Министерство образования Республики Беларусь
Учреждение образования
«Белорусский государственный университет
информатики и радиоэлектроники»

СТАТИСТИЧЕСКИЕ ОСНОВЫ ИНДУКТИВНОГО ВЫВОДА

УЧЕБНОЕ ПОСОБИЕ

Допущено Министерством образования Республики Беларусь
в качестве учебного пособия для студентов высших учебных заведений
по специальности «Искусственный интеллект»

Минск БГУИР 2009

УДК (519.23+519.816):004.8(075.8)
ББК 22.17+32.813я73
С78

Р е ц е н з е н т ы:

кафедра теории вероятностей и математической статистики
Белорусского государственного университета;

заведующий кафедрой информатики и вычислительной техники
Высшего государственного колледжа связи, кандидат технических наук,
доцент Е. В. Новиков

Авторы:
В. В. Голенков, М. Д. Степанова, С. А. Самодумкин, Н. А. Гулякина

Статистические основы индуктивного вывода : учеб. пособие /
С78 В. В. Голенков, М. Д. Степанова, С. А. Самодумкин, Н. А. Гулякина. –
Минск : БГУИР, 2009. – 202 с. : ил.
ISBN 978-985-488-313-7

Учебное пособие предназначено для студентов, магистрантов и аспирантов, изучающих проблемы индуктивного вывода как вида правдоподобных рассуждений; модели, методы и алгоритмы распознавания и обучения, а также статистические методы индуктивного вывода, необходимые для переработки и получения знаний в интеллектуальных системах.

Пособие будет полезно специалистам в области информационных технологий.

УДК (519.23+519.816):004.8(075.8)
ББК 22.17+32.813я73

ISBN 978-985-488-313-7

ОГЛАВЛЕНИЕ

Список основных обозначений и сокращений 7

Введение 8

Часть I. Общие сведения об индуктивном выводе 10

1. Умозаключение и его виды. Понятие индукции 11

1.1. Определение умозаключения и его виды 11
1.2. Дедуктивные умозаключения 12
1.3. Индуктивные выводы 13
1.3.1. Индуктивные умозаключения 13
1.3.2. Основные виды индукции 16
1.3.3. Индуктивная логика как инструмент получения нового знания 24
1.4. Традуктивные умозаключения 26

2. Индуктивный вывод 28

2.1. Индуктивные рассуждения как вид правдоподобных рассуждений 28
2.2. Правило правдоподобного вывода 30
2.3. Модели и задачи индуктивного вывода 33
2.4. Способы оценки рациональности гипотез 35
2.5. Задачи индуктивного вывода 36

3. Модели и методы обобщения 38

3.1. Модели обобщения 38
3.1.1. Модели обобщения по выборкам 38
3.1.2. Модели обобщения по данным 39
3.2. Методы обобщения 39
3.2.1. Методы обобщения по признакам 39
3.2.2. Структурно-логические методы обобщения 42
3.3. Индуктивное описание фактов 47
3.4. Правила индуктивного обобщения 48

Часть II. Автоматическое порождение гипотез для получения новых знаний 50

4. Формирование гипотез 50

4.1. Общие методы выдвижения гипотез 50
4.2. Виды обоснованности гипотез свидетельствами 51
4.3. Модели подтверждения и принятия гипотез 53

5. Имитация индуктивных рассуждений 55

5.1. Автоматическое порождение гипотез 55
5.2. Основы ДСМ-метода порождения гипотез 55
5.2.1. Общая характеристика метода 55
5.2.2. Основные этапы ДСМ-метода 57
5.2.3. Определение гипотез 57

5.2.4. Выявление сходства 58
5.2.5. Правила порождения гипотез 58
5.2.6. Рассуждения в ДСМ-методе 61

Часть III. Статистические основы индуктивного вывода 63

6. Вероятность и индукция 63

6.1. Вероятностный характер индуктивных рассуждений 63
6.2. Статистическая индукция 67
6.2.1. Умозаключение в статистической индукции 68
6.2.2. Виды статистической индукции 68
6.3. Наиболее распространенные типы вероятностно-статистических моделей, используемых в индуктивном выводе 71
6.4. Правила действий со случайными событиями и вероятностями их осуществления 72
6.4.1. Случайные события и правила действий с ними 72
6.4.2. Основные правила действий с вероятностями 73
6.4.3. Вероятность и индукция 76

7. Статистический вывод, основанный на проверке гипотез 77

7.1. Необходимость формулировки и проверки гипотез 77
7.2. Статистическая гипотеза 78
7.3. Процедура проверки гипотезы. Область принятия и отклонения гипотезы 80
7.4. Ошибки первого и второго рода 86
7.5. Мощность критерия 87
7.6. Решающее правило и статистика критерия 89

8. Общая схема статистической проверки гипотез 90

8.1. Общая схема статистической проверки гипотез 90
8.2. Понятие *P*-значения 92

9. Основные типы статистических гипотез 96

9.1. Гипотезы о типе закона распределения исследуемой случайной величины 96
9.2. Гипотезы об однородности двух или нескольких выборок и характеристик анализируемых совокупностей 97
9.3. Гипотезы о числовых значениях параметров исследуемой генеральной совокупности 97
9.4. Гипотезы о типе зависимости между компонентами исследуемого многомерного признака 98
9.5. Гипотезы независимости и стационарности ряда наблюдений 98

10. Проверка соответствия выбранной модели распределения исходным данным (критерии согласия) 99

10.1. Выбор модели распределения для описания данных 99

10.2. Критерий χ^2 Пирсона 100
10.3. Критерий Колмогорова 102

11. Гипотезы однородности 104

11.1. Критерии однородности распределений 105
11.2. Однородность математических ожиданий 106
11.3. Однородность дисперсий 110
11.4. Равенство параметров двух биномиальных распределений 112

12. Гипотезы о числовых значениях параметров 114

12.1. Гипотеза о значении математического ожидания 114
12.2. Гипотеза о значении дисперсии 118
12.3. Гипотеза о значении параметра биномиального распределения 119

Часть IV. Индуктивный вывод в машинном обучении и распознавании 121

13. Машинное обучение 121

13.1. Понятие обучения 121
13.2. Способы представления исходной информации 124
13.3. Модели обучения 125
13.4. Обучение на примерах 127
13.5. Задача обучения «без учителя» 128
13.6. Задача обучения «с учителем» 128
13.6.1. Задача обобщения понятий по признакам 129
13.6.2. Алгоритм ДРЕВ 131

14. Обучение «без учителя» 132

14.1. Общая постановка задачи распознавания в условиях отсутствия обучающих выборок 133
14.2. Меры близости 134
14.2.1. Коэффициенты корреляции как меры близости 135
14.2.2. Меры близости между объектами, описываемыми бинарными переменными 135
14.2.3. Меры расстояния 137
14.3. Алгоритмы классификации методом кластерного анализа 144

15. Деревья решений 146

15.1. Характеристики дерева решений 146
15.2. Структура дерева классификации 148
15.3. Вычислительные задачи древообразных классификаторов 151
15.3.1. Определение качества предсказания 151
15.3.2. Выбор разбиений 152
15.3.3. Определение правила прекращения разбиения 153
15.3.4. Нахождение дерева «правильного размера» 153
15.4. Построение дерева решений 155

16. Сущность задач распознавания. Классификация посредством задания границы разделения 159
16.1. Понятия классификации и распознавания 159
16.2. Математическая постановка задачи распознавания 161
16.3. Основные задачи классификации (распознавания) 162
16.4. Детерминированные системы 166
16.4.1. Алгоритм распознавания, основанный на принципе разделения 166
17. Статистические алгоритмы распознавания 169
17.1. Вероятностные системы распознавания 169
17.2. Правила классификации при известных плотностях распределения 171
17.2.1. Правило классификации максимального правдоподобия 171
17.2.2. Байесовское правило классификации 175
17.2.3. Вероятность ошибочной классификации для правила максимального правдоподобия 175
17.2.4. Классификация при различных ковариационных матрицах 176
17.2.5. Линейные дискриминантные функции Фишера 176
17.3. Классификация при наличии обучающих выборок 177
17.3.1. Подстановочное правило классификации 177
17.3.2. Оценка вероятности ошибочной классификации 179
17.3.3. Основные этапы решения задачи классификации 180
18. Логические и структурные методы распознавания 185
18.1. Логические методы распознавания 185
18.2. Структурные методы распознавания 188
18.2.1. Структурные (лингвистические) системы 188
18.2.2. Структурные (лингвистические) модели распознавания 189
18.2.3. Реализация процесса распознавания на основе структурных методов 191
Литература 193
Приложение 1. Критерии проверки гипотез 195
Приложение 2. Основные вероятностные распределения 198

СПИСОК ОСНОВНЫХ ОБОЗНАЧЕНИЙ И СОКРАЩЕНИЙ

АПГ – автоматическое порождение гипотез
ВПП – вероятностное правило принятия
ВПО – вероятностное правило отклонения
ИС – интеллектуальная система
КАТ – квазиаксиоматическая теория
МП – правило максимального правдоподобия
ПВ – правдоподобный вывод
ППВ – правило правдоподобного вывода
СВ – случайная величина
A^T – транспонированная матрица A
cov (ξ, η) – ковариация случайных величин ξ и η
D{·} – символ дисперсии
E{·} – символ математического ожидания
det (·) – определитель матрицы
E(λ) – экспоненциальный закон распределения
$F_\xi(\cdot)$ – символ функции распределения СВ ξ
$f_\xi(\cdot)$ – символ плотности распределения вероятностей СВ ξ
$F_{m,\,n}$ – закон распределения Фишера с (m, n) степенями свободы
$N(\mu, \sigma^2)$ – одномерный нормальный (гауссовский) закон распределения с математическим ожиданием μ и дисперсией σ^2
$N_p(\mu, \Sigma)$ – p-мерный нормальный закон распределения с математическим ожиданием μ и ковариационной матрицей Σ
$P(\cdot)$ – символ вероятности
t_k – закон распределения Стьюдента с k степенями свободы
u_p – квантиль порядка p закона $N(0, 1)$
$\Phi(z)$ – функция распределения закона $N(0, 1)$
$\varphi(z)$ – плотность распределения вероятностей для закона $N(0, 1)$
$\hat{q}$ – статистическая оценка параметра θ
χ^2_k – закон распределения χ^2 Пирсона с k степенями свободы
H Þ F – H объясняет F
$\vdash$ – символ выводимости

ВВЕДЕНИЕ

Данное учебное пособие – одно из серии учебных пособий, разрабатываемых на кафедре интеллектуальных информационных технологий Белорусского государственного университета информатики и радиоэлектроники. Курс «Статистические основы индуктивного вывода» изучают студенты специальности 1-40 03 01 «Искусственный интеллект».

Цели и задачи изучения курса:

- ознакомиться с основными моделями и методами индуктивного вывода, распознавания и обучения;
- ознакомиться с вероятностно-статистическими методами переработки знаний и проверки гипотез;
- изучить вероятностно-статистические методы реализации механизмов индуктивного вывода и формирования гипотез;
- освоить методы решения задач распознавания и машинного обучения в интеллектуальных системах различного назначения, базирующихся на индуктивном выводе.

Материал курса основывается на знаниях, полученных студентами при изучении дисциплин «Теория вероятностей и математическая статистика», «Математические основы интеллектуальных систем», «Проектирование баз знаний».

В теории и практике искусственного интеллекта методы индуктивного вывода используются при решении задач индуктивного формирования понятий, которые являются составной частью задач машинного обучения и распознавания образов (классификации). Целью индукции является вывод знания общего, теоретического характера исходя из наблюдений и другой известной информации, состоящей обычно из знания предметной области и примеров. Эти эмпирические знания обобщаются в новое знание.

Индуктивное обобщение наиболее часто исследуется в машинном обучении, когда имеют место наблюдения неизвестного понятия, взятые в форме описания примеров и контрпримеров. Необходимо определить понятие, которое правильно отличает примеры от контрпримеров. Аналогичные задачи возникают и при распознавании образов.

В настоящее время большое значение приобретают методы индуктивного вывода, основанные на вероятностно-статистическом подходе. Это связано

прежде всего с недостоверным характером самого индуктивного вывода и необходимостью оценки правдоподобия получаемых заключений.

В пособии материал изложен в следующей последовательности.

Часть I посвящена общим сведениям об индуктивном выводе (индуктивные рассуждения как вид правдоподобных рассуждений, модели индуктивного вывода, модели и методы индуктивного обобщения, способы оценки рациональности гипотез, автоматическое порождение гипотез).

В части II рассмотрены модели подтверждения и принятия гипотез в индуктивном выводе, а также автоматическое порождение гипотез (на примере ДСМ-метода, GUHA-метода и метода Плоткина).

Часть III содержит материал, связанный со статистическими основами индуктивного вывода (вероятность и индукция, статистическая индукция, принципы статистического вывода, основанные на проверке гипотез, основные типы статистических гипотез, методы и критерии статистической проверки гипотез).

Часть IV посвящена применению индуктивного вывода в машинном обучении и распознавании (модели обучения, обучение на примерах, задачи обучения «без учителя» и задачи обучения «с учителем», статистические алгоритмы распознавания, логические и структурные методы распознавания). Необходимость включения данной части продиктована широкой практикой применения вышеописанных методов в прикладных интеллектуальных системах.

Данное учебное пособие подготовлено при поддержке Белорусского республиканского фонда фундаментальных исследований (грант Ф08Р-137 «Проектирование интеллектуальных обучающих систем на основе модульных объектов и гибких концептуальных атласов»).

Авторы признательны рецензентам – коллективу кафедры теории вероятностей и математической статистики Белорусского государственного университета, доктору физико-математических наук, профессору Н. Н. Трушу, кандидату технических наук, доценту Е. В. Новикову.

ЧАСТЬ I. ОБЩИЕ СВЕДЕНИЯ ОБ ИНДУКТИВНОМ ВЫВОДЕ

Как показали исследования в области интеллектуальных систем (ИС), их дальнейшее совершенствование связано с автоматизацией извлечения знаний. Одним из основных направлений решения сложных проблем в этой области является использование формальных моделей различных видов рассуждений.

Составной частью интеллектуальной системы является компонент, называемый рассуждателем ИС [2]. Он состоит из генератора гипотез, доказателя теорем и вычислителя. Именно в рассуждателе ИС осуществляется реализация различных видов рассуждений.

Под ***рассуждением*** понимается построение последовательности аргументов, приводящих к принятию некоторого утверждения, которое и является целью рассуждения. Особенностями рассуждения, отличающими его от логического вывода, являются:

- открытость множества возможных аргументов;
- использование метасредств, с помощью которых осуществляется управление логическими выводами, применяемыми в процессе рассуждений;
- использование правил не только достоверного вывода, но и правдоподобного вывода.

Логический вывод в математической логике – это случай рассуждений, когда множество аргументов фиксировано, метасредства (например проверка на непротиворечивость) не используются и применяются только правила достоверного вывода, по которым из истинных аргументов (посылок) можно получить лишь истинные заключения.

В широком смысле к достоверному выводу относится дедуктивный вывод. Дедукция – это в высшей степени идеализированная форма рассуждений. Для моделирования таких аспектов человеческих рассуждений, как неопределенность, противоречивость информации и т. п., дедукции недостаточно, поэтому необходимо использовать недедуктивные или правдоподобные формы рассуждений, к которым относятся индукция и аналогия.

1. УМОЗАКЛЮЧЕНИЕ И ЕГО ВИДЫ. ПОНЯТИЕ ИНДУКЦИИ

1.1. Определение умозаключения и его виды

Ключевые понятия: умозаключение, вывод, структура умозаключения

Умозаключение – это форма мышления, посредством которой из одного и более суждений, используя определенные правила рассуждений, выводится новое суждение. ***Процесс получения суждения, являющегося необходимым или возможным следствием других суждений, называется выводом***. Умозаключение является логической операцией по получению нового знания. Новое знание может оказаться как истинным, так и ложным. Это зависит от двух причин: 1) истинности исходных суждений; 2) правильной связи исходных суждений между собой.

Структура умозаключения. Формальный анализ умозаключения начинается с выявления его структуры. Любое умозаключение состоит из **посылок** и **заключения** (или **вывода** из посылок). Посылки – это исходные суждения, а заключение – новое суждение.

По направленности логического вывода умозаключения делятся на дедуктивные, индуктивные и традуктивные.

По качеству вывода или по степени достоверности умозаключения делятся на достоверные (строгие) и вероятностные (правдоподобные, нестрогие). Достоверные умозаключения гарантируют вывод заключения из посылок, а правдоподобные дают его лишь с некоторой вероятностью.

Для формализации умозаключений используют понятие ***формальной системы***. Формальная система задается четверкой: Ф = < T, P, A, B> [20].

Здесь T – некоторое множество базовых элементов;

P – множество синтаксических правил, позволяющих строить из элементов T синтаксически правильные выражения;

A – множество априорно истинных выражений, называемых аксиомами;

B – семантические правила или правила вывода, позволяющие расширять множество истинных выражений.

В данном случае правила вывода рассматриваются как формализация правил рассуждений, используемых человеком. Формальное представление прие-

мов умозаключений часто называют *схемами выводов*. Приведенное выше определение формальной системы выделяет класс моделей рассуждений, где из истинности посылок следует истинность заключений. Такие модели называются *точными* (*дедуктивными* в широком смысле).

1.2. Дедуктивные умозаключения

Ключевые понятия: дедуктивное умозаключение (дедукция), схема дедуктивных рассуждений

В ***дедуктивных*** умозаключениях (от лат. deduction – выведение) рассуждение идет от общего к частному, т. е. в посылках содержится знание более общее, чем в заключении.

Пример 1.1

Все люди имеют родителей
N – человек

N имеет родителей

Дедукция – это цепь умозаключений (рассуждение), звенья которой (высказывания) связаны отношением логического следования. Началом (посылкой) дедукции являются аксиомы, постулаты или просто гипотезы, имеющие характер общих утверждений, а концом – следствие из посылок (частное). Если посылки дедукции истинны, то истинны и ее следствия.

Схема дедуктивных рассуждений. Выдвигается общее положение, относящееся к классу однородных объектов или явлений. Устанавливается принадлежность какого-либо объекта или явления к этому классу. Далее общий для всего класса признак переносится и на отдельный элемент класса.

Пример 1.2. «Все люди смертны», «Сократ есть человек», следовательно, «Сократ смертен».

В этом примере утверждение «Все люди смертны» является общим, относящимся к классу однородных объектов (людей). Утверждение «Сократ есть человек» устанавливает принадлежность объекта к классу. Общим признаком

для рассматриваемого класса объектов является смертность. В заключении «Сократ смертен» общий признак переносится на отдельный элемент класса.

Дедуктивный вывод представляет собой результат распространения некоторого общего знания на частный или отдельный случай. Таким образом, в процессе дедуктивного вывода происходит переход от знаний определенной степени общности к знаниям меньшей степени общности.

1.3. Индуктивные выводы

Ключевые понятия: правдоподобный вывод, индуктивное умозаключение, схема индуктивных рассуждений, виды индукции, методы индуктивного рассуждения, функции индукции

Помимо точных моделей рассуждений, существуют модели, которые необязательно являются точными. Наиболее известными среди таких выводов являются выводы из истинных посылок заключений, которые могут и не быть истинными высказываниями. Такие формализации рассуждений получили название *правдоподобных выводов*. Заключение, получаемое при использовании правил правдоподобного вывода, может быть в некоторых случаях ложным или приводить к ложным утверждениям, если его применить в качестве посылки.

1.3.1. Индуктивные умозаключения

В *индуктивных* умозаключениях (от лат. induction – наведение) рассуждение идет от частного к общему, т. е. в посылках содержится знание менее общее, чем в заключении, а переход от посылок к заключению называется обобщением.

Пример 1.3

Иванов имеет родителей
Петров имеет родителей
Сидоров имеет родителей
Иванов, Петров, Сидоров – люди

Все люди имеют родителей

Определение индуктивного умозаключения. ***Индуктивное умозаключение*** можно определить ***в узком и широком смысле***. ***В узком смысле*** под ***индуктивным умозаключением*** понимают логический переход от частных посылок к общему заключению.

В широком смысле под ***индуктивным умозаключением*** понимают определенный метод научного исследования и теоретического обобщения эмпирического опыта (наблюдений, измерений, экспериментов). В силу незавершенности человеческого опыта и нерегистрируемости класса предметов, выступающих объектом индуктивного анализа, индуктивные выводы относятся к числу правдоподобных или вероятностных умозаключений.

В индуктивной (в широком смысле) модели рассуждений среди множества априорно истинных утверждений выделяют два подмножества: элементарных утверждений (аксиомы-факты) и неэлементарных утверждений (выражения с кванторами), так называемых аксиом-правил. Аксиомы-факты и аксиомы-правила, введенные таким образом, получили название *частных* и *общих суждений*, или *фактов* и *законов*. На индуктивном выводе основана работа интеллектуальных систем, использующих методы построения общих закономерностей и понятий на базе наблюдения их проявлений в частных случаях. При этом обобщение является переходом от рассмотрения данного множества наблюдаемых фактов к большему множеству, включающему данное. Примерами таких систем являются:

- обучающие системы;
- системы распознавания и классификации;
- системы, осуществляющие извлечение знаний из баз данных;
- интеллектуальные системы различного вида, реализующие правдоподобный вывод;
- системы с автоматическим выдвижением гипотез.

Особенностью индуктивных выводов является то, что они основываются на таких правилах рассуждения, которые не гарантируют получение из истинных посылок всегда истинных заключений. Поэтому они являются ***правдоподобными*** и имеют ***вероятностный характер***. Иначе говоря, индукция подчиняется содержательным (в отличие от формально-дедуктивных) правилам вывода. Одно и то же правило индуктивного обобщения в каждом случае требует семантической интерпретации.

Схема индуктивных рассуждений. Выдвигается несколько утверждений об отдельных представителях класса, существование которого признается истинным. Каждое такое утверждение должно касаться одного и того же признака, связанного с элементами этого класса. Далее надо совершить индуктивный шаг, перейдя к утверждению о классе.

Пример 1.4. «Гомер смертен», «Фидий смертен», «Эзоп смертен». Следовательно, все люди смертны.

В этом примере каждая частная посылка является истинной. Кроме того, принадлежность элементов (человек) к одному классу (люди) также признается истинной. Поэтому можно сформулировать общее утверждение о классе: «Все люди смертны».

Запишем общую схему индуктивного умозаключения. Каждая посылка умозаключения будет соответствовать отдельному эмпирическому случаю.

Схема индуктивного вывода

1-й случай. S_1 обладает (или не обладает) свойством P

2-й случай. S_2 обладает (или не обладает) свойством P

3-й случай. S_3 обладает (или не обладает) свойством P

. . . . . .

n-й случай. S_n обладает (или не обладает) свойством P

S_1, S_2, S_3, …, S_n принадлежат к классу S

Все S обладают (или не обладают) свойством P

Если бы вышеприведенная схема была схемой дедуктивного вывода, то при наличии истинных посылок и, следуя этой схеме, мы всегда получали бы истинные умозаключения. Однако в случае индуктивного умозаключения это невозможно. Беря истинные посылки и следуя этой схеме, мы в одних случаях будем получать истинный результат, а в других – ложный. Та или иная точка зрения на заключение зависит от степени субъективной уверенности в достаточности посылок для получения заключения. Поэтому

вместо оценки истинности или ложности заключения используется ***оценка правдоподобности* (*достоверности*)**.

Подобным образом получено огромное количество научных выводов, с которыми наука обращается как с истинными знаниями (законы Ньютона, закон всемирного тяготения и др.). Однако в тех случаях, когда обобщение делается не по существенному признаку или на основе недостаточного числа наблюдений, индуктивная схема вывода приводит к ложному заключению. Подтвердим это примером.

Пример 1.5. Рассмотрим многочлен $y = x^2 - x + 41$. Подставляя вместо x значения 0, 1, …, 5, получим соответственно следующие значения многочлена: 41, 41, 43, 47, 53 и 61. Нетрудно увидеть, что все эти числа – простые. Используя индукцию, мы пришли бы к выводу, что значением этого многочлена всегда оказывается простое число. Проведем еще несколько опытов: $x = 6, 7, \ldots, 20$. Все полученные значения многочлена опять оказываются простыми числами. Этот результат еще раз показывает, что наши предположения правильны. Продолжим опыты дальше. До $x = 40$ значение многочлена всякий раз оказывается простым числом. Если же $x = 41$, то $y = x^2 - x + 41 = 41^2$. Это число уже не является простым. Следовательно, встречаются случаи, когда некоторое утверждение оказывается верным для большого количества натуральных чисел (в приведенном примере до $x = 40$). Но для еще большего натурального числа ($x = 41$) утверждение уже неверно.

При включении индуктивных правил вывода в интеллектуальную систему необходимо выбрать модель правдоподобных рассуждений, семантически адекватную данной предметной области и требуемой точности заключений. Поэтому необходимо знать различные схемы индуктивных рассуждений.

1.3.2. Основные виды индукции

Объектом индуктивного анализа выступает класс отдельных случаев, фактов, процессов. В зависимости от того, в каком объеме исследован данный класс, различают ***полную*** и ***неполную индукцию***, а по степени вероятности заключения выделяют ***перечислительную индукцию*** (или через простое перечисление при отсутствии противоречащего случая) и ***научную индукцию***.

Полная индукция. ***Полная индукция*** – такой вид индуктивного умозаключения, в котором вывод о принадлежности признака, характеризующего отдельный объект, всему классу исследуемых объектов делается на основании установленных фактов о принадлежности данного признака каждому элементу класса.

Схема полной индукции

1-й элемент класса S (S_1) обладает (или не обладает) свойством P
2-й элемент класса S (S_2) обладает (или не обладает) свойством P
3-й элемент класса S (S_3) обладает (или не обладает) свойством P
. . . . . .
n-й элемент класса S (S_n) обладает (или не обладает) свойством P
S_1, S_2, S_3, …, S_n образуют весь класс S

Все S обладают (или не обладают) свойством P

Рассуждения по методу полной индукции применимы лишь к конечным множествам фактов, поэтому с обобщением такого вида в дальнейшем обращаются как с дедуктивным результатом. Например, теорема о сумме внутренних углов треугольника доказывается при помощи метода полной индукции путем последовательного рассмотрения трех видов треугольников (остроугольного, прямоугольного, тупоугольного). Использование свода законов, каталогов моделей автомобилей дает возможность получения истинного заключения по интересующему вопросу с помощью полной индукции. Однако число таких случаев невелико. Чаще всего человек сталкивается с классом предметов, полный анализ элементов которых невозможен. В таком случае заключение делается по неполной индукции.

Неполная индукция. ***Неполная индукция*** – это такой вид индуктивного умозаключения, в ходе которого на основании принадлежности признака части элементов класса делается заключение о принадлежности признака всему классу элементов.

В неполной индукции обобщение относится к бесконечному или конечно-необозримому множеству фактов. Неполная индукция дает заключение, к которому следует относиться лишь с определенной степенью доверия. Ее выводы основываются на многократно повторяющейся повседневной и научной практике, устанавливающей причинно-следственную взаимосвязь явлений и про-

цессов и позволяющей зафиксировать существенные, повторяющиеся свойства предметов. Физические, математические, технические, социальные и другие законы являются обобщением различных научных данных.

Схема неполной индукции

1-й элемент класса S (S_1) обладает (или не обладает) свойством P
2-й элемент класса S (S_2) обладает (или не обладает) свойством P
3-й элемент класса S (S_3) обладает (или не обладает) свойством P
. . . . . .
n-й элемент класса S (S_n) обладает (или не обладает) свойством P
$S_1, S_2, S_3, \ldots, S_n$ принадлежат классу S

Все элементы класса S обладают (или не обладают) свойством P

Истинность индуктивного вывода зависит от полноты и законченности нашего опыта. Поэтому наименее достоверными (а значит наиболее ошибочными) являются выводы, полученные при помощи перечислительной индукции (*через простое перечисление*) и индукции *через отбор фактов*. Наиболее достоверной формой неполной индукции является научная индукция, которая повышает степень доверия к выводам с помощью различных методов выявления сходных и различных признаков объектов.

Перечислительная индукция. ***Перечислительная индукция*** – это такое индуктивное умозаключение, в котором на основании простого перечисления фактов по случайному признаку и при отсутствии явления, противоречащего остальным из числа перечисленных, заключают о принадлежности этого признака всему классу явлений. Перечислительная индукция исследует отношение следования между высказываниями о фактах.

Пусть имеются подмножества: базисных фактов $e_1, e_2, \ldots, e_k$ и фактов $e_{k+1}, \ldots, e_n$, эквивалентных с ними по классообразующему свойству. Рассматривается некоторое свойство h. Возможны две постановки задачи, связанной с перечислительной индукцией:

1) свойство h неизвестно, требуется сконструировать свойство, являющееся общим для двух подмножеств;

2) свойство h известно и имеет место для первого подмножества, необходимо установить, обладают ли этим свойством факты второго подмножества.

Научная индукция. ***Научная индукция*** – это вид умозаключения, в котором отбор существенных признаков класса предметов происходит через установление причинно-следственных связей между элементами данного класса. Поэтому научную индукцию называют также каузальной (от лат. causa – причина). Научная индукция включает ***элиминативную индукцию*** и ***индукцию как обратную дедукцию***.

Элиминативная индукция так же, как и перечислительная индукция, исследует отношение следования между высказываниями о фактах. Но в ней рассматривается дизъюнкция исключающих друг друга гипотез, из которой элиминируются, отвергаются все гипотезы, кроме одной.

В методе ***индукции как обратной дедукции*** полученный индуктивный вывод затем рассматривается как посылка суждения. Из этого суждения дедуктивно получают посылки индуктивного вывода, которые должны быть истинными. Например, открытые И. Кеплером (1571 – 1630) законы движения планет были получены с помощью индукции как обратной дедукции на основе анализа тщательных астрономических наблюдений Марса датским астрономом Т. Браге (1546 – 1601). В свою очередь кеплеровские законы выступили в качестве одного из индуктивных оснований механики И. Ньютона (1643 – 1727). Будучи выводимы из динамики И. Ньютона, они явились блестящим ее подтверждением.

Научная индукция рассматривает предмет не только с точки зрения повторяющихся признаков, но и условий обстоятельств и причин, при которых эти признаки возникают, а при каких – нет. Это делается при помощи различных методов.

Методы индуктивного рассуждения. Английским логиком Д. С. Миллем в конце 19 в. были сформулированы основные принципы индуктивного рассуждения в процессе опытного исследования: ***методы сходства, различий, сопутствующих изменений и остатков*** [14, 15]. Он поставил перед собой задачу нахождения связи между фактами и явлениями на основе их совместного появления или непоявления в последовательности экспериментов. Принципы установления причинно-следственных отношений, которые предложил Милль, основываются на идеях выделения сходства и различия в наблюдаемых ситуациях внешнего мира.

Для описания перечисленных выше методов введем следующие обозначения. Пусть известно, что некоторое множество сущностей x_i ($i = 1, \ldots, n$) принад-

лежит классу сущностей X. Каждая сущность из такого множества характеризуется набором свойств (Properties) P_i I U, где U – множество всех свойств сущностей $U = \bigcup_{i=1}^{n} P_i$. Будем обозначать свойства сущностей строчными латинскими буквами $a, b, c\ldots$.

Метод сходства – метод научной индукции и вид умозаключения, в котором устанавливается *сходная причина* для появления одного и того же признака у разных *явлений*.

Схема умозаключения по методу сходства

Посылка 1: $\{a, b, c\}$ I Properties $x_1 \to x_1$ I X

Посылка 2: $\{a, e, f\}$ I Properties $x_2 \to x_2$ I X

Посылка 3: $\{a, f, g\}$ I Properties $x_3 \to x_3$ I X

. . . . . .

Посылка n: $\{a, e, h\}$ I Properties $x_n \to x_n$ I X

Заключение: " x a I Properties $x \to x$ I X

На том основании, что причина характеризуется повторяемостью и во всех n случаях общим обстоятельством появления сущности x было обстоятельство a, заключаем, что a является причиной x.

В этой схеме все примеры являются положительными. Из нее в силу метода единственного сходства вытекает, что a и x связаны причинно-следственным отношением.

Количество посылок для получения хорошего обобщения не определено. Их должно быть достаточно для того, чтобы свойство a представляло характеристику класса X. Вывод по методу сходства не является безусловным, но он кажется более обоснованным, чем другие предположения, например " x b I Properties $x \to x$ I X.

Пример 1.6 [14]. Установить причину плохой посещаемости студентами лекций. Обозначим признак «плохая посещаемость» – P, а причину – S. В результате трех проверок при разных обстоятельствах получили:

• первая проверка посещаемости лекций была на первой «паре» (S) в субботу (A) в первую учебную неделю (B);

• вторая проверка была на первой «паре» (*S*) в среду (*D*) во вторую учебную неделю (*F*);

• третья проверка была на первой «паре» (*S*) в четверг (*K*) в третью учебную неделю (*M*).

Вывод: во всех трех случаях проверки общим является обстоятельство первая «пара».

Метод различия – метод научной индукции и вид умозаключения, в котором *причина признака* устанавливается на основании сравнения только двух случаев – когда данный признак возникает и когда не возникает. То, чем эти случаи различаются, считается причиной данного признака.

Схема умозаключения по методу различия

Посылка 1: $\{a, b, c\}$ I Properties $x_1 \rightarrow x_1$ I X

……

Посылка k: $\{a, e, h\}$ I Properties $x_k \rightarrow x_k$ I X

Посылка $k+1$: $\neg(\{b, c\}$ I Properties $x_{k+1} \rightarrow x_{k+1}$ I $X)$

Заключение: $"x\ a$ I Properties $x \rightarrow x$ I X

Метод различия использует посылки двух типов, которым соответствует множество ***положительных примеров*** (реализаций, где появляется свойство сущностей, принадлежащих классу *X*) и ***отрицательных примеров***, ***или контрпримеров*** (свойств сущностей, не принадлежащих этому классу).

Чтобы исключить случаи, когда *x* появляется случайным образом, не будучи связанным с *a*, необходимо повторить ситуацию *n* раз. Если число повторений (наблюдений) *n* с точки зрения экспериментатора достаточно для уверенного вывода, то, используя метод единственного различия, можно утверждать, что *a* является причиной, а *x* – следствием, т. е. что между *a* и *x* имеет место причинно-следственное отношение.

Пример 1.7 [14]

1-й случай. Простуда, переутомление, высокое давление вызывают бронхит

2-й случай. Переутомление и высокое давление не вызывают бронхита

Простуда часто бывает причиной бронхита

Метод остатков – метод научной индукции и вид умозаключения, в котором причина интересующего признака устанавливается через исключение случаев, когда данный признак не возникает. Оставшееся обстоятельство, т. е. обстоятельство, вызывающее явление, считается причиной появления признака. В этой схеме для некоторых сущностей априори известна их принадлежность нескольким классам.

Схема умозаключения по методу остатков

Посылка 1: $\{a, b, c\}$ I Properties $x_1 \rightarrow x_1$ I X_1 Ç X_2 Ç X_3

Посылка 2: $\{a\}$ I Properties $x_2 \rightarrow x_2$ I X_1

Посылка 3: $\{b\}$ I Properties $x_3 \rightarrow x_3$ I X_2

Заключение: " x { c} I Properties $x \rightarrow x$ I X_3

Вывод, получаемый этим методом, не абсолютно достоверен.

Метод сопутствующих изменений – метод научной индукции и вид умозаключения, при котором устанавливается причинная связь между изменением обстоятельств и изменением признака, возникающего при данных обстоятельствах. Данный метод позволяет сделать вывод о наличии родового отношения, связывающего два класса сущностей, если известны характеристики обоих классов. Этот метод описывается следующующей схемой:

Схема умозаключения по методу сопутствующих изменений

Посылка 1: $\{a_1, b, c\}$ I Properties $x_1 \rightarrow p_1$ I Properties y_1

Посылка 2: $\{a_2, b, c\}$ I Properties $x_2 \rightarrow p_2$ I Properties y_2

……

Посылка n: $\{a_n, b, c\}$ I Properties $x_n \rightarrow p_n$ I Properties y_n

Посылка $n + 1$: $\{a_{n+1}, b, c\}$ I Properties $x_{n+1} \rightarrow p_{n+1}$ I Properties y_{n+1}

Посылка $n + 2$: $\{a_{n+2}, b, c\}$ I Properties $x_{n+2} \rightarrow p_{n+2}$ I Properties y_{n+2}

……

Заключение: именно обстоятельство a является причиной признака P

Очевидно, что метод сопутствующих изменений не дает ответ на вопрос, какое родовое отношение связывает классы сущностей. Обычно данный метод применяется на первом этапе исследований. Вывод, получаемый этим методом, абсолютно достоверен.

Пример 1.8. Имеется последовательность из четырех (n = 4) чисел x_i:

8, 10, 16, 34, (1.1)

образованная по определенному правилу. Требуется установить вид этого правила и найти значение (n + 1)-го члена последовательности.

