алгоритм содержится.

**Теорема 3.3** [31, 33]. Каждый алгоритм $A\in\{A\}$ представим как последовательное выполнение алгоритмов $B$ и $C$, где $B\left(I,\widetilde{S}_q\right)=\left\|a_{ij}\right\|_{q\times\ell}$, $a_{ij}$ - действительные числа, $C\left(\left\|a_{ij}\right\|\right)=\left\|\beta_{ij}\right\|_{q\times\ell}$; $\beta_{ij}\in\{0,1,\Delta\}$; $B=B(A)$; $C=C(A)$.

Обычно используют запись $A=B\circ C$ и называют $B$ - *распознающим оператором*, а $C$ - *решающим правилом*.

**Определение 3.7** [33]. Решающее правило $C$ называется *корректным* на $\{S\}$, если для всякого конечного набора $\widetilde{S}_q\subset\{S\}$ существует хотя бы одна числовая матрица $\left\|a_{ij}\right\|_{q\times\ell}$, такая что $C\left(\left\|a_{ij}\right\|\right)=\left\|\alpha_{ij}\right\|_{q\times\ell}$.

Множество $\{A\}$ порождает множества $\{B\}$ - распознающих операторов и множество $\{C\}$ решающих правил. На множестве $\{B\}$ вводятся операции сложения и умножения на скаляр. Пусть $B'$, $B''\in\{B\}$, $b_0$ - число,

$$B'\left(I,\widetilde{S}_q\right)=\left\|a'_{ij}\right\|_{q\times\ell};\qquad B''\left(I,\widetilde{S}_q\right)=\left\|a''_{ij}\right\|_{q\times\ell};\qquad b_0B'\left(I,\widetilde{S}_q\right)=\left\|b_0a'_{ij}\right\|_{q\times\ell};$$

$$\left(B'+B''\right)\left(I,\widetilde{S}_q\right)=\left\|a'_{ij}+a''_{ij}\right\|_{q\times\ell}\ (3.1).$$

*Линейным* называется *замыкание* $L\{B\}$ множества $\{B\}$ относительно операций (3.1).

**Определение 3.8** [33]. Если множество матриц $\left\{B\left(I,\widetilde{S}_q\right)\right\}$ ($B$ пробегает множество $\widetilde{B}$) содержит базис в пространстве числовых матриц размера $q\times\ell$, то задача распознавания $Z\left(I,\widetilde{S}_q,\widetilde{P}_\ell\right)$ называется *полной относительно* $\widetilde{B}$.

**Теорема 3.4** [31, 33]. Если множество $\{Z\}$ состоит лишь из задач полных относительно $\widetilde{B}$, то линейное замыкание $L\left\{\widetilde{B}C^*\right\}$ ($C^*$ - произвольное фиксированное корректное решающее правило) является корректным относительно $\{Z\}$.

*Корректность* $L\{T\}$ [19]. Будем рассматривать только задачи $\left\{Z=Z\left(I,\widetilde{S}_q,\widetilde{P}_\ell\right)\right\}$, удовлетворяющие следующим условиям:

1) $\widetilde{S}'_m\cap\widetilde{S}_q=\varnothing$; $I$ - непротиворечивая стандартная начальная информация; все объекты в $\widetilde{S}_q$ различны;

2) существует полное для $\{Z\}$ множество $\widetilde{p}$ признаковых предикатов.

**Определение 3.9** [16, 19]. Распознающим оператором (РО) $B_T$ типа решающего дерева называется РД, каждый лист которого помечен вектором $\widetilde{\theta}^k=\left(\theta_1^k,\theta_2^k,\ldots,\theta_\ell^k\right)$, $k=\overline{1,\mu}$, где $\theta_j^k$ - число объектов из обучающей информации, принадлежащий классу $K_j, j=\overline{1,\ell}$, и соответствующих по значениям признаков листу (ветви) $k$.

**Теорема 3.5** [16]. Линейное замыкание $L\{T\}$ класса алгоритмов вида $B_T\cdot C^*$, где $C^*$- произвольное решающее правило, является корректным на множестве задач с непротиворечивой начальной информацией и полным множеством признаковых предикатов.

Рассмотрим алгоритм, основанный на построении линейного замыкания множества некорректных алгоритмов распознавания из класса РД-моделей [33], которое является корректным на множестве задач с непротиворечивой начальной информацией и полным множеством признаковых предикатов.

Этот алгоритм следует применять, когда эмпирический лес $D_r = (d_1, d_2, \ldots, d_q)$ не является *r*-корректным.

Невозможность построения *r*-корректного ЭРЛ во многих случаях определяется тем, что обучающая выборка является короткой, и, требуется вводить ограничение на допустимый ранг ветвей $r$ так, что $r \le 5$.

Для каждого дерева некорректного леса $D_r$ каждый лист в соответствии с определением 3.9 дополняется информацией $\tilde{\theta}^k(d_s), s = \overline{1,q}$; $k = \overline{1,\mu(d_s)}$, где $\mu(d_s)$-число листьев дерева $d_s$.

Для контрольной выборки $Y_h = \{\tilde{y}_1, \tilde{y}_2, \ldots, \tilde{y}_h\}$ вычисляется $q$ матриц $B(d_s, Y_q) = \left\| b_{ij}^s \right\|_{h \times \ell}$, где $b_{ij}^s = \theta_j^{k_i}(d_s)$; $k_i$- ветвь дерева $d_s$, классифицирующая объект $\tilde{y}_i, i = \overline{1,q}; j = \overline{1,\ell}$; $\ell$ - число классов. Обозначим $\tilde{\tau} = \tau_1, \tau_2, \ldots, \tau_q$ - набор скаляров. Линейная комбинация $B^A = \left\| b_{ij}^A \right\|_{h \times \ell} = \sum_{s=1}^{q} \tau_s \left\| b_{ij}^s \right\|_{h \times \ell}$ определяет *распознающий оператор эмпирического леса*.

Сначала полагается $\tau_s = 1, s = \overline{1,q}$. Пусть объект $\tilde{y}_i$ контрольной выборки принадлежит классу $j^*$. Если для всех $i = \overline{1,h}$ выполняется

$$\max_{1 \le j \le \ell} b_{ij}^A = b_{ij^*}^A \qquad (3.2),$$

то распознающий оператор эмпирического леса обеспечивает безошибочную классификацию всей контрольной выборки. Иначе условие (3.2) нарушается $\varphi^A(Y_h) = \delta$ раз. Тогда решается задача минимизации эмпирического функционала $\varphi^A(Y_h)$: $\min_{\tilde{\tau} \in W} \varphi^A(Y_h)$ (3.3), где $W$ – область допустимых значений скаляров. Если минимум (3.3) достигается на наборе $\tilde{\tau}^*$, то оператор $B_*^A = \sum_{s=1}^{q} \tau_s^* \left\| b_{ij}^s \right\|_{h \times \ell}$ называется *скорректированным* (по контрольной выборке).

**Пример 3.1.** Пусть в задаче с двумя классами ($\ell = 2$) и лесом из двух деревьев ($q = 2$) предъявлена контрольная выборка из трех объектов $Y_3 = \{\tilde{y}_1, \tilde{y}_2, \tilde{y}_3\} (h = 3)$; принадлежность классам указана в таблице 3.1.

Таблица 3.1. Принадлежность объектов контрольной выборки к классам

| Объект | $\widetilde{y}_1$ | $\widetilde{y}_2$ | $\widetilde{y}_3$ |
|---|---|---|---|
| номер класса | 2 | 1 | 2 |

Распознающие операторы двух деревьев $d_1$ и $d_2$ леса имеют вид:

$$B_1 = \begin{Vmatrix} 0 & 2 \\ 5 & 0 \\ 4 & 2 \end{Vmatrix} = \left\| b_{ij}^1 \right\|_{3\times 2};\ B_2 = \begin{Vmatrix} 0 & 2 \\ 4 & 2 \\ 2 & 3 \end{Vmatrix} = \left\| b_{ij}^2 \right\|_{3\times 2}.$$

Для начальных значений скаляров $\tau_1 = \tau_2 = 1$ РО эмпирического леса имеет вид:

$$B^A = B_1 + B_2 = \begin{Vmatrix} 0 & 4 \\ 9 & 2 \\ 6 & 5 \end{Vmatrix} = \left\| b_{ij}^A \right\|_{3\times 2}.$$

Его использование при классификации объектов контрольной выборки $Y_3$ приводит к ошибке: $b_{31}^A = 6 > 5 = b_{32}^A$, но $\widetilde{y}_3$ должен быть отнесен к классу с номером 2 и должно выполняться $b_{31}^A < b_{32}^A$.

Система неравенств для произвольных $\tau_1, \tau_2$ имеет вид:

$$\begin{cases} \tau_1 b_{11}^1 + \tau_2 b_{11}^2 < \tau_1 b_{12}^1 + \tau_2 b_{12}^2 \\ \tau_1 b_{21}^1 + \tau_2 b_{21}^2 > \tau_1 b_{22}^1 + \tau_2 b_{22}^2 \\ \tau_1 b_{31}^1 + \tau_2 b_{31}^2 < \tau_1 b_{32}^1 + \tau_2 b_{32}^2 \end{cases} \qquad (3.4).$$

Подставляя коэффициенты из $B_1$ и $B_2$, получаем:

$$\begin{cases} 0 < 2\tau_1 + 2\tau_2 \\ 5\tau_1 > -2\tau_2 \\ 2\tau_1 < \tau_2 \end{cases}.$$

Поскольку первые два неравенства истинны при любых $\tau_1, \tau_2 > 0$, то взяв $\tau_1 = 1$, получаем $\tau_2 > 2$. Пусть $\tau_1^* = 1, \tau_2^* = 3$. Тогда скорректированный РО имеет вид:

$$B_*^A = \begin{Vmatrix} 0 & 8 \\ 17 & 6 \\ 10 & 11 \end{Vmatrix}_{3\times 2}$$

и обеспечивает правильную классификацию всех объектов выборки $Y_3$. □

Подчеркнём, что методы алгебраической коррекции следует применять только тогда, когда *r*-корректный лес построить не удаётся. В этом случае все листья редуцированных эмпирических деревьев леса, включая те, которым соответствовали ссылки на другие деревья, заменяются специальными пометками – числовыми векторами.

Алгебраическая коррекция является мощным математическим аппаратом, применимым к любому семейству эвристических алгоритмов, но в настоящей диссертации главное внимание уделено не этому вопросу. Коррекция эвристических деревьев, как показано в работе, может достигаться путем построения специальной последовательной процедуры синтеза эмпирического леса. При этом сохраняются полезные структурно-логические свойства алгоритмов распознавания, основанных на построении РД.

## 3.5. Оценка *VCD* класса решающих функций, представленных *r*-редуцированным эмпирическим лесом

В задачах поиска наилучшего решающего правила из некоторого класса РП при ограниченной выборке большую роль играет связь между сложностью выбранного класса и объемом выборки. Из теоретических исследований [3, 41] следует, что чем ỳже класс, тем меньше оценка вероятности ошибки распознавания новых объектов при использовании РП, выбираемого из этого класса. Проблема построения РП при большой размерности векторов, описывающих объекты и ограниченной выборке сводится к выбору класса правил, обладающего малой мерой сложности. В то же время класс должен быть

достаточно “богатым”, чтобы можно было достаточно точно решать прикладные задачи.

Получим оценку VCD *r*-редуцированного эмпирического леса. Заметим, что совокупность ветвей ЭРЛ с пометками одного и того же класса эквивалентна заданию некоторой ДНФ, и число различных решающих правил, определяемых *r*-редуцированным эмпирическим лесом, не больше числа различных ДНФ, состоящих из $\mu \cdot q$ конъюнкций ($\mu$ - наибольшее число листьев по всем РД, входящим в эмпирический лес, а $q$ - число деревьев леса), каждая из которых имеет ранг не более *r*.

Число различных конъюнкций, составленных из переменных $x_1, x_2, \ldots, x_n$ ранга не более *r*, равно $\sum_{i=1}^{r} 2^i C_n^i$. Покажем, что $\frac{2^r (n-r)^r}{r!} < \sum_{i=1}^{r} 2^i C_n^i < \frac{(2n)^{r+1}-1}{2n-1}$.