Для решения задачи используем метод сопутствующих изменений. Определим изменения, сопутствующие переходу от одного члена последовательности к другому в виде: а) разности $d_i = x_{i+1} - x_i$; б) отношения двух последовательных элементов d_{i+1} и d_i числового ряда $\frac{d_{i+1}}{d_i}$, i = 1, 2, …, n.

Сопутствующие изменения в виде разностей элементов ряда (1.1) равны

2, 6, 18. (1.2)

Сопутствующие изменения в виде отношений элементов ряда (1.2) равны

3, 3. (1.3)

Из сходства элементов ряда (1.3) следует, что x_5 в (1.1) равен $x_5 = 34 + 18 \cdot 3 = 88$. Правило для вычисления любого члена x_{i+1} последовательности имеет вид: $x_{i+1} = x_i + d_{i-1}\frac{d_{i-1}}{d_{i-2}}$, $i = 3, 4, \ldots, n$.

Рассматривая интерпретацию законов Милля в рамках исчисления высказываний, можно получить следующие правила вывода:

1) метод сходства:

$$\frac{\begin{array}{c} a \wedge x \rightarrow y \\ b \wedge x \rightarrow y \end{array}}{x \rightarrow y}$$

2) метод различия:

$$\frac{\begin{array}{c} a \wedge x \rightarrow y \\ a \wedge \neg x \rightarrow \neg y \end{array}}{x \rightarrow y}$$

3) метод остатков:

$$\frac{\begin{array}{c} z \wedge x \rightarrow y \wedge w \\ x \rightarrow w \end{array}}{z \rightarrow y}$$

В приведенных выше выражениях a, b, x, y, z, w могут обозначать не только переменные, но и любые формулы.

Выводы, полученные методами сходства и различия, имеют разную логическую структуру. Вывод по методу сходства является одной из гипотез (другая гипотеза $(a \to y) \vee (b \to y)$), из которой следуют посылки индукции. Вывод же по методу различия является дедуктивным следствием исходных посылок. Следует отметить, что применение приведенных выше правил может приводить к противоречиям, т. е. из некоторых наборов данных могут выводиться как утверждения $x \to y$, так и $x \to \neg y$. Необходимо лишь дополнительно оценивать достоверность индуктивных методов. В этом плане все выводы, полученные по приведенным выше правилам, являются гипотезами, достоверность которых должна быть оценена статистически.

1.3.3. Индуктивная логика как инструмент получения нового знания

Основные функции индукции. С точки зрения познавательных задач выделяют две основные функции индукции: 1) ***открытие нового знания***; 2) ***обоснование гипотез и теорий.*** Особое значение при этом придается изучению указанных функций индукции в процессе научного познания, поскольку здесь движение от опыта (данные наблюдения и эксперимента) к теории (открытие новых научных обобщений и их обоснование) является необходимым условием функционирования и развития знания.

Роль индуктивных рассуждений в получении нового знания. Одни индуктивные способы рассуждения (***полная перечислительная индукция***, ***математическая индукция***, ***схемы элиминативной индукции***) могут служить ***методами доказательства научных обобщений***, другие индуктивные способы рассуждения (***неполная индукция через перечисление***, ***индукция как обратная дедукция***) выступают ***методами подтверждения общих законов и принципов***. ***Назначение индукции*** – выдвижение и частичное обоснование гипотез.

Большую роль индукция играет на эмпирическом уровне познания, где она выступает:

1) одним из методов образования эмпирических понятий;

2) основой построения различного рода естественных классификаций;

3) одним из методов открытия эмпирических гипотез (эмпирических обобщений, причинно-следственных законов и т. п.);

4) одним из методов подтверждения и обоснования эмпирических законов.

Область применения индукции. Индукция широко применяется в повседневной деятельности, науке, культуре и образовании, технике и производстве. Например, выборочный контроль качества, проверка соответствия стандартам ведется индуктивно. Экологические тесты, статистический анализ геонаблюдений, распознавание образов в различных приложениях также основаны на индуктивной форме вывода.

Индуктивные обобщения, индуктивные экстраполяции и методы элиминативной индукции, рассчитанные на оценку и исключение конкурирующих гипотез, фактически работают почти во всех областях современной науки.

В качестве примера использования индуктивных рассуждений для получения нового знания рассмотрим применение методов Милля. Эти методы трактуются в широком смысле как некоторые общие принципы поиска закономерностей, ибо поиск закономерностей включает в себя действия по выявлению сходства, различия, сопутствующих изменений. Так, в основе многих алгоритмов классификации и распознавания образов лежит поиск общности между объектами одного класса и различий между объектами разных классов (объекты одного класса должны быть «близки» друг к другу и «удалены» от остальных). Для этой цели могут быть использованы методы разделяющих функций, двоичные деревья, минимальные тесты, кластеры и т. д. Идея поиска общностей и различий в методе разделяющих функций выражается в том, что объекты одного класса лежат по одну сторону, а разных – по различные стороны разделяющей поверхности.

Другие алгоритмы распознавания образов и формирования понятий основаны на обобщении информации об объектах, осуществляемом путем пересечения описаний объектов одного класса. Пересечение описаний объектов с целью формирования признаков класса или понятий фактически представляет собой следование методу сходства.

Существуют и алгоритмы, явно следующие методу различия. Так, формирование понятий о ситуации производится путем сравнения графа описания ситуации, соответствующей понятию, с графом описания ситуаций, являющихся контрпримерами к понятию. В результате такого сравнения в граф описания ситуации вводится описание различий сравниваемых ситуаций. На методе сопутствующих изменений основан ряд алгоритмов восстановления функций и отбора признаков изменения явлений.

1.4. Традуктивные умозаключения

Ключевые понятия: традуктивное умозаключение, умозаключение по аналогии

В ***традуктивных умозаключениях*** объем знаний в посылках и заключении одинаков, поэтому принято говорить, что в них рассуждают от частного к частному или от общего к общему. Традуктивными считаются умозаключения об отношении и по аналогии. Приведем пример умозаключения об отношении.

Умозаключение сложнее суждения
Суждение сложнее понятия

Умозаключение сложнее понятия

Умозаключение по аналогии – традуктивное умозаключение, в котором на основании сходства одних признаков предмета делается вывод о сходстве других признаков. Рассуждение по аналогии можно представить следующей схемой:

Схема умозаключения по аналогии
Предмет *M* обладает признаками *A, B, C, P*
Предмет *S* обладает признаками *A, B, C*
Вероятно, предмет *S* обладает также признаком *P*

Основные принципы рассуждения по аналогии. В основе рассуждения по аналогии лежит сходство между предметами и метод переноса признаков с одного предмета на другой. При этом степень достоверности вывода по аналогии будет зависеть как от числа сходных признаков (чем больше, тем лучше), так и от их существенности (чем существеннее признак, тем вероятней правильный вывод).

Аналогия сходна с индукцией, во-первых, по результату полученного из посылок заключения, которое имеет не достоверный, а только вероятный характер. Во-вторых, она также связана с переносом знания с одного явления или предмета, предварительно исследованного, на другие неисследованные случаи, вследствие чего и возникает неопределенность результата, оцениваемая веро-

ятностью. Разница между подходами состоит в применяемой операции порождения нового знания. В индуктивных рассуждениях – это операция индуктивного обобщения; в рассуждениях по аналогии – некоторое отображение из известной области знания в исследуемую область.

В то же время аналогия существенно отличается от индукции по своей структуре. Если при индукции речь идет о переносе знания от исследованной части к неисследованной и ко всему классу *однородных* вещей, явлений и событий, то при аналогии, как правило, устанавливается сходство между *разнородными* явлениями.

В логике принято различать аналогию, основанную на сходстве свойств, структурных признаков, отношений, функционирования, принципов действия. Кроме того, существует причинно-следственная аналогия, когда у сравниваемых предметов есть одна и та же причина появления интересующего нас свойства.

В правдоподобных рассуждениях (см. подразд. 2.1) известны два возможных принципа вывода по аналогии [2]:

1) предположение становится более правдоподобным, когда оказывается истинным аналогичное предположение;

2) предположение становится несколько более правдоподобным, когда становится более правдоподобным аналогичное предположение.

Сформулированным принципам отвечают соответственно следующие схемы правдоподобных выводов (по аналогии), в которых для обозначения аналогичных предположений использованы символы φ и ψ:

Схема 1

φ аналогично ψ

ψ – истинно

φ более правдоподобно

Схема 2

φ аналогично ψ

ψ – более правдоподобно

φ несколько более правдоподобно

Область применения вывода по аналогии. Вывод по аналогии лежит в основе моделирования и технического творчества, широко используется в литературе, исторических описаниях, философских исследованиях, юридической, педагогической и управленческой практике.

В научных исследованиях и технике широкое применение находит аналогия между моделью и ее прототипом. Она строится таким образом, чтобы модель отражала все наиболее существенные свойства и отношения своего прототипа. Идея моделирования основывается на аналогии или сходстве модели и прототипа. Теоретическое обоснование методам моделирования дает теория подобия. В последние годы все шире стало применяться математическое моделирование и основанный на нем вычислительный эксперимент. При математическом моделировании исследуются модель-аналог зависимостей, отображающих количественные связи и отношения между параметрами реальных процессов.

Подход на основании аналогии применяется и в машинном обучении. Суть его заключается в следующем. На основании имеющихся фактов и знаний о предметной области порождается некоторая новая информация путем переноса знаний из известной предметной области в исследуемую.

2. ИНДУКТИВНЫЙ ВЫВОД

2.1. Индуктивные рассуждения как вид правдоподобных рассуждений

Ключевые понятия: правдоподобное рассуждение, правдоподобный вывод, вероятностная и логическая процедуры в индуктивных рассуждениях

При получении некоторых общих выводов из совокупности имеющихся частных утверждений нужны процедуры, отличающиеся от дедуктивного вывода. Для этого можно привлекать недедуктивные или правдоподобные формы рассуждений. Под ***рассуждениями*** понимаются выводы, в которых одновременно присутствуют как точные дедуктивные умозаключения, так и эвристические соображения, которые с точки зрения классической логики не являются справедливыми.

Термин «***правдоподобное рассуждение***» принадлежит известному венгерскому математику Д. Пойа. Примерами правдоподобных рассуждений являются индукция, аналогия и различные схемы недостоверных выводов.

Выводы из истинных посылок, заключения которых необязательно будут истинными высказываниями, принято называть ***правдоподобными выводами*** (ПВ). В правилах правдоподобного вывода (ППВ) можно использовать различные типы рассуждений. Собственно индуктивные выводы являются подклассом класса правдоподобных выводов.

В индуктивных рассуждениях используются вероятностные и логические процедуры. Тип применяемой процедуры определяется спецификой изучаемого явления.

Вероятностные процедуры в индуктивных рассуждениях. Теория вероятностей изучает такие множества событий, для которых характерна повторяемость. Поведение, соответствующее принципу повторяемости, называют индуктивным поведением. В качестве математической модели индуктивного поведения используются вероятностно-статистические модели, основными отношениями которых являются отношения между частотами повторяющихся событий. В статистической индукции от посылок рассуждения в виде числовых данных, характеризующих некоторое наблюдаемое подмножество объектов (т. е. от выборки), необходимо перейти к заключению в виде статистической гипотезы, касающейся свойств генеральной совокупности.

В качестве примера можно рассмотреть схемы рассуждений, использующие свойства вероятностей.

Схема 1

Вероятность $A \rightarrow B$ больше q

Вероятность A больше r

Вероятность B больше $\max(0, q + r - 1)$

Схема 2

Вероятность $A \rightarrow B$ больше q

Вероятность B меньше r

Вероятность A меньше $\min(1, 1 - q + r)$

Схема 3

Вероятность A & B больше q

Вероятность A больше r

Вероятность B больше $q \cdot r$

Схема 4

Вероятность A & B больше q

Вероятность A меньше r

Вероятность B меньше $\min(1, q / r)$

Логические процедуры в индуктивных рассуждениях. Не всегда предсказание свойств некоторого еще не охарактеризованного события может быть сделано на основании знания частоты появления этого события. Существует широкий класс рассуждений о сходных и различных событиях, результатом которых являются прогностические высказывания о ранее не охарактеризованных событиях. Эти высказывания выводятся на основе информации о сходствах и различиях, обнаруженных у изучаемых событий. Результирующие высказывания могут иметь различную квантификацию.

Пример 2.1. Утверждение «для любого x справедливо $\varphi(x)$» записывается как $\forall x\, \varphi(x)$, где $\forall$ – квантор всеобщности; утверждение «для большинства x справедливо $\varphi(x)$» имеет вид $Mx\, \varphi(x)$, где M – квантор большинства. Индивидуальные высказывания представляются как $\varphi(a)$, где a – индивид.

2.2. Правило правдоподобного вывода

Ключевые понятия: правило правдоподобного вывода, критерий корректности или рациональности

Для каждого вида модели представления знаний [4, 18, 20] установлены процедуры вывода. Рассмотрим правдоподобный вывод для логической модели представления знаний. Для того чтобы описать такую модель, необходимо использовать соответствующие языковые средства, в качестве которых могут быть применены:

1) язык исчисления предикатов (ИП) первого порядка ***L***;

2) модификация языка исчисления предикатов, полученная посредством введения обобщенных кванторов, кванторов по кортежам, использования многозначных логик.

Вывод в исчислении предикатов. Выводом в ИП из множества гипотез Г называется последовательность $A_1, \ldots, A_n$ формул такая, что для любого i фор-

мула A_i есть либо частный случай какой-либо аксиомы, либо элементарное множество формул Г, либо непосредственное следствие каких-либо предыдущих формул по одному из правил вывода. Вывод с пустым множеством гипотез Г называется просто выводом.

Правило правдоподобного вывода. Пусть A, B – формулы $\boldsymbol{L}$, тогда правило вывода I

$$\frac{A_1, \ldots, A_n}{B}$$

будем называть ***правилом правдоподобного вывода***, если из истинности посылок $A_1, \ldots, A_n$ не всегда следует истинность заключения B.

Здесь важно отметить необязательную истинность заключений, которые получаются при использовании правил правдоподобного вывода. Это означает, что такое заключение может быть в некоторых случаях ложным или приводить (если его использовать в качестве посылки) к ложным утверждениям. Таким образом, мы предполагаем наличие множества ситуаций, когда заключение может быть как истинным, так и ложным при истинных посылках вывода. Очевидно, что правила правдоподобного вывода не являются корректными в точном логическом смысле. ***Критерием корректности или рациональности*** такого правила является количественное соотношение числа ситуаций, в которых заключение правила истинно, и числа ситуаций, в которых оно ложно.

В общем случае правило правдоподобного вывода можно назвать ***корректным*** или ***рациональным***, если заключение, построенное при помощи такого правила, истинно в большинстве случаев. Существуют различные системы правил правдоподобных рассуждений, зависящих от способа интерпретации понятия «в большинстве». С этой целью можно использовать вероятностные оценки, значение функции принадлежности или субъективную вероятность. В этом особенность индукции.

Обобщение определения ППВ. Пусть дана некоторая многозначная логика, являющаяся средством формализации ППВ. Множество всех ее истинностных значений, выделенных истинностных значений, невыделенных истинностных значений обозначим посредством V, V_1, V_2 соответственно, где

$$V = V_1 \cup V_2 \text{ и } V_1 \cap V_2 = Æ.$$

Правило вывода I будем называть ***правилом правдоподобного вывода***, если существуют такие $A_1, \ldots, A_n, B$, что $A_1, \ldots, A_n \vdash B$, $Val(A_i) \in V_1$, $i = 1, ..., n$, но $Val(B) \in V_2$, где Val – функция оценки, а символ $\vdash$ обозначает выводимость; выводимость A из множества гипотез Г обозначается через $\Gamma \vdash A$.

Правдоподобным выводом будем называть последовательность формул $A_1, \ldots, A_n$ такую, что A_i ($i = 1, ..., n$) либо гипотезы, либо формулы, истинность которых установлена, либо A_i получены применением правил вывода $I_1, ..., I_m$ ($m < n$) таких, что среди них найдется по крайней мере одно правило правдоподобного вывода I_j ($1 \le j \le m$).

Особенности применения правдоподобного вывода. Правдоподобный вывод можно разделить на статистический и нестатистический. Статистический ПВ применяется при наличии множества однородных событий, для которых характерна повторяемость.

Нестатистический ПВ применяют в ситуациях, когда исходная информация характеризуется следующими особенностями:

1) наличие нечисловых данных (т. е. структура этих данных не исчерпывается числовыми значениями, а включает также характеристики, измеренные в номинальной шкале);

2) наличие частично определенных предикатов, что дает возможность использовать при формализации описания предметной области логик неполную информацию;

3) возможность формулирования нелогических аксиом, система которых заведомо не полна;

4) наличие открытого множества эмпирических (экспериментальных) высказываний, расширяющегося в ходе развития исследований в данной науке.

Последняя особенность означает, что не все высказывания выводимы из сформулированных аксиом, т. е. существует открытое подмножество множества эмпирических высказываний, такое, что каждое высказывание этого множества не выводимо из аксиом.

В качестве примеров наук, использующих информацию, которая характеризуется указанными особенностями, можно привести биохимию, биологию, медицину, психологию, социологию.

Индуктивные методы лежат в основе таких задач, как задачи обучения, распознавания образов, восстановления функций, прогнозирования и обнаружения закономерностей.

2.3. Модели и задачи индуктивного вывода

Ключевые понятия: индуктивная модель, этапы индуктивного вывода

Модель индуктивного вывода. Под ***индуктивной моделью*** (в широком смысле) понимаются системы формальной имитации рассуждений с правилами правдоподобного вывода, которые могут быть реализованы на ЭВМ.

Как следует из определений, приведенных в подразд. 2.2, индуктивный вывод есть правдоподобный вывод с обобщением.

Пусть мы имеем множество гипотез H, свидетельства e обоснованности гипотез, критерий рациональности p (степень правдоподобности утверждения). Необходимо осуществить индуктивный вывод, т. е. синтез гипотезы h^* из некоторого множества H, удовлетворяющего критерию рациональности p, при наличии свидетельств e.

Тогда ***индуктивная модель*** может быть записана в виде следующих соотношений [21]:

$$H_c = \underset{h \in H}{\mathrm{Arg}}\,(h \to e \wedge q);$$

$$H_n = \underset{h \in H_c}{\mathrm{Arg}}\,(p(h \mid e \wedge q) > p(h));$$

$$h^* = \arg \max_{h \in H_n}\,(f(p(h \mid e \wedge q), p(h))),$$

где H, H_c, H_n – соответственно множества исходных, согласованных и подтвержденных гипотез; h – гипотеза (элемент множества H); Arg(…) и arg(…) – аргументные функции, выделяющие соответственно множество и элемент, удовлетворяющий условию, стоящему в скобках; $p(h|e \wedge q)$ – мера рациональности гипотезы при наличии свидетельства e в контексте q; f(…) – монотонная функция (обычно используется один из трех вариантов этой функции: $f = x$; $f = x - y$; $f = (x - y) / (x + y)$).

Отобранная в результате индуктивного вывода гипотеза h^* расширяет знание предметной области, представленное в рамках логической модели в виде существующей формальной теории T, на новую теорию $T' = T \cup h^*$. В конкрет-

ных интеллектуальных системах полученная гипотеза h^* поступает в базу знаний, пополняя хранящуюся в ней информацию.

Этапы индуктивного вывода. Из приведенных выше соотношений следуют четыре этапа индуктивного вывода:

1) *конструирование* множества исходных гипотез H;

2) *порождение* всех совместных со свидетельством e и контекстом q гипотез H_c;

3) *генерация* всех подтверждаемых гипотез H_n из числа совместных (гипотеза h подтверждается, если выполнено условие $p(h|e \wedge q) > p(h)$, т. е. наличие положительного примера e должно увеличить меру рациональности гипотезы);

4) *принятие* гипотезы (из всех подтверждаемых гипотез принимается та, для которой мера рациональности или монотонная функция от такой меры максимальна).

Рассмотрим примеры образования гипотез в соответствии логикой индуктивного вывода Гаека – Гавранека [5], для которой существенным является различение теоретических и эмпирических высказываний.

Пример 2.2

Эта ворона – черная

Та ворона – черная

Заключение: Все наблюдаемые вороны – черные

или

Заключение: Все вороны черные

Пример 2.3.

Эта ворона – черная

Та ворона – черная

Заключение: Произведены наблюдения над многими воронами; относительная частота встречаемости черных ворон высокая

или

Заключение: Наблюдается преобладающая тенденция в сторону черной окраски ворон.

Выделим некоторые важные особенности этих примеров индуктивного вывода. Каждый пример состоит из трех частей. Первая часть описывает имеющиеся у нас фактические данные; она может иметь форму простых предложений («Эта ворона – черная»), или, что равносильно первой, форму таблицы или какие-нибудь подобные формы. Вторая часть есть эмпирическое утверждение: оно может быть более или менее сложным предложением, утверждаемым на основе имеющихся данных. Третья часть есть теоретическое утверждение, индуктивное обобщение. Теоретическое утверждение не является следствием эмпирического; в примере 2.2 эмпирическое утверждение является логическим следствием теоретического, но в примере 2.3 ситуация более сложная. Эмпирические и теоретические вычисления связаны правилами индуктивного вывода: из теоретического предположения (основание знания) и эмпирического утверждения (описание данных) можно вывести теоретические гипотезы. Переход от эмпирического утверждения к теоретическому обосновывается некоторыми правилами рационального индуктивного вывода, даже если они явно не формулируются.

Схема индуктивного вывода такова:

$$\frac{\text{Теоретические допущения, эмпирические утверждения}}{\text{Теоретическое утверждение}}$$

или

$$\frac{\Gamma, \Delta}{\Psi},$$

где Γ – множество теоретических допущений; Δ, Ψ – соответственно множество эмпирических утверждений и теоретическое утверждение.

Эта схема означает, что если мы примем данные теоретические допущения и верифицируем данные эмпирические утверждения, то мы примем также теоретические утверждения, образующее заключение.

2.4. Способы оценки рациональности гипотез

Ключевые понятия: оценки рациональности индуктивного вывода

Гипотезы, полученные в процессе индуктивного вывода, принято оценивать с точки зрения их «разумности», «рациональности», «интересности». Рассмотрим некоторые способы оценки рациональности индуктивного вывода.

Так, в GUHA-методе [5] ***рациональность*** индуктивного вывода понимается следующим образом. Пусть Ф – истинные теоретические допущения, φ – эмпирические данные, ψ – выводимое теоретическое утверждение. Процедура вывода в рассматриваемом методе такова. Если мы примем теоретические допущения и верифицируем эмпирические утверждения, то мы примем также теоретическое утверждение, образующее заключение ψ: Ф, φ ├ ψ. Если окажется, что ψ – ложно, то вероятность появления такого наблюдения, что φ – истинно, должна быть мала (меньше 0,05).

Для оценки обоснованности гипотезы в ДСМ-методе [2, 12] (см. разд. 5) используется квантор J_m, m I {0, 1/(n – 1), 2/(n – 1), …, 1}. Значение m = 1/(n – 1) соответствует гипотезам, достоверность которых неизвестна, значение m = 1 – истинным гипотезам. Если применяемое правило подтверждает гипотезу, то значение m возрастает; если не подтверждает, то уменьшается.

В алгоритмах обобщения, использующих аппарат покрытий, принято оценивать гипотезы с точки зрения мощностей подмножеств, покрываемых ими элементов обучающей выборки.

В ряде систем для подтверждения или отрицания выдвигаемой гипотезы используются методы автоматического порождения новых элементов в виде экзаменационной (тестовой) выборки, которые выдаются для классификации человеку-эксперту или для автоматической классификации. В зависимости от результатов классификации, оцениваемых с помощью ошибки классификации, можно принять следующие решения: 1) ошибка не превышает заданной, т. е. решающее правило или обобщающее понятие построено; 2) решающее правило должно быть уточнено. Во втором случае производится корректировка обучающей выборки, на ее основе переопределяются параметры, и заново производится классификация.

В статистической индукции обоснованность вывода при проверке гипотез оценивается с помощью уровня значимости.

2.5. Задачи индуктивного вывода

Основные задачи индуктивного вывода. Среди главных задач индуктивного вывода следует отметить:

1) возможность направленного генерирования осмысленных гипотез, позволяющих переходить от наблюдаемых фактов к объясняющим их законам;

2) распространение вероятностной схемы на процедуры индуктивного вывода с целью оценки истинности индуктивного заключения по отношению к посылкам, изучения степени подтверждения гипотез, разработки критериев принятия гипотез, полученных применением индуктивных процедур.

Задачи индуктивного вывода можно сформулировать в виде пяти вопросов [5]. Эти вопросы относятся к логике открытия гипотез H при данном знании k, делающих возможным объяснение I некоторого изучаемого явления.

1. Каковы синтаксис и семантика языков, на которых сформулированы эмпирические и теоретические утверждения? Каково их отношение к исчислению предикатов первого порядка классической логики?

2. Каковы правила рационального индуктивного вывода, связывающие эмпирические и теоретические предложения? Является ли гипотеза H обоснованной знанием k?

3. Имеются ли методы для обоснования теоретического утверждения H на основе данных теоретических допущений и эмпирических утверждений k?

4. Каковы условия, при которых теоретическое утверждение или множество теоретических утверждений H при данном знании k дает наиболее разумное и интересное объяснение?

5. Существуют ли методы формирования множества гипотез H на основании данного знания k, дающих наиболее разумное и интересное объяснение?

Ответы на вопросы 1 – 3 относятся к логике индукции; ответы на вопросы 4 – 5 относятся к логике выдвижения гипотез. Построение математических моделей логики открытия (методов обнаружения и оценки некоторых высказываний) является одной из важнейших проблем искусственного интеллекта, т. к. только математические модели могут быть использованы в качестве основы для компьютерных процедур.

3. МОДЕЛИ И МЕТОДЫ ОБОБЩЕНИЯ

3.1. Модели обобщения

Ключевые понятия: обобщение, виды обобщения, методы обобщения

В результате обобщения получаются новые знания, т. е. знания, непосредственно не следующие из ранее известных. В интеллектуальных системах ***обобщение*** понимают как процесс получения знаний, объясняющих имеющиеся факты и способных объяснять, классифицировать или предсказывать новые. В общем виде задача обобщения формулируется следующим образом [2, 5].

Имеются: 1) совокупность наблюдений (фактов) *F*; 2) совокупность требований и допущений к виду результирующей гипотезы *H*; 3) совокупность базовых знаний и предположений, включающих знания об особенностях предметной области, выбранном способе представления знаний, допустимых операторов и эвристик и др. Требуется сформировать (выдвинуть) гипотезу *H*: H Þ F (*H* объясняет *F*).

Форма представления и общий вид гипотезы *H*, а также выбранные модели обобщения зависят от цели обобщения и выбранного способа представления знаний. Модели обобщения включают модели классификации, формирования понятий, распознавания образов, обнаружения закономерностей. Для каждого из перечисленных видов моделей характерны свои собственные цели обобщения, способы представления знаний, средства описания фактов, критерии оценки гипотез. Можно выделить модели ***обобщения по выборкам*** и модели ***обобщения по данным***.

3.1.1. Модели обобщения по выборкам

В ***моделях обобщения по выборкам*** совокупность фактов *F* имеет вид обучающей выборки – множества объектов e_{ij}, $i \in I$, каждый из которых сопоставляется с именем некоторого класса K_j, $j \in J$. Целью обобщения в этом случае может быть:

- формирование понятий, т. е. построение по данным обучающей выборки для каждого класса максимальной совокупности его общих характеристик;

• классификация – построение по данным обучающей выборки минимальной совокупности характеристик, которая отличала бы элементы одного класса от элементов других классов;

• определение закономерности последовательного появления событий.

К моделям обобщения по выборкам относятся лингвистические модели, модели автоматического синтеза алгоритмов и программ по примерам, модели распознавания и классификации при наличии обучающих выборок, модели обучения с учителем.

3.1.2. Модели обобщения по данным

В ***моделях обобщения по данным*** априорное разделение фактов на классы отсутствует. Здесь могут ставиться такие цели:

- получение гипотезы, обобщающей данные факты;
- выделение образов на множестве наблюдаемых данных, группировка данных по признакам;
- установление закономерностей, характеризующих совокупность наблюдаемых данных.

Такие модели обобщения используются, например, при статистической проверке гипотез, в кластерном анализе, регрессионном анализе и т. д.

3.2. Методы обобщения

По способу представления знаний и допущений о свойствах объектов, входящих в обучающую выборку, методы обобщения делятся на ***методы обобщения по признакам и структурно-логические*** (концептуальные) методы.

3.2.1. Методы обобщения по признакам

На признаковых структурах решаются обычно задачи ***классификации*** и ***формирования понятий*** [9, 23]. Признаки подразделяются на детерминированные, вероятностные, логические и структурные [23].

Задача классификации ставится следующим образом. Объекты обучающей выборки представляются в виде совокупности значений признаков. Пусть дано множество M объектов ω; на этом множестве существует разбиение на ко-

нечное число подмножеств (классов) Ω_i, $i = 1, \ldots, m$, $M = \prod_{i=1}^{m} \Omega_i$. В общем случае для каждого класса Ω_i задана информация в виде обучающей выборки $< w_i^+, w_i^- >$ множеств положительных (w_i^+) и отрицательных (w_i^-) примеров: $w_i^+ \text{ I } \Omega_i$; $w_i^- \cap \Omega_i = \varnothing$. В ряде методов классификации обучающая выборка может состоять только из положительных примеров. Пусть результатом наблюдения над объектом ω является реализация p-мерного случайного вектора $X = (x_1, \ldots, x_p)^{\mathrm{T}}$.

Задача классификации формулируется следующим образом: требуется построить ***решающее правило*** $ш_i(x)$, согласно которому по наблюденному значению вектора X произвольный объект $\omega \in M$ относится к одному из возможных классов Ω_i, $i = 1, \ldots, m$.

Задача формирования понятий. В ***задаче формирования понятий*** по обучающей выборке $W \text{ I } M$ требуется по тем или иным критериям выделить совокупность классов $\{\Omega_i\}$ и построить для них решающие правила $ш_i$.

Понятие можно представить в виде логической функции, в которой булевы переменные, соответствующие значениям признаков, соединены логическими операциями конъюнкции, дизъюнкции и отрицания. В зависимости от типа логической операции можно выделить конъюнктивные, простые и дизъюнктивно-конъюнктивные понятия [2].

Введем в рассмотрение булеву переменную h_{ij}, принимающую значения [0, 1]:

$$h_{ij} = \begin{cases} 1, & \text{если } i\text{-й признак принял } j\text{-е значение из множества} \\ & \text{допустимых значений;} \\ 0, & \text{в противном случае.} \end{cases}$$

Тогда конъюнктивным понятием назовем понятие, которое можно описать выражением вида $\&_i h_{ij}$, i – индекс признака, j – индекс его значения.

Простым понятием назовем понятие вида $\&_i \vee_j h_{ij}$. Дизъюнктивно- конъюнктивные понятия описываются выражениями вида $\vee_k \alpha_к$, где $\alpha_к = \&_i h_{ij}$.

Предложенный способ определения значения h_{ij} используется только для качественных признаков. Если признак принимает количественные значения, можно задать h_{ij} следующим образом:

$$h_{ij} = \begin{cases} 1, \text{ если значение } i\text{-го признака лежит в } j\text{-м интервале } \{a_j, b_j\}; \\ 0, \text{ в противном случае.} \end{cases}$$

В этом случае множество всех допустимых значений для признака образуется объединением таких интервалов: $\mathrm{U}_j\{a_j, b_j\}$.

Решение задач классификации и формирования понятий осуществляется в признаковом пространстве. Обозначим R^p пространство возможных значений наблюдаемых векторов признаков x. Тогда классы являются областями $R_1, \ldots, R_m$ пространства R^p, и решающие правила $ш_i$, $i = 1, \ldots, m$, можно представить в виде разделяющих поверхностей в пространстве признаков. В процессе обобщения происходит оценка коэффициентов разделяющих функций, которые, как правило, выбираются линейными, квадратичными или кусочно-линейными. Охарактеризуем ряд математических моделей, используемых в методах обобщения по признакам.

В зависимости от того, какого рода признаки используются в методах обобщения, можно выделить несколько типов математических моделей. Это ***детерминированные, вероятностные, логические, структурные и комбинированные модели***.

Статистические модели. Этот тип моделей основан на использовании аппарата математической статистики. Данные модели применяются в тех случаях, когда известны или могут быть просто определены вероятностные характеристики классов, например соответствующие функции распределения [17, 23].

Модели, построенные на основе так называемого «метода потенциальных функций» [17, 23]. В основе этих моделей лежит заимствованная из физики идея потенциала, определенного для любой точки пространства и зависящего от расположения источника. В качестве функции принадлежности объекта классу используется потенциальная функция – всюду положительная и монотонно убывающая функция расстояния. Потенциальная функция $ш_k$ для класса K строится так, чтобы ее значение на множестве объектов ω I K было максимальным.

Модели вычисления оценок (метод голосования) [9, 23]. В этих моделях анализируется «близость» между частями описаний (подмножеств) ранее классифицированных объектов и объекта ω, который надо распознать. Близость объекта ω к каждому из классов определяется на основании сравнения значений

признаков объекта ω из заданных подмножеств с соответствующими значениями признаков эталонных объектов. По набору оценок близости вырабатывается общая оценка принадлежности распознаваемого объекта классу.

Модели, основанные на исчислении высказываний, в частности, на аппарате алгебры логики. В этих моделях классы и признаки объектов рассматриваются как логические переменные, а описание классов на языке признаков представляется в форме булевых соотношений.

Лингвистические (синтаксические) модели. В этих моделях для построения алгоритмов распознавания используются специальные грамматики, порождающие языки. Языки состоят из предложений (цепочек символов-слов), каждое из которых описывает объекты, принадлежащие конкретному классу. Решающее правило ищется в виде грамматики *G*, такой, что предложения из множества положительных примеров w^+ порождаются *G*, а предложения из множества контрпримеров w^- не порождаются.

Применение синтаксических методов требует наличия совокупностей предложений для описания множества объектов, принадлежащих всем классам. При этом множество предложений должно быть подразделено на подмножества по числу классов. Элементами подмножеств являются предложения, описывающие объекты, принадлежащие данному подмножеству (классу).

3.2.2. Структурно-логические методы обобщения

Структурно-логические методы обобщения используются в формально-логических системах. Эти методы предназначены для решения задачи обобщения на множестве объектов, имеющих логическую структуру (последовательности событий, иерархически организованные сети, алгоритмические и программные схемы).

Особенность индуктивных формальных систем правдоподобного вывода состоит в том, что аксиомы-факты в такой системе являются истинными утверждениями, а законы, выводимые по правилам индуктивного вывода, не абсолютно истинны. Чтобы подчеркнуть это обстоятельство, в формальной системе выделяются два языка: язык описания теории и язык описания фактов или наблюдений. Для связи между этими языками в язык описания теории вводятся специальные кванторы, имеющие статистическую интерпретацию. В настоящее

время известен целый ряд формальных систем индуктивного вывода. Основой каждой из них является набор правил (схем) индуктивного вывода.

Далее будет дана краткая характеристика схем вывода Плоткина, GUHA-метода и метода обобщения на семантических сетях [2]. ДСМ-метод [2, 12, 23] будет рассмотрен в разд. 5. Методы Плоткина, GUHA, ДСМ относятся к методам автоматического порождения гипотез, реализующих автоматическое извлечение знаний из фактов.

Схема индуктивного вывода Плоткина. В основу этого метода положена схема ***индуктивной резолюции*** (процедура ***обратной дедукции***). Имея множество фактов, выраженных на языке логики предикатов первого порядка, данный ***метод строит индуктивную гипотезу (в виде формулы того же языка), из которой факты могут быть выведены дедуктивно.*** Каждый факт представляется в виде совокупности условий (конъюнкции литералов) e_i и результата f_i – дизъюнкции литералов. Исходя из такого описания фактов, а также из набора общих утверждений (*Th*-аксиом) предметной области, представленных в виде конъюнкции литералов, с использованием метода Плоткина генерируется формула *H*.

Формула *H*, называемая объяснением результата f_i при заданных условиях e_i и общих утверждениях *Th*, является непротиворечивым расширением множества аксиом и обеспечивает

$$\vdash (e_i \,\&\, Th \,\&\, H) \rightarrow f_i,\ i = 1, \ldots, n.$$

При построении индуктивного обобщения *H* метод Плоткина использует алгоритм антиунификации. Этот алгоритм строит «наименее общее обобще-ние» двух предложений на основе замены констант в этих предложениях переменными.

GUHA-метод автоматического порождения гипотез [5] – это общий метод автоматического выдвижения гипотез. Он представляет собой ***систему процедур порождения и оценки элементарных высказываний-гипотез об одновременной выполнимости заданного набора свойств на имеющихся эмпирических данных***. Такие гипотезы называются эмпирическими высказываниями. Гипотезы, порожденные GUHA-методом, имеют вид $\psi \sim \varphi$, где ψ и φ – утверждения, имеющие единственную свободную переменную *x*, а символом ~ обозначен связывающий *x* ассоциативный или индуктивный квантор: $(\sim x)(\psi, \varphi)$. Выражение $\psi \sim \varphi$ в зависимости от выбранного квантора может служить для представления широкого диапазона зависимостей между ψ и φ (от логического следования до статистической корреляции).