Получим оценку снизу. $\sum_{i=1}^{r} 2^i C_n^i > 2^r C_n^r = 2^r \cdot \frac{n \cdot (n-1) \cdot \ldots \cdot (n-r+1)}{r!} > 2^r \cdot \frac{(n-r)^r}{r!}$.

Получим оценку сверху. $\sum_{i=1}^{r} 2^i C_n^i = 2C_n^1 + \ldots + 2^r C_n^r = 2n + \ldots + 2^r \frac{n(n-1)\ldots(n-r+1)}{r!} <$

$< 2n + \ldots + 2^r n^r = \frac{(2n)^{r+1} - 2n}{2n-1} < \frac{(2n)^{r+1}-1}{2n-1}$.

Оценим VCD конечного класса $DNF(n, \mu, r)$ решающих правил, образованных дизъюнктивными нормальными формами, содержащими не более $\mu$ конъюнкций ранга не более *r*, состоящими из литералов *n* переменных.

**Теорема 3.6.** При $r \le \frac{n}{2}$ справедливо неравенство

$$\frac{\left(\left(\frac{2n}{r} - 2\right)^r - \mu\right)^\mu}{\mu!} < |DNF(n, \mu, r)| < \frac{1{,}5^\mu n^{r\mu}}{\mu!}.$$

Доказательство. Оценка мощности класса решающих правил $DNF(n, \mu, r)$ снизу получается, если рассмотреть только одинаковые по рангу $r > 1$ конъюнкции с одинаковым числом инверсий равным $\left[\frac{r}{2}\right]$. Обозначим множество

таких конъюнкций $K\left(n,r,\left[\frac{r}{2}\right]\right)$. Каждые две различные составленные из этих конъюнкций ДНФ определяют две различные функции из $P_2(n)$. Это следует из того, что для любой конъюнкции из $K\left(n,r,\left[\frac{r}{2}\right]\right)$ можно указать набор значений переменных, на котором она обращается в единицу, а любая отличная от нее конъюнкция обращается в нуль. Действительно, пусть $L=\bar{x}_{i_1}\cdot\bar{x}_{i_2}\cdot\ldots\cdot\bar{x}_{i_{[r/2]}}\cdot x_{i_{[r/2]+1}}\cdot\ldots\cdot x_{i_r}$ - произвольная конъюнкция ранга $r$ с ровно $\left[\frac{r}{2}\right]$ отрицательными литералами. Пусть $H$ - любая конъюнкция из множества $K\left(n,r,\left[\frac{r}{2}\right]\right)$, отличная от $L$. Если $H$ состоит из тех же переменных, что и $L$, то хотя бы одна из переменных $x_{i_1},x_{i_2},\ldots,x_{i_{[r/2]}}$ войдет в конъюнкцию $H$ без инверсии. Не теряя общности, пусть это будет переменная $x_{i_1}$. Тогда для любой точки $\tilde{\alpha}=\alpha_1,\alpha_2,\ldots,\alpha_n$ такой, что $\alpha_{i_1}=0,\alpha_{i_2}=0,\ldots,\alpha_{i_{[r/2]}}=0,\alpha_{i_{[r/2]+1}}=1,\ldots,\alpha_{i_r}=1$, будет иметь место $L(\tilde{\alpha})=1$, а $H(\tilde{\alpha})=0$. Пусть теперь в $H$ содержится хотя бы одна переменная $x_p$, не содержащаяся в $L$. Если переменная $x_p$ входит в $H$ без инверсии, то для точки $\tilde{\alpha}$ такой, что $\alpha_{i_1}=0,\alpha_{i_2}=0,\ldots,\alpha_{i_{[r/2]}}=0,\alpha_{i_{[r/2]+1}}=1,\ldots,\alpha_{i_r}=1,\alpha_p=0$, будет иметь место $L(\tilde{\alpha})=1$, а $H(\tilde{\alpha})=0$. Если же переменная $x_p$ входит в $H$ с инверсией, то для точки $\tilde{\alpha}$ такой, что $\alpha_{i_1}=0,\alpha_{i_2}=0,\ldots,\alpha_{i_{[r/2]}}=0,\alpha_{i_{[r/2]+1}}=1,\ldots,\alpha_{i_r}=1,\alpha_p=1$, будет иметь место $L(\tilde{\alpha})=1$, а $H(\tilde{\alpha})=0$.

Учитывая установленное свойство конъюнкций множества $K\left(n,r,\left[\frac{r}{2}\right]\right)$, получаем: $\left|K\left(n,r,\left[\frac{r}{2}\right]\right)\right|=C_r^{\left[\frac{r}{2}\right]}C_n^r$, $|DNF(n,\mu,r)|>C^{\mu}_{C_r^{\left[\frac{r}{2}\right]}C_n^r}$, $C_r^{\left[\frac{r}{2}\right]}C_n^r>\left(\frac{2n}{r}-2\right)^r$,

$$\left|DNF\left(n,\mu,r\right)\right| > C^{\mu}_{\left(\frac{2n}{r}-2\right)^r} > \frac{\left(\left(\frac{2n}{r}-2\right)^r-\mu\right)^{\mu}}{\mu!}.$$

Для оценки мощности класса $DNF\left(n,\mu,r\right)$ решающих правил сверху рассмотрим сумму

$$\sum_{i=1}^{r} 2^i C_n^i < 2n + \frac{(2n)^2}{2!} + \frac{(2n)^3}{3!} + \ldots + \frac{(2n)^r}{r!} < 2n + \frac{(2n)^2}{2!} + \frac{(2n)^3}{3!} + \frac{(2n)^4}{4!} + n^5 + \ldots + n^r.$$

Легко проверить, что при $n > 3$

$$\sum_{i=1}^{r} 2^i C_n^i < n + n^2 + \ldots + n^r = \frac{n^{r+1}-n}{n-1} < \frac{n^{r+1}}{n-1} = \frac{n}{n-1}\cdot n^r < 1{,}5n^r.$$

$$C^{\mu}_{1{,}5n^r} = \frac{\left(1{,}5n^r\right)!}{\mu!\left(1{,}5n^r-\mu\right)!} < \frac{\left(1{,}5n^r\right)^{\mu}}{\mu!} = \frac{1{,}5^{\mu}n^{r\cdot\mu}}{\mu!}. \ \square$$

**Следствие 3.1**. $\left|DNF\left(n,\mu,r\right)\right| = \Theta\left(n^{r\cdot\mu}\right)$ при любых заданных константах $r$, $\mu$ и $n \to \infty$.

Доказательство. По теореме 3.6 $\dfrac{\left(\left(\frac{2n}{r}-2\right)^r-\mu\right)^{\mu}}{\mu!} < \left|DNF\left(n,\mu,r\right)\right| < \dfrac{1{,}5^{\mu}n^{r\mu}}{\mu!}$, следовательно, поскольку $r$ и $\mu$ являются константами, $\left|DNF\left(n,\mu,r\right)\right| = O\left(n^{r\cdot\mu}\right)$.

Получим оценку снизу. $\left|DNF\left(n,\mu,r\right)\right| > \dfrac{\left(\left(\frac{2n}{r}-2\right)^r-\mu\right)^{\mu}}{\mu!} \geq \dfrac{\left(\left(\frac{2n}{r}-\frac{n}{r}\right)^r-\mu\right)^{\mu}}{\mu!}$ при $\dfrac{n}{r} \geq 2$. Это условие всегда выполняется для моделей с ограниченным рангом конъюнкций, а с учетом того, что $r$ - константа, $n > 2r$ тем более выполняется при $n \to \infty$.

$$|DNF(n,\mu,r)| > \frac{\left(\left(\frac{n}{r}\right)^r - \mu\right)^\mu}{\mu!} \geq \frac{\left(\left(\frac{n}{r}\right)^r - \frac{1}{2}\left(\frac{n}{r}\right)^r\right)^\mu}{\mu!} \quad \text{при} \quad \left(\frac{n}{r}\right)^r \geq 2\mu,$$ что при $n \geq 2r$ обеспечивается условием $2^{r-1} \geq \mu$ по смыслу задачи ($r$ - ранг, $\mu$ - число листьев)

$|DNF(n,\mu,r)| > \frac{\left(\frac{1}{2}\left(\frac{n}{r}\right)^r\right)^\mu}{\mu!} \geq \frac{n^{r\mu}}{\mu!\,2^\mu\, r^{r\mu}}$. Обозначая $\gamma = \mu!\,2^\mu\, r^{r\mu} = const$, получаем $n^{r\mu} = \gamma|DNF(n,\mu,r)|$, поэтому при $n \to \infty$ $n^{r\mu} = O(|DNF(n,\mu,r)|)$. □

**Следствие 3.2**. $VCD(DNF(n,\mu,r)) = O(\log_2 n)$ при заданных константах $\mu, r$ и $n \to \infty$.

Доказательство. С учетом теоремы 3.6 и неравенства (2.8), получим $VCD(DNF(n,\mu,r)) = \mu \log_2 \frac{3}{2} + r\mu \log_2 n - \log_2 \mu! < \mu \log_2 \frac{3}{2} \log_2 n + r\mu \log_2 n$ при $n > 2$. Поэтому $VCD(DNF(n,\mu,r)) < c \log_2 n$, где $c = \mu \log_2 \frac{3}{2} + r\mu = const$. □

**Следствие 3.3**. $VCD(DNF(n,\mu,r)) = \Theta(\log_2 n)$ при $\mu, r = const$ и $n \to \infty$.

Доказательство. В следствии 2.2 получена оценка $\log_2 n = O(VCD(BDT(\mu,n)))$ для любого возможного ранга $r \leq \mu - 1$, поэтому $\log_2 n = O(VCD(BDT(\mu,n,r)))$. Очевидно, что $BDT(\mu,n,r) \subset DNF(n,\mu,r)$, поэтому $\log_2 n = O(VCD(DNF(n,\mu,r)))$ и $VCD(DNF(n,\mu,r)) = O(\log_2 n)$ по следствию 3.2.□

Обозначим $BDF(n,\mu,r,q)$ - класс решающих правил, порождаемых эмпирическим решающим лесом.

**Теорема 3.7.** $\max(\mu q, \log_2 n) < VCD(BDF(n,\mu,r,q)) < r\mu q \log_2 n - \mu q \log_2 \frac{\mu q}{2}$ (3.5)

Доказательство. Оценка сверху. Пусть $F$ - $r$-редуцированный эмпирический лес. Для любой точки $\tilde{x} \in B^n$ классифицирующая ее конъюнкция, соответствующая некоторой ветви одного из деревьев леса $F$, определяется структурой леса однозначно. Поэтому решающее правило, определяемое лесом $F \in BDF(n,\mu,r,q)$, соответствует некоторой зафиксированной ДНФ.

Следовательно, если решающее правило $\gamma \in BDF(n,\mu,r,q)$, то $\gamma \in DNF(n,\mu q,r)$, и тогда

$$VCD(BDF(n,\mu,r,q)) < VCD(DNF(n,\mu q,r)) < r\mu q \log_2 n - \mu q \log_2 \frac{\mu q}{2}.$$

Оценка снизу. Используем оценку для одного дерева с $\mu q$ листьями, учитывая, что она была получена при рассмотрении случая лишь одной внутренней вершины. Очевидно, что $BDT(\mu,n,1) \subset BDT(\mu,n,r) \subset BDF(n,\mu,r,q)$, откуда получаем $\max(\mu q, \log_2 n) < VCD(BDF(n,\mu,r,q))$ (см. доказательство теоремы 2.6). □

**Следствие 3.4**. $VCD(BDF(n,\mu,r,q)) = \Theta(\log_2 n)$ при любых заданных константах $r,\mu,q$ и $n \to \infty$.

Доказательство. Очевидно, что $\log_2 n \le \max\{\mu, \log_2 n\}$. Следовательно, с учетом (3.5) $\log_2 n \le VCD(BDF(n,\mu,r,q))$ и $\log_2 n = O(VCD(BDF(n,\mu,r,q)))$. Из неравенства (3.5) легко получить оценку $VCD(BDF(n,\mu,r,q)) = O(\log_2 n)$. Поэтому $VCD(BDF(n,\mu,r,q)) = \Theta(\log_2 n)$. □

Таким образом, построенный по ***DFBSA*** алгоритму скорректированный эмпирический лес не приводит к значительному усложнению порождаемого им класса решающих правил по сравнению с классом РП, порожденных одним РД. Между тем модель эмпирического решающего леса, построенная согласно стратегии ***DFBSA***, в определенном смысле является компромиссом между усложняющей РП переподгонкой и ухудшением качества классификации объектов обучающей таблицы, а также позволяет выявлять новые закономерности между признаками.