В GUHA-методе предложено индуктивное исчисление, оперирующее статистическими высказываниями. Принцип метода состоит в автоматическом образовании и отборе всех заслуживающих внимания эмпирических гипотез. Индуктивное обобщение эмпирических гипотез производится по критерию рациональности, гарантирующему истинность соответствующей теоретической гипотезы с высокой вероятностью.

Для исчисления используется модификация исчисления предикатов первого порядка. Модели исчисления имеют вид

$M = < Mod,\ f_1, \ldots,\ f_n >$,

где Mod – конечное и непустое базовое множество; $f_i : Mod \rightarrow V\ (i = 1, ..., n)$ – унарные отображения, $V_i = \{1, 2, ..., k_i, Un\}$, Un – символ неопределенной информации.

Язык эмпирического исчисления состоит из функциональных символов F_i, выделенных переменных x, символов логических операций &,V, $\rightarrow$, $\neg$, (K), кванторов q. Атомарная формула, или литерал, имеет вид $(K)F_i(x)$, где K – подмножество V_i.

Интерпретация атомарной формулы $(K)F_i(x)$ осуществляется следующим образом: для любого объекта o из Mod

$|(K)\ F_i\ (x)|\ [o] = 1$, если $f_i\ (o)$ принадлежит K;

$|(K)\ F_i\ (x)|\ [o] = U$, если $f_i\ (o) = Un$;

$|(K)\ F_i\ (x)|\ [o] = 1$, если $f_i\ (o)$ не принадлежит K.

Открытые формулы строятся из литералов с помощью символов логических операций «конъюнкция» &, «дизъюнкция» $\vee$, «следование» $\rightarrow$, «отрицание» $\neg$, имеющих для значений 1 (истина), 0 (ложь) обычную интерпретацию. Квантор q связывает две открытые формулы и имеет обобщенную интерпретацию статистического характера.

Пусть предметная область представлена обучающей выборкой. Формально обучающая выборка приводится в виде модели M. Пусть цель исследования представлена в виде открытой формулы $P1(x)$. Принцип GUHA-метода состоит в переборе всех открытых формул вида $P(x)$ и вычислении значения кванторной формулы $(qx)(P1(x), P(x))$ в модели M. Истинные модели считаются обобщенными с конечной модели M на всю предметную область. Квантор q воплощает индуктивный шаг. Так как задача научного исследования заключается в выяв-

лении взаимосвязей, квантор q должен образовывать высказывания с соответствующей интерпретацией. Так как открытая формула либо истинна, либо ложна, либо неопределена на любом объекте модели, она может рассматриваться как признак. Тогда квантор q соответствует некоторому критерию проверки независимости двух признаков. Если принимается гипотеза о зависимости двух признаков, то формула $(qx)(P1(x), P(x))$ считается истинной. Корректность индуктивного шага обеспечивается корректностью соответствующего критерия проверки гипотез.

Формулы, составляющие с $P1(x)$ истинную кванторную формулу (прошедшие по критерию), образуют множество, называющееся решением. В GUHA-методе решение представлено в виде двух множеств – базиса и дополнения. Каждой формуле $P(x)$ из базиса соответствует одна формула $L(x)$ из дополнения. Формулы множества решения выводятся из формулы $P(x)$ присоединением любой (в том числе пустой) подформулы $L(x)$.

Формула $P(x)$ относится к базису, если формула $(qx)(P1(x), P(x))$ истинна, а все формулы вида $(qx)(P1(x), Q(x))$ ложны в модели M. Здесь $Q(x)$ – любая непустая подформула $P(x)$. Иными словами, для любой формулы базиса множество решения не содержит никакой подформулы этой формулы. Формула дополнения образуется из соответствующей формулы базиса по определенному правилу. Поэтому в процессе вычисления сначала ищутся формулы базиса, а затем из них определяются формулы дополнения.

В индуктивном выводе используются статистические процедуры проверки порожденных гипотез, однако этот метод содержит также правила правдоподобного вывода, которые могут быть отнесены к структурной индукции. Рациональные правила вывода позволяют с помощью имеющихся эмпирических и теоретических высказываний (входящих в посылки вывода) выводить новые теоретические высказывания относительно обрабатываемых данных.

Перечислим, как отражается последовательность шагов научного исследования в процедуре GUHA-метода.

1. Автоматически порождаются и оцениваются все гипотезы в языке исчисления предикатов первого порядка.

2. Выделение базиса соответствует выделению существенных знаний, так как базис определяет все решение. Определение базисной формулы соответствует определению сущности явления как такой совокупности признаков, что удаление любого из них разрушает (элиминирует) само явление.

По аналогии с ассоциативными знаниями человека знания в GUHA-методе также имеют характер эмпирических ассоциаций, а в основу проверки гипотез положена определенная мера зависимости. Знания GUHA-метода формулируются в языке исчисления предикатов первого порядка, позволяющем отобразить любые понятия. Поэтому GUHA-метод предметно-независим.

Методы обобщения на сетях. Применение аппарата семантических сетей для решения задач классификации и формирования понятий позволяет получать обобщенные представления множеств объектов, имеющие семантическую трактовку, и использовать в процессе обобщения семантику таких понятий, как «признак», «имя», «класс», «отношение». Язык представления наблюдений и классов (понятий) основан на формализме семантической сети.

Рассмотрим суть метода обобщения на сетях [2]. Пусть M – множество объектов ω. Каждый объект ω I M представляется сетью, называемой семантическим графом, который включает вершины двух типов: объектные и предикатные. Объектной вершине приписывается имя объекта, имя базового класса и вектор его признаков. Предикатной вершине приписывается имя отношения (возможно, с отрицанием). Семантический граф, предназначенный для представления объекта ω I M, распадается на иерархически упорядоченное множество p-подграфов, служащих для представления объекта ω, его частей и т. п.

Обобщенный семантический граф также представляется в виде обобщенных p-подграфов (op-подграфов). Каждый op-подграф предназначен для представления множества K объектов. Объектной вершине первого уровня op-подграфа приписывается имя множества K, имя базового класса T (K I T) и совокупность ограничений на изменения значений признаков объектов из K. Объектной вершине второго уровня приписывается имя множества K_v, которое может быть получено применением конечного числа операций объединения и пересечения к множествам, представимым на op-подграфах. Предикатным вершинам op-подграфов приписаны выражения логики исчисления высказываний, в которых роль высказываний играют имена отношений между объектами.

Обобщения g на семантической сети получаются применением к элементам g операторов обобщения, таких как удаление вершины из графа, замена значения признака множеством значений, замена имени отношения Q на предикатной вершине выражением $L = Q \vee R$.

3.3. Индуктивное описание фактов

<u>Ключевые понятия</u>: индуктивное описание фактов, способы индуктивного описания фактов

При построении компьютерных интеллектуальных систем, основанных на переработке знаний в соответствии с моделью индуктивного вывода, применяются индуктивное описание фактов и правила индуктивного обобщения. Рассмотрим некоторые положения на основании работ В. К. Финна и В. Г. Ивашко.

Индуктивное описание фактов. Под ***индуктивным описанием*** мы будем понимать построенное обобщение на основе установления сходства в рассматриваемом множестве примеров. Тип индуктивного описания фактов зависит от назначения интеллектуальной системы. Наиболее часто используются ***характеристические описания*** класса объектов, т. е. индуктивные описания, которые содержат все признаки, общие для объектов данного класса. Такому индуктивному описанию должны удовлетворять все известные объекты из исследуемого класса (условие полноты) и только они; на любом известном отрицательном примере класса характеристическое описание выполняться не должно (условие непротиворечивости). Цель построения характеристического индуктивного описания – выделить исследуемый класс объектов среди всех возможных других. Примером описания такого типа в языке логики первого порядка может служить конъюнкция предикатов-признаков, которая удовлетворяет следующим условиям: 1) непротиворечивости; 2) полноты. Добавление любого другого конъюнкта нарушает полноту.

Когда необходимо лишь различать объекты выделенного класса среди объектов конечного и фиксированного числа других классов, используется ***отличительное описание***. Оно содержит минимальное множество признаков, достаточное, чтобы выделить исследуемый класс объектов среди фиксированного числа других. Все алгоритмы распознавания образов, которые строят разделяющие поверхности в пространстве признаков, ведут поиск именно отличительного описания. Есть еще один тип индуктивных описаний – таксономия. В этом случае все известные объекты разделяются на классы в соответствии с различными понятиями, которые не заданы заранее, а вырабатываются в процессе обучения. Число различных классов не задается и не фиксируется. Задача автоматического построения таксономических описаний – обобщение задачи кластерного анализа в случае нечисловых объектов.

3.4. Правила индуктивного обобщения

Ключевые понятия: правило опускания несущественных условий, правило замены констант, правило замыкания интервалов, правило обобщения «вверх по дереву», правило «конструктивной» индукции, стратегия обобщения

Каждая интеллектуальная система, автоматически извлекающая знания из фактов, использует свои специфические правила индуктивного обобщения, записанные на соответствующем формальном языке. Арсенал используемых для этого средств не исчерпывается логикой предикатов первого порядка. Здесь можно назвать методы распознавания образов, сети, фреймы и т. д. Можно, однако, выделить некоторые типы правил, общих для таких ИС. Различия будут заключаться в условиях применимости правил.

Правило опускания несущественных условий. При индукции через простое перечисление существенными являются те условия, которым удовлетворяет некоторое множество наблюдаемых объектов. Выделенные существенные условия затем распространяются в силу индуктивного предположения на любые объекты соответствующего класса. В индуктивных методах Д. С. Милля (см. п. 1.3.2) существенные признаки выделяются путем установления максимального сходства и различия наблюдаемых явлений, причем максимальное множество общих признаков считается обнаруженной причиной принадлежности объекта к исследуемому классу. В методе различия Д. С. Милля удаление множества существенных признаков сразу нарушает принадлежность объекта к соответствующему классу. Правила опускания несущественных условий наиболее употребительны в системах машинного обучения.

Правило замены констант в описании объектов на переменные. Примерами правил такого типа являются правила расширения области определения переменных и области действия кванторов.

Правило замыкания интервала для числовых данных. Это правило можно сформулировать следующим образом. Пусть задан числовой интервал $[a, b]$ и условие $\varphi(x)$, содержащее единственный числовой параметр x; пусть $\varphi(a)$ и $\varphi(b)$ истинны, тогда будем считать $\varphi(x)$ истинным для всех значений параметра x, таких, что $a \le x \le b$.

Правило обобщения «вверх по дереву» классификации структурированных фактов. В этом случае для некоторого параметра x имеется дерево родовидовой классификации и задано условие $\varphi(x)$ со структурным параметром x. Если для значения s параметра x имеются непосредственные потомки a, b, c и $\varphi(a)$, $\varphi(b)$, $\varphi(c)$ – истинны, то будем считать истинным и утверждение $\varphi(s)$.

Правило «конструктивной» индукции. Под конструктивной индукцией понимают процедуры порождения новых параметров, характеризующих наблюдаемые объекты, из исходных параметров. Фактически это правило состоит в расширении пространства признаков на основании изучения соответствующей выборки объектов. В частности, может быть построена числовая функция от исходных параметров, характеризующих объекты выборки. Следует отметить, что правила «конструктивной» индукции являются скорее практическими и не имеют достаточной формализации.

Стратегия обобщения. В литературе выделяют методы, обобщающие исходные факты «снизу – вверх» (стратегия, определяемая данными), и методы обобщения «сверху – вниз» (стратегия, определяемая моделью). В первом случае упорядочивается множество положительных примеров $\{e_1, ..., e_n\}$ и строится множество индуктивных обобщений $\{G_1, ..., G_n\}$, таких, что G_1 совпадает с описанием e_1. Для всякого i, $1 \leq i \leq n-1$ строится G_{i+1}, которое обобщает G_i и e_{i+1}. При этом подразумевается, что ни один контрпример обобщению G_{i+1} не удовлетворяет. G_n является индуктивным обобщением всего множества положительных примеров класса. Примерами применения стратегии «снизу – вверх» служат индуктивные методы Г. Плоткина.

При стратегии «сверху – вниз» исследуется возможность применения некоторых стандартных обобщений-шаблонов на всем массиве исходных фактов. Задается критерий предпочтения, который отбирает среди множества допустимых индуктивных описаний «лучшие». Такая стратегия более устойчива против «зашумления» данных, но ее вычислительная сложность больше вычислительной сложности стратегии «снизу – вверх».

ЧАСТЬ II. АВТОМАТИЧЕСКОЕ ПОРОЖДЕНИЕ ГИПОТЕЗ ДЛЯ ПОЛУЧЕНИЯ НОВЫХ ЗНАНИЙ

4. ФОРМИРОВАНИЕ ГИПОТЕЗ

4.1. Общие методы выдвижения гипотез

Ключевые понятия: отношения предпочтения и совместимости

Методы выдвижения гипотез включают в себя процедуры сравнения гипотез: лучшей считается та гипотеза, которая «проще» и «более совместима» с исходными данными. Содержание этих понятий определяется самим существом используемого метода. Например, можно задать два отношения α и β, называемые отношениями предпочтения и совместимости, такие что

$h_1 \alpha h_2$ U «h_1 проще h_2»;

$h \beta V$ U «h совместима с обучающей выборкой V».

Тогда можно полагать, что гипотеза h_1 лучше гипотезы h_2 при выборке V, если $h_1 \alpha h_2$ и $h_1 \beta V$, $h_2 \beta V$.

Например, в задачах синтеза языков часто применяются следующие отношения [11]:

$h_1 \alpha h_2$ U «$L(h_1)$ I $L(h_2)$»;

$h \beta V$ U «V^+ I $L(h)$ и $V^- \cap L(h) = \varnothing$»,

где $L(h)$ – язык, определяемый гипотезой h;

V^+ и V^- – множества положительных и отрицательных примеров из V.

Для любых h_1 и h_2 из $h_1 \alpha h_2$ следует, что если h_1 лучше h_2 при любой выборке V, с которой h_1 и h_2 совместимы, то отношение предпочтения α называется независимым от выборки.

Если независимое от выборки отношение предпочтения вычислимо, то можно привести простую схему выдвижения гипотез, применяемую во многих приложениях.

Шаг 0. Вычислить наилучшую при данной выборке гипотезу. «Потребовать» новый пример. Перейти к шагу 1.

Шаг *n* (*n* > 0). Если текущая гипотеза совместима с примерами, то «потребовать» новый пример и перейти к шагу $n + 1$. В противном случае перейти к шагу 0.

Отношение предпочтения можно сформулировать и для задачи синтеза языков по положительному представлению: гипотеза h_V I H считается наилучшей для обучающей выборки V, если V I $L(h_V)$ и для всех h I H, таких, что V I $L(h)$, выполнено $L(h_V)$ I $L(h)$, где $L(h)$ означает язык, определяемый гипотезой H.

Большое распространение получили отношения предпочтения, основанные на теореме Байеса. Пусть $P(h)$ есть вероятность появления гипотезы в вероятностном пространстве H; $P(V)$ – вероятность появления выборки в пространстве примеров; $P(V|h)$ – вероятность того, что h совместима с данными V. Тогда вероятность

$$P(h|V) = \frac{P(h)P(V|h)}{P(V)}$$

может быть использована для определения отношения α: $h_1 \alpha h_2$ тогда и только тогда, когда $P(h_1|V) \geq P(h_2|V)$. Задача состоит в максимизации $P(h|V)$.

Отношение предпочтения можно использовать для структуризации гипотез, что позволяет исключить не одну, а группу неудачных гипотез. Это отношение часто применяется в системах формирования понятий при индуктивном выводе. На множестве всех конъюнкций атомарных формул вводится отношение предпочтения $K_1 \alpha K_2$ для двух множеств атомов. Задача формирования понятия заключается в нахождении минимальной общей конъюнкции для заданного множества фактов, представленных атомарными формулами, не содержащими переменных. Если некоторая конъюнкция не согласована с каким-либо фактом, то и все менее общие конъюнкции можно исключить.

4.2. Виды обоснованности гипотез свидетельствами

Ключевые понятия: совместимость, подтверждение, приемлемость гипотез

Имеются три вида обоснованности гипотез свидетельствами: ***совместимость, подтверждение*** **и** ***приемлемость*** (т. е. возможность принять гипотезу на основе свидетельства). Первый из этих видов обоснованности имеет дедуктивную природу, два других вида – индуктивные.

Совместимость. Свидетельство *E* ***совместимо*** с гипотезой *H*, если и только если из *E* не следует отрицание *H*. Если гипотеза несовместима со свидетельством, то свидетельство дедуктивно опровергает эту гипотезу. Одним из способов дедуктивного опровержения гипотез является правило *модус толленс* $A \supset \overline{B}$, $B \vdash \overline{A}$. Однако существуют и недедуктивные способы опровержения гипотез.

Подтверждение. ***Подтверждение*** является более сильным видом обоснованности гипотез, чем совместимость. Свидетельство подтверждает гипотезу, если гипотеза совместима со свидетельством и если она обоснована при наличии истинного свидетельства в большей степени, чем при его отсутствии (т. е. если свидетельство вносит положительный вклад в обоснование гипотезы).

Помимо гипотез *H* и свидетельств *E*, при рассмотрении процедур подтверждения и принятия гипотез будем использовать также исходное знание *W*, на базе которого выдвигаются гипотезы и ставятся эксперименты. В качестве *W* могут выступать также совокупности граничных и начальных условий, вспомогательные гипотезы, с помощью которых осуществляется выведение из гипотезы эмпирических следствий. Для рассмотрения процедуры принятия *H*, помимо свидетельств и исходного знания *W*, нужно также иметь множество альтернативных гипотез, к которому принадлежит данная гипотеза *H*.

Приемлемость. Наиболее сильным видом индуктивной обоснованности является ***приемлемость***. Гипотеза может быть принята на основе свидетельства, если она подтверждается этим свидетельством в большей степени, чем любая из известных ее альтернатив (конкурирующих гипотез). Приемлемость гипотезы зависит не только от нее самой и имеющихся свидетельств, но и от совокупности альтернатив.

Различие между процедурами подтверждения и принятия гипотезы. Между индуктивными процедурами подтверждения и принятия гипотезы существует следующее различие. Каждая гипотеза подтверждается своими истинными эмпирическими следствиями, но не каждая подтвержденная гипотеза может быть принята. Одними и теми же данными одновременно могут подтверждаться многие (но не все) конкурирующие гипотезы, а принятой может быть только одна из них. При этом различные исследователи на основе одинаковых данных могут принимать в одно и то же время различные гипотезы.

Отношение между приемлемостью гипотезы и ее истинностью. Между приемлемостью гипотезы и ее истинностью существует сложное отношение:

принимаемая гипотеза необязательно истинна (поскольку принятие является индуктивной, а не дедуктивной процедурой и не исключает ошибок), а истинная гипотеза (на тех же основаниях) не всегда принимается. Иногда гипотеза принимается не потому, что она истинна, а в силу того, что неизвестны лучшие альтернативы. Уточнение процедур подтверждения и принятия гипотезы производят в вероятностной теории индукции.

4.3. Модели подтверждения и принятия гипотез

Ключевые понятия: вероятностная модель подтверждения гипотезы, вероятностная модель приемлемости гипотезы, правило принятия и отклонения гипотезы

Модель подтверждения гипотезы. Рассмотрим некоторые модели подтверждения и принятия (опровержения) гипотез, использующие полученные значения вероятностей. Простейшая ***вероятностная модель подтверждения гипотезы*** H свидетельством E при данном исходном знании W характеризуется вероятностным неравенством

$$P(H, E \wedge W) > P(H, W). \qquad (4.1)$$

Главная особенность этого определения состоит в том, что подтверждение понимается не как условная вероятность гипотезы относительно свидетельства, а как увеличение вероятности гипотезы в случае истинности свидетельства. Однако поскольку числовые значения $P(H, E \wedge W)$ и $P(H, W)$ обычно неизвестны, определение (4.1) не может быть применено для установления факта подтверждения конкретной гипотезы конкретным свидетельством. Поэтому помимо определения (4.1) имеет место следующая теорема подтверждения (Т1), дающая качественный критерий подтверждения гипотез свидетельствами:

Т1. Если $H, W \vdash E$, $W \vdash E$ и $0 < P(W) < 1$, то E подтверждает H независимо от конкретных значений вероятностей H, E и W. (***Каждая гипотеза подтверждается своими эмпирическими свидетельствами.***)

Следствие E в **Т1** должно быть эмпирическим. При $P(E) = 1$ выполнение неравенства (4.1) невозможно: никакое свидетельство с априорной вероятностью 1 не может подтвердить никакой гипотезы.

Неравенство (4.1) и теорема **Т1** не эквивалентны. Подтверждение **Т1** транзитивно влево: если E подтверждает H и $H_1 \vdash H$, то E подтверждает H_1. Но подтверждение на основе (4.1) необязательно транзитивно (влево или как-то иначе).

Теорема **Т1** отражает также и неоднозначность процедуры подтверждения: одно и то же свидетельство может подтверждать различные (и даже альтернативные) гипотезы. Эта теорема позволяет также понять принцип, согласно которому гипотеза не может подтверждаться теми фактами, которые она собирается объяснить (эти факты содержатся в исходном знании W).

Модель приемлемости гипотезы. Простейшая ***вероятностная модель приемлемости гипотезы*** H исследователем a на основе свидетельства E и исходного знания W выражается неравенством

$$P_a(H, E \wedge W) > P_a(H_j, E \wedge W), \qquad (4.2)$$

которое должно выполняться для любой гипотезы H_j из множества альтернатив, рассматриваемых субъектом a. Вероятность P_a в (4.2) выражает зависимость принятия гипотезы от субъекта познания a.

Правила принятия и отклонения гипотезы. На практике условие (4.2) записывается в следующем виде: гипотеза H может быть принята исследователем на основе свидетельства E при исходном знании W, если

$$P_a(H, E \wedge W) \geq 1 - \varepsilon,\ 0 \leq \varepsilon < 0{,}5. \qquad (4.3)$$

Если $P_a(H, E \wedge W) \leq \varepsilon$, то H можно отклонить.

Соответствующие (4.3) ***индуктивные правила принятия и отклонения*** можно назвать вероятностным правилом принятия (ВПП) и вероятностным правилом отклонения (ВПО). Для ВПП верна транзитивность вправо: если принимается некоторая гипотеза, то принимаются и любые ее следствия. Полагая $\varepsilon < 1/k$, можно распространить эти правила на k различных альтернатив, дизъюнкция которых истинна.

Особенности правил принятия и отклонения. Рассмотренные правила имеют и такую особенность: ВПП не замкнуто относительно конъюнкции (из приемлемости H_1 и приемлемости H_2 на основе E не следует приемлемость $H_1 \wedge H_2$ на той же основе), а ВПО не замкнуто относительно дизъюнкции (из отклонения H_1 и отклонения H_2 свидетельством E не следует отклонение $H_1 \cup H_2$ тем же свидетельством).

При дедуктивном выводе справедливо следующее **правило принятия и отклонения гипотез**:

из приемлемости каждой из $n \geq 2$ гипотез следует и приемлемость их конъюнкции, а из отклонения каждой из $n \geq 2$ гипотез следует и отклонение их дизъюнкции.

5. ИМИТАЦИЯ ИНДУКТИВНЫХ РАССУЖДЕНИЙ

5.1. Автоматическое порождение гипотез

Интеллектуальные системы обладают важным свойством – способностью извлекать знания, в частности, порождать индуктивное обобщение из имеющихся фактов, выдвигать предположения о неизвестных закономерностях или виде некоторой неизвестной функции в частично неопределенной информационной среде. Такие предположения носят название гипотез. Автоматическое порождение (выдвижение) гипотез, или автоматическое гипотезирование в интеллектуальных системах происходит различными методами. Наибольшую известность приобрели ***методы обучения***, ***GUHA-метод*** и ***ДСМ-метод***. Характеристика GUHA-метода была дана в подразд. 3.2.2, основные принципы методов обучения будут изложены в разд. 15. Ниже мы рассмотрим характеристики ДСМ-метода порождения гипотез.

5.2. Основы ДСМ-метода порождения гипотез

5.2.1. Общая характеристика метода

Назначение ДСМ-метода. На основе ДСМ-метода можно строить интеллектуальные системы, способные к поиску скрытых закономерностей в процессе обучения. В **ДСМ**-методе формализованы и программно реализованы методы сходства и различия **Д**жона **С**тюарта **М**илля, в честь которого и назван рассматриваемый метод автоматического порождения гипотез. В отличие от GUHA-метода ДСМ-метод не опирается на идеи, характерные для теории вероятностей и математической статистики. ***ДСМ-метод*** *предназначен для выявления причинно-следственных эмпирических зависимостей на множестве нечи-*

словых фактов в условиях неполноты информации. ДСМ-метод представляет собой систему автоматического порождения гипотез (АПГ) с целью поиска закономерностей. Эта система может быть использована для прогнозирования свойств объектов различной природы и анализа причин этих свойств. Задача прогнозирования в ДСМ-методе решается как задача распознавания того, обладает ли объект некоторым свойством, характерным для заданного класса.

Формирование гипотез в ДСМ-методе, как и в системах автоматического обучения, происходит на основе имеющихся в распоряжении системы наборов примеров, подтверждающих и отрицающих формируемую гипотезу (положительные и отрицательные примеры). Гипотезы о закономерностях выдвигаются в результате таких обобщений положительных примеров, которые не являются обобщениями отрицательных примеров.

Допущения для применимости. Допущениями для применимости ДСМ-метода являются следующие утверждения:

- исходные события представляют собой множества двух сортов;
- исследуемые явления представляют собой отношения между множествами;
- в данных существуют эмпирические корреляции;
- существуют как причины наличия корреляций (положительные причины), так и причины отсутствия корреляций (отрицательные причины).

Область применения. ДСМ-метод применим в тех предметных областях, в которых данные хорошо структурированы, но плохо формализованы. Под хорошей структурированностью понимается возможность выразить алгебраическими средствами операции пересечения («локального сходства»), объединения, отношения вложения и разности для объектов определенной структуры. Примерами таких данных могут служить:

1) объекты-множества, причины-множества;

2) объекты-кортежи, причины-кортежи;

3) объекты-слова, причины-слова;

4) объекты-графы, причины-графы.

5.2.2. Основные этапы ДСМ-метода

Автор ДСМ-метода В. К. Финн выделяет три основных этапа рассматриваемого метода [12]. На первом этапе, или **этапе выявления сходства**, задаются:

1) структура данных, характеризующих объекты предметной области;

2) операция сходства, сопоставляющая двум объектам из предметной области третий объект, выражающий сходство первых двух.

На втором этапе, **этапе правил**, объекты из предметной области делятся на положительные примеры (объекты, вызывающие некоторый интересующий нас эффект W) и отрицательные примеры (объекты, не вызывающие эффект W). Правила находят сходства положительных примеров и проверяют ряд условий, позволяющих называть найденные свойства гипотезами о структурных причинах эффекта W.

На третьем этапе, **этапе рассуждений**, составляются последовательности применения правил правдоподобного вывода и проверок некоторых условий на множестве всех исходных данных и полученных гипотез. Это необходимо для получения вывода об обоснованности ДСМ-вывода.

5.2.3. Определение гипотез

Для определения гипотез воспользуемся обозначениями, введенными в [2]. Введем три множества: **причины** $A = (a_1, a_2, \ldots, a_p)$, **следствия** $B = (b_1, b_2, \ldots, b_m)$ и множество **оценок** $Q = (q_1, q_2, \ldots, q_l)$. Выражение вида

$a_i \Rightarrow b_i ; q_k$

будем называть ***положительной гипотезой***. Такая гипотеза может быть выражена утверждением: «a_i является причиной b_j с оценкой достоверности q_k».

Выражение вида

$a_i^{\ 1} > b_j ; q_k$

будем называть ***отрицательной гипотезой***. Выразим ее утверждением: «a_i не является причиной b_j с оценкой достоверности q_k». Далее будем положительные гипотезы обозначать $n^+ i\ j\ k$, а отрицательные – как $n^- i\ j\ k$.

Среди значений q_k выделим два специальных, которые можно обозначить как 0 и 1. Значение 0, приписанное положительной или отрицательной гипотезе, означает, что соответствующее утверждение является ложным. Приписывание гипотезам значения оценки, равного 1, означает, что данная гипотеза яв-

ляется истинной. Все остальные оценки, отличные от 0 и 1, будут представляться рациональными числами вида s/n, где величина n характеризует «дробность» используемых оценок достоверности. Чем больше n, тем с большей точностью оценивается степень достоверности гипотез.

5.2.4. Выявление сходства

Исходные данные для ДСМ-метода. Структура данных, характеризующих объекты предметной области, должна содержать фрагменты, ответственные за наличие или отсутствие соответствующего свойства. Для работы ДСМ-метода требуются следующие виды наборов объектов:

1) набор объектов, про которые известно, что они обладают исследуемым свойством: (+)-объекты (***положительные примеры***);

2) набор объектов, про которые известно, что они не обладают исследуемым свойством: (–)-объекты (***отрицательные примеры***);

3) набор объектов, наличие свойства в которых требуется определить: τ-объекты (***недоопределенные примеры***).

Определение сходства. Формально сходство определяется как некоторое двухместное отношение или метрика, сопоставляющая паре аргументов значение их сходства (близости). Понятие сходства, используемое в ДСМ-методе, задается через операцию П, которая двум объектам сопоставляет третий, выражающий сходство первых двух. Эта операция обладает алгебраическими свойствами, позволяющими однозначно определять сходство объектов независимо от порядка их расположения. Такая операция позволяет также задавать сходство через отношение.

Простейшим способом определения операции сходства служит теоретико-множественная операция пересечения $\cap$. Другим способом задания сходства, использующимся в ДСМ-методе, является задание сходства на кортежах фиксированной длины с таблично-упорядоченными значениями компонент.

5.2.5. Правила порождения гипотез

Формирование списка гипотез на основании сходства [2]. Предположим, что исследуется некоторое свойство объектов W. Тогда исходные данные представляются множествами положительных, отрицательных и недоопределенных примеров. ***Положительные примеры*** суть объекты, о которых известно,

что они обладают свойством W, об ***отрицательных примерах*** известно, что они не обладают этим свойством, а о ***недоопределенных примерам*** не известно ни то, ни другое. На основании этой информации необходимо сформировать список гипотез о причинах, ответственных за наличие положительного свойства (положительные гипотезы), и причинах, ответственных за его отсутствие (отрицательные гипотезы).

Рассматривается группа положительных примеров, в которых ищем некоторую часть описания k_+ объектов, общую для определенной совокупности примеров e_+ из группы. Если такие фрагменты описаний удалось выделить, они могут считаться кандидатами в причины. Таких кандидатов может быть несколько. Поис происходит следующим образом.

Если описание k_+ группы e_+ не совпадает с содержанием какой-либо группы отрицательных примеров, а число таких примеров $|e_+| \geq 2$ (требование, чтобы у гипотезы было не менее двух подтверждающих примеров), то пара множеств (e_+, i_+) называется ***положительной гипотезой*** относительно свойства W. Отрицательные гипотезы определяются аналогично.

Образуем матрицу M^+, в которой строки соответствуют выделенным кандидатам a_i, а столбцы – следствиям b_j. На пересечении строк и столбцов будут записываться оценки достоверности q_k для гипотез $n^+ i\, j\, k$. Для множества отрицательных примеров аналогично строится матрица M^-, содержащая оценки достоверности отрицательных гипотез $n^- i\, j\, k$. Кандидаты в причины в матрицах M^+ и M^- могут частично совпадать.

Классификация недоопределенных примеров. Полученные гипотезы можно применять для классификации недоопределенных примеров из множества недоопределенных примеров G_τ (т. е. распознавания того, обладают они на самом деле свойством W или нет). Здесь возможны три ситуации:

1) вынесение решения в пользу положительной гипотезы;
2) вынесение решения в пользу отрицательной гипотезы;
3) отказ от принятия решения.

Решение принимается на основании следующего правила. Если для недоопределенного примера $g_\tau \in G_\tau$ и содержания i_+ некоторой положительной гипотезы (e_+, i_+) имеет место соотношение $i_+ \subset g_\tau$ и g_τ не обладает сразу всеми признаками из содержания ни одной отрицательной гипотезы, то делается

положительный прогноз (т. е. утверждается, что скорее всего пример g_τ обладает свойством W). При этом (e_+, i_+) называется гипотезой в пользу положительного прогноза для g_τ.

Если недоопределенный пример g_τ обладает всеми признаками из содержания некоторой отрицательной гипотезы и не включает содержания ни одной положительной гипотезы, то делается ***отрицательный прогноз***. При этом (e_-, i_-) называется гипотезой в пользу отрицательного прогноза для g_τ.

Если набор с i_- не включает содержания ни одной гипотезы или включает содержания гипотез разных знаков, то прогноз не осуществляется. Пи этом первая из указанных ситуаций называется ***недоопределенным прогнозом***, а вторая – ***противоречивым***.

Оценка обоснованности гипотез. Для оценки обоснованности гипотез при ДСМ-методе используется квантор I_m, m I $\left\{0, \frac{1}{n-1}, \frac{2}{n-1}, ..., 1\right\}$. Значение $m = \frac{1}{n-1}$ соответствует гипотезам, достоверность которых неизвестна, значение $m = 1$ – истинным гипотезам. Если применяемое правило вывода подтверждает гипотезу, то значение m возрастает; если не подтверждает, то значение m уменьшается.

На каждом шаге работы ДСМ-метода множества положительных и отрицательных примеров могут пополняться. Новые наблюдения будут либо подтверждать, либо опровергать сформированные гипотезы. В этом случае оценки достоверности построенных гипотез надо либо увеличивать, либо соответственно уменьшать. Таким образом, в процессе накопления новой информации оценки гипотез либо приближаются к 0 или 1, либо ведут себя неустойчиво. Гипотезы, достоверность которых стала меньше некоторого нижнего порога, могут исчезать из матриц M^+ и M^-. При достижении некоторого верхнего порога достоверности гипотеза может быть признана в системе как установленный факт.

Другие способы формирования гипотез. Новые гипотезы формируются не только на основании выделения в примерах определенного сходства. Они могут также использовать ***метод различия***. Различие в описаниях групп примеров может порождать гипотезы, включаемые в M^+ и M^-. Кроме того, в ДСМ-методе, помимо реализации принципов Милля, используются некоторые выводы по ***аналогии***. Для этого на множестве описаний объектов вводится тем

или иным способом понятие сходства. Если, например, речь идет о структурных формулах химических соединений, то мерой сходства для них может быть совпадение самих структур при различных химических элементах в их позициях, либо, наоборот, наличие в некоторых фиксированных позициях структуры одинаковых элементов. Существует в этом методе и правило отрицательной аналогии, а также градации тех и других правил по существенности сходства. Таким образом, ДСМ-метод представляет возможность проведения рассуждений весьма широкого спектра.

Для формализации схем Милля в ДСМ-методе используются ***предикаты сходства, различия и аналогии***. Эти предикаты применяются как к множеству положительных примеров, на основании которых формируются гипотезы, так и к множеству отрицательных примеров.

5.2.6. Рассуждения в ДСМ-методе

Вывод правдоподобных гипотез в ДСМ-методе [12] осуществляется в рамках квазиаксиоматической теории (КАТ). Квазиаксиоматическая теория есть тройка $T = < \Sigma, \Sigma', R >$, где Σ – множество аксиом, заведомо неполно описывающих предметную область; Σ' – множество элементарных высказываний об объектах из предметной области (фактов). Обычно они соответствуют утверждениям вида «Объект *C* обладает свойством *W*». Множество Σ' открыто и может пополняться путем проведения новых экспериментов, наблюдений и т. п.; R – множество правил вывода, которое содержит правила как достоверного, так и правдоподобного вывода: $R = R' \cup R^0$, где R' – множество правил правдоподобного вывода; R^0 – множество правил достоверного вывода.

Рассуждение в КАТ есть построение цепочки формул $y_1, \ldots, y_m$, заканчивающейся целью рассуждения $y_m = \varphi$, где каждая $ш_i$ есть либо аксиома из Σ, либо фактическое высказывание из Σ' (соответствующее положительному или отрицательному примеру), либо формула $ш_i$ получена из цепочки $y_1, \ldots, y_m$ применением правила *R*, как, например, гипотезы и прогнозы, описанные в подразд. 5.2.5. Определение рассуждения в КАТ отличается от определения логического вывода тем, что:

1) Σ' – открытое множество, элементы которого, относящиеся к цели φ, вставляются в цепочку, если имеет место отношение сходства цели;

2) среди правил R, применяемых при построении указанной цепочки формул, имеются правила из множества правил правдоподобного вывода R', т. е. правил порождения и отбора гипотез и прогнозов;

3) в процессе построения цепочки y_1, …, y_m могут использоваться метасредства, например, проверка на логическую непротиворечивость различных утверждений, невыводимость, выполнимость некоторых условий и т. д.

Множество аксиом Σ состоит из множества процедурных аксиом Σ_p и множества декларативных аксиом Σ_d. Аксиомы из Σ_p выражают собой применение правил порождения гипотез и прогнозов. Часть декларативных аксиом Σ_d^0 описывает структуру данных рассматриваемой предметной области. Другая часть декларативных аксиом Σ_d^1 описывает некоторые свойства причинных отношений и гипотез о них: правила комбинации следствий одних и тех же причин, а также принципы каузальной непротиворечивости и полноты.

Требование аксиомы каузальной непротиворечивости заключается в том, чтобы положительный пример не мог стать отрицательным прогнозом на каком-либо шаге порождения гипотез. Аксиома каузальной полноты требует того, чтобы все положительные примеры становились положительным прогнозом на каком-либо шаге порождения гипотез. При выполнении аксиом каузальной непротиворечивости и полноты ДСМ-вывод может быть назван произведенным на достаточном основании. Если эти аксиомы не выполняются, то необходимо пополнение системы новыми примерами, позволяющими делать вывод на достаточном основании.

ЧАСТЬ III. СТАТИСТИЧЕСКИЕ ОСНОВЫ ИНДУКТИВНОГО ВЫВОДА

6. ВЕРОЯТНОСТЬ И ИНДУКЦИЯ

6.1. Вероятностный характер индуктивных рассуждений

Ключевые понятия: оценка правдоподобности, вероятность

Новые знания, получаемые в процессе деятельности человека, в своем происхождении базируются на индуктивных предположениях, поскольку эти предположения и гипотезы конструируются на основе конечного числа реальных наблюдений. Эти малочисленные наблюдения объективно отражают реальный мир, и в них содержится вся информация о пока непознанных закономерностях. Цель индукции – приспособить наши представления к полученному опыту в такой степени, в какой это возможно.

Логика индукции обращается к понятию вероятности, так как результаты наблюдений не гарантируют истинности гипотез и эмпирических теорий. В этом и состоит ***особенность индукции***: исходя из истинных посылок и следуя схеме индуктивных рассуждений, в одних случаях мы получим истинный результат, а в других – ложный. Поэтому вместо оценки истинности или ложности заключения используют ***оценку правдоподобности*** или ***достоверности*** (см. подразд. 2.4). Одним из видов оценки достоверности является вероятность.

Решение об истинности или ложности заключения зависит от вероятностного характера посылок для получения заключения. Эти посылки связаны с определенными допущениями, касающимися реального мира. Считается, что фактические допущения, положенные в основу посылок индуктивных рассуждений, также имеют вероятностный характер.

Таким образом, учитывая, во-первых, характер посылок индукции и, во-вторых, саму схему индуктивных рассуждений, приводящую к правдоподобному заключению, можно утверждать, что ***индуктивные выводы достаточно вероятны***.

В некоторых работах по индукции ***вероятность*** интерпретируется ***как логическое отношение***, существующее между посылкой конкретного индуктивного вывода и его заключением: между посылками и заключением должно воз-

никать такое отношение, которое гарантирует истинность заключения при истинности посылок или по меньшей мере делает вероятным это заключение.

Если статистическая интерпретация вероятности основывается на анализе объективных свойств массовых случайных событий, таких как их относительная частота, то логическая интерпретация определяется как степень разумной веры в гипотезу при имеющихся данных.

Статистическое понятие вероятности характеризует численное значение степени возможности появления массового случайного события при длительных эмпирических испытаниях и тем самым является объективным по своему содержанию. Оно отображает то, что происходит в реальном мире, и не зависит от мнения субъекта. Именно поэтому статистическая вероятность получила такое широкое распространение в естествознании, технических и социально-гуманитарных науках.

В теории аргументации статистическая вероятность может быть использована в случаях, когда делаются предсказания на основе установления относительных частот появления событий, и в разнообразных статистических выводах. Одним из важных способов аргументации в статистике является умозаключение от выборки, или образца к генеральной совокупности, или популяции. Если выборка из совокупности сделана с соблюдением необходимых требований, т. е. является репрезентативной, то на основании тщательного исследования можно с той или иной степенью вероятности утверждать, что выдвинутая гипотеза о генеральной совокупности будет справедливой. Такой подход аналогичен обычному индуктивному рассуждению, в котором на основе анализа некоторых членов класса предметов делается вероятностное заключение, что свойство, которым обладают исследованные члены класса, будет присуще всем членам класса. Однако для характеристики индуктивных рассуждений обращаются к другой интерпретации вероятности, которую называют логической и индуктивной, а иногда, чтобы охватить все способы недемонстративных рассуждений, – просто правдоподобной, противопоставляя ее тем самым достоверно истинной аргументации.

Во всяком рассуждении существует определенная логическая связь между посылками и заключением. В дедуктивных умозаключениях она выступает в виде логического следования заключения из посылок. Другими словами, если эти посылки истинны, а отношение между ними и заключением удовлетворяет

правилам дедукции, то рассуждение и основанная на ней аргументация считаются полностью обоснованными.

Совершенно другой характер имеет отношение между посылками и заключением индукции. Если в посылках содержится информация о некоторых исследованных членах определенного класса явлений, то она переносится на другие неисследованные члены, их группу или весь класс в целом. Ясно, что эта информация может оказаться неверной относительно непроверенных членов класса и тем более всего класса. Таким образом, известная нам информация может служить только в качестве частичного обоснования индуктивного заключения.

Одинаково важными и самостоятельными следует считать как логическую, так и статистическую интерпретацию вероятности. В то время как статистическая интерпретация дает нам знание о реальных процессах, происходящих в мире, и рассматривает вероятность как относительную частоту при длительных наблюдениях массовых, повторяющихся, случайных событий, логическая вероятность характеризует мир нашего знания, в котором мы используем для подтверждения одних утверждений (заявлений, предположений, предсказаний, гипотез и решений) другие утверждения (эмпирические свидетельства, факты, показания и т. п. доводы).

Таким образом, логическая структура не только индуктивных рассуждений, но и умозаключений по аналогии, статистических выводов и других недедуктивных рассуждений в общей форме может быть охарактеризована с помощью понятия логической вероятности. В свою очередь логическая вероятность объясняется посредством понятия степени подтверждения предлагаемого утверждения или гипотезы всеми имеющимися в наличии релевантными высказываниями (посылками или аргументами). Подобно тому, как в дедуктивных рассуждениях мы непосредственно имеем дело не с реальными вещами и процессами, а высказываниями о них, так и в недедуктивных заключениях речь должна идти о двух видах высказываний. Высказывания первого – это гипотеза, мнение или решение, второго вида – это совокупность высказываний, которые в той или степени подтверждают первые. Поскольку всякое заключение недедуктивного рассуждения можно рассматривать как гипотезу, постольку ее можно представить в виде следующей формулы:

$P(H|E) = c$,

где P – вероятность; H – гипотеза; E – свидетельства, подтверждающие гипотезу; c – степень подтверждения, выраженная в виде числа.

Модель реальной ситуации состоит из утверждений о возможных исходах (результатах) и определений вероятностного механизма, обусловливающего тот или иной тип результата. Модель является идеализацией реальной ситуации. Ее адекватность зависит от того, насколько обоснованы и соответствуют модели предположения, на которых модель базируется.

Когда модель построена, теоретические положения могут быть применены, чтобы путем дедуктивных рассуждений определить характеристики данных, вытекающие из модели, а в конечном итоге – из предположений относительно реальной ситуации. Логическая дедукция через теорию приводит к описанию вероятностных свойств данных при условии, что модель адекватна реальной ситуации. Здесь вероятность является средством связи модели с данными. Эта связь дедуктивная.

Статистическая теория действует в противоположном направлении. Она берет реальные данные, порожденные практической ситуацией (или спланированным экспериментом) и использует их, чтобы обосновать определенную модель, сделать разумные предположения, оценить числовые значения существенных параметров или предложить модель. Этот обратный, индуктивный процесс возможен только потому, что язык вероятностно-статистической теории пригоден для формирования дедуктивной связи.

Таким образом, мы имеем схему рассуждений в виде индукции как обратной дедукции. Цель таких рассуждений – обеспечить возможность вывода относительно модели на основании информации, содержащейся в выборочных данных, или построить процедуру принятия решения, соответствующего реальной ситуации.

На рис. 6.1 представлены различные компоненты, участвующие в статистических рассуждениях (реальная ситуация, модель, информация), и связи между ними.

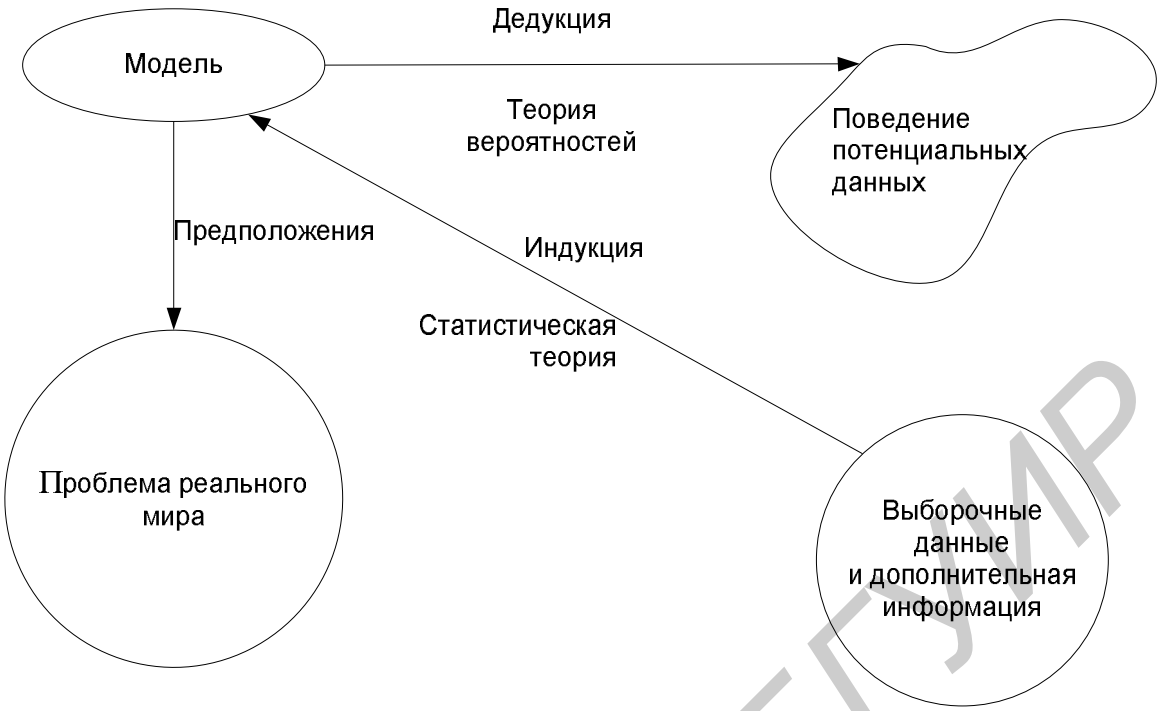


Рис. 6.1. Статистические рассуждения и их компоненты

Научный подход к индуктивным методам состоит в изучении и обосновании тех условий, которые обеспечивают гарантированную степень успеха индуктивного вывода. Применяемые теоремы и формулы теории вероятностей и математической статистики не служат обоснованием индуктивной процедуры, а используются для нахождения наилучшего решения при выборе гипотезы из множества альтернатив.

6.2. Статистическая индукция

<u>Ключевые понятия</u>: статистическая гипотеза, выборка, генеральная совокупность, проверка гипотезы, оценивание параметров, статистическая оценка, свойство оценки

Теория статистической индукции является единственной теорией индукции, которая применяется на практике. ***К статистической индукции относятся рассуждения, вывод (заключение) которых есть какая-либо статистическая гипотеза.***

6.2.1. Умозаключение в статистической индукции

Структура умозаключения в статистической индукции. ***Заключением*** **(*выводом*)** в статистической индукции является статистическая гипотеза, т. е. гипотеза, касающаяся распределения вероятностей некоторой случайной переменной (скалярной или векторной) в генеральной совокупности. ***Посылки рассуждения*** в этом случае содержат числовые данные, характеризующие некоторое наблюдаемое подмножество объектов.

Информация, используемая в статистическом выводе. Информация, необходимая для выполнения статистического вывода, представляется двумя совокупностями объектов:

1) реально наблюдаемая, статистически представленная рядом из *n* наблюдений в виде последовательности чисел $x_1, x_2, \ldots, x_n$ (т. е. ***выборка***);

2) теоретически домысливаемая (так называемая ***генеральная совокупность***).

Основные свойства и характеристики выборки, называемые эмпирическими (или выборочными), могут быть проанализированы и вычислены по имеющимся данным. Основные свойства и характеристики генеральной совокупности, называемые теоретическими, неизвестны исследователю. Получить знания об этих теоретических свойствах и характеристиках можно по соответствующим свойствам и характеристикам выборок в процессе индуктивных рассуждений. Это означает, что суждение о гипотезе приходится выносить по результатам выборочных реализаций случайных величин. Случайность наблюдаемых величин порождает некоторую неопределенность наших заключений о гипотезах.

6.2.2. Виды статистической индукции

Различают два основных вида статистической индукции: ***проверка гипотез*** и ***оценивание параметров***.

Проверка гипотез. С ***проверкой гипотез*** мы имеем дело тогда, когда среди гипотез, являющихся возможными ответами на данную проблему, заранее выделяется некоторая гипотеза. Эта гипотеза носит название нулевой (H_0). В процессе проверки (тестирования) гипотезы мы должны принять решение об отклонении или принятии этой гипотезы. Гипотеза отклоняется, когда по выборке получен результат, маловероятный при истинности выдвинутых гипотез. Числовое значение этой малой вероятности называется уровнем значимости.

Уровень значимости характеризует достоверность принятого решения. Причины для выдвижения гипотезы H_0 могут быть различными, но всякий раз требуется, чтобы для H_0 была вычислена вероятность ошибочного решения. Привычная классическая формулировка «*результаты наблюдений подтверждают выдвинутую гипотезу*» должна быть заменена на заключение «*результаты наблюдений при заданном уровне значимости не противоречат выдвинутой гипотезе*».

Оценивание параметров. В математической статистике индуктивным правилам философов и логиков соответствуют оценочные функции, или ***оценки*** [19, 24]. С ***оцениванием параметров*** мы имеем дело, когда нужно решить следующую задачу: на основании имеющихся статистических данных необходимо вычислить как можно более точные ***приближенные значения*** (***статистические оценки***) для одного или нескольких параметров, характеризующих функционирование некоторой реальной системы.

Статистическая оценка. Любая функция от результатов наблюдений исследуемой случайной величины называется ***статистикой***. Статистика $\hat{}_n$, используемая в качестве приближенного значения неизвестного параметра θ, называется ***статистической оценкой***. Все статистики и статистические оценки являются случайными величинами. При повторении выборки из той же самой генеральной совокупности и подстановке новых выборочных значений в ту же самую «функцию-оценку» мы получаем другое число в качестве оценки интересующего нас параметра, т. е. имеется неконтролируемый разброс в значениях оценки при повторениях эксперимента. Оценки должны обладать желательными для нас свойствами.

Свойства оценок. В качестве основной меры точности статистической оценки $\hat{}_n$ неизвестного параметра θ используется дисперсия $D(\hat{}_n)$ данной оценки, которая должна быть минимальной на множестве всех возможных оценок. Здесь дисперсия выступает как мера случайного разброса относительно истинного значения оцениваемого параметра. Это свойство оценки носит название ***эффективности***: $D(\hat{}_n) = \min$.

Свойство ***состоятельности*** оценки $\hat{}_n$ обеспечивает ее сходимость (по вероятности) к истинному значению оцениваемой величины θ по мере роста объема выборки (числа наблюдений): $\hat{}_n \to \theta$ при $n \to \infty$.

Свойство ***несмещенности*** оценки заключается в том, что результат усреднения всевозможных значений этой оценки, полученных по различным выборкам заданного объема, дает в точности истинное значение оцениваемого параметра: E (ˆ „) = θ (E – математическое ожидание).

Задачи оценивания и проверки гипотез связаны с уменьшением неопределенности наших знаний об истинных значениях некоторой переменной. На рис. 6.2 приведена схема вывода, применяемая при оценивании. Его суть составляет такое отображение выборочного пространства K на параметрическое пространство Ω, что если мы наблюдаем X и вычисляем оценку $\tilde{\theta}(X)$, то делаем заключение о том, что истинное значение параметра равно θ или оно принадлежит области ω(X) в Ω. Это два случая, которые соответствуют точечному и интервальному оцениванию.

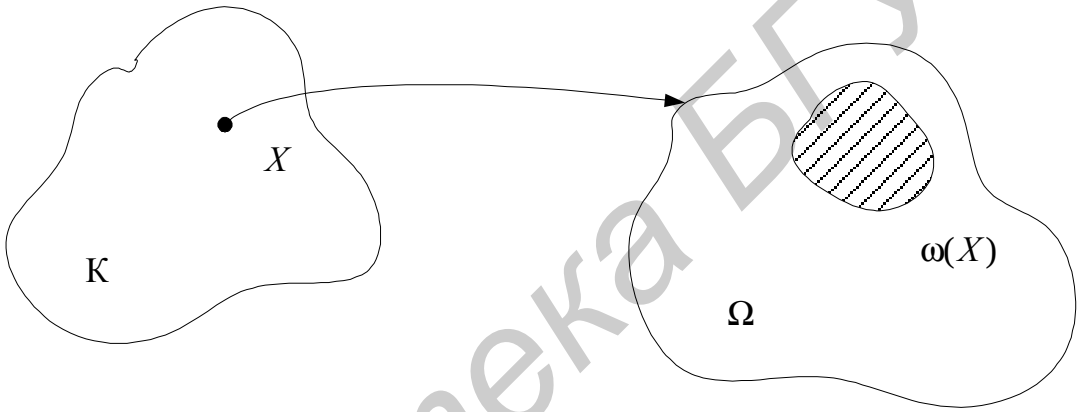


Рис. 6.2. Вывод в задаче оценивания

При проверке гипотез выборочное пространство разделяется на две области {S_0, S_1}. Область S_1 – это критическая область, где гипотеза H_0 отвергается (принимается гипотеза H_1), если статистика критерия $t(X)$ I S_1. Область S_0 – это не критическая область, где нет оснований отвергнуть H_0, если $t(X)$ I S_0.

Если гипотеза имеет форму

H_0: θ I ω I Ω,

то статистический вывод относительно θ можно представить рис. 6.3.

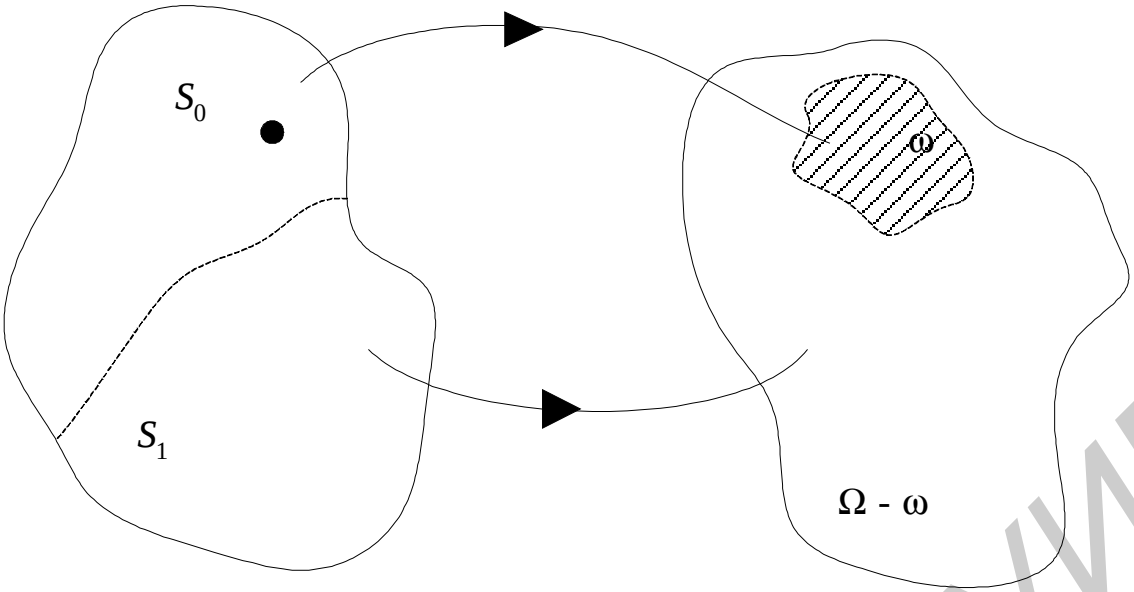


Рис. 6.3. Процесс вывода при статистической проверке гипотез

6.3. Наиболее распространенные типы вероятностно-статистических моделей, используемых в индуктивном выводе

Понятие математической модели. Теория вероятностей предоставляет исследователю набор математических моделей, воспроизводящих механизмы функционирования гипотетических реальных явлений или систем случайной природы. ***Математическая модель – это абстракция реального мира, в которой интересующие исследователя отношения между реальными элементами заменены подходящими отношениями между математическими объектами*** [19]. Математические модели, в описании которых используются случайные величины, называются ***вероятностными***. Задачей математической статистики является обоснованный выбор среди множества возможных моделей той, которая наилучшим образом соответствует имеющимся статистическим данным, характеризующим реальное поведение изучаемой системы.

Типы вероятностно-статистических моделей. К наиболее распространенным типам вероятностно-статистических моделей относятся модели законов распределения вероятностей; линейные модели; модели марковского типа; геометрические модели.

Модели законов распределения вероятностей случайной величины. Понятие закона распределения вероятностей случайной величины является базовым в теории вероятностей и математической статистике [3, 8]. Знать ***закон***

распределения вероятностей случайной величины – значит уметь поставить в соответствие любому ее возможному значению (или любой области ее возможных значений) величину вероятности появления именно этого значения (или вероятности попадания случайной величины в заданную область ее возможных значений). Наиболее распространенные в статистических исследованиях модели законов распределения описаны в [1, 3, 8, 19] и приведены в прил. 2.

Линейные вероятностные модели. ***Линейные вероятностные модели*** описывают характер и структуру взаимосвязей между случайными величинами (регрессионные модели, модели дисперсионного анализа, модели факторного анализа и временных рядов) [1, 10].

Геометрические модели [10]**.** ***Геометрические модели*** позволяют осуществлять удобную визуализацию исходных многомерных данных. В этих моделях каждый объект, характеризуемый вектором признаков $X = (X_1, X_2, \ldots, X_p)$, представляется в виде точки в p-мерном пространстве. С помощью геометрических моделей можно выделить поверхность существенно меньшей размерности p.

Модели марковского типа. ***Модели марковского типа*** описывают закономерности случайных переходов объектов из одного состояния в другое путем прямого задания вероятностей таких переходов. Подобные модели используются в социологии, медицине, демографии, экономике, теории надежности.

6.4. Правила действий со случайными событиями и вероятностями их осуществления

Ключевые понятия: элементарное событие, пространство элементарных событий, правила действий с событиями и вероятностями, условная вероятность, формула полной вероятности, априорная и апостериорная вероятности

6.4.1. Случайные события и правила действий с ними

Случайные события. Множество всех возможных исходов эксперимента (или ***элементарных событий***) называется ***пространством элементарных событий*** **Ω**. Рассмотрим случай, когда пространство Ω состоит из конечного или счетного числа элементарных событий $\omega_1, \omega_2, \ldots, \omega_n, \ldots$. Факт «пространство Ω состоит из элементарных событий $\omega_1, \omega_2, \ldots, \omega_n, \ldots$» обозначается

$$\Omega = \{\omega_1, \omega_2, \ldots, \omega_n, \ldots\} \text{ или } \Omega = \{\omega_i\},\ i = 1, 2, \ldots . \qquad (6.1)$$

Приведем пример случайного эксперимента и соответствующего ему пространства элементарных событий.

Пример 6.1. Выбрасывание одной игральной кости:

$\Omega = \{\omega_1 = 1, \omega_2 = 2, \omega_3 = 3, \omega_4 = 4, \omega_5 = 5, \omega_6 = 6\}$.

Кроме элементарных событий, имеются так называемые составные (или разложимые) события. Событие С называется составным, если можно указать по меньшей мере два таких элементарных события ω1 и ω2, что из осуществления каждого из них в отдельности следует факт осуществления события С.

Тогда случайным событием А назовем любое подмножество пространства элементарных событий (6.1), т. е. $A = \{\omega_{i1}, \omega_{i2}, \ldots, \omega_{in}\}$. Это определение следует интерпретировать так: осуществление любого из элементарных событий $\omega_{i1}, \omega_{i2}, \ldots, \omega_{in}$, входящих в А, влечет за собой осуществление события А.

Пример 6.2. В примере 6.1 события A_1 = {выпадет четное число очков} и A_2 = {выпавшее число очков не превзойдет 3} запишутся соответственно

$A_1 = \{\omega_2, \omega_4, \omega_6\}$ и $A_2 = \{\omega_1, \omega_2, \omega_3\}$.

Изложим правила действий с событиями.

Правила действий с событиями. ***Сумма*** **(*объединение событий*)** $A_1, A_2, \ldots, A_k$ – это такое событие $A = A_1 + A_2 + \ldots + A_k$, которое заключается в наступлении хотя бы одного из событий $A_1, A_2, \ldots, A_k$.

Произведение **(*пересечение*)** ***событий*** $A_1, A_2, \ldots, A_k$ – это такое событие $A = A_1 \cdot A_2 \cdot \ldots \cdot A_k$, которое заключается в обязательном наступлении всех событий $A_1, A_2, \ldots, A_k$.

Разность событий A_1 и A_2 – это такое событие $A = A_1 - A_2$, которое заключается в одновременном осуществлении двух фактов: событие A_1 произошло, а событие A_2 не произошло. В терминологии элементарных событий разность $A = A_1 - A_2$ определяется как событие, состоящее из всех тех элементарных событий, которые входят в A_1, но не входят в A_2.

6.4.2. Основные правила действий с вероятностями

Исчисление вероятностей базируется на системе аксиом, предложенной *А. Н. Колмогоровым*.

Аксиома 1. Каждому событию A из пространства событий Ω приписывается неотрицательное число $P(A)$, называемое вероятностью A.

Аксиома 2. Вероятность достоверного события A равна 1.

Аксиома 3 (аксиома сложения). Пусть A_1, A_2, ..., A_n – взаимно исключающие события. Тогда

$$P(A_1 + A_2 + ... + A_n) = P(A_1) + P(A_2) + ... + P(A_n).$$

Отметим, что когда Ω бесконечно, третья аксиома справедлива для бесконечного числа взаимно исключающих событий.

Теорема сложения вероятностей (вероятность суммы событий). ***Формула сложения вероятностей*** для двух событий A_1 и A_2 имеет вид

$$P(A_1 + A_2) = P(A_1) + P(A_2) - P(A_1 A_2). \quad (6.2)$$

Если A_1 и A_2 – несовместные (взаимно исключающие) события, то формула (6.2) дает

$$P(A_1 + A_2) = P(A_1) + P(A_2).$$

Формула (6.2) может быть обобщена на случай произвольного числа слагаемых.

Теорема умножения вероятностей (вероятность произведения событий).

Пусть B – непустое событие, считающееся уже состоявшимся («условие»), а A – событие, вероятность $P(A|B)$ которого требуется вычислить. ***Формула умножения вероятностей*** для двух событий A и B имеет вид

$$P(A \cdot B) = P(A|B) \cdot P(B).$$

Условная вероятность. ***Условная вероятность*** $P(A \mid B)$ события A при условии, что уже имеет место событие B, определяется с помощью формулы $P(AB) = P(A \mid B) \cdot P(B)$, откуда

$$P(A|B) = \frac{P(AB)}{P(B)}. \quad (6.3)$$

Независимость событий. Два события A и B называют независимыми, если $P(A \cdot B) = P(A) \cdot P(B)$.

Если $P(A \cdot B) > P(A) \cdot P(B)$, то зависимость событий считается положительной, если $P(A \cdot B) < P(A) \cdot P(B)$ – зависимость отрицательная. В первом случае также справедливы неравенства $P(A \mid B) > P(A)$ или $P(B \mid A) > P(B)$, а во втором, наоборот, $P(A|B) < P(A)$ или $P(B|A) < P(B)$.

Формула полной вероятности. В случае, когда условия B_1, B_2, …, B_k образуют полную группу событий, для вычисления вероятности $P(A)$ можно использовать соотношение

$$P(A) = P(A|B_1)\,P(B_1) + P(A|B_2)\,P(B_2) + \ldots + P(A|B_k)\,P(B_k),$$

которое называется ***формулой полной вероятности***.

Формула Байеса. Пусть интересующее нас событие A может произойти с одним из событий B_1, B_2, …, B_k, образующих полную группу. Известны вероятности $P(B_i)$, $P(A|B_i)$, $i = 1, \ldots, k$. Формулу

$$P(B_i \mid A) = \frac{P(B_i)P(B_i \mid A)}{\sum_{i=1}^{k} P(B_i)P(B_i \mid A)} \qquad (6.4)$$

называют ***формулой Байеса***.

Вероятности $P(B_i)$ называются ***априорными вероятностями***, а вероятности $P(A|B_i)$ – ***апостериорными вероятностями***. Теорема Байеса определяет вероятность события B_i, если событие A известно.

Приведем ряд **элементарных утверждений** исчисления вероятностей, которые будем использовать далее.

(1) $P(Æ) = 0$; вероятность невозможного события равна нулю. Например, $P(A \cdot \overline{A}) = 0$, так как невозможно, чтобы событие A произошло и не произошло одновременно.

(2) $P(A) + P(\overline{A}) = 1$.

(3) $P(A_1 + A_2) = P(A_1) + P(A_2) - P(A_1 A_2)$; вероятность альтернативы двух произвольных событий (или появления хотя бы одного из них) равно сумме вероятностей этих событий минус вероятность их одновременного появления.

(4) $P(A \cdot B) = P(A \mid B) \cdot P(B)$; вероятность конъюнкции двух событий равна вероятности одного из них, умноженной на условную вероятность другого.

(5) Если A_1 I A_2, то: а) $P(A_2) \geq P(A_1)$, а также $P(A_2 \mid A_1) = 1$; б) $P(A_2|B) \geq$ $\geq P(A_1|B)$, а также $P(A_2 \mid (A_1 \cdot B)) = 1$ для произвольного B. Отношение I для множеств означает включение, а для высказываний – следование. Если A_1 и A_2 являются множествами, то выражение A_1 I A_2 означает, что A_1 содержится в A_2 (каждый элемент множества A_1 является также элементом множества A_2). Если A_1 и A_2 являются высказываниями, то

A_1 I A_2 означает, что высказывание A_2 логически вытекает из A_1 или A_2 есть следствие (результат) A_1.

Сформулируем аналогичные утверждения для условных вероятностей:

(1а) $P(Æ|A) = 0$;

(2а) $P(A|B) + P(\overline{A}|B) = 1$;

(3а) $P(A_1 + A_2|B) = P(A_1|B) + P(A_2|B) – P(A_1 \cdot A_2|B)$;

(4а) $P(A_1 \cdot A_2|B) = P(A_1|B) + P(A_2|(B \cdot A_1))$.

Для обозначения операции логического сложения используются символы «$\vee$», «$\cup$» или «+», а для обозначения операции умножения – «$\wedge$», «$\cap$» или «$\cdot$».

6.4.3. Вероятность и индукция

Правильность индуктивного вывода находится в зависимости от величины вероятности заключения по отношению к посылкам. В разд. 6 приведена интерпретация вероятности как логического отношения, существующего между посылкой индукции и ее заключением. Рассмотрим некоторые ***свойства вероятности, которые обеспечивают правильность индуктивных выводов.*** Эти свойства выражаются через утверждения исчисления вероятностей (см. подразд. 6.4). Указанные утверждения можно относить к высказываниям (полагаем, что A и B обозначают высказывания).

Свойство 1. В силу утверждения 1 условная вероятность суждения имеет следующее свойство. ***Если суждение x логически следует из суждения y, то*** $\boldsymbol{P(x|y) = 1}$.

Если принять, что величина вероятности $P(x|y)$ должна представлять оценку заключения x на основе посылки y, то самую высокую оценку получат дедуктивные заключения, в которых вывод логически следует из посылок.

Свойство 2. Из утверждения 1 и определения (6.3) условная вероятность суждения имеет свойство 2: ***если x и y логически несовместимы, то*** $\boldsymbol{P(x|y) = 0}$. Конъюнкция таких высказываний является логически ложной. Поэтому наихудшую оценку должны получать такие суждения, заключения которых несовместимы с посылками.

Вероятностный подход к проверке гипотез. Одним из типов индуктивного рассуждения является такой вид рассуждения, заключением которого является некоторая гипотеза H, а посылками служат: 1) некоторое суждение, относящееся к знаниям W; 2) суждение E, представляющее собой следствие гипо-

тезы H и знания W. Заключение, или гипотеза H, не вытекает из посылок. Типичная процедура подобного рода есть процедура тестирования гипотез через наблюдение следствий, которые оказались истинными.

Допустим, что мы проверяем гипотезу H с помощью какого-либо эксперимента, в результате которого истинным оказывается суждение E, являющееся следствием гипотезы H и знания W. Рассмотрим, как сохраняются свойства вероятности гипотезы H с учетом свидетельства E и знания W. На основании ***формулы Байеса*** (6.4) выполняется равенство

$$P(H \mid E \wedge W) = \frac{P(E \mid H \wedge W) P(H \mid W)}{P(E \mid W)}.$$

Допустим, что E есть следствие H и W, поэтому на основании утверждения 5 имеем $P(E \mid H \wedge W) = 1$, а отсюда

$$P(H \mid E \wedge W) = \frac{P(H \mid W)}{P(E \mid W)}.$$

Если бы свидетельство E было следствием первоначального знания W, то $P(E \mid W) = 1$, и тогда $P(H \mid E \wedge W) = P(H \mid W)$. Это равенство выполнялось бы в случае, если $P(H \mid W) = 0$ (если бы гипотеза H была несовместима со знанием W). Однако в случаях, когда вероятность $P(H \mid W) > 0$ или $P(E \mid W) \leq 1$, происходит увеличение вероятности гипотезы H:

$$P(H \mid E \wedge W) > P(H \mid W).$$

Этот результат можно интерпретировать так: свидетельство E увеличивает в определенной степени знание W. Это увеличение будет тем больше, чем более неправдоподобно было свидетельство E с точки зрения знания W.

7. СТАТИСТИЧЕСКИЙ ВЫВОД, ОСНОВАННЫЙ НА ПРОВЕРКЕ ГИПОТЕЗ

7.1. Необходимость формулировки и проверки гипотез

Проблемы, возникающие при принятии решений в различных областях деятельности, часто сводятся к формулировке и оценке истинности одного или нескольких предположительных утверждений (гипотез) на основе результатов анализа накопленной информации об изучаемом явлении.

В зависимости от специфики и свойств предметной области, особенностей решаемых задач собранный материал может иметь своеобразную природу, структуру представления и хранения в памяти компьютера, а его обработка и анализ могут осуществляться специальными методами.

В слабоформализованных предметных областях для принятия решений, кроме экспериментальных данных, могут использоваться также накопленные профессиональные знания (знания эксперта). Для получения обоснованного заключения об истинности той или иной гипотезы или их совокупности применяются специальные методы, имеющие необязательно статистическую основу [13].

В формализованных предметных областях, где каждая проблема может быть описана аналитической моделью, а для ее решения используются строгие математические методы и алгоритмы, проверка гипотез опирается, как правило, на экспериментальные данные и осуществляется в основном статистическими методами с привлечением различных критериев, например критериев значимости. Сущность такого рода задач состоит в проверке истинности некоторого предположительного утверждения (нулевой гипотезы) H_0 о характере некоторого массового явления по отношению к конкурирующей гипотезе H_1.

В практической деятельности подобные задачи возникают при следующих обстоятельствах:

1) верификации допущений, предположений изучаемых закономерностей, относящихся к определенным процессам или генеральной совокупности;

2) на этапах принятия решений о выбираемых моделях данных и их зависимостях.

7.2. Статистическая гипотеза

Ключевые понятия: статистическая гипотеза, простая и сложная статистические гипотезы

Понятие статистической гипотезы. ***Гипотеза*** – это предположительное утверждение, которое может быть или не быть истинным.

Статистическая гипотеза – это гипотеза, которая допускает наблюдения статистической природы. Такие наблюдения могут возникать в различных областях деятельности человека. Вот некоторые примеры.

1. Эта игральная кость правильная.
2. Сведения поставщика деталей о своей продукции ложны.
3. Новый метод обучения лучше, чем старый.
4. Требования стандарта на чистоту воздуха в нашем городе не соблюдаются.
5. Данные о занятости населения предполагают наличие дискриминационной политики при найме на работу.

Так как эти гипотезы являются статистическими, то они могут быть сформулированы более точно.