### 3.6. Выводы

Отметим главные результаты, изложенные в разделе 3. Описана новая классифицирующая модель - эмпирический решающий лес. Построение эмпирического решающего леса направлено на поиск системы ветвей леса, правильно классифицирующей все объекты непротиворечивой обучающей выборки на основе конъюнктивной закономерности заранее ограниченного ранга. Разработан алгоритм синтеза эмпирического решающего леса. Идея синтеза совокупности деревьев $d_1, d_2, \ldots, d_q$ основана на том, что при попадании классифицируемого объекта, не участвовавшего в обучении, в область отказа дерева $d_1$, классификация этого объекта должна осуществляться другим деревом $d_2$ на основе конъюнктивной закономерности ограниченного ранга.

Предложены алгоритмы принятия решений эмпирическим решающим лесом: последовательный алгоритм с переходами по ссылкам; алгоритм принятия решений на основе наиболее “компетентной” ветви эмпирического леса; алгоритм принятия решений на основе “голосования” ветвей эмпирического решающего леса.

Исследования эмпирического решающего леса, проведенные в разделе 3, позволяют сделать следующие выводы.

1. Построение эмпирического решающего леса представляет собой процедуру последовательной коррекции решающих деревьев, каждое из которых, вообще говоря, может допускать ошибки на некоторых объектах обучающей выборки. Возникновение этих ошибок связано с заданием ограничения на ранги ветвей деревьев, входящих в лес. Тем не менее, сохраняется возможность получения *r*-корректного на обучающей информации эмпирического решающего леса при выполнении условия: для каждого объекта $\tilde{x}$ из таблицы обучения должен существовать интервал $N_{\tilde{x}}^r$ ранга не больше заданной величины *r* такой, что $\tilde{x} \in N_{\tilde{x}}^r$ и множество $N_{\tilde{x}}^r \cap T_{mn\ell}$ содержит объекты только одного и того же

класса. Это условие, как доказано выше, является необходимым и достаточным для существования *r*-корректного леса.

2. Если в силу заданных ограничений корректный лес построить не удаётся, то возможно использование алгебраических корректоров. На основе алгебраического подхода академика Журавлёва построен алгоритм коррекции *r*-некорректного эмпирического леса. Этот алгоритм основан на линейном замыкании множества некорректных алгоритмов распознавания из класса РД-моделей, которое, как известно, является корректным на множестве задач с непротиворечивой начальной информацией и полным множеством признаковых предикатов.

3. Класс решающих правил, порожденных одним деревом и класс решающих правил, порожденных эмпирическим лесом – совокупностью решающих деревьев – имеют один и тот же порядок *VCD*, равный $\Theta(\log_2 n)$, где *n*–размерность признакового пространства, иначе говоря – один порядок сложности. В то же время обеспечивается коррекция, позволяющая настроиться по обучающей выборке на правильную классификацию как можно большего числа и, возможно, даже всех объектов. Такая возможность объясняется, прежде всего, тем, что эмпирический лес, вообще говоря, использует различные подсистемы признаков для синтеза отдельных деревьев, не увеличивая при этом рангов решающих конъюнкций.

4. В результате применения процедур синтеза *r*-редуцированного эмпирического леса получается решающее правило, представимое в виде ДНФ, что позволяет использовать его для логической интерпретации закономерностей и синтеза индуктивных оптимизационных моделей в канонической форме.

# Раздел 4
# ПРОГРАММНАЯ РЕАЛИЗАЦИЯ И АПРОБАЦИЯ АЛГОРИТМА *DFBSA* СИНТЕЗА ЭМПИРИЧЕСКОГО РЕШАЮЩЕГО ЛЕСА

## 4.1. Программная реализация алгоритма синтеза эмпирического решающего леса

Среди методов распознавания, активно используемых в современных информационных системах, наиболее распространены методы, основанные на решающих деревьях и нейросетевых технологиях. Нейросетевые алгоритмы менее приспособлены для решения задач с качественными переменными и не позволяют получать логические описания классов объектов в явном виде.

В мировой практике широко применяются программные комплексы, основным инструментом для принятия решений в которых служат решающие деревья. Наиболее известными среди них являются CART [44], C4.5[130], “Дуэль”[14] и многие другие. Однако, несмотря на ряд преимуществ, РД склонны к “перенастройке” на обучающие данные (overfitting), что привело к появлению различных методов *коррекции структуры решающих деревьев*, в частности, редукции ветвей (pruning) [57, 60, 61, 75, 76, 78, 84, 112], наращивания отдельных вершин РД (grafting) [156, 157]. Предложенный в разделе 3 диссертации алгоритм синтеза эмпирического решающего леса ***DFBSA***, в значительной степени устраняет эти недостатки и удачно сочетает как стратегию редукции, так и теоретическую возможность возврата назад с целью выбора “информативных” признаков во внутренние вершины каждого следующего РД леса.

Описываемая в этом разделе диссертации программная реализация – *информационно-распознающая система Forest Based Learning (FBL)* – предназначена для решения задач обучения и распознавания объектов на основе технологии решающих деревьев и принципиально новых подходов к оцениванию и редукции отдельных классификаторов и организации последовательных процедур построения решающей среды.

*Актуальность разработанного программного комплекса* подтверждается востребованностью высокоточных алгоритмов обучения распознаванию для широко класса приложений; его *новизна* заключается в принципиально новом подходе к построению распознающей процедуры, неизвестной ранее в мировой литературе и обеспечивающей, как показали экспериментальные данные, снижение ошибки распознавания.

Следует отметить, что алгоритм синтеза эмпирического решающего леса и процедуры принятия решений ЭРЛ удовлетворяют всем требованиям, предъявляемым специалистами в различных прикладных областях к алгоритмам обучения ЭВМ [38]:

- возможность анализа больших объемов данных;
- возможность определения качественных периодических классифицирующих закономерностей при анализе малых обучающих выборок;
- автоматическое исключение неинформативных свойств;
- возможность работы в условиях слабого выполнения гипотезы компактности;
- быстрое обучение;
- быстрое прогнозирование;
- возможность работы с пропусками в некоторых значениях свойств;
- возможность работы с качественными свойствами;
- высокая точность при решении практических задач.

Алгоритм ***DFBSA*** синтеза эмпирического решающего леса удачно сочетает стратегию редукции, отсекая ветви, ранг которых превышает заданное пороговое значение и возможность возврата (отката) для построения нового разбиения по системе признаков, вообще говоря, не участвовавших ранее в обучении. Процесс синтеза очередного дерева осуществляется, прежде всего, на тех объектах, для которых предыдущее дерево не способно сформировать простое правило, описывающее их класс; затем происходит достройка РД на оставшихся объектах таблицы обучения.

Алгоритм ***DFBSA*** допускает построение областей отказа для каждого РД, входящего в лес. Область отказа представляет собой интервал, соответствующий редуцированной ветви РД.

Система ***FBL*** позволяет по обучающей информации построить как одно корректное решающее дерево, так и эмпирический решающий лес, ранг которого может быть задан пользователем. Существенным обстоятельством является то, что и дерево и лес используют в процессе синтеза один и тот же критерий ветвления. В системе ***FBL*** в качестве критерия ветвления был выбран и реализован *D*-критерий максимальной отделимости пар разных классов. Алгоритмы синтеза РД и ЭРЛ реализованы с использованием инструментального средства Delphi в среде Windows. Система ***FBL*** предоставляет информацию для сравнения характеристик и качества распознавания корректным РД и корректным ЭРЛ. В частности, можно получить информацию о числе контрольных объектов, распознанных решающим деревом и эмпирическим решающим лесом, числе объектов из совокупной области отказа на каждом этапе синтеза очередного дерева леса, числе листьев в каждом дереве леса, ранге ветвей РД, на котором были допущены ошибки при распознавании объектов контрольной выборки и описания классов объектов в виде ДНФ.

В программном комплексе ***FBL*** реализована процедура распознавания с переходами по ссылкам.

Структурно программный комплекс состоит из следующих подсистем:

1) управление базой экспериментальных данных;

2) синтеза решающего дерева и эмпирического решающего леса;

3) оценивания качества решающих правил;

4) визуализации решающего дерева и эмпирического решающего леса и синтеза логических описаний классов в виде ДНФ.

## 4.2. Апробация алгоритма DFBSA синтеза эмпирического решающего леса

Для экспериментальной проверки теоретических выводов, методов и алгоритмов синтеза логических правил распознавания, полученных в диссертации, был использован массив данных наблюдений за патогенными микроорганизмами, предоставленный доктором медицинских наук Хайтовичем А.Б.

Вкратце опишем проблемную область и параметры решавшейся задачи.

Задача исследования заключалась в проведении анализа базы данных патогенных вибрионов и аэромонад; выявлении эмпирических информативных систем признаков; построении алгоритмов распознавания классов объектов. База данных патогенных вибрионов и аэромонад представлена в виде таблицы обучения, содержащей 365 153-мерных булевых векторов, разделенных на 5 классов:

Vibrio cholerae 01 серогруппы, биовар eltor серовар Огава ($\omega_1$);

Vibrio cholerae non 01 ($\omega_4$);

Aeromonas ($\omega_6$);

Vibrio alginolyticus ($\omega_{10}$);

Vibrio parahaemolyticus ($\omega_{14}$).

По мнению специалистов, в рассматриваемой проблемной области актуальными являются вопросы изучения биологических свойств этих классов в сравнительном аспекте, что в значительной степени должно способствовать уточнению особенностей этих групп микробов, определению диагностической ценности признаков. Особое место в сравнительном изучении биологических свойств вибрионов и аэромонад занимает чувствительность к антимикробным препаратам. Определение уровня чувствительности к антибиотикам необходимо для выбора наиболее эффективных препаратов для лечения и профилактики заболеваний. Вместе с тем остается дискуссионным вопрос о возможности

использования антибиотиков для дифференциации вибрионов и аэромонад. В качестве дифференциального признака было предложено использовать чувствительность к новобиоцину, полимиксину, вибриостатику О129. Перечислим признаки, составляющие описание вибрионов и аэромонад: рост на щелочном агаре, рост на среде Касаткина, рост на 1,5% агаре для обнаружения роения, способность светиться в темноте, тирозиназная активность, толерантность к хлориду натрия в пептонной воде, посев при температуре $+4^0$С и $+42^0$С, выращивание на щелочном агаре при pH 11 и при pH12, каталазная активность, оксидазная активность, реакция "холерарот", тест тяжа, образование индола, определение протеолитической активности, образование сероводорода, определение диастатической активности, уреазная активность на среде Христенсена, расщепление глюкозы в среде Hugh и Zeitson по типам О и по типам F, образование газа из глюкозы, образование газа из глицерина, окисление глюкозы, окисление глицерина, окисление арабинозы, окисление дульцита, окисление инозита, окисление лактозы, окисление 10% лактозы, окисление мальтозы, окисление маннита, окисление маннозы по типу F и по типу О, окисление рамнозы, окисление сорбита, окисление салицина, окисление ксилозы, окисление сахарозы, изучение декарбоксилазы лизина, изучение декарбоксилазы орнитина, изучение дегидролазы аргинина, фенилаланиновый тест, реакция Фогес-Проскауэра при $+20^0$С и при $+37^0$С, реакция Метилрот при $+20^0$С и при $37^0$С, лецитиназная активность, реакция гемагглютинации, фибринолитическая активность, плазмокоагулазная активность, гемолитическая активность по методу Грейга с 1% бараньими эритроцитами, гемолитическая активность на среде с глицерином, гемолитическая активность с 3% человеческой кровью в агаре, гемолитическая активность на среде Wagatsuma, способность лизировать культуры холерными фагами С, способность лизировать культуры холерным фагом eltor, способность лизировать культуры холерным фагом ХДФ-3, ХДФ-4, ХДФ-5, способность лизировать культуры холерным авторским фагом Ф-93, Ф-94, Ф-95, Ф-96, способность лизировать культуры поливалентным холерным фагом, отношение к вибриостатику О129 в концентрации 10мкг и 150 мкг, РНГА

с эритроцитарным иммуноглобулинновым холерным диагностикумом, объемная реакция живых и убитых кипячением культур в течение 2-3 ч с холерной сывороткой О1 в 1:100, О Р А с RO - 1: 100, О Р А Inaba 1:100, Ogava 1:100, О РА 1:100 О и ОН кроличьи холерные сыворотки, схемы Дрожевкиной-Арутюнова, ТЭПВ, рост микробов на среде А-2, для выделения аэромонад, в-галактозидазная активность, способность ассимиляции в синтетической среде ограниченного минерального состава без NaCl арабинозы, рамнозы, глюкозы, мальтозы, лактозы, сахарозы, крахмала, салицина, глицерина, маннита, сорбита, аспарагина, цитрата, глутамина, орнитина, оргинина, гистидина, аланина, серина, на среде Козера с хлоридом натрия у арабинозы, рамнозы, глюкозы, мальтозы, лактозы, сахарозы, крахмала, салицина, глицерина, маннита, сорбита, аспарагина, цитрата, глутамина, орнитина, аргинина, гистидина, аланина, серина.

Результаты работы программного комплекса ***FBL*** с базой данных патогенных вибрионов наглядно представлены выходными формами на рисунках. Рисунки 4.1 - 4.2 иллюстрируют корректный эмпирический решающий лес, состоящий из трех деревьев, ранг каждой ветви которых не превышает 5. В таблицах 4.1 – 4.2 для сравнения приведены некоторые сведения единичном корректном РД (максимальный ранг ветви 7) и корректном ЭРЛ, полученные в одном из 150 проведенных экспериментов (310 объектов было выбрано случайным образом на обучение, 55 объектов на контроль).

Рисунок 4.3 демонстрирует процесс обучения корректного решающего дерева и средний процент ошибок на контроле в 150 экспериментах. Как видно из рисунка 4.3, средний процент ошибок на контроле, совершаемых корректным РД, не ниже 6%.

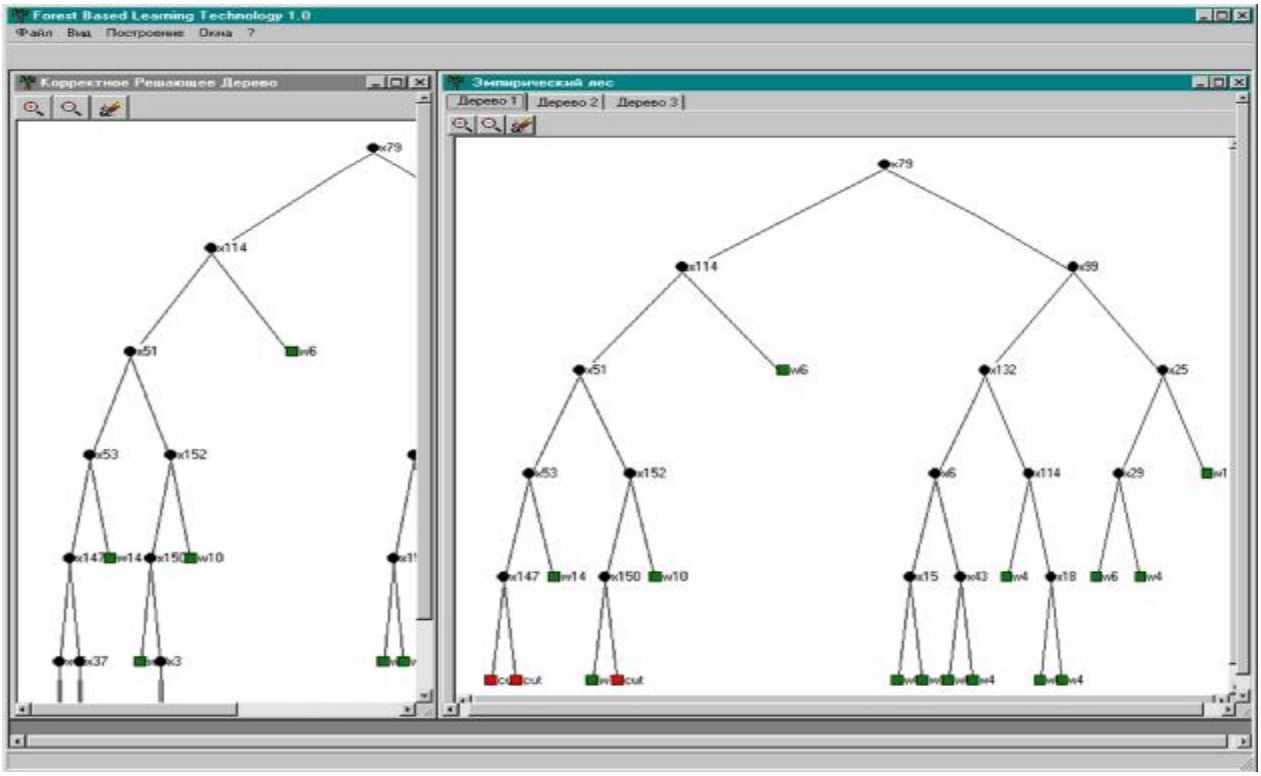


Рис. 4.1 Построение РД1 и редукция ветвей, ранг которых больше 5

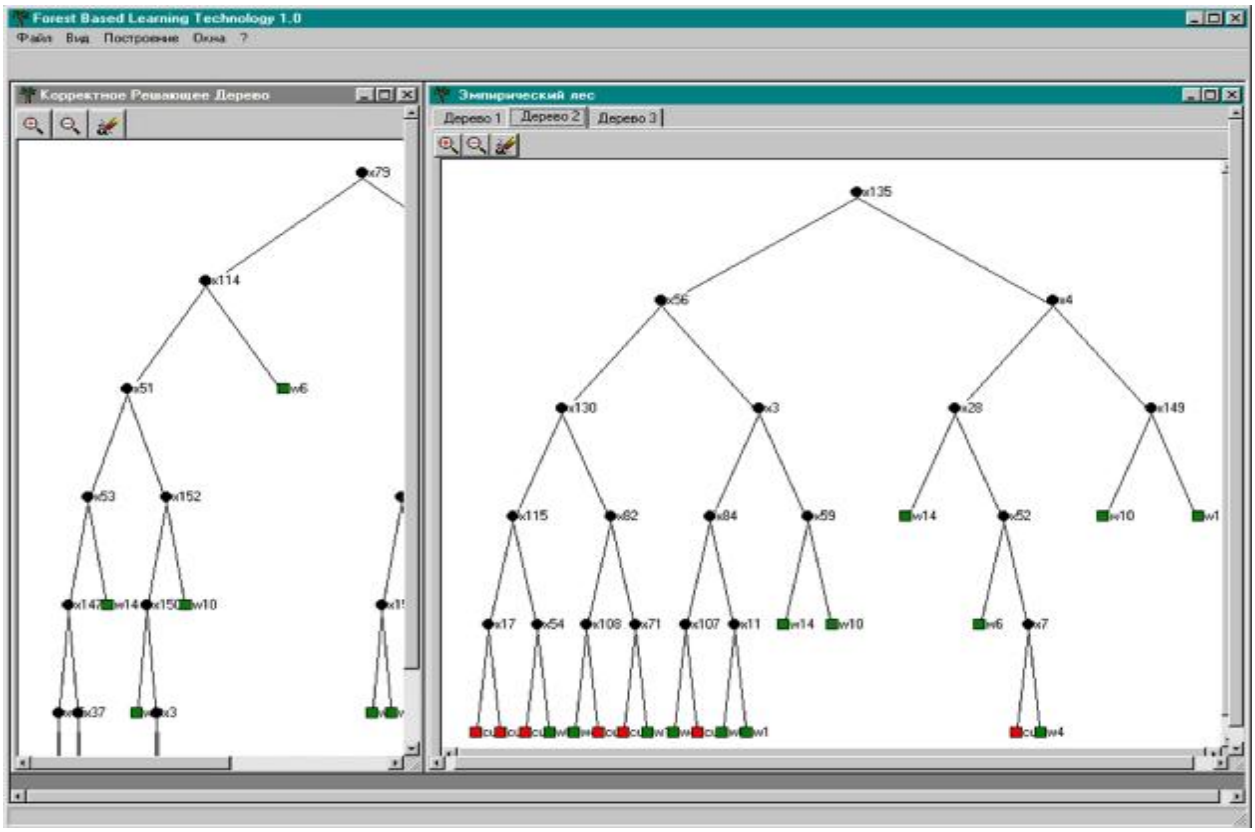


Рис. 4.2 Построение РД2 на объектах из области отказа РД1 и его достройка на оставшихся обучающих объектах

Таблица 4.1. Сравнение характеристик корректного РД и корректного ЭРЛ

| | число ошибок на контроле | число использованных при синтезе признаков | номера использованных при синтезе признаков | число листьев |
|---|---|---|---|---|
| корректное РД | 5 из 55 | 19 | 3,6,15,18,25,29,37, 40,43,51,53,59 | 21 |
| корректный ЭРЛ | 4 из 55 | 34 | | 28 |
| РД1 | | 15 | 6,15,18,25,29,43, 51,53,79,99,114, 132,147,150,152 | 14 |
| РД2 | | 19 | 3,4,7,11,17,28,52, 54,56,59,71,82,84, 107,108,115,130, 135,149 | 13 |
| РД3 | | - | - | 1 |

Таблица 4.2. Отличительные характеристики корректного ЭРЛ

| корректный ЭРЛ | число объектов в области отказа | номера объектов из области отказа |
|---|---|---|
| РД1 | 14 | 111,240,242,243,246,247,248,261,272, 286,290,298,301,310 |
| РД2 | 5 | 240,243,246,248,272 |
| РД3 | 0 | - |

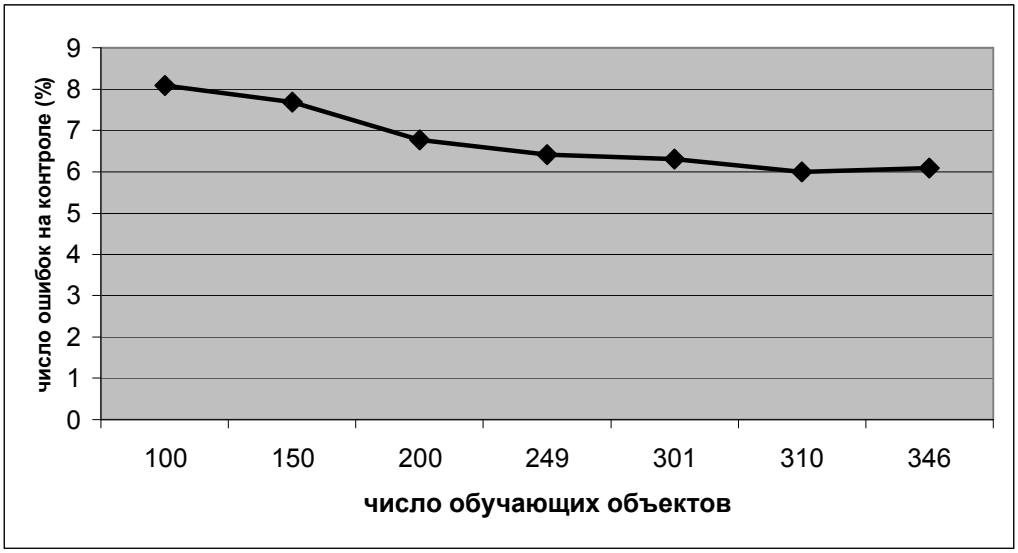


Рис.4.3. Средний процент ошибок на контроле в 150 экспериментах с ростом числа обучающих правил приближается к 6%

В таблице 4.3 проводится сравнительный анализ качества распознавания корректным решающим деревом и корректным эмпирическим решающим лесом при длинах обучающей и контрольной выборки соответственно 310 и 55.