1. Вероятности выпадения каждой грани такой игральной кости равны. Это означает, что мы имеем равномерное распределение случайной переменной, которая представляет собой число точек на лицевой поверхности каждой грани данной кости.
2. Средняя длина детали, поступившей от поставщика, больше, чем он заявлял (или меньше, или отличается от заявленной длины).
3. Средняя оценка стандартного тестирования у обучавшихся по новому методу выше, чем у обучавшихся по старому методу.
4. Значения параметров, характеризующих загрязнение воздуха в городе, больше, чем установлено стандартом.
5. Для некоторых работодателей переменная «принятие на работу» не будет независимой от переменной «пол» (или «этническая принадлежность», «вероисповедание» и т. д.).

Из приведенных примеров следует, что ***статистическая гипотеза – это утверждение относительно характера или неизвестных параметров распределения случайных величин.*** Гипотеза называется ***простой***, если она полностью определяет распределение случайной величины. Например, значение некоторого параметра Θ ***в точности*** равно заданной величине Θ_0. В других случаях гипотеза называется ***сложной***.

Гипотеза 1 является гипотезой относительно абстрактной модели вероятностного распределения случайной переменной, которая описывает игральную кость. Данная гипотеза может быть проверена с помощью наблюдений, например ста бросаний кости. Следующие три гипотезы касаются параметра. Они проверяются с помощью выборочных данных. Эти данные рассматриваются как выборочное распределение, необходимое для оценки параметра. Последняя гипотеза утверждает независимость двух качественных (нечисловых) перемен-

ных. Она проверяется путем сравнения наблюдаемых данных с истинными данными, ожидаемыми в случае независимости переменных. Отметим, что только три из пяти гипотез включают параметр, две другие представляют собой статистические утверждения другого рода.

Для каждого из этих примеров на практике невозможно непосредственно определить истинность гипотезы. Например, вероятностное распределение для кости является моделью всех возможных бросаний, которые мы не можем наблюдать. Практически невозможно измерить длину каждой из сотен, а может быть и тысяч поступающих деталей. Для гипотезы 3 невозможно протестировать сегодня всех обучающихся по новому методу в ближайшие 15 лет. Конечно, можно протестировать студентов через год после окончания обучения. Но оценку эффективности нового метода нужно делать до его реализации, а не после. Для гипотезы 4 можно ли проверить каждый кубический метр воздуха в городе? Наконец, что означает в условиях гипотезы 5 непосредственная верификация для каждого человека?

7.3. Процедура проверки гипотезы. Область принятия и отклонения гипотезы

Ключевые понятия: область принятия и отклонения гипотезы, статистический критерий, уровень значимости, мощность критерия, виды альтернативных гипотез

Из-за невозможности определить истинность гипотезы прямым путем мы «проверяем» гипотезу, т. е. устанавливаем, не противоречит ли выдвинутая гипотеза имеющимся выборочным данным. Целью проверки гипотезы является оценка правомочности статистической гипотезы. ***Процедура обоснованного сопоставления высказанного предположительного утверждения (гипотезы) с результатами наблюдения*** носит название ***статистической гипотезы*** [19].

Результат сопоставления высказанной гипотезы с выборочными данными может быть либо ***отрицательным*** (данные наблюдения противоречат высказанной гипотезе, а поэтому гипотезу надо ***отклонить***), либо ***неотрицательным*** (данные наблюдения не противоречат высказанной гипотезе, и поэтому ее можно ***принять*** в качестве одного из возможных решений). При этом неотрицательный результат статистической проверки гипотезы не означает, что вы-

сказанное предположительное утверждение является наилучшим: просто гипотеза не противоречит имеющимся выборочным данным, однако таким же свойством наряду с H могут обладать и другие гипотезы.

Для того чтобы применить вероятностно-статистические принципы к задаче проверки гипотез, необходимо, чтобы гипотеза была сформулирована в виде утверждений, имеющих отношение к характеристикам вероятностного пространства, а именно, относительно природы рассматриваемого случайного процесса или величины неизвестных параметров.

Пусть проверяется гипотеза H_0, при справедливости которой случайная величина (или наблюдение) X имеет плотность $f(x; q_0)$. В качестве альтернативы предлагается гипотеза H_1, при справедливости которой это же наблюдение имеет плотность $f(x; q_1)$, $q_0 \neq q_1$. На практике проверка гипотезы начинается с получения подлежащей анализу случайной выборки из процесса или генеральной совокупности. По наблюденным значениям случайной величины X (исходам) необходимо сделать вывод об одном из вероятностных распределений, которое могло бы охарактеризовать поведение выборки: подчиняется ли она закону $f(x; q_0)$ или $f(x; q_1)$.

Область принятия и отклонения гипотезы. Решение об отклонении или принятии гипотезы H_0 выносится с помощью правила или процедуры, которая делит диапазон возможных значений исходов в выборке на два множества. Первое из них – это ***множество принятия гипотезы*** H_0 (**область принятия**). Второе множество называется **множеством отклонения нулевой гипотезы (областью отклонения)** или ***критическим множеством*** (*множеством принятия альтернативой гипотезы* H_1). Множество отклонения является дополнительным к множеству принятия нулевой гипотезы. Обозначим C_a и C_r соответственно как множества принятия и отклонения гипотезы (рис. 7.1). Тогда $C_a = \overline{C_r}$, $C_a \cap C_r = Æ$.

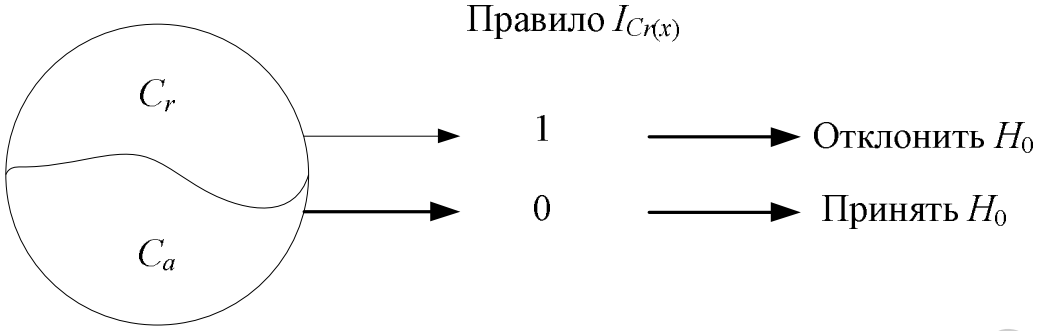


Рис. 7.1. Множество значений статистики критерия
для построения области принятия и отклонения гипотезы

Критическое множество для простых гипотез H_0, H_1 строится на основании леммы Неймана – Пирсона, утверждающей, что критическое множество должно включать те значения x случайной величины X, для которых

$$\frac{L_{H_1}(X_1, \ldots, X_n; q)}{L_{H_0}(X_1, \ldots, X_n; q)} = \frac{L(X_1, \ldots, X_n; q_1)}{L(X_1, \ldots, X_n; q_0)} > K, \qquad (7.1)$$

где L_{H_1} и L_{H_0} – значения функций правдоподобия наблюдений X_1, ..., X_n, вычисленные в предположении справедливости соответственно гипотез $H_1: q = q_1$ и $H_0:$ $= q_0$; $L(X_1, \ldots, X_n) = p(X_1; q) \cdot p(X_2; q) \cdot \ldots \cdot p(X_n; q)$.

Критическое значение $K > 0$ выбирается так, чтобы уровень значимости критерия равнялся ранее заданному числу $\alpha (0 < \alpha < 0{,}5)$. Другими словами, отношение правдоподобия (так называется в математической статистике отношение конкурирующих плотностей) определяет такое множество R на числовой прямой, что для всех x выполняется неравенство (7.1), где постоянная K выбирается согласно условию

$$\mathrm{P}\{X \in R \mid H_0\} = \mathrm{P}\{R \mid H_0\} = \int_R p(x; q_0)\, dx = \quad . \qquad (7.2)$$

Граница критической области W определяется путем решения уравнения (7.2) относительно x_k при фиксированном уровне значимости α.

Лемма Неймана – Пирсона утверждает, что критерий, построенный по этому правилу, будет иметь наибольшую мощность (наименьшую вероятность ошибки второго рода) среди всех критериев заданного уровня α.

Стандартная процедура проверки нулевой гипотезы состоит в том, что мы наблюдаем реализацию x случайной величины X и смотрим, какому множеству принадлежит значение x: множеству принятия нулевой гипотезы или критиче-

скому множеству. В первом случае принимается H_0, во втором случае – H_1. Эта процедура выполняется с помощью ***статистического критерия***, который позволяет по наблюденному значению x случайной величины X сделать выбор между нулевой и альтернативной гипотезами (рис. 7.2).

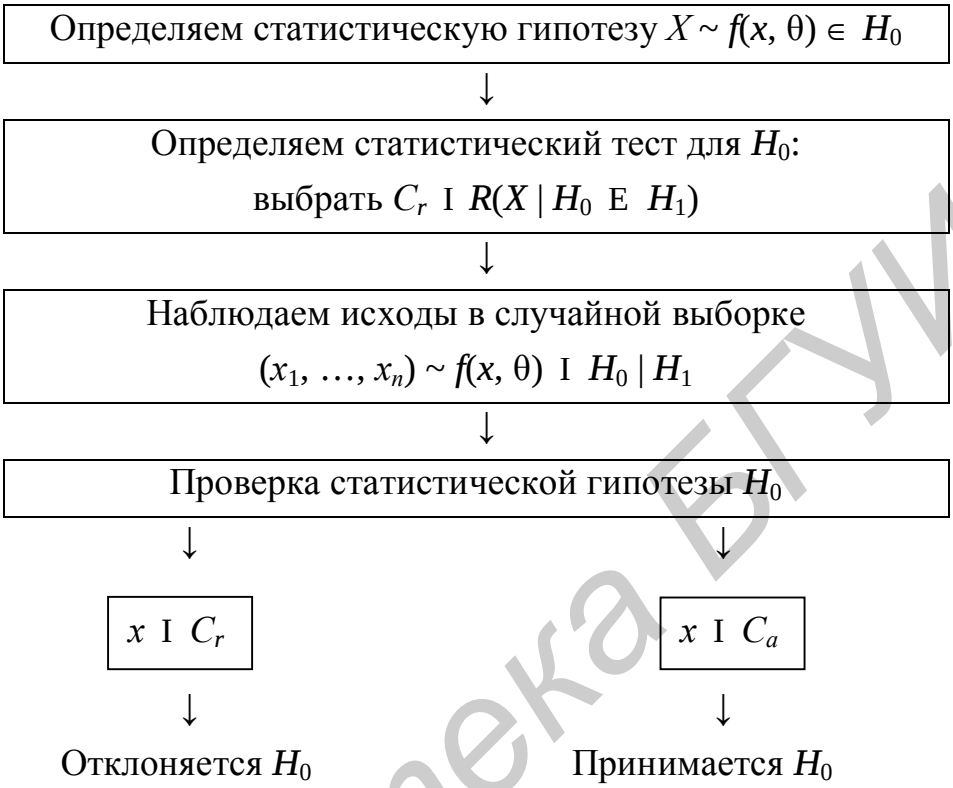


Рис. 7.2. Процедура проверки статистической гипотезы

Эти два множества (множество принятия нулевой гипотезы и критическое множество) выбираются на основании следующих принципов.

1. Зададим малую вероятность α совершить ***ошибку первого рода***, т. е. отвергнуть H_0, когда она верна. Численное значение α называется ***уровнем значимости*** критерия. Обычно значение α выбирают из интервала $0 < \alpha \leq 0{,}1$. Особенно распространенной является величина уровня значимости 0,05. Она означает, что в среднем в пяти случаях из ста мы будем ошибочно отклонять высказанную гипотезу при использования данного статистического критерия.

2. Зададим также малую вероятность β совершить ***ошибку второго рода,*** т. е. принять H_0, когда она неверна. Вероятность дополнительного события, т. е. правильного отклонения нулевой гипотезы, называется ***мощностью критерия.***

3. Критическое множество строится на основании принципа отношения правдоподобия [19, 24].

Виды альтернативных гипотез. С понятием области принятия или отклонения гипотезы тесно связаны виды альтернативных гипотез (альтернатив). В статистической проверке гипотез рассматриваются три вида альтернатив: ***двусторонняя***, ***правосторонняя***, ***левосторонняя альтернативные гипотезы***.

Утверждение, представляющее собой двустороннюю альтернативу, записывается математически с использованием отношения ≠ («не равно»). В этом случае вся область возможных значений статистики критерия разделяется на три части: 1 – область неправдоподобно малых значений, соответствующих H_1; 2 – область значений, соответствующих справедливости H_0; 3 – неправдоподобно больших значений, соответствующих H_1. В этом случае критическая область находится по обе стороны от области принятия (рис. 7.3).

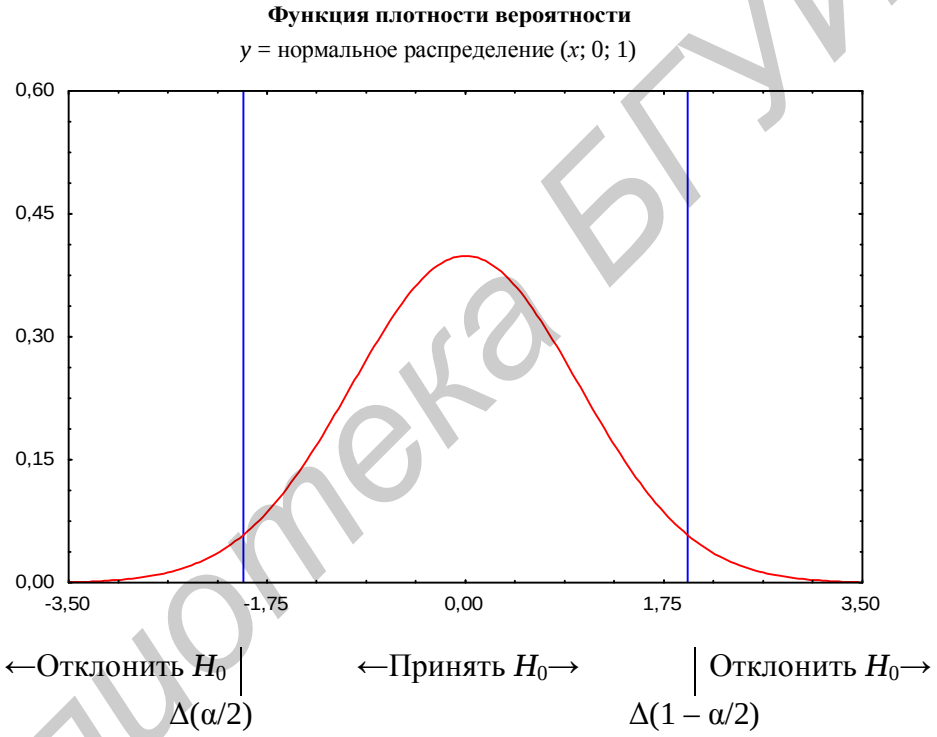


Рис. 7.3. Область принятия нулевой гипотезы H_0 при двусторонней альтернативе (α – уровень значимости)

Утверждение, представляющее собой правостороннюю альтернативу, записывается математически с использованием отношения > («больше»). В этом случае вся область возможных значений статистики критерия разделяется на две части: 1 – область значений, соответствующих справедливости H_0; 2 – область значений, соответствующих H_1. В этом случае критическая область находится справа от области принятия (рис. 7.4).

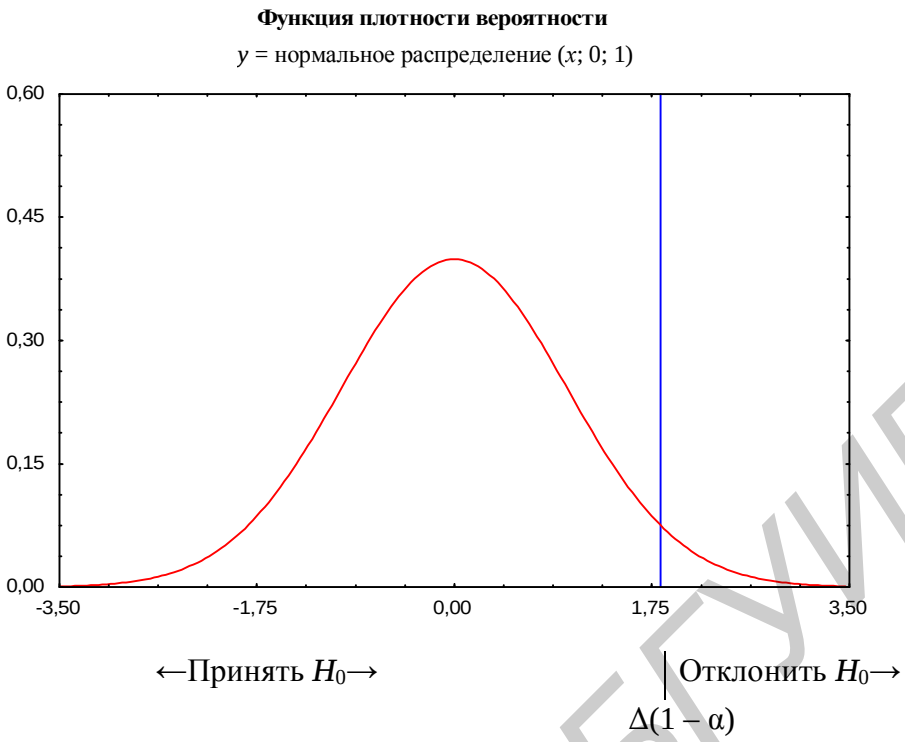


Рис. 7.4. Область принятия нулевой гипотезы H_0 при правосторонней альтернативе (α – уровень значимости)

Рассмотрим левостороннюю альтернативу (рис. 7.5).

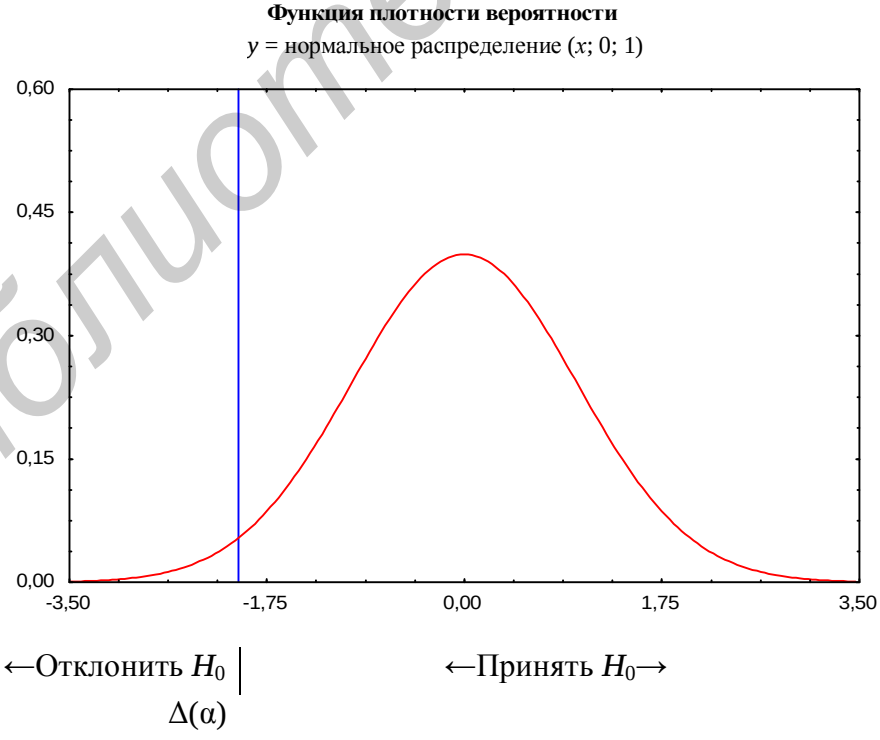


Рис.7.5. Область принятия нулевой гипотезы H_0 при левосторонней альтернативе (α – уровень значимости)

Утверждение, представляющее собой левостороннюю альтернативу, записывается математически с использованием отношения < («меньше»). В этом случае вся область возможных значений статистики критерия разделяется на две части: 1 – область значений, соответствующих H_1; 2 – область значений, соответствующих справедливости H_0. В этом случае критическая область находится слева от области принятия, как показано на рис. 7.5.

7.4. Ошибки первого и второго рода

Ключевые понятия: ошибка I рода, ошибка II рода

С точки зрения статистической проверки гипотез существуют два вида ошибок, называемых ***ошибкой I рода*** и ***ошибкой II рода***.

Ошибка I рода – это неправильное действие в соответствии с H_1: действовать в соответствии с H_1, если справедлива H_0 (принять H_1, если верна H_0). Ошибка II рода – это неправильное действие в соответствии с H_0: действовать в соответствии с H_0, если справедлива H_1 (принять H_0, если верна H_1). Вероятность ошибки P интерпретируется как условная вероятность. Условные вероятности этих двух типов ошибок обозначаются соответственно α и β:

$\alpha =$ **P** (ошибка I рода) = **P** (действие в соответствии с H_1| H_0 истинна);

$\beta =$ **P** (ошибка II рода) = **P** (действие в соответствии с H_0| H_1 истинна).

В табл. 7.1 показаны возможности принятия решения и ошибки двух типов по отношению к гипотезе H_0. Отметим, что если гипотеза H_0 справедлива и она принимается, то в таблице указано, что решение принято правильно. Если справедлива гипотеза H_1, а принимается H_0, то при решении допущена ошибка II рода. Если справедлива гипотеза H_0, а принимается гипотеза H_1, то при решении допущена ошибка I рода.

Таблица 7.1

Решения и ошибки при статистической проверке гипотез

Наше решение, основанное на данных	Состояние реального мира (неизвестное нам)	
	H_1 ложна (H_0 истинна)	H_1 истинна
Действие в соответствии с H_0	Правильное решение	Ошибка II рода
Действие в соответствии с H_1	Ошибка I рода	Правильное решение

Пусть $f(x; \theta_0)$, $f(x; \theta_1)$ – плотности распределения статистики критерия соответственно при справедливости нулевой гипотезы H_0 и альтернативной гипотезы H_1; θ_0, θ_1 – параметры распределения при H_0 и H_1. Тогда ошибки I и II рода определяются выражениями

$$\alpha = \int_{x_k}^{\infty} f(x;\ q_0)dx$$

и

$$\beta = \int_{-\infty}^{x_k} f(x;\ q_1)dx,$$

где x_k – граница критической области $\boldsymbol{W}$.

7.5. Мощность критерия

Ключевые понятия: мощность критерия

Мощность критерия π – это вероятность различения альтернативной гипотезы H_1, если она верна (или вероятность правильного отклонения нулевой гипотезы H_0):

$$\pi = 1 - \beta = \mathbf{P}\{\text{действия в соответствии с } H_1 \mid H_1 \text{ верна}\}. \quad (7.3)$$

$$\pi = \int_{x_k}^{\infty} f(x;\ q_1)dx, \quad (7.4)$$

где x_k – граница критической области $\boldsymbol{W}$.

Это понятие имеет важное теоретическое и практическое значение. С теоретической точки зрения мощность критерия является одной из главных характеристик, на основании которой выбирается критерий проверки гипотезы среди возможных. В математической статистике существует развитая теория оптимальных критериев, с помощью которой идентифицируются свойства различных критериев.

Практическое использование мощности критерия рассмотрим на примерах.

Пример 7.1. По техническим условиям в изготовленной продукции доля дефектных (бракованных) изделий p не должна превышать 0,01. На предприятии организован контроль качества продукции, в ходе которого проверяется гипотеза H_0: $p = 0{,}01$, H_1: $p > 0{,}01$.

В этом случае мощность критерия представляет собой вероятность того, что мы правильно отследим плохую ситуацию:

P{действовать в соответствии с $H_1 \mid p > 0{,}01$}.

Это вероятность того, что мы остановим производство и примем меры для устранения причин брака, если в контролируемой партии продукции число дефектных изделий слишком велико. В данном примере область отклонения гипотезы при уровне значимости α определяется множеством значений $\{\hat{p} \mid \hat{p} > 0{,}0162\}$, а мощность критерия $\pi = P\{\hat{p} > 0{,}0162 \mid p > 0{,}01\}$.

Пример 7.2. Рассмотрим гипотезы о значении математического ожидания μ нормального распределения: 1) H_0: $\mu = \mu_0 = 22$; 2) H_0: $\mu = \mu_0 = 24$ и альтернативную гипотезу H_1: $\mu > \mu_0$. Примем, что величина дисперсии σ^2 известна и равна 16, объем выборки $n = 9$. Граница критической области x_k равна 22,2. Определим мощность критерия относительно каждой из этих гипотез согласно (7.3):

$\pi_1 = P\{\overline{X} > 22{,}2 \mid \mu = 22\}$, $\pi_2 = P\{\overline{X} > 22{,}2 \mid \mu = 24\}$.

Эти вероятности можно записать в виде интегралов (7.4):

$$\pi_1 = \int_{22{,}2}^{\infty} f_{\overline{X}}(\bar{x}; 22; \frac{4}{3})\, d\bar{x},$$

$$\pi_2 = \int_{22{,}2}^{\infty} f_{\overline{X}}(\bar{x}; 24; \frac{4}{3})\, d\bar{x}.$$

Решив уравнения, получим мощности, соответствующие каждой из гипотез:

$\pi_1 = 0{,}44038$, $\pi_2 = 0{,}91149$.

Рис. 7.6 помогает понять принцип вычисления мощности критерия с уровнем значимости α. Критическое множество, соответствующее уровню значимости α, расположено под правым «хвостом» распределения $f(x, \mu_0)$ (нулевая альтернатива). Мощность критерия есть площадь под альтернативной плотностью распределения $f(x, \mu_1)$, лежащая справа от критического значения $x_к$ и равная $1 - \beta$.

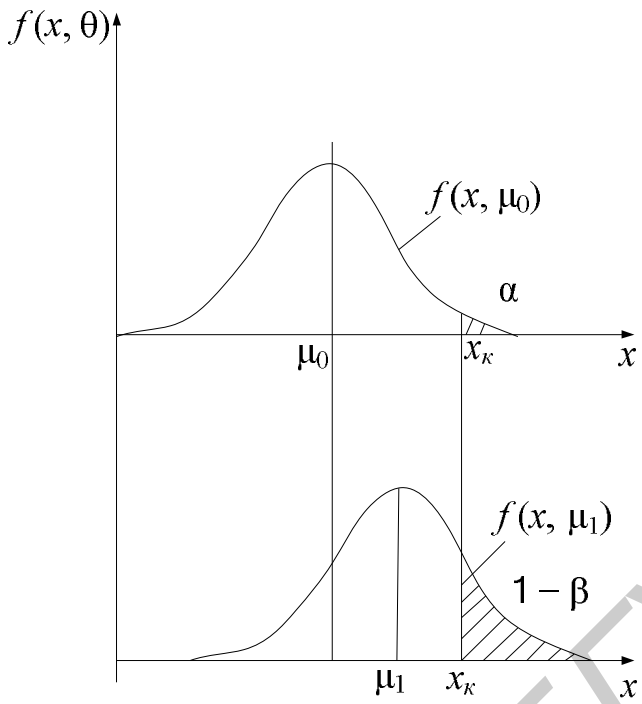


Рис. 7.6. Критическое множество и мощность критерия

7.6. Решающее правило и статистика критерия

Ключевые понятия: решающее правило принятия гипотезы, решающее правило отклонения гипотезы

Статистика критерия $g_k(x_1, x_2, ..., x_n)$ есть некоторая функция от результатов наблюдений. Эта критическая статистика g_k сама является случайной величиной и в предположении справедливости нулевой гипотезы H_0 подчинена некоторому хорошо изученному закону распределения.

Существуют два возможных вывода при проверке гипотезы: либо мы отклоняем нулевую гипотезу («отклонить H_0»), либо мы отказываемся отклонить нулевую гипотезу («отказ отклонить H_0»).

В общем виде решающее правило формулируется следующим образом:

Если значение статистики критерия, вычисленное по выборке, попадает в область отклонения гипотезы, то следует действовать в соответствии с H_1 (принять H_1).

Запишем **решающее правило принятия или отклонения** нулевой гипотезы, основанное на критическом значении статистики критерия γ_k:

$$\text{принимается гипотеза } \begin{cases} H_0, & \text{если} \quad g_k < \Delta(e), \\ H_1, & \text{если} \quad g_k \geq \Delta(e), \end{cases}$$

где Δ(ε) – порог теста размером ε (критическое значение статистики критерия).

Порог теста Δ(ε) определяется:

1) для двусторонней альтернативы как квантили уровня $\varepsilon_1 = 1 - \alpha/2$ и $\varepsilon_2 = \alpha/2$ распределения статистики критерия (α – уровень значимости);

2) для левосторонней альтернативы как квантиль уровня $\varepsilon = \alpha$ распределения статистики критерия;

3) для правосторонней альтернативы как квантиль уровня $\varepsilon = 1 - \alpha$ распределения статистики критерия.

8. ОБЩАЯ СХЕМА СТАТИСТИЧЕСКОЙ ПРОВЕРКИ ГИПОТЕЗ

Ключевые понятия: схема статистической проверки гипотезы, *p*-значение

8.1. Общая схема статистической проверки гипотез

Задачи, сводящиеся к оценке истинности нулевой гипотезы H_0 по отношению к альтернативной гипотезе H_1, могут быть решены с помощью различного рода статистических критериев.

Несмотря на разнообразие самих гипотез и применяемых статистических критериев, их можно объединить в следующую общую логическую схему [19].

1. Выдвижение гипотез H_0, H_1.

2. Выбор ***уровня значимости*** α – вероятности ошибочного отклонения нулевой гипотезы. Эту величину называют также ***размером критерия*** (теста). Выбор величины уровня значимости α зависит от размера потерь, которые мы понесем в случае ошибочного решения. В большинстве практических задач пользуются стандартными значениями уровня значимости:

α = 0,1; 0,05; 0,025; 0,01; 0,005; 0,001.

3. Выбор ***критической статистики*** (критерия) $g_k(x_1,\ x_2,\ ...,\ x_n)$ – некоторой функции от результатов наблюдений. Эта критическая статистика g_k сама является случайной величиной и в предположении справедливости нулевой гипотезы H_0 подчинена некоторому хорошо изученному закону распределения с плотностью $f_g(u)$. Критическая статистика строится на основании ***принципа отношения правдоподобия*** [19].

4. Определение ***критической области W*** (множества значений критической статистики, при которых гипотеза отклоняется) исходя из следующего условия: $P(g_k(x_1,\ x_2,\ ...,\ x_n) \in W \mid H_0) = \alpha$. Из таблиц распределения $f_g(u)$ находятся квантили уровня $\alpha/2$ и уровня $1 - \alpha/2$, соответственно равные $\Delta(\alpha/2)$ и $\Delta(1 - \alpha/2)$. Они разделяют всю область возможных значений случайной величины $г_k$ на три части: 1 – ***область неправдоподобно малых*** $(-\infty,\ \psi_{\alpha/2}]$; 2 – ***правдоподобных*** $(\psi_{\alpha/2},\ \psi_{1-\alpha/2})$; 3 – ***неправдоподобно больших*** $[\psi_{1-\alpha/2}, \infty)$ значений ***в условиях справедливости нулевой гипотезы H_0*** (двусторонняя альтернативная гипотеза H_1).

В тех случаях, когда опасными для нашего утверждения являются только односторонние отклонения (т. е. только «слишком маленькие» или только «слишком большие» значения критической статистики g_k), находят лишь одну квантиль. Для первого случая определяется квантиль уровня $\Delta(\alpha)$, которая будет разделять весь диапазон значений $г_k$ на две части: область неправдоподобно малых и область правдоподобных значений (левосторонняя альтернатива). Во втором случае вычисляется квантиль уровня $\Delta(1 - \alpha)$; она будет разделять весь диапазон значений $г_k$ на область правдоподобных и область неправдоподобно больших значений (правосторонняя альтернатива).

5. Определение на основе выборочных данных $x_1,\ x_2,\ ...,\ x_n$ численной величины статистики g_k.

6. Выработка решения. Если $g_k \in W$, то гипотезу H_0 рекомендуется отклонить, в противном случае ее можно принять, так как имеющиеся данные не противоречат высказанной гипотезе. Однако для большей уверенности, прежде чем принять гипотезу, ее желательно подвергнуть проверке с помощью других критериев или повторить эксперимент, увеличив объем выборки.

Решение, принимаемое на основе статистического критерия, может оказаться ошибочным в двух случаях: когда ошибочно отклоняется гипотеза H_0 (с вероятностью α) и когда ошибочно принимается гипотеза H_0 (с вероятностью β), где вероятности ошибочных решений α и β – ошибки первого и второго рода, а величина 1 – β – мощность соответствующего критерия [8, 19].

8.2. Понятие *P*-значения

Стандартная процедура проверки нулевой гипотезы состоит в том, что мы наблюдаем реализацию случайной величины X и смотрим, попадает ли значение x в область принятия (отклонения) нулевой гипотезы. Однако часто более удобно выполнять другую процедуру, являющуюся в некотором смысле обратной по отношению к описанной. Вместо того чтобы работать с фиксированным уровнем значимости α и только соглашаться с принятием или отклонением H_0, можно найти уровень значимости α, соответствующий реализации случайной величины X. Число α, полученное таким образом, называется ***P-значением***.

Сформулируем решающее правило принятия или отклонения нулевой гипотезы, основанное на P-значении.

Пусть X – результат наблюдения, $Y = Y(X)$ – статистика критерия (скалярная случайная величина), с помощью которой проверяется нулевая гипотеза H_0, $\Delta(\alpha)$ – критическое значение (порог теста) уровня α, т. е. $\mathbf{P}\{Y > \Delta(\alpha) \mid H_0\} = \alpha$, $F(Y)$ – функция распределения статистики Y при нулевой гипотезе H_0. Возможны следующие три ситуации.

1. Проверяется нулевая гипотеза H_0: $\xi = \alpha$ против правосторонней альтернативы H_1: $\xi > \alpha$. В этом случае критическая область $\boldsymbol{W}$ определяется как область таких возможных значений статистики критерия $Y(X)$, для которых значение функции распределения $F(Y(X))$ превышает 1– α:

$W = \{Y(X) : F(Y(X)) > F(\Delta(\alpha)) = 1 - \alpha\}$.

Тогда P-значение есть величина $P = 1 - F(Y(X))$, которая получается в результате подстановки значения статистики критерия $Y(X)$ в ее функцию распределения. Решающее правило, основанное на P-значении, имеет вид

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } P \geq a, \\ H_1, \text{ если } P < a. \end{cases}$$

2. Проверяется нулевая гипотеза H_0: $\xi = \alpha$ против левосторонней альтернативы H_1: $\xi < \alpha$. В этом случае критическая область $\boldsymbol{W}$ определится как область таких возможных значений статистики критерия $Y(X)$, для которых значение функции распределения $F(Y(X))$ не превышает α:

$W = \{Y(X) : F(Y(X)) < F(\Delta(\alpha) = \alpha\}$.

В этом случае P-значение есть величина $P' = F(Y(X))$. Решающее правило, основанное на P-значении, имеет вид

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } P' \geq a, \\ H_1, \text{ если } P' < a. \end{cases}$$

3. Проверяется нулевая гипотеза H_0: $\xi = \alpha$ против двусторонней альтернативы H_1: $\xi \neq \alpha$. В этом случае критическая область задается как

$\boldsymbol{W} = \{Y(X) : F(Y(X)) > F(\Delta(\alpha)) = 1 - \alpha/2$ и $F(Y(X)) < F(\Delta(\alpha)) = \alpha/2\}$.

В рассматриваемом случае P-значение есть величина P'', определяемая из уравнения $1 - P''/2 = F(Y(X))$. Решающее правило имеет вид

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } P'' \geq a, \\ H_1, \text{ если } P'' < a. \end{cases}$$

Во всех трех рассмотренных случаях решающее правило, использующее P-значение, имеет одинаковое представление.

Пример 8.1. Пусть выборка X (n = 25) представляет собой независимые одинаково распределенные случайные величины, имеющие нормальное распределение $N(\mu, \sigma^2)$ с неизвестным математическим ожиданием μ и известной дисперсией σ^2. Выборочные оценки среднего значения и среднего квадратического отклонения равны соответственно $\overline{X} = 4{,}26$, $\sigma = 0{,}47$. Рассмотрим три случая.

1. При уровне значимости $\alpha = 0{,}05$ необходимо проверить гипотезу о числовом значении математического ожидания H_0: $\mu = 4{,}5$ против альтернативы H_1: $\mu \neq 4{,}5$ (см. подразд. 7.3). Для этого используем статистику критерия Z, имеющую стандартное нормальное распределение:

$Z = |\overline{X} - \mu_0| / (\sigma / \sqrt{n}) = |4{,}26 - 4{,}5| / (0{,}48/\sqrt{25}) = 2{,}50$.

Решаем уравнение $1 - P/2 = \Phi(2{,}50)$, где $\Phi(\cdot)$ – функция стандартного нормального распределения. Получаем $1 - P/2 = 0{,}9938$, откуда $P = 0{,}0124 < \alpha = 0{,}05$. Таким образом, гипотеза H_0 отклоняется (рис. 8.1).

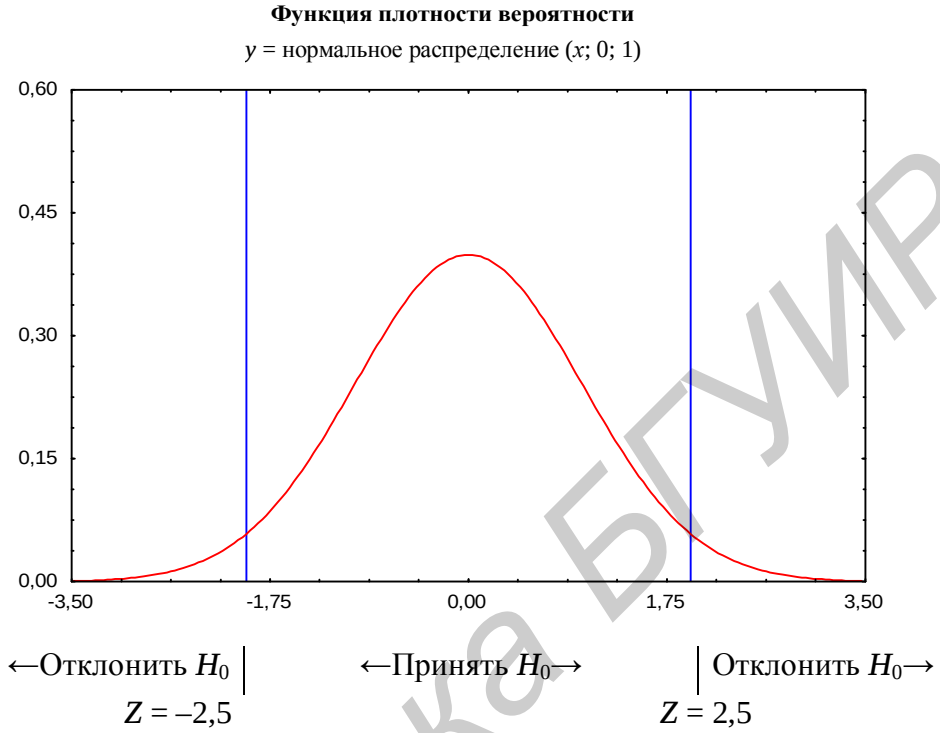


Рис. 8.1. Область принятия нулевой гипотезы H_0: $\mu = 4{,}5$; H_1: $\mu \neq 4{,}5$

2. Проверим гипотезу о значении математического ожидания при правосторонней альтернативе H_0: $\mu = 50$, H_1: $\mu > 50$ (см. подразд. 7.3). Пусть $n = 36$, $\overline{X} = 50{,}50$; $\sigma = 1{,}64$; $\alpha = 0{,}05$. Статистика критерия

$$Z = (50{,}50 - 50{,}0)/(1{,}64/\sqrt{36}) = 1{,}829.$$

Так как статистика критерия Z имеет стандартное нормальное распределение $\Phi(x)$, находим P-значение: $P = 1 - \Phi(1{,}829) = 1 - 0{,}9663 = 0{,}0337$. Следовательно, гипотеза H_0 отклоняется, так как $P = 0{,}0337 < \alpha = 0{,}05$ (рис. 8.2).

3. Проверим гипотезу о значении математического ожидания при левосторонней альтернативе H_0: $\mu = 19{,}5$; H_1: $\mu < 19{,}5$ (см. подразд. 7.3).

Пусть $n = 64$, $\overline{X} = 18{,}90$; $\sigma = 2{,}4$; $\alpha = 0{,}05$. Статистика критерия

$$Z = (18{,}90 - 19{,}50)/(2{,}4/\sqrt{64}) = -2{,}0.$$

Из стандартного нормального распределения находим P-значение: $P = \Phi(-2{,}0) = 0{,}0228$. Следовательно, гипотеза H_0 отклоняется, так как $P = 0{,}0228 < \alpha = 0{,}05$ (рис. 8.3).

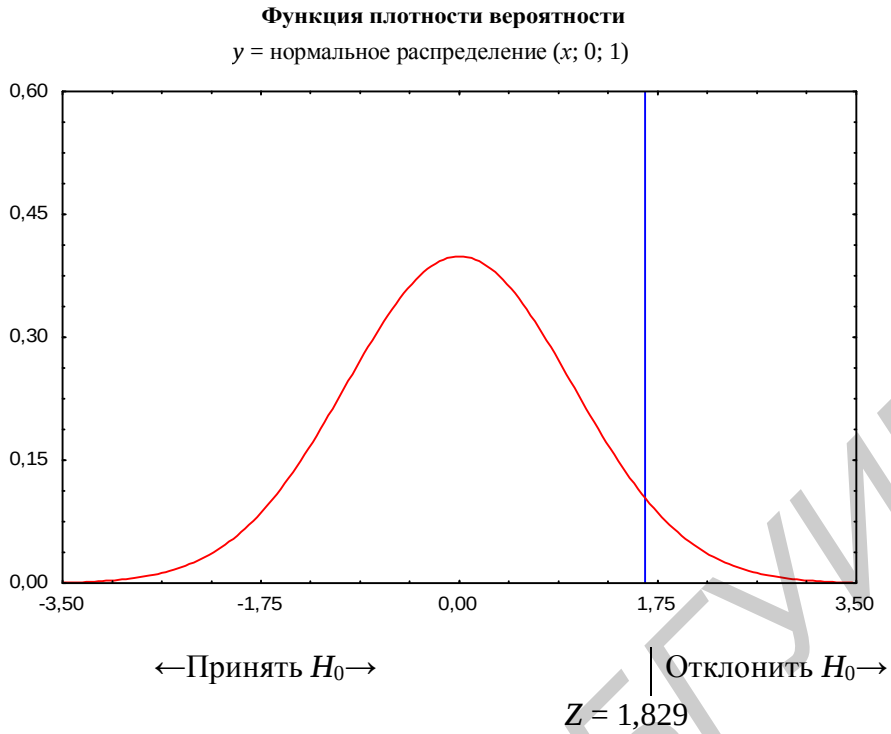


Рис. 8.2. Область принятия нулевой гипотезы H_0: μ = 50; H_1 : μ >50

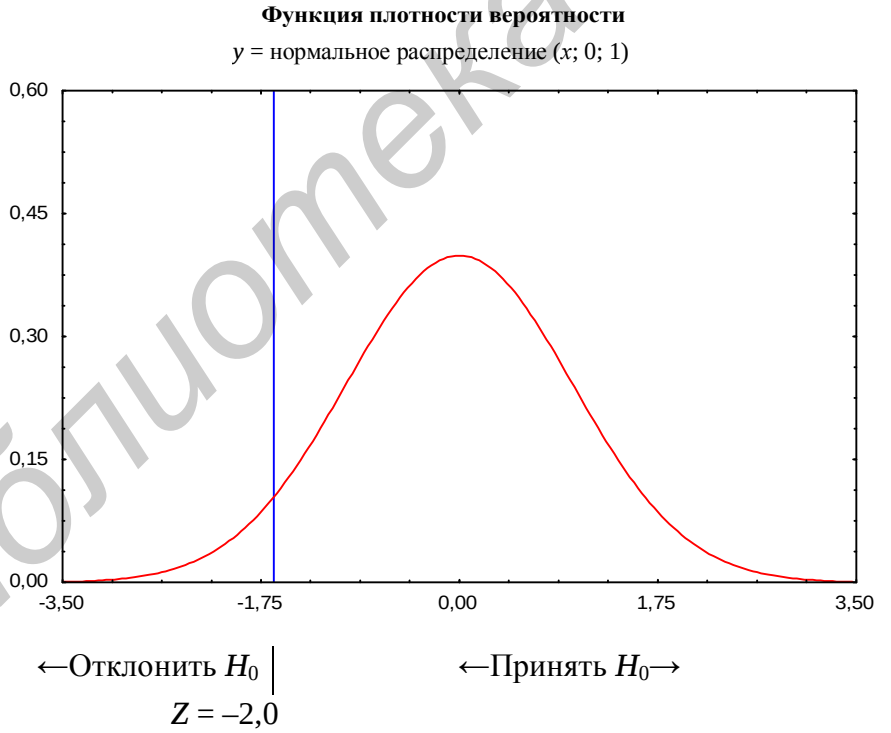


Рис. 8.3. Область принятия нулевой гипотезы H_0: μ = 19,5; H_1: μ < 19,5

При выработке решения можно воспользоваться традиционной интерпретацией *P*-значений:

1) $P > 0{,}1$– имеем хорошее согласие с H_0;

2) $P = 0{,}05$ – есть некоторые сомнения в истинности H_0;

3) $P = 0{,}02$ – довольно сильный довод против H_0;

4) $P \leq 0{,}01$ – гипотеза H_0 почти наверняка не подтверждается.

9. ОСНОВНЫЕ ТИПЫ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Ключевые понятия: критерии согласия, критерии адекватности

По схеме, приведенной в подразд. 8.1, можно решить большое количество практических задач, применив различные статистические критерии. Высказываемые в ходе решения задач гипотезы можно подразделить на следующие типы [19]:

- об общем виде закона распределения исследуемой случайной величины;
- об однородности двух или нескольких выборок;
- о числовых значениях параметров исследуемой генеральной совокупности;
- об общем виде зависимости, существующей между компонентами исследуемого многомерного признака;
- о независимости и стационарности ряда наблюдений.

9.1. Гипотезы о типе закона распределения исследуемой случайной величины

Имеется ряд наблюдений $X_1, X_2, \ldots, X_n$ исследуемой случайной величины ξ. Задача состоит в подборе некоторой модельной функции распределения $F_{\text{mod}}(X)$, с помощью которой можно описать функцию распределения $F_\xi(X)$ случайной величины ξ. Проверяемая гипотеза имеет вид

$$H\colon F_{\text{o}}(X) \equiv F_{\text{mod}}(X), \qquad (9.1)$$

где гипотетическая модельная функция может быть: 1) заданной *однозначно* (тогда $F_\xi(X) = F_0(X)$, где $F_0(X)$ – полностью известная функция); 2) *заданной с точностью до принадлежности к некоторому параметрическому семейству* (тогда $F_{\text{mod}}(X) = F(X; \theta)$, где θ – некоторый параметр, значения которого неизвестны, но могут быть оценены по выборке $X_1, X_2, \ldots, X_n$).

Проверка гипотез об общем виде закона распределения осуществляется с помощью ***критериев согласия***, которые основаны на использовании мер раз-

личия между анализируемой эмпирической функцией распределения и гипотетической модельной $F_{\text{mod}}(X)$ (см. разд. 10).

9.2. Гипотезы об однородности двух или нескольких выборок и характеристик анализируемых совокупностей

Исходная информация для проверки гипотезы однородности представлена m выборками объема n_i, $i = 1, \ldots, m$:

1-я выборка: $X_{11}, X_{12}, \ldots, X_{1n_1}$;

2-я выборка: $X_{21}, X_{22}, \ldots, X_{2n_2}$;

. . . . . .

m-я выборка: $X_{m1}, X_{m2}, \ldots, X_{mn_m}$.

Обозначим вероятностный закон распределения, которому подчиняются наблюдения j-й выборки, средние значения и дисперсии соответственно $F_j(X)$, a_j, s_j^2. Эмпирические характеристики этого закона обозначим $\hat{a}_j$ и $\hat{s}_j^2$. Тогда основные гипотезы однородности можно записать в виде

$$H_F: F_1(X) \equiv F_2(X) \equiv \ldots \equiv F_m(X); \quad (9.2)$$

$$H_a: a_1 = a_2 = \ldots = a_m; \quad (9.3)$$

$$H_S: s_1^2 = s_2^2 = \ldots = s_m^2. \quad (9.4)$$

В случае неотрицательного результата проверки гипотез (9.3) и (9.4) говорят, что соответствующие выборочные характеристики (например, $\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_m$) различаются ***статистически незначимо***.

Особо выделяется случай гипотез типа (9.2), когда число выборок $m = 2$, а одна из выборок содержит малое количество наблюдений (в частном случае – одно). В таком виде проверка гипотез типа (9.2) означает ***проверку аномальности одного или нескольких резко выделяющихся наблюдений***. Описание критериев проверки гипотез типа (9.3) и (9.4) приводится в подразд. 11.2 и подразд. 11.3.

9.3. Гипотезы о числовых значениях параметров исследуемой генеральной совокупности

Задачи этого типа возникают, например, когда нужно проверить точность некоторого прибора, инструмента, устойчивость определенного технологического процесса и т. п. Они сводятся к проверке гипотезы о параметрах иссле-

дуемой генеральной совокупности. Формально гипотезы подобного рода приобретают вид

$$H_0: \theta = \Delta_0, \qquad (9.5)$$

где θ – некоторый параметр исследуемого распределения, а Δ_0 – область его конкретных значений. В качестве таких параметров часто рассматриваются наиболее информативные характеристики распределения: среднее значение (медиана в непараметрическом случае) и дисперсия.

Гипотеза аналогичного типа рассматривается, если необходимо проверить статистическую незначимость отличия от нуля выборочного коэффициента корреляции [1, 19]. В этом случае гипотеза записывается как предположительное утверждение H_0: $r_{xy} = 0$.

9.4. Гипотезы о типе зависимости между компонентами исследуемого многомерного признака

При исследовании статистической зависимости, например, компоненты $x^{(2)}$ от компоненты $x^{(1)}$ анализируемого двумерного признака $X = (x^{(1)}, x^{(2)})$, бывает необходимо проверить *гипотезу об общем виде этой зависимости*. Например, проверить гипотезу о том, что $x^{(2)}$ и $x^{(1)}$ связаны линейной регрессионной связью [1, 19]:

$$H: \mathbf{E}\,(x^{(2)} | x^{(1)}) = b_0 + b_1 x,$$

где b_0 и b_1 – некоторые неизвестные параметры модели.

Статистические критерии, с помощью которых проверяются гипотезы данного типа, называют ***критериями адекватности***.

9.5. Гипотезы независимости и стационарности ряда наблюдений

Принятие решения о независимости или стационарности наблюдений (сохранения закона распределения) основывается на результатах проверки соответствующих гипотез, например:

$$H: \mathbf{E}\,x_i = a = \text{const}, \; i = 1, 2, \ldots, n;$$

$$H: r\,(x_i, x_{i+1}) = 0, \; i = 1, 2, \ldots, n - 1,$$

где **E**(·) – математическое ожидание случайной переменной; $r\,(x_i,\ x_{i+1})$ – коэффициент корреляции, построенный по совокупности двумерных наблюдений $X = (x^{(1)},\ x^{(2)})$.

Гипотезы этого типа являются частным случаем гипотез типа (9.5).

10. ПРОВЕРКА СООТВЕТСТВИЯ ВЫБРАННОЙ МОДЕЛИ РАСПРЕДЕЛЕНИЯ ИСХОДНЫМ ДАННЫМ (КРИТЕРИИ СОГЛАСИЯ)

Ключевые понятия: критерий согласия ч^2 Пирсона, критерий Колмогорова

10.1. Выбор модели распределения для описания данных

При анализе данных весьма часто возникает необходимость определения вида функции распределения, пригодной для описания выборочных данных. При этом подбирается соответствующая модель и проверяется ее пригодность для описания распределения результатов наблюдений. Выбор этой модели обычно основан на априорной информации о физическом характере изучаемых явлений или на статистической проверке предположения о согласованности модели с наблюдаемыми данными.

Пусть нами высказано предположение, что ряд наблюдений $x_1, x_2, \ldots, x_n$ исследуемой случайной величины ξ образует случайную выборку из распределения с некоторой модельной функцией $F_0(x;\ q^{(1)}, \ldots, q^{(s)})$, где общий вид функции F_0 (т. е. тип модели) считается известным, а параметры $q^{(1)}, \ldots, q^{(s)}$ могут быть как известными, так и неизвестными.

Критерии согласия предназначены для проверки гипотезы

$$H_0:\ F_0(x) = F_0(x;\ q^{(1)}, \ldots, q^{(s)}) \qquad (10.1)$$

и основаны на использовании различных мер расстояний между анализируемой эмпирической функцией распределения, определяемой по выборке, и гипотетической модельной $F_0(x;\ q^{(1)}, \ldots, q^{(s)})$.

Такие критерии подразделяются на два класса – общие критерии согласия и специальные критерии согласия. Первые из них применимы к формулировке гипотезы как гипотезы о согласии наблюдаемых результатов с любым априорно

предполагаемым распределением вероятностей. Специальные критерии согласия предполагают специальные нулевые гипотезы, формулирующие согласие с определенной формой распределения вероятностей: нормальной, экспоненциальной и т. д.

Среди общих критериев согласия можно выделить две основные группы:

1) критерии, основанные на изучении разницы между теоретической плотностью распределения и эмпирической гистограммой;

2) критерии, основанные на расстоянии между теоретической и эмпирической функциями распределения вероятностей.

10.2. Критерий ч^2 Пирсона

Критерий основан на сравнении эмпирической гистограммы распределения случайной величины с ее теоретической плотностью. Критерий χ^2 Пирсона позволяет проверить гипотезу (10.1), когда значения параметров $q^{(1)}, ..., q^{(s)}$ неизвестны и данные сгруппированы. Процедура проверки гипотезы состоит из следующих шагов [1, 19, 25].

Диапазон значений исследуемой случайной величины ξ разбивается на k взаимно исключающих и непересекающихся интервалов $I_1, ..., I_k$. Длина интервалов разбиения необязательно одинакова.

На основании выборочных данных $x_1, x_2, ..., x_n$ строятся статистические оценки $\hat{q}^{(1)}, ..., \hat{q}^{(s)}$ неизвестных параметров $\theta^{(1)}, ..., \theta^{(s)}$, от которых зависит закон распределения F.

Подсчитывается число наблюдений n_i, попадающих в каждый интервал группирования I_i, $i = 1, \ldots, k$.

Вычисляются вероятности событий ξ I I_i, т. е. вероятности p_i попадания случайной величины ξ в интервал I_i:

$$p_i = F_0(l_i;\ \hat{q}^{(1)}, ..., \hat{q}^{(s)}) - F_0(l_{i-1};\ \hat{q}^{(1)}, ..., \hat{q}^{(s)}),$$

где l_{i-1} и l_i – левый и правый концы i-го интервала группирования.

Вычисляется ожидаемое число наблюдений n_i в интервале I_i при условии справедливости гипотезы H_0: $n_i = n\,p_i$.

Вычисляется статистика

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i - \nu_i)^2}{\nu_i},$$

которая при верной H_0 имеет χ^2-распределение с $f = k - s - 1$ степенями свободы.

Гипотеза о том, что исследуемая случайная величина ξ подчиняется закону распределения F_0, ***принимается на уровне значимости*** α, если

$$\Delta(\alpha/2) \le \chi^2 < \Delta(1 - \alpha/2),$$

где $\Delta(\varepsilon)$ – квантиль уровня ε имеет χ^2-распределение с $f = k - s - 1$ степенями свободы.

Если $\chi^2 \ge \Delta\,(1 - \alpha/2)$, ***гипотеза*** H_0 ***отклоняется***, так как выполнение неравенства свидетельствует о слишком большом отклонении исследуемого закона распределения от $F_0(x)$. Случай $\chi^2 < \Delta\,(\alpha/2)$ требует дополнительного исследования [19]. Слишком малые значения статистики критерия говорят о неудачном выборе закона F (завышение числа параметров), нарушении технологии выборочного обследования и т. д.

Пример 10.1. Требуется проверить гипотезу H_0 (10.1) о том, что генеральная совокупность имеет стандартное нормальное распределение ($\mu = 0$, $\sigma^2 = 1$), т. е. $H_0 : F(x) = \Phi(x)$. Для проверки этой гипотезы используем критерий согласия χ^2 на уровне значимости $\alpha = 0{,}05$ по результатам наблюдений выборки объемом $n = 60$. Данные сгруппированы в $k = 7$ интервалов, число параметров распределения $s = 2$. В результате вычислений получаем значение статистики $\chi^2 = 5{,}3482$ и число степеней свободы $f = 7 - 2 - 1 = 4$. По таблицам χ^2-распределения [16] определяем границы области принятия H_0: $\chi^2_{0{,}025;\,4} = 0{,}4844$ и $\chi^2_{0{,}975;\,4} = 11{,}1433$. Так как вычисленное значение статистики критерия $\chi^2 = 5{,}3482$ лежит в интервале $0{,}4844 < \chi^2 < 11{,}1433$, то можно сделать вывод о том, что результаты наблюдений не противоречат гипотезе H_0 (рис. 10.1).

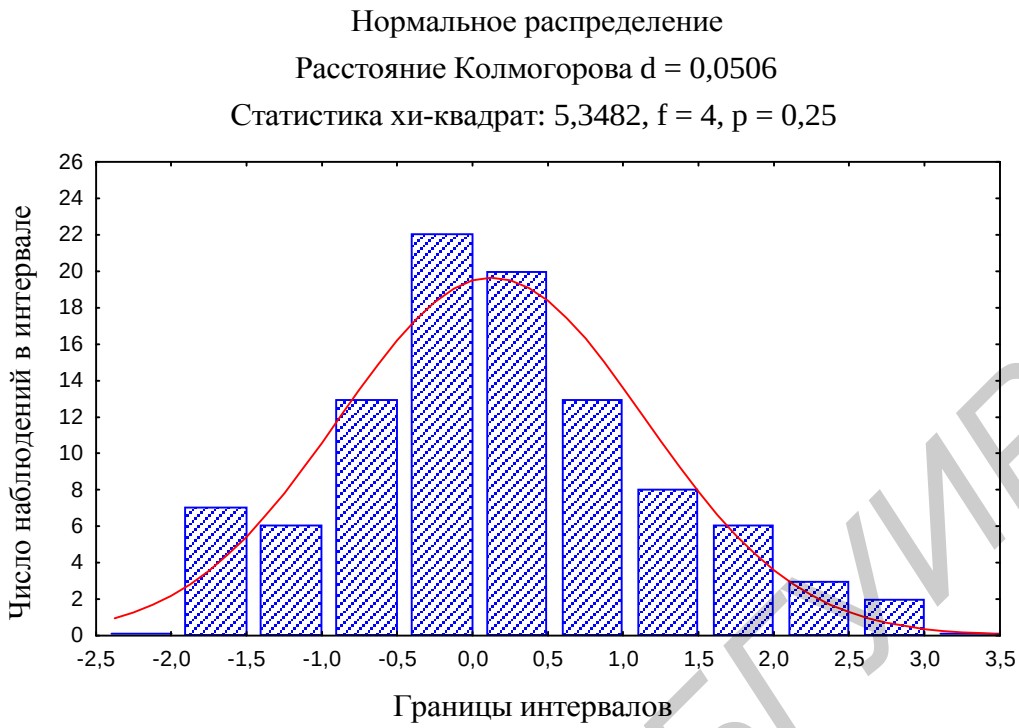


Рис. 10.1. Результаты проверки гипотезы согласия

Проверку гипотезы H_0 можно выполнить, используя P-значение (см. подразд. 8.2). Для этого необходимо сравнить вычисленное P-значение, равное 0,25, с уровнем значимости α. Так как α = 0,05 < 0,25, то принимается H_0. Следовательно, результаты наблюдений могут быть описаны нормальным распределением.

10.3. Критерий Колмогорова

Когда модельное распределение известно полностью и является непрерывным, для проверки гипотезы согласия (10.1) целесообразно использовать критерий Колмогорова.

Определим расстояние Колмогорова между эмпирической $F_э(x)$ и теоретической функциями распределения:

$$D = \max_x |F_э(x) - F_0(x)|.$$

Эмпирическим (или ***выборочным***, т. е. построенным по выборке объема n) аналогом теоретической функции распределения $F(x)$ является функция $\hat{F}^{(n)}(x)$, определяемая соотношением

$$\hat{F}^{(n)}(x) = \nu(x) / n$$

или в случае группированных данных соотношением

$$\hat{F}^{(n)}(x) = (\nu_1 + \nu_2 + \ldots + \nu_{ix}) / n,$$

где $\nu(x)$ – число наблюденных значений исследуемой случайной величины в выборке $x_1, x_2, \ldots, x_n$, меньших x; ν_i – число наблюденных значений в выборке, попавших в i-й интервал группирования, а ix – номер самого правого из интервалов группирования, правый конец которых не превосходит x.

Решающее правило:

$$\text{принимается гипотеза } \begin{cases} H_0, & \text{если} \quad \sqrt{n}D \le \Delta(a), \\ H_1, & \text{если} \quad \sqrt{n}D > \Delta(a). \end{cases}$$

Если верна гипотеза H_0 и $n \ge 20$, то независимо от вида модельного распределения $F_0(x)$ случайная величина $\sqrt{n}\,D$ имеет распределение Колмогорова. Это позволяет определить границу критической области (порог теста) $\Delta(\alpha)$ при уровне значимости α как квантиль распределения Колмогорова. Значения $\Delta(\alpha)$ приведены в табл. 10.1 [25].

Таблица 10.1

Значения квантилей распределения Колмогорова

α	0,01	0,05	0,1	0,2
Δ	1,63	1,36	1,22	1,07

Для приближенного вычисления критического расстояния $d_\alpha(n_1, n_2)$ можно использовать следующее соотношение:

$$d_\alpha(n_1, n_2) = \left(0{,}5 \ln \frac{1}{1-\alpha}\right)^{-1}.$$

Если $D_{n_1, n_2} > d_\alpha(n_1, n_2)$, гипотеза согласия H_0 отклоняется при уровне значимости α. При $n \ge 20$ полезна аппроксимация*

$$\chi^2 = \frac{6nD_{n_1, n_2} + 1}{9n},$$

распределение которой описывается χ^2-распределением с $f = 2$ степенями свободы.

* Большев, Л. Н. Таблицы математической статистики / Л. Н. Большев, Н. В. Смирнов. – М. : Наука, 1965

При $n \geq 10$ необходимо использовать более точное приближение

$$d_\alpha(n_1, n_2) \approx \sqrt{\frac{y}{2n}} - \frac{1}{6n},$$

где $y = -\ln \alpha$ при $0{,}01 \leq \alpha \leq 0{,}2$.

Пример 10.2. Проверить на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 (10.1) о том, что распределение анализируемой случайной величины является стандартным нормальным ($\mu = 0$, $\sigma^2 = 1$). Получена выборка объемом $n = 60$. Для проверки этой гипотезы используем критерий Колмогорова. В результате обработки данных имеем $D = 0{,}0506$. Отсюда значение статистики критерия $\sqrt{n}\,D = \sqrt{60} \cdot 0{,}0506 = 0{,}3919$. По табл. 10.1 при $\alpha = 0{,}05$ находим границу области принятия нулевой гипотезы $\Delta(\alpha) = 1{,}36$. Так как $\sqrt{n}\,D = 0{,}3919 < 1{,}36$, то результаты наблюдений не противоречат гипотезе H_0.

Применим другой способ проверки гипотезы, в котором используется решающее правило, основанное на P-значении:

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } P \geq a, \\ H_1, \text{ если } P < a. \end{cases}$$

Для статистики критерия, равной $c^2 = 3{,}5916$, вычисляется P-значение. Его величина $p = 0{,}61$, поэтому при уровне значимости $\alpha = 0{,}05$ нулевая гипотеза об однородности распределений принимается.

Результаты проверки гипотезы согласия с помощью критериев χ^2 Пирсона и Колмогорова показаны на рис. 10.1.

11. ГИПОТЕЗЫ ОДНОРОДНОСТИ

Ключевые понятия: критерии однородности распределений, гипотезы однородности математических ожиданий и дисперсий, гипотеза о параметрах двух биноминальных распределений

Формулировка гипотез однородности. Общая характеристика гипотез однородности была дана в подразд. 9.2. В данном разделе будут рассмотрены критерии однородности для $m = 2$ выборок.

Гипотезы равенства (однородности) функций распределения H_1, **математических ожиданий** H_2 и **дисперсий** H_3 и их альтернативы $\overline{H}_1$, $\overline{H}_2$ и $\overline{H}_3$ можно записать в виде

$$H_1: F_1(x) = F_2(x),\ \overline{H}_1: F_1(x) \neq F_2(x); \tag{11.1}$$

$$H_2: m_1 = m_2,\ \overline{H}_2: m_1 \neq m_2; \tag{11.2}$$

$$H_3: s_1^2 = s_2^2,\ \overline{H}_3: s_1^2 \neq s_2^2. \tag{11.3}$$

11.1. Критерии однородности распределений

Исходные данные. Наблюдаются две независимые случайные выборки $X^{(1)}$ и $X^{(2)}$ объемом n_1 и n_2 соответственно: $X^{(i)} = (x_{i1}, ..., x_{in_i})$, $i = 1, 2$.

Гипотеза однородности распределений состоит в том, что генеральные совокупности, из которых извлечены выборки, одинаковы, и, следовательно, им соответствуют одинаковые функции распределения.

Нулевая гипотеза (11.1) в случае двух выборок формулируется следующим образом:

H_1: $F_1(x) = F_2(x)$.

Для проверки гипотезы однородности распределений (11.1) можно использовать ***двухвыборочный критерий Колмогорова – Смирнова и критерий однородности*** χ^2.

Двухвыборочный критерий Колмогорова – Смирнова [1, 19]. Пусть $x_{(1)}, \ldots, x_{(n_1)}$ и $y_{(1)}, \ldots, y_{(n_2)}$ – вариационные ряды, состоящие из элементов первой и второй выборок соответственно. Определим две **эмпирические функции распределения** $F_1^{(n_1)}(x)$ и $F_2^{(n_2)}(x)$. Вводится следующая статистика, определяющая разность двух эмпирических функций распределения:

$$D_{n_1, n_2} = \max_{-\infty < x < \infty} \left| F_1^{(n_1)}(x) - F_2^{(n_2)}(x) \right|.$$

Положим $n_2 \le n_1$. При верной нулевой гипотезе H_1 (11.1) случайная величина $\sqrt{n_0}\, D_{n_1, n_2}$ ($n_0 = \sqrt{\dfrac{n_1 n_2}{n_1 + n_2}}$, $n_2 \to \infty$) имеет распределение Колмогорова.

Критической областью для статистики $\sqrt{n_0}\, D_{n_1, n_2}$ является область больших значений. Нулевая гипотеза H_1 отвергается, если $\sqrt{n_0}\, D_{n_1, n_2} > d_\alpha(n_1, n_2)$, где

$d_\alpha(n_1, n_2)$ – критическое значение статистики $\sqrt{n_0}\, D_{n_1, n_2}$ при уровне значимости α. Значения квантилей Δ(α) приведены в табл. 10.1 (см. подразд. 10.3).

Пример 11.1. Проверить гипотезу (11.1) о том, что две выборки объемом $n_1 = 60$ и $n_2 = 80$ принадлежат одной и той же генеральной совокупности. Пусть значение статистики $D_{n_1, n_2} = 0{,}2667$. Для α = 0,05 по табл. 10.1 находим границу критической области Δ(0,05) = 1,36. Так как $\sqrt{n_0}\, D_{n_1, n_2} = 5{,}8554 \cdot 0{,}2667 = 1{,}5616 > 1{,}36$, то нулевая гипотеза отклоняется. Это означает, что генеральные совокупности, из которых извлечены выборки, не одинаковы, и, следовательно, им соответствуют различные функции распределения.

Применим другой способ проверки гипотезы, в котором используется решающее правило, основанное на *P*-значении:

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } P \ge a, \\ H_1, \text{ если } P < a. \end{cases}$$

Для статистики критерия, равной 1,5616, вычисляется *P*-значение. Его величина $p < 0{,}025$, поэтому при уровне значимости α = 0,05 нулевая гипотеза об однородности распределений отклоняется.

11.2. Однородность математических ожиданий

Гипотеза H_2 однородности математических ожиданий (теоретических средних) при двусторонней альтернативе $\overline{H}_2$ записывается как

$$H_2: m_1 = m_2, \quad \overline{H}_2: m_1 \neq m_2.$$

Исходные данные представлены двумя независимыми случайными выборками $X^{(1)}$ и $X^{(2)}$ объемом n_1 и n_2 соответственно:

$$X^{(i)} = (x_{i1}, \ldots, x_{in_i}),\ i = 1, 2.$$

Предполагается, что выборки извлечены из нормальных распределений с равными дисперсиями ($s_1^2 = s_2^2$).

Для проверки гипотезы равенства математических ожиданий в двух выборках в этом случае используется ***t*-критерий Стьюдента**, или **двухвыборочная статистика Стьюдента** [19, 22]:

$$t = (\bar{X}_1 - \bar{X}_2) / \sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}\ . \qquad (11.4)$$

В формуле (11.4) использованы следующие обозначения:

$\bar{X}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}$ – выборочное среднее для выборки $X^{(i)}$;

$s^2 = \frac{\sum_{i=1}^{2}(n_i - 1)s_i^2}{\sum_{i=1}^{2} n_i - 2}$ – объединенная выборочная дисперсия;

$$s_i^2 = \frac{1}{n_i - 1}\sum_{j=1}^{n_i}(x_{ij} - \bar{X}_i)^2\ - \qquad (11.5)$$

выборочная оценка дисперсии для выборки $X^{(i)}$ (i = 1, 2).

Решающее правило имеет вид

$$\text{принимается гипотеза } \begin{cases} H_2, \text{ если } |t| \le \Delta(e), \\ \bar{H}_2, \text{ если } |t| > \Delta(e), \end{cases}$$

где Δ(·) – квантиль уровня 1– α/2 t-распределения Стьюдента с $n_1 + n_2 - 2$ степенями свободы.

При правосторонней альтернативе $\bar{H}_2$: $m_1 > m_2$ **решающее правило** имеет вид

$$\text{принимается гипотеза } \begin{cases} H_2, \text{ если } t \le \Delta(e), \\ \bar{H}_1, \text{ если } t > (e), \end{cases}$$

где Δ(·) – квантиль уровня 1 – α t-распределения Стьюдента с $n_1 + n_2 - 2$ степенями свободы.

При левосторонней альтернативе $\bar{H}_2: \mu_1 < \mu_2$ **решающее правило** имеет вид

$$\text{принимается гипотеза } \begin{cases} H_2, \text{ если } t \geq \Delta(e), \\ \bar{H}_2, \text{ если } t < \Delta(e), \end{cases}$$

где $\Delta(\cdot)$ – квантиль уровня α t-распределения Стьюдента с $n_1 + n_2 - 2$ степенями свободы.

Критерий (11.4) оказывается чувствительным к отклонениям распределений от нормальности, различиям дисперсий s_1^2 и s_2^2 и неравенству объемов выборок n_1 и n_2.

Для обеспечения устойчивости решений к нарушению нормальности распределения и равенства дисперсий разработаны устойчивые критерии, например **t-критерий Уэлча** [19, 25].

Для проверки гипотез

$H_d: m_1 - m_2 = \delta, \qquad \bar{H}_d: m_1 - m_2 \neq \delta$

(значение δ фиксировано, в случае гипотезы равенства средних $\delta = 0$) при возможных нарушениях условия равенства дисперсий ($s_1^2 \neq s_2^2$) используется тест вида

$$\text{принимается гипотеза } \begin{cases} H, \text{ если } |t_1| \geq \Delta(e), \\ \bar{H}, \text{ если } |t_1| > \Delta(e). \end{cases}$$

Здесь $t_1 = (\bar{X}_1 - \bar{X}_2 - \delta) / (s_1^2/n_1 + s_2^2/n_2)^{1/2}$ – **статистика Уэлча**;

$\Delta(\cdot)$ – квантиль уровня $1 - \alpha/2$ t-распределения Стьюдента с n_1 степенями свободы;

$$n_1 = \left(\frac{c_1^2}{n_1 - 1} + \frac{(1 - c_1^2)^2}{n_2 - 1} \right)^{-1}, \quad c_1 = \frac{s_1^2/n_1}{s_1^2/n_1 + s_2^2/n_2}.$$

При правосторонней и левосторонней альтернативах решающие правила аналогичны случаю применения t-критерия Стьюдента.

Пример 11.2. Имеем $\overline{X}_1 = 32$, $s_1^2 = 5{,}9$ ($n_1 = 25$); $\overline{X}_2 = 36{,}2$, $s_2^2 = 11{,}2$ ($n_2 = 25$). Известно, что анализируемые распределения являются нормальными с равными дисперсиями. Проверить гипотезу о равенстве двух математических ожиданий при уровне значимости $\alpha = 0{,}05$ против правосторонней альтернативы.

Решение. Выдвигаем гипотезы

$H_0 : \mu_1 = \mu_2$, $H_1: \mu_1 < \mu_2$.

В качестве критерия используем двухвыборочную статистику Стьюдента (11.4), которая при справедливости H_0, нормальном распределении исходной случайной переменной и равенстве дисперсий имеет t-распределение Стьюдента с $n_1 + n_2 - 2$ степенями свободы. Определим величину статистики критерия: $t = -5{,}08$.