Таблица 4.3. Качество распознавания объектов на контроле корректным РД и корректным ЭРЛ

| | средний процент ошибок на контроле |
|---|---|
| корректное РД | 6,1 |
| ЭРЛ ранга 5 | 6,18 |
| ЭРЛ ранга 6 | 5,5 |

Из таблицы 4.3 видно, что 6-корректный (ранг ветвей не более шести) ЭРЛ позволяет значительно уменьшить процент ошибок распознавания по сравнению с единственным корректным РД, средний ранг ветвей которого не больше 8 и средним числом листьев 20.

Приведем описания классов, полученные по корректному РД, для данных таблицы 4.1.

Описания классов по корректному решающему дереву:

Vibrio cholerae 01 серогруппы, биовар eltor серовар Огава ($\omega_1$):

$x_{25}x_{99}x_{79}$

Vibrio cholerae non 01 ($\omega_4$)

$\bar{x}_3x_{150}\bar{x}_{152}x_{51}\bar{x}_{114}\bar{x}_{79} \vee x_{43}x_6\bar{x}_{132}\bar{x}_{99}x_{79} \vee \bar{x}_{114}x_{132}\bar{x}_{99}x_{79} \vee x_{18}x_{114}x_{132}\bar{x}_{99}x_{79} \vee x_{29}\bar{x}_{25}x_{99}x_{79}$

Aeromonas ($\omega_6$)

$x_{114}\bar{x}_{79} \vee \bar{x}_{15}\bar{x}_6\bar{x}_{132}\bar{x}_{99}x_{79} \vee \bar{x}_{43}x_6\bar{x}_{132}\bar{x}_{99}x_{79} \vee \bar{x}_{18}x_{114}x_{132}\bar{x}_{99}x_{79} \vee \bar{x}_{29}\bar{x}_{25}x_{99}x_{79}$

Vibrio alginolyticus ($\omega_{10}$)

$\bar{x}_{40}\bar{x}_{147}\bar{x}_{53}\bar{x}_{51}\bar{x}_{114}\bar{x}_{79} \vee \bar{x}_{37}x_{147}\bar{x}_{53}\bar{x}_{51}\bar{x}_{114}\bar{x}_{79} \vee \bar{x}_{150}\bar{x}_{152}x_{51}\bar{x}_{114}\bar{x}_{79} \vee x_{59}x_3x_{150}\bar{x}_{152}x_{51}\bar{x}_{114}\bar{x}_{79} \vee$

$\vee x_{152}x_{51}\bar{x}_{114}\bar{x}_{79}$

Vibrio parahaemolyticus ($\omega_{14}$)

$x_{40}\bar{x}_{147}\bar{x}_{53}\bar{x}_{51}\bar{x}_{114}\bar{x}_{79} \vee x_{37}x_{147}\bar{x}_{53}\bar{x}_{51}\bar{x}_{114}\bar{x}_{79} \vee x_{53}\bar{x}_{51}\bar{x}_{114}\bar{x}_{79} \vee \bar{x}_{59}x_3x_{150}\bar{x}_{152}x_{51}\bar{x}_{114}\bar{x}_{79} \vee$

$\vee x_{15}\bar{x}_6\bar{x}_{132}\bar{x}_{99}x_{79}$

Получим описание класса Vibrio cholerae 01 серогруппы, биовар eltor серовар Огава ($\omega_1$) для построенного ЭРЛ:

$D_1(\omega_1) = x_{25}x_{99}x_{79}$,

$R_1 = \bar{x}_{79}\bar{x}_{114}\bar{x}_{51}\bar{x}_{53} \vee \bar{x}_{79}\bar{x}_{114}x_{51}\bar{x}_{152}x_{150}$,

$D_1^0(\omega_1) = x_{25}x_{99}x_{79}$,

$D_2(\omega_1) = x_{11}x_{84}\bar{x}_3x_{56}\bar{x}_{135} \vee x_{71}x_{82}x_{130}\bar{x}_{56}\bar{x}_{135}$,

$D_2^0(\omega_1) = D_1(\omega_1) \vee (R_1 \wedge D_2(\omega_1)) = x_{79}x_{99}x_{25} \vee \bar{x}_{79}\bar{x}_{114}\bar{x}_{51}\bar{x}_{53}x_{11}x_{84}\bar{x}_3x_{56}\bar{x}_{135} \vee$

$\vee \bar{x}_{79}\bar{x}_{114}x_{51}\bar{x}_{152}x_{150}x_{11}x_{84}\bar{x}_3x_{56}\bar{x}_{135} \vee \bar{x}_{79}\bar{x}_{114}\bar{x}_{51}\bar{x}_{53}x_{71}x_{82}x_{130}\bar{x}_{56}\bar{x}_{135} \vee$

$\vee \bar{x}_{79}\bar{x}_{114}x_{51}\bar{x}_{152}x_{150}x_{71}x_{82}x_{130}\bar{x}_{56}\bar{x}_{135}$

Таблица 4.4. Результаты экспериментов

| Число ошибок на контроле РД | Ранг РД | Число листьев РД | Ранг «ошибочных» ветвей | Число ошибок на контроле ЭРЛ ранга 6 | Число деревьев леса | Результаты сравнения ошибок на контроле для ЭРЛ и РД |
|---|---|---|---|---|---|---|
| 2 | 7 | 22 | 4,7 | 2 | 3 | |
| 4 | 6 | 19 | 6 | 4 | 1 | |
| 0 | 7 | 22 | 0 | 0 | 3 | |
| 5 | 7 | 18 | 4,5,6 | 5 | 3 | |
| 1 | 7 | 25 | 4 | 1 | 3 | |
| 4 | 7 | 21 | 4,6 | 4 | 3 | |
| 5 | 7 | 22 | 4,6,7 | 5 | 3 | |
| 3 | 7 | 19 | 4,6,7 | 4 | 4 | - |
| 5 | 7 | 18 | 4,6,7 | 4 | 3 | + |
| 2 | 7 | 22 | 5,7 | 2 | 3 | |
| 6 | 7 | 18 | 2,4,6 | 7 | 5 | - |
| 3 | 8 | 21 | 4,6 | 3 | 3 | |
| 4 | 7 | 19 | 5,7 | 4 | 3 | |
| 3 | 7 | 18 | 4,6 | 3 | 3 | |
| 5 | 7 | 20 | 3,5,7 | 4 | 3 | + |
| 2 | 8 | 25 | 4,6 | 2 | 3 | |
| 3 | 7 | 22 | 4,6 | 3 | 3 | |
| 1 | 7 | 24 | 4 | 1 | 3 | |
| 5 | 7 | 21 | 4,6,7 | 4 | 3 | + |
| 5 | 7 | 18 | 4,6,7 | 4 | 4 | + |
| 2 | 7 | 17 | 4,6 | 2 | 3 | |
| 6 | 7 | 17 | 4,6,7 | 6 | 3 | |
| 4 | 7 | 16 | 2,4,6 | 4 | 3 | |
| 3 | 7 | 18 | 4,6 | 3 | 3 | |
| 3 | 7 | 21 | 4,5,7 | 3 | 4 | |
| 5 | 7 | 16 | 3,4,5,6 | 5 | 3 | |
| 3 | 7 | 20 | 4,7 | 3 | 3 | |
| 1 | 7 | 21 | 6 | 1 | 3 | |
| 3 | 7 | 19 | 5,6,7 | 3 | 3 | |
| 5 | 7 | 18 | 4,5,6,7 | 5 | 3 | |
| 4 | 8 | 19 | 3,4,7 | 4 | 3 | |
| 4 | 8 | 22 | 4,6,7 | 3 | 4 | + |
| 3 | 7 | 20 | 4,5 | 3 | 4 | |
| 6 | 8 | 23 | 4,7 | 6 | 4 | |
| 3 | 8 | 21 | 5,7 | 2 | 4 | + |
| 6 | 7 | 21 | 4,6,7 | 6 | 3 | |
| 4 | 7 | 21 | 3,5,7 | 4 | 3 | |
| 5 | 7 | 18 | 4,5,6,7 | 4 | 3 | + |
| 1 | 7 | 22 | 7 | 0 | 3 | + |
| 6 | 8 | 19 | 4,6 | 6 | 3 | |
| 5 | 7 | 22 | 4,5,7 | 4 | 3 | + |
| 4 | 7 | 21 | 4,6,7 | 3 | 3 | + |
| 5 | 7 | 19 | 2,4,6 | 5 | 3 | |
| 3 | 7 | 18 | 4,6 | 3 | 3 | |

| Число ошибок на контроле РД | Ранг РД | Число листьев РД | Ранг «ошибочных» ветвей | Число ошибок на контроле ЭРЛ ранга 6 | Число деревьев леса | Результаты сравнения ошибок на контроле для ЭРЛ и РД |
|---|---|---|---|---|---|---|
| 5 | 7 | 20 | 4,6,7 | 5 | 3 | |
| 3 | 7 | 22 | 4,6,7 | 2 | 4 | + |
| 4 | 7 | 19 | 4,5,6 | 4 | 3 | |
| 1 | 7 | 22 | 4 | 1 | 3 | |
| 1 | 7 | 22 | 4 | 1 | 3 | |
| 2 | 7 | 20 | 5,7 | 2 | 3 | |
| 4 | 7 | 20 | 4,6,7 | 3 | 3 | + |
| 3 | 7 | 20 | 6,7 | 2 | 3 | + |
| 4 | 7 | 22 | 4,6,7 | 4 | 3 | |
| 1 | 7 | 20 | 6 | 1 | 3 | |
| 4 | 7 | 16 | 4,6 | 4 | 3 | |
| 3 | 7 | 18 | 4,6 | 3 | 3 | |
| 4 | 7 | 23 | 4,6 | 4 | 3 | |
| 2 | 7 | 21 | 6 | 4 | 3 | - |
| 5 | 7 | 23 | 4,7 | 4 | 3 | + |
| 0 | 7 | 23 | 0 | 0 | 3 | |
| 1 | 7 | 22 | 7 | 1 | 3 | |
| 6 | 8 | 21 | 4,6,8 | 6 | 3 | |
| 0 | 7 | 22 | 0 | 0 | 3 | |
| 2 | 7 | 19 | 6,7 | 1 | 3 | + |
| 4 | 7 | 22 | 4,7 | 5 | 3 | - |
| 3 | 7 | 22 | 6 | 3 | 3 | |
| 6 | 7 | 22 | 4,5,6 | 6 | 3 | |
| 0 | 7 | 22 | 0 | 0 | 3 | |
| 3 | 8 | 21 | 4,6 | 3 | 3 | |
| 1 | 7 | 22 | 4 | 2 | 3 | - |
| 3 | 7 | 20 | 4,7 | 2 | 4 | + |
| 4 | 7 | 21 | 3,5,7 | 4 | 3 | |
| 5 | 7 | 18 | 4,5,6,7 | 4 | 3 | + |
| 2 | 7 | 23 | 4,7 | 2 | 3 | |
| 2 | 7 | 20 | 6 | 2 | 3 | |
| 7 | 7 | 16 | 4,5,6,7 | 6 | 3 | + |
| 8 | 7 | 15 | 3,4,6,7 | 8 | 3 | |
| 4 | 7 | 21 | 4,7 | 4 | 3 | |
| 5 | 8 | 21 | 4,6,7 | 5 | 5 | |
| 3 | 7 | 19 | 4,7 | 2 | 3 | + |
| 6 | 7 | 16 | 3,4,6 | 6 | 3 | |
| 3 | 7 | 22 | 5 | 3 | 4 | |
| 2 | 7 | 20 | 4,6 | 2 | 3 | |
| 2 | 7 | 21 | 7 | 1 | 3 | + |
| 3 | 7 | 23 | 6,7 | 2 | 3 | + |
| 4 | 8 | 20 | 2,4,7 | 4 | 4 | |
| 4 | 7 | 24 | 5,6 | 4 | 3 | |
| 2 | 7 | 17 | 4 | 2 | 3 | |
| 1 | 7 | 22 | 4 | 1 | 4 | |
| 0 | 7 | 22 | 0 | 0 | 3 | |

| Число ошибок на контроле РД | Ранг РД | Число листьев РД | Ранг «ошибочных» ветвей | Число ошибок на контроле ЭРЛ ранга 6 | Число деревьев леса | Результаты сравнения ошибок на контроле для ЭРЛ и РД |
|---|---|---|---|---|---|---|
| 4 | 7 | 20 | 4,6,7 | 4 | 3 | |
| 3 | 7 | 21 | 4,6,7 | 3 | 3 | |
| 4 | 7 | 15 | 4 | 4 | 3 | |
| 5 | 7 | 23 | 4,5,7 | 4 | 3 | + |
| 5 | 7 | 24 | 4,5,6,7 | 5 | 3 | |
| 7 | 7 | 16 | 3,4,5,7 | 6 | 5 | + |
| 0 | 7 | 22 | 0 | 0 | 3 | |
| 2 | 7 | 19 | 4,6 | 2 | 3 | |
| 3 | 7 | 22 | 4,7 | 1 | 3 | + |
| 1 | 7 | 19 | 4 | 1 | 3 | |

Для оценки достоверности того, что эмпирический решающий лес в среднем даёт меньшее число ошибок на контроле, чем отдельное, точно настроенное на обучающую выборку БРД, были проведены дополнительные эксперименты. Случайно и независимо выбирались обучающие выборки длины $m = 310$, на которых строились корректное БРД и *r*-корректный ЭРЛ. В каждом таком эксперименте БРД и ЭРЛ оценивались на контрольной выборке из 55 объектов. Эксперимент повторялся $n = 100$ раз (см. таблицу 4.4). В результате экспериментов рассматривались такие исходы: $A$ - меньше ошибок на контроле делает ЭРЛ; $Б$ - ЭРЛ и БРД дают одинаковое число ошибок; $B$ - БРД делает меньше ошибок, чем ЭРЛ. Событие $A$ из 100 экспериментов произошло $m_A = 24$ раза; событие $B$ произошло $m_B = 5$ раз. Частоты $\frac{m_A}{n} = \nu(A)$ и $\frac{m_B}{n} = \nu(B)$ являются, соответственно, оценками вероятностей $P(A)$ и $P(B)$ того, что на произвольной контрольной выборке точнее будет ЭРЛ и БРД.

Гипотеза $H_1$: $P(A) > P(B)$ проверялась против конкурирующей гипотезы $H_0$: $P(A) = P(B)$ с использованием критерия [4]

$$U_H = \frac{m_A/n - m_B/n}{\sqrt{\frac{m_A + m_B}{n+n}\left(1 - \frac{m_A + m_B}{n+n}\right)\left(\frac{1}{n} + \frac{1}{n}\right)}} \approx 3{,}7$$

На уровне значимости $\alpha = 0{,}01$ вычислялось $(1-2\alpha)/2 = 0{,}49$; $\Phi(U_{кр}) = 0{,}49$; $U_{кр} = 2{,}32$; $U_{кр} = 2{,}32 < 3{,}7 = U_H$.

На основании полученных неравенств гипотеза $H_1$ принимается: вероятность того, что ЭРЛ сделает меньше ошибок при распознавании превышает соответствующую вероятность для БРД на уровне значимости 0,01.

### 4.3. Выводы

Программный комплекс ***FBL***, в котором реализован алгоритм синтеза эмпирического решающего леса, был применен при решении практически важной задачи эффективного распознавания и формирования описаний классов патогенных вибрионов. Эмпирический решающий лес повысил точность распознавания объектов, не участвовавших ранее в обучении, по сравнению с одним решающим деревом, при использовании одного и того же критерия ветвления для выбора признаковых предикатов во внутренние вершины решающих деревьев.

Эксперименты ЭРЛ и БРД проводились на тщательно собранном микробиологами эмпирическом материале достаточного объёма. На уровне значимости $\alpha = 0{,}01$ установлено: вероятность того, что ЭРЛ будет точнее, чем отдельное корректное на обучающей выборке БРД, выше, чем вероятность того, что точнее будет БРД.

В среднем, для всех групп проведенных экспериментов, ЭРЛ обеспечивал точность распознавания выше, чем БРД.

Таким образом, дополнительно к теоретически обоснованному и доказанному преимуществу ЭРЛ перед отдельным корректным БРД добавлен факт значимого экспериментального подтверждения того, что ЭРЛ в среднем является более точной распознающей процедурой, чем БРД.

# ЗАКЛЮЧЕНИЕ

В диссертационной работе исследованы и усовершенствованы алгоритмы обучения и распознавания, основанные на построении бинарных решающих деревьев; разработаны обоснованные правила редукции бинарных решающих деревьев, основанные на оценивании конъюнктивных закономерностей; создана последовательная процедура синтеза совокупности решающих деревьев – алгоритм синтеза эмпирического решающего леса – и методы коррекции совокупности редуцированных решающих деревьев как набора эвристических процедур принятия решений.

В работе получены следующие основные **результаты**:

7. Разработан вероятностный критерий отсечения (редукции) ветвей бинарного решающего дерева, имеющих число внутренних вершин, превышающее заданное значение ранга *r*. Обоснована редукция с точки зрения неслучайности (закономерности) обнаружения в эмпирической выборке конъюнктивной закономерности ранга *r*.
8. На основе оценок *VCD* (сложности класса решающих правил по теории Вапника-Червоненкиса) для классов алгоритмов распознавания, определяемых бинарными решающими деревьями с ограничением на число вершин, обоснована целесообразность усложнения правил распознавания и процедур коррекции решений.
9. Разработаны методы построения корректной совокупности решающих деревьев (эмпирического решающего леса), обеспечивающей возможность точной настройки на обучающую выборку с одновременным соблюдением ограничения на ранг ветвей РД.
10. Получена оценка сложности эмпирического решающего леса и изучены другие его свойства как специального семейства алгоритмов распознавания.

11. Разработаны алгоритмы коррекции совокупности некорректных эмпирических решающих деревьев, обеспечивающие повышение точности классификации.
12. Создано необходимое программное обеспечение и проведены эксперименты на реальных данных с целью подтверждения теоретических результатов, полученных в диссертации.

В итоге можно заключить, что *разработанный в диссертации процедурный способ коррекции редуцированных БРД – эмпирический лес – позволяет получить более точный алгоритм распознавания, чем алгоритмы, реализуемые отдельными корректными БРД*. При этом эмпирический решающий лес сохраняет возможность логического описания объектов в виде дизъюнктивных нормальных форм, и его применение в современных интеллектуализированных информационных системах целесообразно, что подтверждается апробированной в работе программной реализацией ***FBL***.

# ЛИТЕРАТУРА


1. Ахо А.В., Хопкрофт Д.Э., Ульман Д.Д. Структуры данных и алгоритмы. : Пер. с англ. : Уч.пос. - М.: Издательский дом “Вильямс”, 2000. – 384 с.
2. Бабич Г.Х. Принятие решений на основе анализа дерева решений в условиях неполноты информации // Кибернетика. – 1986. - №5. – С.113-120.
3. Вапник В.Н. Восстановление зависимостей по эмпирическим данным. – М.: Наука, 1979. – 448 с.
4. Гмурман В.Е. Теория вероятностей и математическая статистика. – М.: Высшая школа, 2001. – 479 с.
5. Гэри М., Джонсон Д. Вычислительные машины и труднорешаемые задачи. – М.: Мир, 1982. – 416 с.
6. Грэхем Р., Кнут Д., Паташник О. Конкретная математика. Основание информатики: Пер. с англ. – М.: Мир, 1998. – 703 с.
7. Гупал А.М. Методы индуктивного вывода и их применение в экспертных системах // Управляющие системы и машины. – 1991. - №7. – С.112-114.
8. Гупал А.М., Цветков А.М. Разработка алгоритмов индуктивного вывода знаний с использованием и листьев и деревьев решений // Управляющие системы и машины. – 1992. - №5/6. – С.21-26.
9. Гупал А.М., Цветков А.М. Разработка алгоритмов индуктивного вывода, основанных на построении деревьев решений // Кибернетика и системный анализ. – 1993. - №3. – С.174-178.
10. Гупал А.М., Пономарев А.А., Цветков А.М. Об одном методе индуктивного вывода с подрезанием деревьев решений // Кибернетика и системный анализ. – 1993. - №5. – С.174-178.
11. Дискретная математика и математические вопросы кибернетики / под ред. С.В. Яблонского. – М.: Наука, 1974. – 312 с.

12. Донской В.И. Алгоритмы обучения, основанные на построении решающих деревьев // Журнал вычислительной математики и математической физики. – 1982, Т.22. - №4. – С.963-974.
13. Донской В.И. Асимптотика числа бинарных решающих деревьев // Ученые записки Таврического национального университета им. В.И. Вернадского. Серия “Математика”. – 2001 г., Т.14(53), №1, С.36-38.
14. Донской В.И. Дуальные экспертные системы // Изв. Российской АН. Техническая кибернетика. – 1993. - №5. – С.17-22.
15. Донской В.И. Исследование алгоритмов распознавания, основанных на построении решающих деревьев. Диссертация на соискание ученой степени кандидата физ.-мат. наук. Рукоп. – 100 с.
16. Донской В.И. О корректности линейного замыкания множества алгоритмов распознавания типа решающих деревьев // Динамические системы. Вып.5. Киев: Вища школа. 1986. – С.91-94.
17. Донской В.И. О локальном подходе к восстановлению пропущенной информации в булевых таблицах обучения // Динамические системы. Киев: Вища школа. 1986. – Вып. 3. – С.85-89.
18. Донской В.И. Об одном алгоритме обучения распознаванию объектов, описанных булевыми признаками // Динамические системы. – Киев: Вища школа. 1986. – Вып. 3. – С.102-108.
19. Донской В.И., Башта А.И. Дискретные модели принятия решений при неполной информации. – Симферополь: Таврия, 1992. – 166 с.
20. Донской В.И., Дюличева Ю.Ю. Алгоритмы синтеза *r*-редуцированного эмпирического леса // ММРО-11: тезисы докладов. – 2003. – Пущино. – С.71-74.
21. Донськой В.Й., Дюлічева Ю.Ю. Бінарні розв’язуючі дерева в задачах інтелектуального аналізу інформації // Наукові вісті Національного технічного університету України “Київський політехнічний Інститут”. – 2001. – Вип.5. – С.12-18.

22. Донской В.И., Дюличева Ю.Ю. Деревья решений с *k*-значными переменными // Труды Международной конференции “Знание – Диалог - Решение”. – Санкт-Петербург: Изд-во “Лань”, 2001. – Т.1. – С.201-207.
23. Донской В.И., Дюличева Ю.Ю. Индуктивная модель *r*-корректного эмпирического леса // Труды международной конференции по индуктивному моделированию: Львов. – 2002. – С.54-58.
24. Дюличева Ю.Ю. Оценка VCD *r*-редуцированного эмпирического леса // Таврический вестник информатики и информатики. – 2003. - №1. – С. 31-42.
25. Дюличева Ю.Ю. Принятие решений на основе индуктивной модели эмпирического леса // Искусственный интеллект. – 2002. – №2. – С.110-115.
26. Дюличева Ю.Ю. Принятие решений на основе индуктивной модели эмпирического леса // Тезисы докладов Международной научной конференции “Интеллектуализация обработки информации”. – Симферополь: КНЦ НАН Украины. – 2002. – С.38-39.
27. Дюличева Ю.Ю. О критерии существования и логическом описании *r*-корректного эмпирического решающего леса // Искусственный интеллект. – 2004. - №1. – С. 167-172.
28. Дюличева Ю.Ю. О программной реализации и апробации алгоритма DFBSA синтеза эмпирического решающего леса // Таврический вестник информатики и математики. – 2003. - №2. – С.35-43.
29. Дюличева Ю.Ю. Стратегии редукции решающих деревьев (обзор) // Таврический вестник информатики и математики. – 2002. - №1. – С.10-17.
30. Дюличева Ю.Ю. DFBSA – алгоритм синтеза эмпирического леса по “ссылкам” // Тезисы докладов Международной научной конференции “On Problems of Decision Making and Control under Uncertainties“. – 2003. – С.95-97.
31. Журавлёв Ю.И. Корректные алгебры над множествами некорректных (эвристические) алгоритмов. I // Кибернетика. – 1977. - №4. – С.14-21.