Определим границу критической области x_k. Значение границы критической области при левосторонней H_1 представляет собой квантиль уровня $p = \alpha$ распределения статистики критерия. Найдем значение квантили $t_{0{,}05;\,48}$ уровня $\alpha = 0{,}05$ t-распределения Стьюдента с $n_1 + n_2 - 2 = 48$ степенями свободы: $t_{0{,}05;\,48} = x_k = -1{,}645$.

Для проверки гипотезы используем решающее правило:

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } t \geq \Delta(e), \\ H_1, \text{ если } t < \Delta(e). \end{cases}$$

При уровне значимости $\alpha = 0{,}05$ гипотеза о равенстве среднего значения $\mu_1 = \mu_2$ отклоняется, так как $t = -5{,}08 < x_k = \Delta(\alpha) = -1{,}645$. Следовательно, при уровне значимости $\alpha = 0{,}05$ принимается гипотеза H_1.

Применим другой способ проверки гипотезы, в котором используется решающее правило, основанное на P-значении:

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } P \geq a, \\ H_1, \text{ если } P < a. \end{cases}$$

Находим значение функции t-распределения с $n_1 + n_2 - 2 = 48$ степенями свободы, соответствующее значению статистики $t = -5{,}08$: $F_t(-5{,}08) = 0{,}000003$. В условиях левосторонней альтернативы получаем $P = F_t(-5{,}08) = 0{,}000003$. Так как $P < \alpha = 0{,}05$, при уровне значимости $\alpha = 0{,}05$ гипотеза H_0 отклоняется.

Многомерный случай. Для нескольких совокупностей ($m > 2$) можно проверить гипотезу о равенстве векторов математических ожиданий – H_0: $\mu_1 = \mu_2 = \ldots = \mu_m$ (предполагается нормальность распределений и равенство ковариационных матриц). Для этой цели используется Λ-статистика Уилкса [1]:

$\Lambda = \det(S) / \det(C)$,

где $S = \frac{1}{n - m} \sum_{i=1}^{m} (n_i - 1) S_i$ – оценка ковариационной матрицы общей генеральной совокупности, $n = \sum_{i=1}^{m} n_i$; элементы матрицы S_i определяются по формуле (17.7); $C = S + \frac{1}{n} \sum_{i=1}^{m} n_i (\overline{X}_i - \overline{X})(\overline{X}_i - \overline{X})^T$ – оценка ковариационной матрицы по выборке, полученной объединением всех m выборок; $\overline{X}$ – вектор средних значений такой объединенной выборки; $\det(\cdot)$ – определитель матрицы.

Распределение статистики Λ очень сложное, поэтому на практике пользуются следующим правилом. Значение Λ заключено в интервале $0 \le \Lambda \le 1$, и если верна нулевая гипотеза, то значение Λ должно быть близко к единице. Малые значения Λ означают несовпадение (различимость) центров классов, т. е. нулевая гипотеза должна быть отклонена.

При ($m = 2$) эта гипотеза эквивалентна гипотезе о равенстве нулю квадрата расстояния Махаланобиса Δ^2 между двумя распределениями. Проверка такой гипотезы рассмотрена в разд. 12 .

11.3. Однородность дисперсий

Если необходимо выяснить, за счет чего обнаружилась неоднородность рассматриваемых в подразд. 11.2 выборок, то следует дополнительно произвести проверку однородности дисперсий:

H_3: $s_1^2 = s_2^2$, $\overline{H}_3$: $s_1^2 \ne s_2^2$.

Статистический критерий однородности двух выборочных дисперсий основан на статистике

$$F = \frac{s_1^2}{s_2^2} = \frac{\sum_{i=1}^{n_1} (x_{1i} - \bar{X}_1)^2 / (n_1 - 1)}{\sum_{j=1}^{n_2} (x_{2j} - \bar{X}_2)^2 / (n_2 - 1)}, \qquad (11.6)$$

где s_i^2 – выборочная оценка дисперсии (11.5), $i = 1, 2$.

Эта статистика при справедливости гипотезы H_3 (11.3) имеет F-распределение Фишера с числами степеней свободы числителя и знаменателя, равными соответственно $n_1 - 1$ и $n_2 - 1$.

Тест уровня значимости α для проверки H_3 (11.3) записывается следующим образом:

принимается гипотеза $\begin{cases} H_3, \text{ если } \Delta(a/2) \le F \le \Delta(1-a/2), \\ \bar{H}_3, \text{ если } F < \Delta(a/2) \text{ или } F > \Delta(1-a/2), \end{cases}$

где Δ(ε) – порог теста, определяемый при верной H_3 (11.3) как квантиль уровня ε F-распределения Фишера с $n_1 - 1$ и $n_2 - 1$ степенями свободы.

Если рассматривается правосторонняя альтернатива $s_1^2 > s_2^2$, то H_3 отклоняется при $F > \Delta(1 - \alpha)$, а для альтернативы $s_1^2 < s_2^2$ гипотеза H_3 отклоняется при $F < \Delta(\alpha)$.

Пример 11.3. При условиях примера 11.2 проверить гипотезу равенства дисперсий.

Решение. Выдвигаем гипотезы H_0, H_1:

H_0: $s_1^2 = s_2^2$, H_1: $s_1^2 \neq s_2^2$

и устанавливаем уровень значимости α = 0,05.

В качестве статистики критерия используем (11.6). Ее значение равно F = 11,2/5,9 = 1,90.

Значение границ критической области при двусторонней H_1 представляет собой квантили уровней $p_1 = \alpha/2$ и $p_2 = 1 - \alpha/2$ распределения статистики критерия. Из таблиц F-распределения для значений функции распределения, равных соответственно 0,025 и 0,975, и степеней свободы $n_1 - 1 = 24$ и $n_2 - 1 = 24$

находим значение квантилей $F_{0,025;\,24} = 0{,}4407$ и $F_{0,975;\,24} = 2{,}2693$. Это означает, что при значениях статистики критерия $0{,}4407 \le F \le 2{,}2693$ принимается гипотеза H_0, а при значениях $F < 0{,}4407$ или $F > 2{,}2693$ – гипотеза H_1.

Для проверки гипотезы используем решающее правило:

$$\text{принимается гипотеза } \begin{cases} H_2, \text{ если } \Delta(a/2) \le F \le \Delta(1-a/2), \\ \bar{H}_2, \text{ если } F < \Delta(a/2) \text{ или } F > \Delta(1-a/2), \end{cases}$$

где $\Delta(\varepsilon)$ – порог теста, определяемый при верной H_0 как квантиль уровня ε F-распределения Фишера с $n_1 - 1$ и $n_2 - 1$ степенями свободы.

При уровне значимости $\alpha = 0{,}05$ гипотеза о равенстве дисперсий принимается, так как $0{,}4407 < F = 1{,}90 < 2{,}2693$.

Применим другой способ проверки гипотезы, в котором используется решающее правило, основанное на P-значении:

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } P \ge a, \\ H_1, \text{ если } P < a. \end{cases}^{\infty}$$

Находим значение функции F-распределения с $n_1 - 1 = 24$ и $n_2 - 1 = 24$ степенями свободы, соответствующее значению статистики $F = 1{,}90$: $F_F(1{,}90) = 0{,}9386$. В условиях двусторонней альтернативы получаем $1 - P/2 = F_F(1{,}90) = 0{,}9386$. Отсюда $P = 0{,}1228$. Так как $P > \alpha = 0{,}05$, то при уровне значимости $\alpha = 0{,}05$ гипотеза H_0 принимается.

11.4. Равенство параметров двух биномиальных распределений

Биномиальное распределение. Случайная величина, имеющая биномиальное распределение, возникает в случайном эксперименте, называемом последовательностью испытаний Бернулли [19].

В схему испытаний Бернулли хорошо укладываются такие случайные эксперименты, как бросание монеты или игральной кости, проверка изделий массовой продукции («годен – брак»), попытка выполнения некоторого задания («выполнено – не выполнено»), результаты медицинских профилактических обследований («здоров – болен») и т. п.

Биномиальное распределение описывает распределение случайной величины $n_p(n)$ – числа появления интересующего нас события в последовательности из n

независимых испытаний, когда вероятность появления этого события в одном испытании равна *p*. Формула биномиального закона распределения имеет вид

$$P\{n_p(m) = x\} = C_n^x \; p^x (1-p)^{n-x}.$$

Гипотеза о параметрах двух биномиальных распределений. Имеются две выборки из биномиального распределения с параметрами p_j, n_j, j = 1, 2. Необходимо проверить гипотезу о том, что эти две выборки взяты из совокупностей, подчиняющихся биномиальному распределению с одинаковыми параметрами p_j. Эту гипотезу можно записать как

$$H_0: \; p_1 = p_2 \text{ или } H_0: \; p_1 - p_2 = 0.$$

Альтернативная гипотеза $H_1: \; p_1 \neq p_2$. Данная задача во многих прикладных областях называется задачей «сравнения двух долей».

Если нулевая гипотеза справедлива, то оценки p_j, относящиеся к каждой из выборок, являются несмещенными оценками одного и того же параметра [19]. Следовательно, эти оценки можно объединить следующим образом:

$$\hat{r} = \frac{X_1 + X_2}{n_1 + n_2},$$

где X_j – число интересующих нас событий в выборке объема n_j, j = 1, 2. Вычислим среднее квадратическое отклонение разности выборочных оценок $\hat{p}_1 - \hat{p}_2$:

$$S_{\hat{p}_1 - \hat{p}_2} = \sqrt{\hat{p} \; (1-\hat{p}) \; (1/n_1 + 1/n_2)}.$$

Статистика критерия имеет вид

$$U = (\hat{p}_1 - \hat{p}_2) / S_{\hat{p}_1 - \hat{p}_2} = (X_1/n_1 - X_2/n_2) / S_{\hat{p}_1 - \hat{p}_2}. \qquad (11.7)$$

При $p_1 = p_2$ и достаточно больших n_1 и n_2 распределение U аппроксимируется стандартным нормальным распределением $N(0, 1)$. Тест уровня значимости α использует статистику U и записывается следующим образом:

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } |U| \le \Delta(e), \\ H_1, \text{ если } |U| > \Delta(e), \end{cases}$$

где Δ(ε) – квантиль уровня $1 - \alpha/2$ стандартного нормального распределения $u_{1-a/2}$.

Если рассматривается правосторонняя альтернатива H_1: $p_1 > p_2$, то H_0 отклоняется при $U > \Delta(1 - \alpha)$, а для альтернативы H_1: $p_1 < p_2$ нулевая гипотеза H_0 отклоняется при $U < \Delta(\alpha)$.

Формулы для проверки рассмотренных в разд. 11 гипотез однородности приведены в табл. П1.1.

Пример 11.4. В двух партиях контролируемых изделий объемом n_1 = 100 шт. и n_2 = 200 шт. обнаружены соответственно X_1 = 3 и X_2 = 5 дефектных изделий. Необходимо проверить гипотезу о равенстве долей дефектных приборов в партиях при двусторонней альтернативе (α = 0,05).

Вычисляем по формуле (11.7) значение U = 0,38. Из статистических таблиц находим границу критической области $\Delta(\varepsilon)$ как квантиль стандартного нормального распределения уровня $1 - \alpha/2 = 1 - 0{,}025 = 0{,}975$: $u_{0{,}975} = 1{,}96$. Так как $U = 0{,}38 < 1{,}96$, то при заданном уровне значимости α = 0,05 у нас нет оснований отклонить нулевую гипотезу.

12. ГИПОТЕЗЫ О ЧИСЛОВЫХ ЗНАЧЕНИЯХ ПАРАМЕТРОВ

Ключевые понятия: гипотезы о значении математического ожидания и дисперсии, гипотеза о значении параметра биноминального распределения

Гипотезы о числовых значениях параметров исследуемой генеральной совокупности имеют вид H_0: $\theta = \Delta_0$, где θ – некоторый параметр исследуемого распределения, а Δ_0 – область его конкретных значений. В качестве таких параметров часто рассматриваются наиболее информативные характеристики распределения: среднее значение и дисперсия [1, 8].

12.1. Гипотеза о значении математического ожидания

Гипотеза о значении математического ожидания нормального распределения может быть проверена в двух случаях: 1) когда дисперсия распределения s^2 известна; 2) когда дисперсия распределения s^2 неизвестна.

Случай 1: дисперсия σ^2 известна

Исходные данные. Выборка $X = (x_1, x_2, ..., x_n)$ представляет собой независимые одинаково распределенные случайные величины. Зададим уровень значимости α.

Предположения. Случайные переменные X имеют нормальное распределение $N(\mu, s^2)$ с неизвестным математическим ожиданием μ и известной дисперсией s^2. Необходимо проверить гипотезу о числовом значении математического ожидания против двусторонней альтернативы:

H_0: $\mu = \mu_0$, H_1: $\mu \neq \mu_0$. (12.1)

Для проверки гипотезы (12.1) используем статистику критерия

$Z = (\overline{X} - m_0)/(s/\sqrt{n})$, (12.2)

подчиняющуюся стандартному нормальному распределению.

Решающее правило имеет вид

$$\text{принимается гипотеза} \begin{cases} H_0, \text{ если } |Z| \le \Delta(e), \\ H_1, \text{ если } |Z| > \Delta(e), \end{cases}$$

где $\Delta(\varepsilon)$ – квантиль уровня $\varepsilon = 1 - \alpha/2$ стандартного нормального распределения, определяющая границу критической области $\boldsymbol{W}$.

Если рассматривается ***правосторонняя альтернатива*** H_1: $\mu > \mu_0$, то H_0 отклоняется при значениях $Z > \Delta(1 - \alpha)$, а для ***левосторонней альтернативы*** H_1: $\mu < \mu_0$ гипотеза H_0 отклоняется при $Z < \Delta(\alpha)$. В этих случаях ***граница критической области*** $\boldsymbol{W}$ определится соответственно как квантили стандартного нормального распределения уровня $1 - \alpha$ и уровня α: $\Delta(1 - \alpha)$ и $\Delta(\alpha)$.

Случай 2: дисперсия σ^2 неизвестна

Исходные данные. Выборка $X = (x_1, x_2, ..., x_n)$ – независимые одинаково распределенные случайные величины. Зададим уровень значимости α.

Предположения. Случайные переменные X имеют нормальное распределение $N(\mu, s^2)$ с неизвестным математическим ожиданием μ и неизвестной дисперсией s^2.

Необходимо проверить гипотезу (12.1). Так как истинное значение дисперсии неизвестно, в качестве оценки s^2 используем выборочную дисперсию

$$s^2 = \sum_{j=1}^{n} (x_j - \bar{X})^2 / (n-1). \qquad (12.3)$$

Статистика критерия

$$t = \frac{\bar{X} - m_0}{s/\sqrt{n}} \qquad (12.4)$$

подчиняется t-распределению Стьюдента с $n_1 = n - 1$ степенями свободы.

Решающее правило имеет вид

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } |t| \le \Delta(a), \\ H_1, \text{ если } |t| > \Delta(a), \end{cases}$$

где $\Delta(\varepsilon)$ – квантиль уровня $\varepsilon = 1 - \alpha/2$ t-распределения Стьюдента с $n - 1$ степенью свободы, определяющая границу критической области $\boldsymbol{W}$.

Если рассматривается правосторонняя альтернатива H_1: $\mu > \mu_0$, то H_0 отклоняется при значениях $t > \Delta(1 - \alpha)$, а для левосторонней альтернативы H_1: $\mu < \mu_0$ гипотеза H_0 отклоняется при $t < \Delta(\alpha)$. В этих случаях граница критической области $\boldsymbol{W}$ определится соответственно как $\Delta(1 - \alpha)$ и $\Delta(\alpha)$ – квантили уровней $1 - \alpha$ и α t-распределения Стьюдента с $n - 1$ степенями свободы.

Пример 12.1. На заводе выпускаются стандартные трубы с внутренним диаметром (ВД), равным 2,40 дюйма, причем дисперсия s^2 этой величины составляет 0,0004 дюйма^2. Известно, что внутренний диаметр является нормальной случайной величиной. Для выборки объемом $n = 25$ произведены измерения ВД. Среднее значение ВД, вычисленное по этой выборке, составляет 2,41 дюйма. Можно ли утверждать, что выборка взята из нормальной совокупности с указанным значением математического ожидания?

Решение. Проверим гипотезу о значении математического ожидания H_0: $\mu = 2{,}40$ против альтернативы H_1: $\mu \neq 2{,}40$.

Вычислим значение статистики критерия Z по формуле (12.2):

$Z = (2{,}41 - 2{,}40) / (0{,}02/\sqrt{25}) = 2{,}50$.

Зададим уровень значимости $\alpha = 0{,}05$. Определим значение квантили $u_{0{,}975}$ уровня $1 - \alpha/2 = 0{,}975$ стандартного нормального распределения: $u_{0{,}975} = 1{,}96$.

При уровне значимости α = 0,05 гипотеза о равенстве среднего значения 2,40 отклоняется, так как $Z = 2{,}50 > u_{0{,}975} = 1{,}96$.

Применим другой способ проверки гипотезы, в котором используется решающее правило, основанное на *P*-значении:

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } P \geq a, \\ H_1, \text{ если } P < a. \end{cases}$$

Для статистики критерия, равной 2,50, вычисляется *P*-значение. С этой целью решаем относительно *P* уравнение $1 - P/2 = \Phi(2{,}5)$, где $\Phi(\cdot)$ – функция стандартного нормального распределения. Получаем $1 - P/2 = 0{,}994$, откуда $P = 0{,}012$. Так как $P = 0{,}012 < \alpha = 0{,}05$, то при уровне значимости α = 0,05 гипотеза H_0 отклоняется.

Если считать, что дисперсия σ^2 неизвестна, а ее выборочная оценка s^2 составляет 0,00048 дюйма^2 (s = 0,024), тогда для проверки гипотезы (12.1) следует воспользоваться статистикой t, которая определяется по формуле (12.4) и равна $t = (2{,}41 - 2{,}40) / (0{,}024/\sqrt{25}) = 2{,}083$.

Вычисленное значение статистики $t = 2{,}083$ сравнивается с квантилью $t_{0{,}975}$ уровня $1 - \alpha/2 = 0{,}975$ t-распределения Стьюдента с $n - 1 = 24$ степенями свободы, которая равна 2,064. При уровне значимости α = 0,05 гипотеза о равенстве среднего значения внутреннего диаметра трубы 2,40 дюйма отклоняется, так как $t = 2{,}083 > t_{0{,}975} = 2{,}064$.

Проверим эту гипотезу с помощью *P*-значения. Находим значение функции t-распределения с $n - 1 = 24$ степенями свободы, соответствующее значению статистики $t = 2{,}083$: $F_t(2{,}083) = 0{,}976$. Получаем $1 - P/2 = 0{,}976$, откуда $P = 0{,}048$. Так как $P = 0{,}048 < \alpha = 0{,}05$, то при уровне значимости α = 0,05 гипотеза H_0 отклоняется.

Многомерный случай. При предположениях нормальности распределений и равенства ковариационных матриц (m = 2) можно проверить гипотезу о равенстве двух векторов математических ожиданий. Эта гипотеза эквивалентна гипотезе о равенстве нулю квадрата расстояния Махаланобиса Δ^2 (17.4) между двумя распределениями:

H_0: $\Delta^2 = 0$, H_1: $\Delta^2 \neq 0$. (12.5)

Гипотеза (12.5) проверяется с помощью статистики

$$T^2 = n_1 n_2 D^2/(n_1 + n_2),$$

где D^2 – выборочная оценка квадрата расстояния Махаланобиса Δ^2,

$$D^2 = (\bar{X}^{(1)} - \bar{X}^{(2)})^T S^{-1} (\bar{X}^{(1)} - \bar{X}^{(2)}), \qquad (12.6)$$

$S = \dfrac{(n_1-1)S_1 + (n_2-1)S_2}{n_1+n_2-2}$ – выборочная оценка ковариационной матрицы.

В случае справедливости H_0 величина

$$Y = \frac{(n_1+n_2-p-1)T^2}{p(n_1+n_2-2)}$$

имеет F-распределение с $\nu_1 = p$ и $\nu_2 = n_1 + n_2 - p - 1$ степенями свободы.

Решающее правило:

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } Y \le \Delta(a), \\ H_1, \text{ если } Y > \Delta(a). \end{cases}$$

Порог теста $\Delta(\alpha)$ представляет собой квантиль уровня $1 - \alpha$ F-распределения Фишера с $\nu_1 = p$ и $\nu_2 = n_1 + n_2 - p - 1$ степенями свободы.

12.2. Гипотеза о значении дисперсии

Исходные данные. Выборка $X = (x_1, x_2, ..., x_n)$ – независимые одинаково распределенные случайные величины. Зададим уровень значимости α.

Предположения. Случайные переменные X имеют нормальное распределение $N(\mu, s^2)$ с неизвестными математическим ожиданием μ и дисперсией s^2.

Необходимо проверить гипотезу о числовом значении дисперсии

$H_0: s^2 = s_0^2$ против двусторонней альтернативы $H_1: s^2 \ne s_0^2$. (12.7)

Статистика критерия c^2 вычисляется по следующей формуле:

$$c^2 = (n-1)\, s^2 / s_0^2, \qquad (12.8)$$

где s^2 – выборочная оценка дисперсии (12.3). Эта статистика имеет c^2-распределение с $n - 1$ степенью свободы.

При уровне значимости α решающее правило имеет вид

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } \Delta(a/2) \le c^2 \le \Delta(1-a/2), \\ H_1, \text{ если } c^2 < \Delta(a/2) \text{ или } c^2 > \Delta(1-a/2), \end{cases}$$

где $\Delta(\varepsilon)$ – граница критической области $\boldsymbol{W}$, представляющая собой квантиль уровня ε c^2-распределения с $n-1$ степенью свободы.

Если рассматривается правосторонняя альтернатива $s^2 > s_0^2$, то гипотеза H_0 (12.7) отклоняется при $c^2 > \Delta(1-\alpha)$, а для левосторонней альтернативы $s^2 < s_0^2$ гипотеза H_0 отклоняется при $c^2 < \Delta(\alpha)$. В этих случаях граница критической области $\boldsymbol{W}$ определится соответственно как $\Delta(1-\alpha)$ и $\Delta(\alpha)$ – квантили уровней $1-\alpha$ и α c^2-распределения с $n-1$ степенью свободы.

Пример 12.2. Стандартная процедура проверки коэффициента упругости образцов резины предписывает, что среднее квадратическое отклонение σ при измерении этого коэффициента составляет $\sigma_0 = 18{,}0$ единиц. Взята выборка объемом $n = 20$ и получено значение выборочного стандартного отклонения $s = 23{,}2$ единицы. Обосновано ли предположение о нестабильности стандартной процедуры проверки коэффициента упругости?

Решение. Проверяемые гипотезы запишутся следующим образом:

H_0: $s^2 = s_0^2 = 324$, H_1: $s^2 \neq s_0^2$.

Пусть уровень значимости $\alpha = 0{,}05$. Вычисленное значение статистики критерия c^2 (12.8) равно $19\cdot(23{,}2)^2/(18{,}0)^2 = 31{,}56$. Критические значения для двустороннего критерия равны $\chi^2_{0{,}025;\,19} = 8{,}907$ и $\chi^2_{0{,}975;19} = 32{,}85$, тогда $8{,}907 < c^2 < 32{,}85$. Следовательно, у нас нет оснований отклонить H_0. Это означает, что процедура проверки коэффициента упругости стабильна.

Критическое значение для левостороннего критерия (также при $\alpha = 0{,}05$) равно $c^2_{0{,}05;\,19} = 10{,}12$, а правостороннего – $c^2_{0{,}95;\,19} = 30{,}14$.

12.3. Гипотеза о значении параметра биномиального распределения

Пусть $x_1, x_2, \ldots, x_n$ – независимые одинаково распределенные случайные величины, имеющие биномиальное распределение с параметрами p и n:

$P\{n_p(n) = x\} = C_n^x\ p^x(1-p)^{n-x}$.

Необходимо проверить гипотезу (при уровне значимости α) о числовом значении p_0 параметра p биномиального распределения

H_0: $p = p_0$ против альтернативы H_1: $p \neq p_0$.

Критическое значение статистики при достаточно большом n вычисляется с применением нормальной аппроксимации по формуле

$$Z = \frac{m - np_0}{\sqrt{p_0(1-p_0)n}},$$

где m – частота появления интересующего нас значения случайной биномиальной величины в последовательности из n испытаний. **Решающее правило** использует статистику критерия Z и имеет вид

$$\text{принимается гипотеза } \begin{cases} H_0, \text{ если } |Z| \le \Delta(e), \\ H_1, \text{ если } |Z| > \Delta(e), \end{cases}$$

где $\Delta(\varepsilon)$ – квантиль уровня $\varepsilon = 1 - \alpha/2$ стандартного нормального распределения.

Если рассматривается правосторонняя альтернатива H_1: $p > p_0$, то гипотеза H_0 отклоняется при $Z > \Delta(1 - \alpha)$, а для левосторонней альтернативы H_1: $p < p_0$ гипотеза H_0 отклоняется при $Z < \Delta(\alpha)$.

Формулы для проверки рассмотренных в разд. 11 гипотез о числовых значениях параметров приведены в табл. П1.2.

Пример 12.3. При 100 бросаниях монеты 60 раз выпадал герб. Необходимо при уровне значимости $\alpha = 0{,}05$ проверить гипотезу о «правильности» монеты, т. е. H_0: $p = 0{,}5$ против альтернативы, что монета асимметрична, из-за чего чаще выпадает герб H_1: $p = p_1 > p_0$. Вычисляем значение статистики критерия:

$$Z = \frac{60 - 100 \cdot 0{,}5}{\sqrt{100 \cdot 0{,}5 \cdot 0{,}5}} = 2{,}00.$$

Подставляем вычисленное значение статистики критерия в функцию стандартного нормального распределения $\Phi(1{,}9) = 0{,}9772$. Решаем относительно P уравнение $1 - P/2 = \Phi(2{,}0)$. Получаем $1 - P/2 = 0{,}9772$, откуда $P = 0{,}0456$. Следовательно, нулевая гипотеза о симметрии монеты должна быть отклонена при уровне значимости $\alpha = 0{,}05$ ($P = 0{,}0456 < \alpha = 0{,}05$).

ЧАСТЬ III. ИНДУКТИВНЫЙ ВЫВОД В МАШИННОМ ОБУЧЕНИИ И РАСПОЗНАВАНИИ

13. МАШИННОЕ ОБУЧЕНИЕ

Ключевые понятия: индуктивное обучение, машинное обучение, обучающая выборка, добыча знаний, типы обучения, модели обучения, обучение на примерах, задача обучения «без учителя», задача обучения «с учителем»

13.1. Понятие обучения

Индуктивное формирование понятий, входящее в круг задач индуктивного вывода, является неотъемлемой частью машинного обучения и распознавания образов (классификации).

К задачам индуктивного формирования понятий прежде всего относятся задачи моделирования способности человека формировать описания, охватывающие множество примеров некоторого понятия. В основе процесса индуктивного формирования понятий лежит умение человека выделять наиболее общие характеристики среди описаний и пренебрегать второстепенными характеристиками, присущими конкретным примерам понятий.

Индуктивное обучение. Создание упрощенной модели окружающей реальности называется ***индуктивным обучением*** [2]. В процессе обучения человек наблюдает окружающую действительность и определяет существующие в ней взаимосвязи между объектами и событиями. Он группирует сходные объекты в классы и строит правила, предсказывающие поведение объектов того или иного класса.

Аналогично обучается и компьютер. Изучение и компьютерное моделирование процесса обучения является предметом исследования в области искусственного интеллекта, называемой ***машинным обучением*** (Machine Learning). Как правило, система машинного обучения пользуется не единичными наблюдениями, а совокупностью наблюдений. Такая совокупность называется ***обучающим множеством,*** или ***обучающей выборкой***.

Одной из разновидностей машинного обучения при поиске закономерностей является извлечение, ***добыча знаний*** из баз данных (БД). Дальнейшим развитием этого направления стала разработка средств обнаружения знаний, пред-

ставленных в базах данных неявным способом. Такие средства применяются в машинном обучении, распознавании образов, при извлечении знаний в экспертных системах и других задачах искусственного интеллекта.

Процесс обнаружения знаний в базах данных (Knowledge Discovery in Databases – *KDD*) включает несколько этапов: накопление первичных данных, преобразование данных, поиск закономерностей в данных, оценку, обобщение и структурирование найденных закономерностей.

Анализ данных с целью обнаружения знаний имеет следующие особенности: 1) данные имеют неограниченный объем; 2) данные являются разнородными (количественные, качественные, категориальные); 3) результаты должны быть легко интерпретируемыми и понятными; 4) средства анализа данных должны быть просты в использовании.

Основные аналитические инструменты, удовлетворяющие перечисленным требованиям, сегодня относят к области технологий добычи данных (Data Mining). В основу этих технологий положена концепция шаблонов (паттернов) и зависимостей, отражающих многообразные взаимосвязи в данных. Поиск паттернов производится автоматическими методами, не ограниченными априорными предположениями о структуре выборки и виде распределений значений анализируемых показателей.

Отыскиваемые паттерны должны отражать неочевидные регулярности в данных, составляющие так называемые скрытые знания.

В «добыче» знаний используются самые различные методы и подходы. Это методы проверки заранее сформулированных гипотез (verification – driven data mining) и разведочного анализа, составляющего основу оперативной аналитической обработки данных (on-line analytical processing – OLAP). Программные продукты, реализующие **нейросетевой подход**, также нередко относят к категории Data Mining.

В последнее время получили распространение системы рассуждений на основе аналогичных случаев (case based reasoning – CBR). Эти системы находят в прошлом близкие аналоги существующей ситуации и выбирают тот же ответ, который был для них правильным. В наибольшей мере требованиям Data Mining удовлетворяют методы поиска логических закономерностей в данных. Их результаты чаще всего выражаются в виде продукционных правил. С помощью таких правил решаются задачи прогнозирования, классификации, распознавания об-

разов, сегментации БД, извлечения из данных «скрытых» знаний, интерпретации данных, установления ассоциаций в БД и др. Логические методы работают в условиях разнородной информации.

Широко применяемым в настоящее время подходом к выявлению и изображению логических закономерностей в данных являются деревья решений. Наиболее известные процедуры CHAID (chi square automatic interaction detection), CART (classification and regression trees) и ID3 (Interactive Dichotomizer – интерактивный дихотомайзер) реализованы во многих пакетах, предназначенных для анализа данных.

Типы обучения. Обучение происходит в частично или полностью неизвестной среде. Целью обучения в любых условиях является нахождение некоторой функции. ***Обучение в частично неизвестной среде*** ведется на основе множества предъявляемых учителем примеров в виде пар (аргумент, значение). Такой вид обучения называется индуктивным обучением, а конкретные предположения о виде неизвестной функции – гипотезой. В зависимости от выбранного представления для нахождения этой неизвестной функции может использоваться математический аппарат булевых функций, логических формул, вероятностно-статистических методов, нейронных сетей, деревьев решений и др.

Существует и другой тип обучения, когда обучаемый не получает примеров (***обучение в полностью неизвестной среде***). Он начинает действовать самостоятельно и временами получает поощрения. Возникающие в таких условиях проблемы автоматического выдвижения гипотез реализуются в области коллективного поведения автоматов и адаптивного управления.

Накопленные в интеллектуальных системах знания (помимо примеров и подкреплений) позволяют улучшить способность к обучению. Эта особенность используется при обучении на основе объяснения примеров. Объяснение примеров – это метод извлечения общих правил (гипотез) из индуктивных наблюдений и некоторой общей теории. Этот способ обучения относится к накопительному (кумулятивному) виду обучения, когда накопление знаний улучшает способность к обучению – с учителем и без учителя (рис. 13.1).

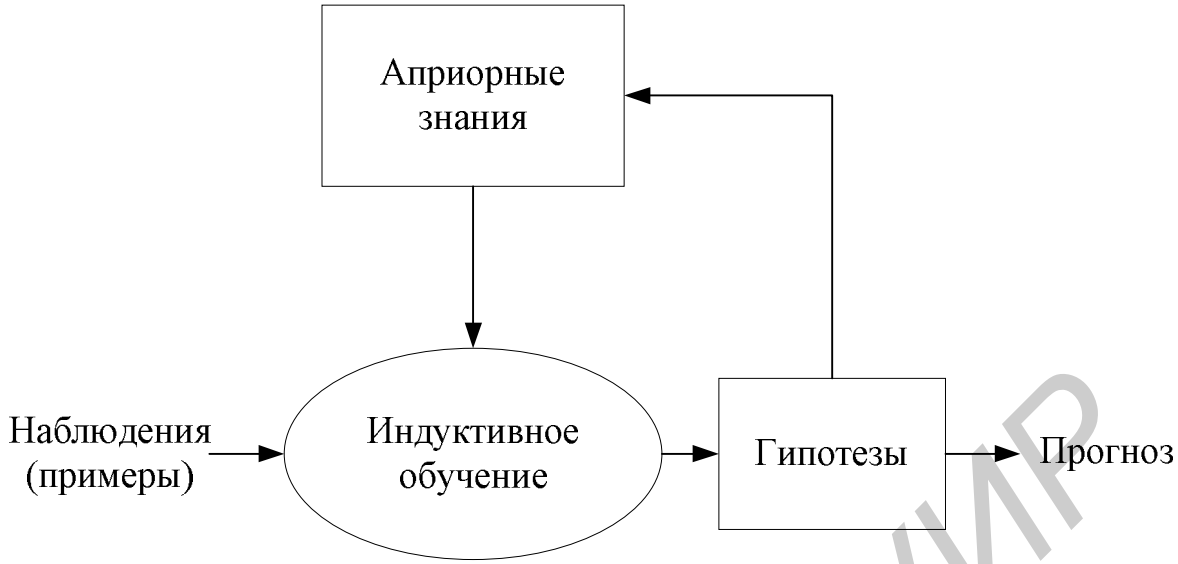


Рис. 13.1. Накопительный (кумулятивный) вид обучения

В пособии будут рассмотрены два метода обучения: ***обучение «с учителем» и обучение «без учителя»***. В данном разделе будет приведена общая характеристика указанных методов. Методу обучения «без учителя» посвящен разд. 14, а методу обучения «с учителем» – разд. 16.

13.2. Способы представления исходной информации

Источники информации. В процессе индуктивного обучения на основе предъявленных примеров необходимо сформировать общее понятие, охватывающее примеры и исключающее отрицательные примеры. Примеры для обучения можно получить из **трех источников** [2]:

от учителя – лица, которое знает содержание формируемого понятия и подбирает наиболее удачные обучающие выборки;

из внешней среды, с которой взаимодействует обучаемый объект; в этом случае обучающие выборки формируются случайным образом в зависимости от внешних факторов;

из самой интеллектуальной системы – действия самой системы приводят к созданию обучающей выборки, которая содержит сведения о сходных ситуациях с известными результатами; эту информацию можно затем обобщить.

Обучающие выборки должны содержать как положительные, так и отрицательные примеры формируемого понятия. Если в выборке будут представлены

только положительные примеры, то обучение может привести к построению чрезмерно общих индуктивных понятий. Такие понятия смогут охватить все предъявленные примеры, но не смогут различить два близких объекта, принадлежащих разным классам.

Признаковое описание. Все примеры понятия в системе представляются в виде ***признакового описания объекта***, т. е. каждый объект исследования (пример) задается набором свойств, или признаков. ***Признаки*** могут быть подразделены на ***детерминированные***, ***вероятностные***, ***логические*** и ***структурные***. Характеристика каждого из видов признаков дается в разд. 15 – 17.

13.3. Модели обучения

Неформальные модели. В психологии под обучением понимают способность к приобретению ранее неизвестных умений и навыков. Аналогичное неформальное понимание обучения имеет место и в интеллектуальных системах. Говорят, что интеллектуальная система обучилась чему-либо, если она стала способной выполнять некоторые процедуры или решать некоторые задачи, которые до этого была неспособна выполнять.

В интеллектуальных системах используются различные модели обучения. Наиболее исследованными являются модели, относящиеся к обучению на примерах.

Формальные модели. Обучение как математическая задача может быть отнесено к классу ***оптимизационных проблем поиска описаний*** [11].

Индивидуальная оптимизационная задача L есть пятерка

$< X_L, Y_L, r_L, F_L, J_L >$,

где X_L и Y_L – множества входных и выходных записей; r_L – отношение между X_L и Y_L (или функция ρ: $X_L \to Y_L$); F_L – множество отношений, называемых описаниями; J_L – мера качества для описания F_L, показывающая для каждого $f_L \in F_L$ степень его близости к ρ_L. Задача состоит в отыскании оптимального по J_L описания f_L^* из F_L.

Спецификация задачи часто оказывается неполной. Например, мера качества J может быть плохо формализуемой, информация об отношении ρ – задаваться только примерами пар $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, для которых x_i и y_i

связаны отношением $x_i r y_i$ и т. д. Если спецификация полная, то обучение не нужно, так как получается традиционная оптимизационная задача. **Для задач обучения характерна неполная спецификация**.