32. Журавлёв Ю.И. Корректные алгебры над множествами некорректных (эвристические) алгоритмов. II // Кибернетика. – 1977. - №6. – С.21-27.
33. Журавлёв Ю.И. Корректные алгебры над множествами некорректных (эвристические) алгоритмов. III // Кибернетика. – 1978. - №2. – С.35-41.
34. Журавлёв Ю.И. Об алгебраическом подходе к решению задач распознавания или классификации // Проблемы кибернетики. Вып.3. – М.: Наука, 1978. – С.5-68.
35. Журавлёв Ю.И. Теоретико-множественные методы в алгебре логики // Проблемы кибернетики. – М.: Наука, 1962. – Вып.8. – С.5-44.
36. Загоруйко Н.Г., Ёлкина В.Н., Лбов Г.С. Алгоритмы обнаружения эмпирических закономерностей. – Новосибирск: Наука, 1985. – 110 с.
37. Закревский А.Д. Логика распознавания. – Минск: Наука и техника, 1988. – 119 с.
38. Казанцев В.С. Задачи классификации и их программное обеспечение (пакет КВАЗАР). – М.: Наука, 1990. – 136 с.
39. Киселёва Н.Н. Компьютерное конструирование неорганических веществ с использованием методов обучения ЭВМ // ММРО-11: тезисы докладов. – 2003. – Пущино. – С.323-326.
40. Колмогоров А.Н. Теория информации и теория алгоритмов. – М.: Наука, 1987. – 305 с.
41. Лбов Г.С. Методы обработки разнотипных экспериментальных данных. – Новосибирск: Наука. – 1981. – 160 с.
42. Люгер Дж.Ф. Искусственный интеллект. – М.: Издательский дом “Вильямс”. – 2003. – 864 с.
43. Нильсон Н. Искусственный интеллект. – М.: Мир. – 1989. – 293 с.
44. Норушис А. Метод повышения вычислительной эффективности классификатора, аппроксимируя области решений локусами // Статистические проблемы управления. – Вып.93. – Вильнюс. – 1990. – С.112-129.

45. Норушис А. Построение логических (древообразных) классификаторов методами нисходящего поиска (обзор) // Статистические проблемы управления. – Вып.93. – Вильнюс. – 1990. – С.131-157.
46. Соловьев Н.А. Тесты (теория, построение, применение). – Новосибирск: Наука. – 1978. – 190 с.
47. Цветков А.М. Разработка алгоритмов индуктивного вывода с использованием деревьев решений // Кибернетика и системный анализ. – 1993. - №1. – С.174-178.
48. Цвєтков О.М. Дослідження індуктивних методів виводу знань в експертних системах: Автореф. дис... к. ф.-м. наук: 05.13.16/ Ін-т кібернетики ім. В.М. Глушкова. – Київ., 1994. – 16 с.
49. Фу К. Структурные методы в распознавании образов . – М.: Мир. – 1977. – 319 с.
50. Шоломов Л.А. Основы теории дискретных логических и вычислительных устройств. – М.: Наука. – 1980. – 400 с.
51. Яблонский С.В. Введение в дискретную математику. – М.: Наука. – 1979. – 272 с.
52. Alsabti K., Ranka S., Singh V. CLOUDS: A Decision Tree Classifier for Large Datasets // Proceedings of 4th International Conference on Knowledge Discovery and Data Mining. – 1998. – P.2-8.
53. Anthony M., Brightwell G., Cooper C. The Vapnik-Chervonenkis Dimension of a Random Graph // DMATH: Discrete Mathematics. – 1995. – Vol.138. – P.43-56.
54. Barron A. The Minimum Description Length Principle in Coding and Modeling // IEEE Transactions on Information Theory. – 1998. – Vol.44, №6. – P. 2743-2760.
55. Bennett K.P. Global Tree Optimization: A Non-Greedy Decision Tree Algorithm // Computing Science and Statistics. – 1994. – Vol.26. – P.156-160.

56. Bensusan H. God Doesn't Always Shave with Occam's Razor – Learning When and How to Prune // Proceedings of the 10th European Conference on Machine Learning. – Berlin, Germany. – 1998. – P.119-124.
57. Bradford J.P., Kunz C., Kohavi R., Brunk C., Brodley C.E. Pruning Decision Trees with Misclassification Costs // European Conference on Machine Learning. – 1998. – P.131-136.
58. Breiman L. Bagging Predictors // Machine Learning. – 1996. – Vol.24, №2. – P.123-140.
59. Breiman L. Random Forests CA 94720: Technical Report / Statistics Department University of California, Berkley – 2001. – 32 p.
60. Breslow L.A., Aha D.W. Comparing Tree-Simplification Procedures // Nave Center for Applied Research in Artificial Intelligence, Technical Report No. AIC-96-015. – 1996. – P.1-10.
61. Breslow L.A., Aha D.W. Simplifying Decision Trees: A Survey // Knowledge Engineering Review 12. – 1997. – P.1-40.
62. Brodley C.E., Utgoff P.E. Multivariate Decision Trees // Machine Learning. – 1995. – Vol. 19(1). – P.45-77.
63. Buntine W. Learning Classification Trees // Statistics and Computing Journal. – 1992. – Vol.2. – P.63-73.
64. Buntine W. Classifiers: A Theoretical and Empirical Study // Proceedings 12th of International Joint Conference on Artificial Intelligence. Sydney. – 1991. – P. 638-644.
65. Chen D., Daescu O., Hu X., Xu J. Finding an Optimal Path without Growing the Tree // Proceedings of 6th Annual European Symposium on Algorithms. - Springer LNCS. – 1998. – P.356-367.
66. Chen H., Ho T.K. Evaluation of Decision Forests on Text Categorization // Proceedings of the 7th (SPIE) Conference on Document Recognition and Retrieval. – San Jose, US. – 2000. – P.191-199.
67. Crèmilleux B., Ragel A., Bossom J.L. An Interactive and Understandable Method to Treat Missing Values: Applications to a Medical Data Set //

Proceedings of 5th International Conference on Information Systems Analysis and Synthesis. – Orlando. – 1999. – P.137-144.

68. Crèmilleux B., Robert C. Use of Attribute Selection Criteria in Decision Trees in Uncertain Domains // Uncertainty in Intelligent and Information Systems, Advances in Fuzzy Systems, Application and Theory. – World Scientific. – 2000. – Vol.20. – P.150-161.
69. Crèmilleux B., Robert C., Gaio M. Uncertain Domains and Decision Trees: ORT versus C.M. Criteria // Proceedings of 7th Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems. – Editions EDK, Paris, France. – 1998. – P.540-546.
70. Devroye L., Gyorfi L., Lugosi G. A probabilistic Theory of Pattern Recognition. Springer-Verlag, NY. – 1996. – 636 p.
71. Dietterich T.G., Hild H., Bakiri G. A Comparative Study of ID3 and Backpropagation for English Text-to-Speech Mapping // Machine Learning. – 1990. – P.24-31.
72. Dong M., Kothari R. Classifiability Based Pruning of Decision Trees // Proc. International Joint Conference on Neural Networks (IJCNN). – Vol. 3. – 2001. – P.1739-1743.
73. Drucker H. Effect of Pruning and Early Stopping on Performance of a Boosted Ensemble // Proceedings of the International Meeting on Nonlinear Methods and Data Mining. – 2000. – P.26-40.
74. Elomaa T. The Biases of Decision Tree Pruning Strategies // Advances in Intelligent Data Analysis, Proceedings of 3rd IDA. Lecture Notes in Computer Science 1642. – Springer. – 1999. – P.63-74.
75. Elomaa T., Kääriäinen An Analyses of Reduced Error Pruning // Journal of Artificial Intelligence Research – 2001. – Vol.15. – P.163-187.
76. Esposito F., Malerba D., Semeraro G.A. Comparative Analysis of Methods for Pruning Decision Trees // IEEE Transactions on Pattern Analyses and Machine Intelligence. – 1997. – Vol.19(5). – P. 476-491.

77. Fern A., Givan R. Online Ensemble Learning: An Empirical Study // Proceedings of 17th International Conference on Machine Learning, Morgan Kaufmann, San Francisco, CA. – 2000. – P.279-286.
78. Frank E. Pruning Decision Trees and Lists // Ph. D. Thesis. University of Waikato. Departament of Computer Science. Hamilton, New Zealand. – 2000. 258 p.
79. Freud Y., Iyer R., Schapire R.E., Singer Y. An Efficient Algorithm for Combining Preferences // Machine Learning: Proceedings of the 15th International Conference. – 1998. – 84 p.
80. Freud Y., Mansour Y., Schapire R.E. Why Averaging Can Protect Against Overfitting? // In Proceedings of the Eight International Workshop on Artificial Intelligence and Statistics. – 2001. – P.1-9.
81. Freud Y., Mason L. The Alternating Decision Tree Learning Algorithm // Proceedings of 16th International Conference on Machine Learning. – Morgan Kaufmann, San Francisco, CA. – 1999. – P.124-133.
82. Freud Y., Schapire R.E. Experiments with a New Boosting Algorithm // International Conference on Machine Learning. – 1996. – P.148-156.
83. Fournier D. UnDeT: a Tool for Building Uncertain Trees // Computing and Information Systems. – 2000. – Vol.7. – P.73-78.
84. Fournier D., Crèmilleux B. A Quality Index for Decision Tree Pruning // Knowledge-Based Systems. – 2002. – Vol.15. – P.37-43.
85. Fournier D., Crèmilleux B. A Trade-Off between Depth and Impurity for Pruning Decision Trees // Proceedings of 4th Australian Knowledge Acquisition Workshop. – Sydney, Australia. – 1999. – P.102-116.
86. Fournier D., Crèmilleux B. Using Impurity and Depth for Decision Trees Pruning // Second International ICSC Symposium on Engineering of Intelligent Systems (EIS 2000). – Pisley, UK. – 2000. – P.320-326.
87. Gehrke J., Ganti V., Ramakrishnan R., Loh W. BOAT – Optimistic Decision Tree Construction // Proceedings of the ACM SIGMOD Conference on Management of Data. – 1999. – P.169-180.

88. Gehrke J., Ganti V., Ramakrishnan R. Rainforest – A framework for Fast Decision Tree Construction of Large Datasets // Data Mining and Knowledge Discovery. – 2000. – Vol.4, №2/3. – P.127-162.
89. Golea M., Bartlett P.L., Lee W.S., Mason L. Generalization in Decision Trees: Does Size Matter? // Advances in Neural Information Processing Systems. – 1998. – Vol.10. – P.259-265.
90. Grove A.J., Schuurmans Boosting in the limit: Maximizing the Margin of Learned Ensembles // Proceedings of the 15th National Conference on Artificial Intelligence. – Madison, WI. – 1998. – P.692-699.
91. Hall L.O., Collins R., Boyew K.W., Banfield Error-Based Pruning of Decision Trees Grown on Very Large Data Sets Can Work! // International Conference on Tools for Artificial Intelligence. – 2002. – P.233-238.
92. Helmbold D.P., Schapire R.E. Predicting Nearly as well as the Best Pruning of a Decision Tree // Machine Learning. – 1997. – Vol. 27(1). – P.51-68.
93. Ho T.K. C4.5 Decision Forests // Proceedings of the 14th International Conference of Pattern Recognition, Brisbane, Australia. – 1998. – P.17-20.
94. Ho T.K. Random Decision Forests // Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, Canada – 1995. – P.278-282.
95. Ho T.K. The Random Subspace Method for Constructing Decision Forests // IEEE Transactions on Pattern Analysis and Machine Intelligence. – 1998. – Vol.20, №8. – P.832-844.
96. Hyafil L. Construction Optimal Binary Decision Trees is NP-Complete // Information Processing Letters. – 1976. – P.15-17.
97. Jiang W. Does Boosting Overfit: Views from an Exact Solution // Technical Report 00-04, IL 60208. – Department of Statistics. – Northwestern University, September. - 2000. – 11 p.
98. Kalles D., Morris T. Efficient Incremental Induction of Decision Trees // Machine Learning. – 1996. – Vol.24, №3. – P.231-242.

99. Kalles D., Papagelis A. Stable Decision Trees: Using Local Anarchy for Efficient Incremental Learning // International Journal on Artificial Intelligence Tools. – 2000. – Vol.9, №1. – P.79-95.
100. Kanal L.N. Problem-Solving Models and Search Strategies for Pattern Recognition // IEEE Trans. on Pattern and Mach, Intel. – 1979. – Vol. PAMI – 1, Number 2. – P. 193-201.
101. Kazuyuki Amano, Tsukuru Hirosawa, Yusuke Watanabe, Akira Maruoka The Computational Power of a Family of Decision Forests // Lecture Notes in Computer Science. – 2001. – Vol.2136. – P.123-133.
102. Kearns M., Mansour Y. On the Boosting Ability of Top-Down Decision Tree Learning Algorithms // Proceedings of the 13th International Conference on Machine Learning. – 1996. – 24 p.
103. Kohavi R. Wrappers for Performance Enhancements and Oblivious Decision Graphs: PhD Thesis, Stanford University, Department of Computer Science – 1995. – 210 p.
104. Kohavi R., Li C. Oblivious Decision Trees, Graphs, and Top-Down Pruning // Proceedings of the 14th International Joint Conference on Artificial Intelligence, Morgan Kaufmann. – 1995. – P.1071-1077.
105. Kohavi R., Sahami M. Error-Based and Entropy-Based Discretization of Continuous Features // Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon. – 1996. – P.114-119.
106. Kothari R., Dong M. Decision Trees for Classification: A Review and Some New Results // Pattern Recognition: From Classical to Modern Approaches, S.R. Pal (Eds.). – Chapter 6, World Scientific. – 2001. – P.169-184.
107. Kulkarny A.V. Optimal and Heuristic Synthesis of Hierarchical Classifiers // IEEE Trans. on Computers. – 1978. – Vol.C27. – Number 8. – P.771-776.
108. Liu H., Setiono R. A Probabilistic Approach to Feature Selection – A Filter Solution // Proceedings of the 13th International Conference on Machine Learning. – Bari, Italy. – 1996. – P. 319-327.

109. Loh W.-Y. Split Selection Methods for Classification Trees // Statistica Sinica. – 1997. – Vol.7. – P.815-840.

110. Mahmoud H.M. On Tree-Growing Search Strategies // The Annals of Applied Probability. – 1996. – Vol.6. – P.1284-1302.

111. Malerba D., Esposito f., Semeraro G. A Further Comparison Methods of Decision Tree Induction // Learning From Data: Artificial Intelligence and Statistics V, D. Fisher and H.Lenz, eds., Lecture Notes in Statistics. Berlin: Springer, No.112. – 1996. – P.365-374.

112. Mansour Y. Pessimistic Decision Tree Pruning Based on Tree Size // Proceedings of 14th International Conference on Machine Learning. – Morgan Kaufmann. – 1997. – P. 195-201.

113. Mansour Y., McAllester D. Generalization Bounds for Decision Trees // Proceedings of 13th Annual Conference on Computing Learning Theory. – Morgan Kaufmann, San Francisco. – 2000. – P.69-80.

114. Margineantu D.D., Dietterich T.G. Pruning Adaptive Boosting // Proceedings of 14th International Conference on Machine Learning. – Morgan Kaufmann. - 1997. – P. 211-218.

115. Mehta M., Rissanen J., Agrawal R. MDL-based decision tree pruning // Proceedings of the 1rd International Conference on Knowledge Discovery and Data Mining. Montreal, Canada. – 1995. – P.216-221.

116. Meir R., Mannor S., Zhang T. The Consistency of Greedy Algorithms for Classification // Proceedings of the Annual Conference on Computational Theory. – Sydney. – 2002. – Vol.2375. – P.319-333.

117. Minos Garofalakis, Dongjoon Hyan, Rajeev Rastogi, Kyuseok Shim Building Decision Trees with Constraints // Data Mining and Knowledge Discovery. – 2003. – Vol.7. – P.187-214.

118. Minos Garofalakis, Dongjoon Hyan, Rajeev Rastogi, Kyuseok Shim Efficient Algorithms for Constructing Decision Trees with Constraints // Proceedings of ACM SIGKDD. Boston, Massachusetts. – 2000. – P.335-355.

119. Murphy P.M. An Empirical Analysis of the Benefit of Decision Tree Size Biases as a Function of Concept Distribution // Technical Report 95-29. – Department of Information and Computer Science. – University of California, Irvine. – 1995. - 13p.
120. Murphy P.M., Pazzani M.J. Exploring the Decision Forest: An Empirical Investigation of Occam's Razor in Decision Tree Induction // Journal of Artificial Intelligence Research. – 1994. – Vol.1. – P.257-275.
121. Murthy S.K. Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey // Data Mining and Knowledge Discovery. – 1998. – Vol.2, №4. – P.345-389.
122. Murthy S.K., Kasif S., Salzberg S. A System for Induction of Oblique Decision Trees // Journal of Artificial Intelligence Research. – 1994. – Vol.2. – P.1-32.
123. Murthy S.K., Kasif S., Salzberg S. OC1: Randomized of Oblique Decision Trees // Proceedings of the 11th National Conference on Artificial Intelligence. – 1993. – P.322-327.
124. Murthy S.K., Salzberg S. Decision Tree Induction: How Effective is the Greedy Heuristic? // Proceedings of the 1st International Conference on Knowledge Discovery in Databases. – Canada. – 1995. – P. 222-227.
125. Murthy S.K., Salzberg S. Lookahead and Pathology in Decision Tree Induction // Proceedings of the 14th International Joint Conference on Artificial Intelligence. – Morgan Kaufmann. – P. 1025-1031.
126. Oates T., Jensen D. The Effects of Training Set Size on Decision Tree Complexity // Proceedings 14th International Conference on Machine Learning. Morgan Kaufmann. – 1997. – P.254-262.
127. Oates T., Jensen D. Toward a Theoretical Understanding of Why and When Decision Tree Pruning Algorithms Fail // Proceedings of 16th National Conference on Artificial Intelligence. – 1999.
128. Oliver J.J. On Pruning and Averaging Decision Trees // Proceedings of the 12th International Conference on Machine Learning. – 1995. – P.430-437.

129. Quinlan J.R. Bagging, Boosting, and C4.5 // Proceedings of 13th National Conference on Artificial Intelligence. – 1996. – P.725-730.
130. Quinlan J.R. Boosting First-Order Learning // Algorithmic Learning Theory, 7th International Workshop. – Springer, Sydney, Australia. – 1996. – Vol.1160. – P.143-155.
131. Quinlan J.R. Improved Use of Continuous Attributes in C4.5 // Journal of Artificial Intelligence Research. – 1996. – Vol.4. – P.77-90.
132. Quinlan J.R. MDL and Categorical Theories (continued) // Proceedings of 12th International Conference on Machine Learning. – Morgan Kaufmann. – 1995. – P.464-470.
133. Quinlan J.R. MiniBoosting Decision Trees // Journal of Artificial Intelligence Research. – 1998. – P.1-15.
134. Quinlan J.R. Unknown Attribute Values in Induction // Proceedings of the 6th International Workshop on Machine Learning, San Mateo, CA – 1989. – P. 164-168.
135. Ragel A., Crèmilleux B. MVC – A Preprocessing Method to Deal with Missing Values // Knowledge-Based Systems. – Elsevier. – 1999. – Vol.12. – P.285-291.
136. Rastogi R. PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning // Data Mining and Knowledge Discovery. – 2000. – Vol.4, №4. – P.315-344.
137. Rounds E.M. A Combined Nonparametric Approach to Feature Selection and Binary Decision Tree Deign // Pattern Recognition. – 1980. – Vol.12, №5. – P.313-317.
138. Safavian S.R., Landgrebe D. A Survey of decision Tree Classifier Methodology // IEEE Transactions on Systems, Man, and Cybernetics. – 1991. – Vol. 3. – P.660-674.
139. Schaffer C. When Does Overfitting Decrease Prediction Accuracy in Induced Decision Trees and Rule Sets? // In Proceedings of the European Working Session on Learning (EWSL-91), Berlin. – 1991. – P. 192-205.

140. Schaffer C. Overfitting Avoidance as Bias // Machine Learning. – 10. – 1993. – P.153-178.

141. Schapire R.E. A Brief Introduction to Boosting // Proceedings of the 16th International Joint Conference on Artificial Intelligence. – 1999. – P.1401-1406.

142. Schapire R.E. Theoretical Views of Boosting // Lecture Notes in Computer Science. – 1999. – Vol.1572. – P.1-10.

143. Schapire R.E., Freud Y., Barlett P., Lee W.S. Boosting the margin: A new explanation for the Effectiveness of Voting Methods // Proceedings of 14th International Conference of Machine Learning. – Morgan Kaufmann. – 1997. – P.322-330.

144. Schapire R.E., Singer Y. Improved Boosting Algorithms Using Confidence-rated Predictions // Computational Learning Theory. – 1998. – P. 80-91.

145. Simon H.U. The Vapnik-Chervonenkis Dimension of Decision Trees with Bounded Rank // Information Processing Letters. – 1991. – 39. – P.137-141.

146. Shawe-Taylor J., Anthony M., Biggs N.L. Bounding Sample Size with the Vapnik-Chervonenkis Dimension // Discrete Applied Mathematics. – 1993. – Vol.42, №1. – P.65-73.

147. Shih Y.-S. Families of Splitting Criteria for Classification Trees // Statistics and Computing. – 1999. – Vol.9. – P.309-315.

148. Shih Y.-S. Selecting the Best Splits for Classification Trees with Categorical Variables // Statistics and Probability Letters. – 2001. – Vol.54. – P.341-345.

149. Utgoff P.E. Incremental Induction of Decision Trees // Machine Learning. – Vol.4. – P.161-186.

150. Utgoff P.E., Berkman N.C., Clouse J.A. Decision Tree Induction Based on Efficient Tree Restructuring // Machine Learning. – 1997. – Vol.29, №1. – P.5-44.

151. Utgoff P.E., Clouse J.A. A Kolmogorov-Smirnoff Metric for Decision Tree Induction MA 01003: Technical Report / Department of Computer Science University of Massachusetts, Amherst – 1996. – 10 p.

152. Uther William T.B., Veloso Manuela M. The Lumberjack Algorithm for Learning Linked Decision Forests // In Symposium on Abstraction, Reformulation and Approximation (SARA-2000), Lecture Notes in Artificial Intelligence. Springer Verlag. – 2000. – Vol.1864. – P.219-230.

153. Wallace C.S., Boulton D.M. An Information Measure for Classification // Computer Journal. – 1968. – Vol.11, №2. – P.185-194.

154. Wang H., Zaniolo C. CMP: A Fast Decision Tree Classifier Using Multivariate Predictions // Proceedings of 16th International Conference on Data Engineering. – San Diego, USA. – 2000. – P.1-12.

155. Wang Y., Witten I. Inducing Model Tree for Continuous Classes // Proceedings of Poster Papers, 9th European Conference on Machine Learning. – Prague, Czech. – 1997. – P.1-10.

156. Webb G.I. Decision Tree Grafting // Proceedings of the 15th International Joint Conference on Artificial Intelligence. – Morgan Kauffman. – Nagoya, Japan. – 1997. – P.846-851.

157. Webb G.I. Decision Tree Grafting From the All-Tests-But-One Partition // Proceedings of 16th International Joint Conference on Artificial Intelligence. – Morgan Kauffman. – San Francisco, CA. – 1999. – P.702-707.

158. Webb G.I. Further Experimental Evidence Against the Utility of Occam's Razor // Journal of Artificial Intelligence Research. – 1996. – Vol. 4. – P.397-417.

159. Webb G.I. The Problem of Missing Values in Decision Tree Grafting // Proceedings of the 10th Australian Joint Conference on Artificial Intelligence. – 1998. – P.273-283.

160. Zheng Z. Constructing New Attributes for Decision Tree Learning: A Thesis…for degree of Doctor of Philosophy. – The University of Sydney NSW 2006, Australia., 1996. – 256 P.

161. Zheng Z. Naive Bayesian Classifier Committees // Proceedings of the 10th European Conference on Machine Learning. Berlin: Springet-Verlag. – 1998. – P.196-207.