Известные методы решения задачи поиска описаний можно классифицировать по способу спецификации проблем, типу разрешенных алгоритмов, классу исследуемых проблем и т. д. Решение задач обучения характеризуется следующими подходами.

В теории статистических гипотез (разд. 7 – 8) рассматривается множество *M* реализаций некоторого случайного объекта с распределением вероятностей *p*(*x*) на *M*. Пусть W – произвольное подмножество *M*, {*H*} – некоторое множество гипотез фиксированного типа, связанных с вероятностью *p*(*x* I *W*) и характеризующих ее. Требуется на основании выборки (обучающей последовательности) из *M*, полученной в соответствии с *p*(*x*), выбрать наиболее подходящую гипотезу из {*H*}. В GUHA-методе (см. п. 3.2.2) объединяются методы статистической теории на стадии выдвижения гипотез с логическими методами на стадии обоснования и построения следствий из этих гипотез.

В теории параметрической адаптации предполагается, что множество *F* описаний, среди которых ищется оптимальное описание f_L^*, может быть охарактеризовано вектором параметров. Тогда выбор f_L^* сводится к поиску экстремума меры качества, задаваемому функционалом, в который входит плотность распределения *p*(*x*) случайного процесса. Если *p*(*x*) неизвестна и ее нельзя предварительно восстановить, то для решения задачи используются алгоритмы, называемые адаптивными, или обучающимися. Алгоритмы параметрического обучения находят применение в системах классификации, обучающихся системах, антенных и кодирующих устройствах и т. д.

Теория индуктивного вывода представляет собой дискретную математическую модель обучения по примерам. Множества X и Y – счетные, искомое описание ρ в общем случае задается с помощью последовательности троек вида $(x_i y_i a_i)$, таких, что a_i Î {0, 1} и $x_i r y_i$ тогда и только тогда, когда a_i = 1. Это означает, что тройки $(x_i y_i a_i)$ представляют примеры и контрпримеры ρ. В качестве *F* выбирается множество процедур, например формальные грамматики и т. д.

13.4. Обучение на примерах

Типы задач. Целью обучения на примерах является получение новых знаний (в виде общих закономерностей) из имеющихся примеров. Такой способ получения новой информации применяется при решении задач прогнозирования, идентификации (синтеза) функций, распознавания, индуктивного вывода.

Спецификация задач обучения на примерах. Для спецификации задач обучения на примерах необходимо определить следующие их характеристики.

1. Класс искомых описаний (обычно это множества и функции).
2. Пространство гипотез, т. е. множество формальных выражений, соответствующих возможным описаниям. Каждое из искомых описаний должно иметь в пространстве гипотез хотя бы одного своего представителя. Обратное неверно – пространство гипотез может представлять более широкий класс описаний, чем искомый.
3. Множество примеров для каждого описания, а также разрешенные последовательности этих примеров, называемые допустимыми представлениями этого описания.
4. Критерий успешности вывода, т. е. определение того, в каком случае гипотеза, на которой установился процесс решения задачи, считается приемлемой.

Например, для задачи расшифровки языков могут быть заданы следующие спецификации. Класс описаний – множество формальных языков в фиксированном алфавите; пространство гипотез – конкретный способ представления этих языков. Допустимым представлением языка является любая бесконечная последовательность слов из этого языка. Такое представление называется позитивным. Если вместе с примерами слов из языка в последовательности встречаются помеченные контрпримеры, то такое представление называется позитивно-негативным. Критерием успешности является точное соответствие найденного описания множеству позитивных примеров. Другими критериями могут быть совпадение с точностью до заданного числа слов и т. п.

13.5. Задача обучения «без учителя»

Известны два метода обучения: обучение «с учителем» и обучение «без учителя».

Метод обучения «без учителя» предполагает, что системе предъявляется совокупность объектов, не разделенных на классы. Количество классов также может быть неизвестным. Система должна сама определить классы объектов, основываясь на общих свойствах объектов из данного множества примеров. При этом задача ставится как разбиение исследуемой совокупности объектов на классы при известном либо неизвестном заранее числе классов.

Алгоритмы, решающие задачу обучения «без учителя», многообразны [17]. К ним относятся алгоритмы расщепления смесей вероятностных распределений; эвристические алгоритмы; алгоритмы, использующие понятие центра тяжести; алгоритмы иерархической классификации; алгоритмы автоматической классификации и т. д.

В разд. 14 будут рассмотрены два наиболее часто используемых метода обучения без учителя: 1) агломеративная процедура иерархической классификации объектов или признаков с графическим представлением результатов классификации; 2) метод k-средних [1, 17, 25].

13.6. Задача обучения «с учителем»

В ***обучении «с учителем»*** учитель определяет классы и предлагает примеры объектов каждого класса. Система должна найти общие свойства объектов каждого класса, сформировав тем самым ***описания класса***. Описание класса вместе с самим классом дает нам классифицирующее правило:

если <*описание*> то <*класс*>.

Это решающее правило для любого примера из обучающей выборки указывает, принадлежит этот пример понятию или нет. Построенное правило может быть использовано для предсказания класса объектов, неизвестных системе ранее. Решающее правило является корректным, если оно в дальнейшем успешно распознает объекты, не вошедшие в обучающую выборку.

В зависимости от того, какого рода информацией мы располагаем при решении задачи обучения «с учителем», используются определенные математические методы, реализованные в виде алгоритмов (см. разд. 16). Наиболее широко ис-

пользуются алгоритмы дискриминантного анализа [1, 6, 17], а также древообразные классификаторы [2, 17] и индукция решающих деревьев [2]. Краткая характеристика этих алгоритмов приведена в разд. 13, 15.

Алгоритмы классификации и распознавания, рассмотренные в разд. 16, 17, предназначены для числовой информации (количественных признаков). При исследовании объектов, для описания которых используются качественные признаки, близость или сходство объектов определяется на основании совпадения качественных значений. Это необходимо для нахождения правильных сочетаний значений некоторых признаков. Поиск наиболее существенных сочетаний признаков проводится с помощью аппарата логических функций. Алгоритмы, использующие такой подход, относятся к алгоритмам обучения «с учителем».

13.6.1. Задача обобщения понятий по признакам

Пусть имеется множество объектов, состоящее из положительных и отрицательных примеров формируемых понятий. Такое множество называется обучающей выборкой. На основании обучающей выборки необходимо построить понятие, разделяющее объекты, относящиеся к множествам положительных и отрицательных примеров.

Под обобщением понимается переход от рассмотрения единичного объекта o или некоторой совокупности объектов T к рассмотрению множества объектов V такого, что $o \in V$ или $T \in V$ [2].

Пусть $O = \{o_1, o_2, \ldots, o_n\}$ – множество объектов. Каждый объект o_i описывается p признаками $X_1, X_1, \ldots, X_p$ и может быть представлен множеством значений признаков $o_i = \{x_{kj}\}$, $k = 1, \ldots, p$, $j = 1, \ldots, n$.

В основе процесса обобщения лежит сравнение описаний исходных объектов, заданных совокупностью значений признаков, и выделение наиболее характерных фрагментов этих описаний. В зависимости от того, входит или не входит объект в объем некоторого понятия, назовем его *положительным* или *отрицательным* объектом для этого понятия.

Пусть O – множество всех объектов, представленных в некоторой системе знаний, V – множество положительных объектов и W – множество отрицательных объектов. Будем рассматривать случай, когда $O = V \cup W$, $V \cap W = \varnothing$, $W = \bigcup_i W_i$ и $W_i \cap W_j = \varnothing$ $(i \neq j)$. Пусть K – непустое множество объектов, такое, что

$K = K^+ \cup K^-$, где $K^+ \subset V$ и $K^- \subset W$. Будем называть K обучающей выборкой. На основании обучающей выборки надо построить правило, разделяющее положительные и отрицательные объекты.

Понятие считается сформированным, если удалось построить решающее правило, которое для любого примера из обучающей выборки указывает, принадлежит этот пример понятию или нет. Правило имеет вид «ЕСЛИ <условие> ТО <искомое понятие>». Условие представляется в виде логической функции, в которой булевы переменные, отражающие значения признаков, соединены логическими операциями конъюнкции, дизъюнкции и отрицания. Решающее правило является корректным, если оно в дальнейшем успешно распознает объекты, не вошедшие первоначально в обучающую выборку.

После того как распознающее правило построено, проводится проверка его качества. Для этого проводится разделение объектов новой, экзаменационной выборки на примеры и контрпримеры с помощью правила классификации.

Пример 13.1 [2]. В табл. 13.1 приведен пример обучающей выборки. Каждый объект имеет четыре признака: класс, рост, цвет волос и цвет глаз.

Таблица 13.1

Класс	Рост	Волосы	Глаза
–	Низкий	Светлые	Карие
–	Высокий	Темные	Карие
+	Высокий	Светлые	Голубые
–	Высокий	Темные	Голубые
–	Низкий	Темные	Голубые
–	Высокий	Светлые	Карие
+	Низкий	Светлые	Голубые

Рассмотрев в таблице примеры, можно установить следующую закономерность: все голубоглазые объекты со светлыми волосами относятся к классу «+»; все темноволосые либо светловолосые объекты, но с карими глазами относятся к классу «–». Признак «рост» не влияет на номер класса. Для задания решающего правила можно применить ряд способов: 1) использование логических функций; 2) продукционные правила; 3) деревья решений.

Использование логических функций. Две функции, задающие классы «+» и «–», могут иметь вид:

$P^+(X)$ = (*глаза* = *голубые*) & (*волосы* = *светлые*),

$P^-(X)$ = (*волосы* = *темные*) ∪ (*волосы* = *светлые*) & (*глаза* = *карие*).

Использование продукционных правил. Решающее правило может быть задано следующим набором продукционных правил:

если *волосы* = *светлые* & *глаза* = *голубые* то *класс* = «+»,

если *волосы* = *светлые* & *глаза* = *карие* то *класс* = «–»,

если *волосы* = *темные* то *класс* = «–».

Далее рассмотрим модель описания классов с помощью дерева решений.

На рис. 13.2 приведен пример представления решающего правила для обучающей выборки из табл. 13.1.

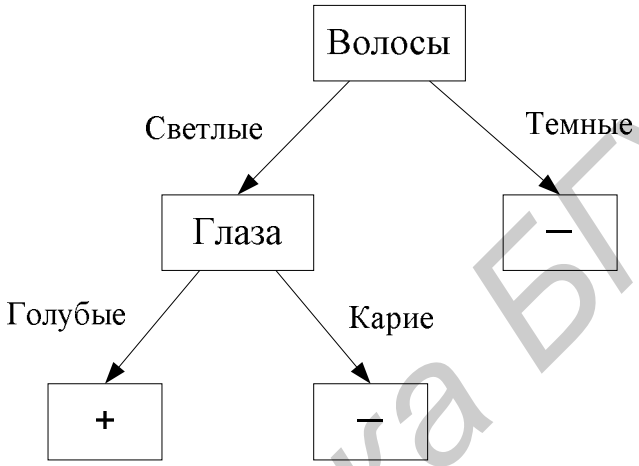


Рис. 13.2. Решающее правило в виде дерева решений

13.6.2. Алгоритм ДРЕВ

Данный алгоритм является методом качественного обобщения по признакам [2]. Необходимо построить обобщенное понятие на основе обучающей выборки, содержащей примеры K^+ и контрпримеры K^-. Для этого формируется логическая функция принадлежности к обобщенному понятию, используемая в дальнейшем как правило классификации. В этой логической функции булевы переменные, представляющие значения признаков, соединены операциями конъюнкции, дизъюнкции и отрицания.

На первом этапе работы алгоритма формируется обобщенное конъюнктивное понятие на основе поиска признаков, значения которых являются общими для всех объектов выборки K^+ и не встречаются среди контрпримеров K^-. В результате находится логическая функция $П_K$, значение которой равно 1 на всех примерах из выборки K^+ и равно нулю на всех контрпримерах K^-.

Затем формируется дизъюнктивное понятие $П_Д$. Его построение начинается с выбора среди объектов K^+ такого признака A_i, который является наиболее существенным для обобщенного понятия. Для выбранного признака ищется значение c′, которое называют разделяющим значением, так как на его основе происходит разбиение выборок K^+ и K^- на две пары подвыборок K_1^+ и K_1^-, K_{-1}^+ и K_{-1}^-. K_1^+ и K_1^- содержат примеры со значениями c′, а K_{-1}^+ и K_{-1}^- – примеры со значениями ¬ c′.

Наиболее существенное значение признака определяется с помощью одного из критериев потенциальной важности $Ф^k$, $k = 1, \ldots, q$. В качестве критерия может быть применен следующий:

$$Ф = \max_{i,\,j} \left(m_{ij} - \frac{1}{a_i}\right),$$

где m_{ij} – частота появления j-го значения i-го признака в примерах и контрпримерах, а a_i – число различных значений i-го признака в примерах и контрпримерах.

После разбиения выборок на подвыборки на основе c′ к каждой паре подвыборок применяется аналогичная процедура. В результате работы алгоритма формируется дерево решений, конечным вершинам которого либо сопоставлены подвыборки, для которых существует обобщенное конъюнктивное понятие, либо подвыборка обратилась в пустое множество.

14. ОБУЧЕНИЕ «БЕЗ УЧИТЕЛЯ»

Ключевые понятия: задача классификации, кластер, кластерный анализ, меры близости, алгоритмы классификации методом кластерного анализа

В этом разделе будут рассмотрены алгоритмы, решающие задачу обучения «без учителя». Эта задача ставится как задача классификации, т. е. задача разбиения исследуемой совокупности объектов на классы при известном либо неизвестном заранее числе классов, когда *отсутствуют обучающие выборки*. Установление естественных групп объектов, или классов основано на поиске максимального *сходства* (общности) между объектами одного класса и *различий* между объектами разных классов.

В зависимости от наличия априорных сведений о природе искомых классов и их характера, а также от конечных целей исследования используются различные методы [2, 17]. Далее будет описан метод кластерного анализа, целью которого является выделение естественных подмножеств объектов, или кластеров. Это достигается группировкой подобных (по определенным критериям) объектов.

Кластерный анализ применяется во многих областях деятельности: в медицине, биологии, экономике, маркетинге, психологии и т. д. В маркетинге, например, полезно построить и описать различные сегменты рынка на основании результатов изучения потенциальных потребителей. Страховые компании также может интересовать выделение различных классов потребителей, так как от этого зависят цены на предоставляемые услуги. Приведем примеры других приложений кластерного анализа: классификация компаний в соответствии с их типами организационной структуры, используемыми технологиями; определение психологических типов личности на основе опросников. Во всех рассмотренных примерах исходная совокупность объектов должна быть разбита на классы таким образом, чтобы объекты, принадлежащие одному классу, были в определенном смысле близкими, похожими друг на друга.

14.1. Общая постановка задачи распознавания в условиях отсутствия обучающих выборок

Дадим общую постановку задачи распознавания в условиях отсутствия обучающих выборок [1, 17, 25]. Наблюдается неклассифицированная случайная выборка $X = \{ X_j, j = 1, \ldots, n\}$, где X_j – p-мерное наблюдение из некоторого класса Ω_k, $k = 1, \ldots, m$. Номера классов k, из которых зарегистрированы наблюдения, неизвестны.

Информация о совокупности объектов, подлежащих классификации, может быть задана следующими способами:

1) ($n \times p$)-матрицей X, содержащей n наблюдений p признаков (переменных);

2) ($n \times n$)-матрицей D попарных расстояний (близостей) между объектами:

$$D = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \dots & \dots & \dots & \dots \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{pmatrix}.$$

Матрица D содержит меры сходства или меры различия n объектов. Если d_{ij}, $(i, j = 1, \dots, m)$ – расстояния, то они являются мерами различия. Чем больше расстояние, тем меньше сходство объектов. Если значения d_{ij} представляют собой меры сходства, то справедливо обратное утверждение: чем больше значения сходства, тем объекты более похожи, более близки. Расстояние и сходство являются двойственными характеристиками наблюдений в кластерном анализе. Если d_{ij} – расстояние, то в качестве меры близости, сходства можно использовать $d'_{ij} = \max_{i,j} d_{ij} - d_{ij}$.

Задача классификации заключается в том, чтобы всю анализируемую совокупность объектов $O = \{O_1, O_2, \dots, O_n\}$, статистически представимую в виде матриц X или D, разбить в соответствии с заданным критерием на сравнительно небольшое число однородных в определенном смысле групп или классов. Число классов может быть заранее как неизвестно, так и известно. Полученные в результате разбиения классы называют ***кластерами*** (образами), а методы их нахождения соответственно ***кластерным анализом, распознаванием образов «без учителя»***.

14.2. Меры близости

Понятие ***однородности объектов*** определяется заданием правила вычисления величины ρ_{ij}, характеризующей либо ***расстояние*** $d(O_i, O_j)$ между объектами O_i и O_j, либо ***степень близости*** (или ***сходства***) $r(O_i, O_j)$ тех же объектов. При выборе меры близости важную роль играет природа исследуемых данных, тип признаков, описывающих объекты. Для номинальных признаков целесообразно использовать меры сходства, а для количественных непрерывных признаков – расстояния. В алгоритмах кластерного анализа наиболее

часто рассматриваются следующие группы мер близости (сходства) [1, 17]: 1) ***коэффициенты корреляции***; 2) меры сходства*; 3) ***меры расстояния***.

Количественное оценивание близости (сходства) основано на понятии ***метрики***. При этом подходе к сходству объекты представляются точками координатного пространства, причем замеченные сходства и различия между точками находятся в соответствии с метрическими расстояниями между ними. Существуют четыре критерия, которым должна удовлетворять мера сходства, чтобы быть метрикой: 1) симметрия; 2) неравенство треугольника; 3) различимость нетождественных объектов; 4) неразличимость идентичных объектов.

14.2.1. Коэффициенты корреляции как меры близости

Коэффициенты корреляции r, часто называемые угловыми мерами ввиду их геометрической интерпретации, – самый распространенный тип сходства между признаками.

Свойства коэффициентов корреляции, формулы для их оценки по исходным выборочным данным, процедуры проверки гипотез относительно коэффициентов корреляции рассмотрены в подразд. 7.1 [1].

Недостатком коэффициента корреляции Пирсона является то, что он не представляет собой метрику: он не удовлетворяет третьему критерию, а во многих приложениях может не выполняться второй критерий (неравенство треугольника).

14.2.2. Меры близости между объектами, описываемыми бинарными переменными

Для установления сходства между объектами, описываемыми бинарными переменными, вводят так называемые коэффициенты ассоциативности. Рассмотрим эти коэффициенты, обратившись к 2×2-таблице, в которой 0 и 1 представляют собой значения бинарной переменной, a, b, c, d – число соответствующих пар значений двух переменных: (1, 1), (1, 0), (0, 1) и (0, 0).

* Факторный, дискриминантный и кластерный анализ / Дж.-О. Ким [и др.]. – М. : Финансы и статистика, 1989

Значения бинарных переменных	1	0
1	*A*	*B*
0	*C*	*d*

Существуют две наиболее часто используемые меры сходства: простой коэффициент совстречаемости и коэффициент Жаккара[*].

Простой коэффициент совстречаемости имеет вид

$S = (a + d)/(a + b + c + d)$,

где S – сходство между двумя объектами, которое изменяется в пределах от 0 до 1.

Коэффициент Жаккара определяется следующим образом:

$J = (a + d)/(a + b + c)$.

Подобно простому коэффициенту совстречаемости он изменяется от 0 до 1.

Пример 14.1. Определить меру сходства $n = 3$ объектов, каждый из которых описывается векторной бинарной переменной, имеющей $p = 8$ компонент x_i, $i = 1, \ldots, 8$.

Номер объекта	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
1	0	1	1	0	1	1	0	0
2	1	0	0	1	0	0	1	1
3	0	0	1	0	1	0	1	0

Решение. В качестве меры сходства используем коэффициент Жаккара. Матрица сходства размерностью (3×3) будет иметь вид

$$D = \begin{pmatrix} 1,0 & 0 & 0,333 \\ & 1,0 & 0,250 \\ & & 1,000 \end{pmatrix}.$$

[*] Факторный, дискриминантный и кластерный анализ / Дж.-О. Ким [и др.]. – М. : Финансы и статистика, 1989

14.2.3. Меры расстояния

Широкий спектр ***мер расстояния*** можно представить с помощью L_r-норм:

$$d(x_i, x_j) = \left(| \sum_{k=1}^{p} x_i^{(k)} - x_j^{(k)} |^r \right)^{1/r}. \qquad (14.1)$$

Ниже приведены наиболее известные расстояния.

1. **Евклидово расстояние** $d_1(x_i, x_j) = \sqrt{\sum_{k=1}^{p} (x_i^{(k)} - x_j^{(k)})^2}$.

2. **Расстояние Минковского** $d_2(x_i, x_j) = (\sum_{k=1}^{p} w_k^l \, | x_i^{(k)} - x_j^{(k)} |^l)^{1/l}$, где w_k – неотрицательные веса, $k = 1, \ldots, p$.

3. **Расстояние городских кварталов** (city block)

$$d_3(x_i, x_j) = \sum_{k=1}^{p} w_k \, | x_i^{(k)} - x_j^{(k)} |.$$

Для объектов, задаваемых дихотомическими (бинарными) признаками при $w_k = 1$, это расстояние является хеммингловым расстоянием.

Пример 14.2. Заданы три объекта ($n = 3$), описываемых двумя переменными: $X_1 = (0, 0)$, $X_2 = (1, 0)$ и $X_3 = (5, 5)$. Матрица расстояний для L_1-нормы (14.1) имеет вид

$$D_1 = \begin{pmatrix} 0 & 1 & 10 \\ 1 & 0 & 9 \\ 10 & 9 & 0 \end{pmatrix},$$

для квадратичной (евклидовой) L_2-нормы

$$D_2 = \begin{pmatrix} 0 & 1 & 50 \\ 1 & 0 & 41 \\ 50 & 41 & 0 \end{pmatrix}. \qquad (14.2)$$

Пример 14.3. Вычислить евклидово расстояние и расстояние городских кварталов между двумя объектами $x_1(x_1^{(1)}, x_1^{(2)})$ и $x_2(x_2^{(1)}, x_2^{(2)})$, используя метрику Минковского при $w = 1$. Пусть заданы разности $x_1^{(1)} - x_2^{(1)} = 4$ и $x_1^{(2)} - x_2^{(2)} = 2$

между соответствующими компонентами векторов признаков, описывающими эти объекты (рис. 14.1).

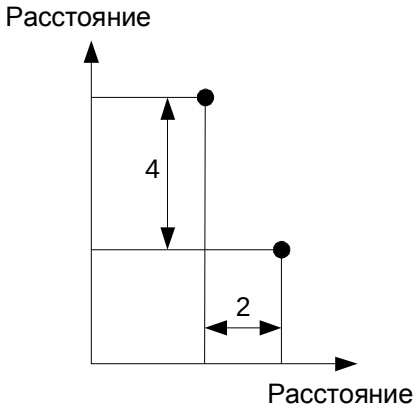


Рис. 14.1. Разности значений координат объектов
для вычисления расстояния

Решение.

Евклидово расстояние (λ = 2):

$((x_1^{(1)} - x_2^{(1)})^2 + (x_1^{(2)} - x_2^{(2)})^2)^{1/2} = (4^2 + 2^2)^{1/2} = 4,472.$

Расстояние городских кварталов (λ = 1):

$|x_1^{(1)} - x_2^{(1)}| + |x_1^{(2)} - x_2^{(2)}| = 4 + 2 = 6.$

Для характеристики взаимного расположения групп объектов наиболее часто используются следующие виды расстояний.

4. Расстояние между двумя группами S_i и S_j равно расстоянию между ближайшими объектами из этих групп (**расстояние «ближайшего соседа»**):

$$\rho_1(S_i, S_j) = \min \{d(x_k, x_m)\},\ x_k \in S_i,\ x_m \in S_j.$$

5. Расстояние между двумя группами S_i и S_j равно расстоянию между их математическими ожиданиями (**расстояние, измеряемое по «центрам тяжести» групп**):

$$\rho_2(S_i, S_j) = d(\overline{X}_i, \overline{X}_j),$$

где $\overline{X}_i$ – вектор математических ожиданий для i-й группы.

6. Расстояние между двумя группами S_i и S_j равно расстоянию между самыми дальними объектами этих групп (**расстояние «дальнего соседа»**):

$$\rho_3(S_i, S_j) = \max \{d(x_k, x_m)\},\ x_k \in S_i,\ x_m \in S_j.$$

7. Расстояние между двумя группами S_i и S_j равно среднему арифметическому всевозможных попарных расстояний между объектами рассматриваемых групп (**расстояние, измеряемое по принципу «средней связи»**):

$$\rho_4(S_i,\ S_j) = (n_i n_j)^{-1} \sum_{x_k} \sum_{x_m} d(x_k, x_m),\ x_k \in S_i,\ x_m \in S_j,$$

где n_i – число объектов в группе S_i.

8. Обобщенное (по Колмогорову) расстояние между классами [17], или обобщенное *K*-расстояние, вычисляется по формуле

$$\rho_\tau^K(S_l, S_m) = \left(\frac{1}{n_l n_m} \sum_{X_i \in S_l} \sum_{X_j \in S_m} d^t(X_i,\ X_j) \right)^{1/t}.$$

В частности, при $\tau \to \infty$ имеем расстояние «дальнего соседа», при $\tau \to -\infty$ – расстояние «ближайшего соседа». Очевидно также, что при $\tau = 1$ $r_1^K(S_l,\ S_m)$ является расстоянием, измеряемым по принципу «средней связи».

Понятие расстояния между группами объектов особенно важно в *агломеративных иерархических кластер-процедурах*, так как принцип работы таких алгоритмов состоит в последовательном объединении объектов и групп объектов, сначала самых близких, а потом все более отдаленных друг от друга.

В [17] предлагается следующая общая формула для вычисления расстояния между классом S_l и классом $S(m, q)$:

$$r_{l(m,\,q)} = \rho\,(S_l,\, S(m,\, q)) = \delta_1 \rho_{lm} + \delta_2 \rho_{lq} + \delta_3 \rho_{mq} + \delta_4\,|\rho_{lm} - \rho_{lq}|,$$

где δ_j – числовые коэффициенты, значения которых определяют агломеративный алгоритм (табл. 14.1). В этой таблице n_l – число объектов из класса S_l. Аналогичный смысл имеют n_m и n_q.

Правило объединения на основании единственной связи формирует кластеры, используя расстояние «ближайшего соседа», правило полной связи работает с расстоянием «дальнего соседа». Правило «средней связи» (невзвешенной и взвешенной) является компромиссом между двумя указанными алгоритмами. Центроидный метод подобен правилу средней связи и использует геометрическое расстояние между кластером S_l и взвешенным центром тяжести кластеров S_m и S_q.

Таблица 14.1

Правило объединения	δ_1	δ_2	δ_3	δ_4
Единственная связь	1/2	1/2	0	–1/2
Полная связь	1/2	1/2	0	1/2
Средняя связь (невзвешенная)	1/2	1/2	0	0
Средняя связь (взвешенная)	$\frac{n_m}{n_m + n_q}$	$\frac{n_q}{n_m + n_q}$	0	0
Центроидное	$\frac{n_m}{n_m + n_q}$	$\frac{n_m}{n_m + n_q}$	$-\frac{n_m n_q}{(n_m + n_q)^2}$	0
Медианное	1/2	1/2	–1/4	0
Уорда (Ward)	$\frac{n_l + n_m}{n_l + n_m + n_q}$	$\frac{n_m + n_q}{n_l + n_m + n_q}$	$-\frac{n_l}{n_l + n_m + n_q}$	0

В методе Уорда для определения присоединяемого к кластеру объекта вычисляется внутриклассовое рассеяние

$$I(S_l) = \frac{1}{n_l} \sum_{i=1}^{n_l} d^2(x_i - \bar{X}_l),$$

где $d^2(\cdot)$ – расстояние; $\bar{X}_l$ – центр тяжести (среднее значение) кластера S_l.

Если используется евклидово расстояние, то $I(S_l)$ представляет собой сумму квадратов внутриклассовых дисперсий p компонент. При объединении двух объектов m и q приращение рассеяния может быть определено как

$$\Delta(S_{m,q}) = \frac{n_m}{n_m + n_q} d^2_{m,q}.$$

К кластеру присоединяется тот объект или группа, которая дает наименьшее приращение внутригруппового рассеяния $\Delta(S_{m,q})$.

От выбора вида расстояния зависит окончательное разбиение объектов на классы при данном алгоритме разбиения. В каждом конкретном случае этот выбор должен производиться по-своему в зависимости от целей исследований, характера решаемой задачи, априорной информации и т. д. Существуют практические рекомендации по выбору расстояния, которые даются исходя из формы кластеров, образуемых при разбиении на основании определенного расстояния.

Особенностью кластеров, полученных с применением расстояния «ближайшего соседа», является наличие цепочки близких друг другу объектов, соединяющих две группы данных. Если кластеры на самом деле имеют выпуклую форму, то разбиение оказывается неудачным. Это расстояние рекомендуется использовать при сложной форме естественной группировки, так как «цепочный» эффект алгоритма «ближайшего соседа» обеспечивает верную классификацию.

Расстояние «дальнего соседа» ориентировано на поиск скоплений, близких к сферическим.

Расстояния, измеряемые по «центрам тяжести» групп и принципу «средней связи», используются при поиске групп не очень сложной, но и не сферической формы (гиперэллипсоидальная форма). Последнее расстояние может находить группы не выпуклой формы, как и расстояние «ближайшего соседа».

Проиллюстрируем эти рекомендации двумя рисунками. На рис. 14.2 и 14.3 представлена выборка, имеющая два кластера выпуклой формы. Если использовать для разбиения объектов алгоритм «ближайшего соседа», то выделятся кластеры, изображенные на рис. 14.2. Этот алгоритм оказывается непригодным, так как в результате образовалась цепочка близких объектов, соединяющих первый и второй кластер. На рис. 14.3 показан результат классификации методом «дальнего соседа». Этот результат соответствует первоначальным предположениям о форме кластеров.

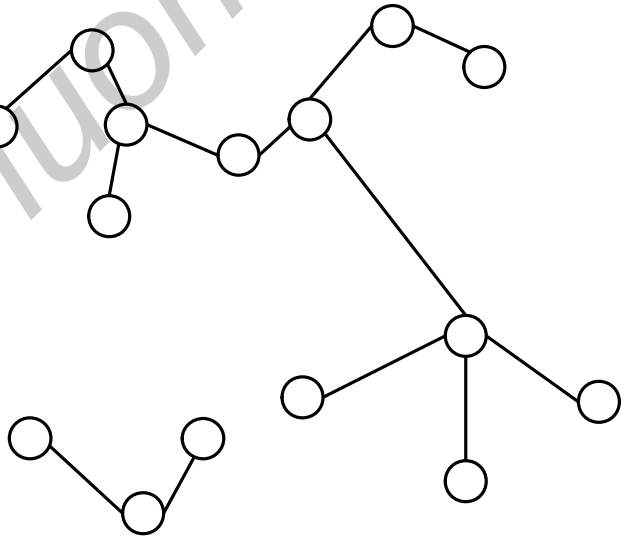


Рис. 14.2. Результат разбиения объектов методом «ближайшего соседа» для кластеров выпуклой формы

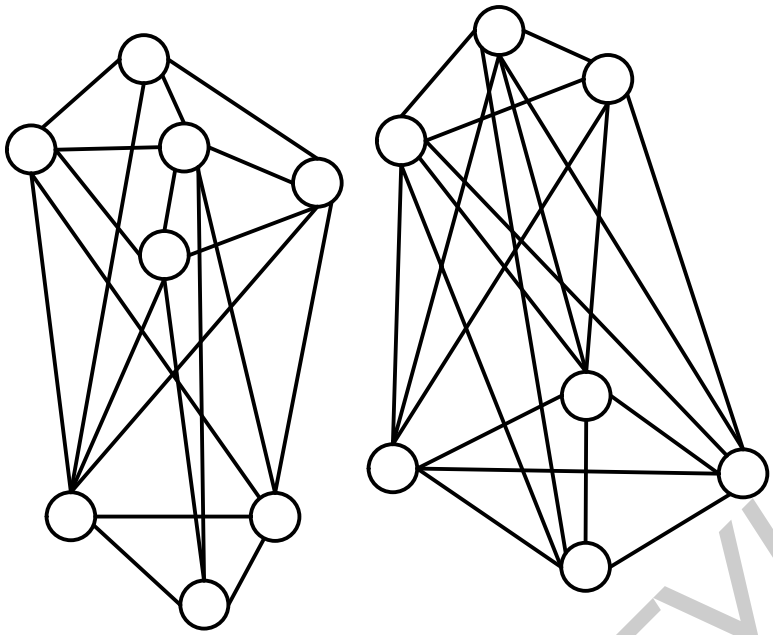


Рис. 14.3. Результат разбиения объектов методом «дальнего соседа» для кластеров выпуклой формы

На рис. 14.4 и 14.5 изображена противоположная ситуация: в выборке имеются кластеры более сложной формы. Применение расстояния «ближайшего соседа» позволило получить эти группы (см. рис. 14.4), в то время как метод «дальнего соседа» сформировал кластеры в форме шаровых скоплений (см. рис. 14.5).

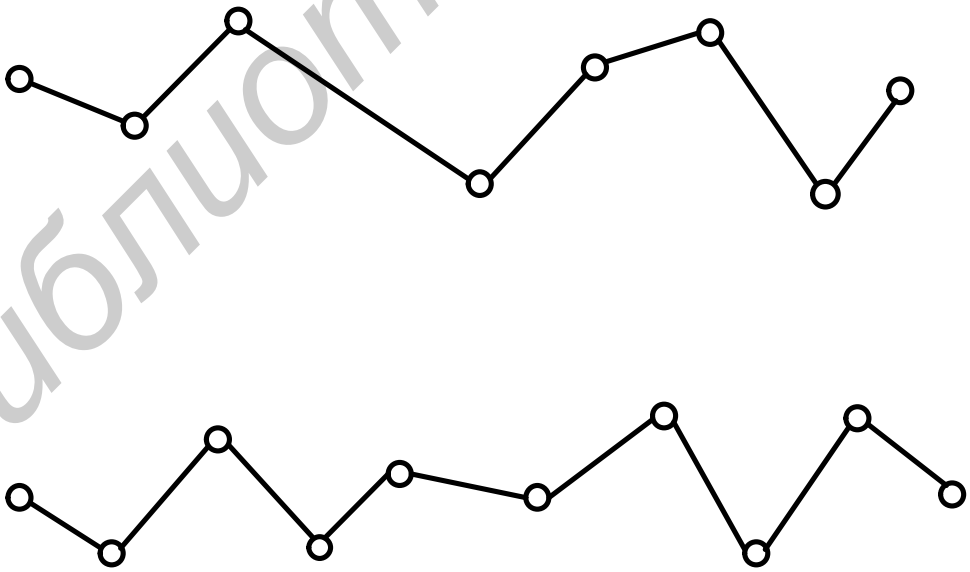


Рис. 14.4. Результат разбиения объектов методом «ближайшего соседа» для кластеров в виде цепочек

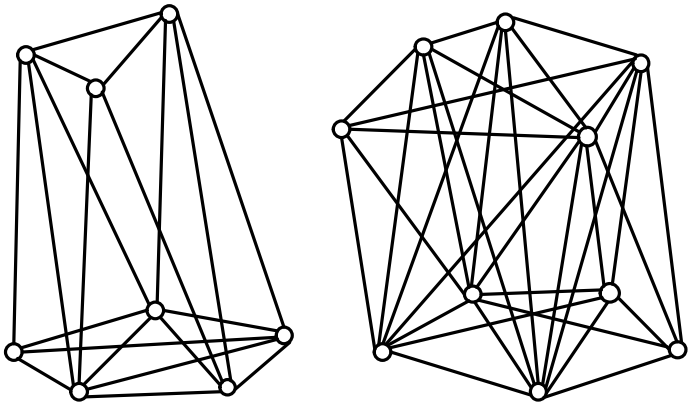


Рис. 14.5. Результат разбиения объектов методом «дальнего соседа» для кластеров в виде цепочек

Пример 14.4. Вычислить евклидово расстояние для правила объединения по принципу единственной связи. Данные представлены тремя объектами: $X_1 = (0, 0)$, $X_2 = (1, 0)$, $X_3 = (5, 5)$. Матрица исходных данных имеет вид

$$X = \begin{pmatrix} 0 & 0 \\ 1 & 0 \\ 5 & 5 \end{pmatrix}.$$

Вычисляем матрицу квадратов евклидовых расстояний

$$D = \begin{pmatrix} 0 & 1 & 50 \\ 1 & 0 & 41 \\ 50 & 41 & 0 \end{pmatrix}.$$

Агломеративный алгоритм начинает работу с $N = 3$ кластеров: $S_l = \{X_1\}$, $S(m) = \{X_2\}$, $S(q) = \{X_{23}\}$. Наименьшее расстояние существует между кластерами S_l и S_m, поэтому они объединяются, образуя новую группу $S(l, m) = \{X_1, X_2\}$. Вычислим расстояние между двумя оставшимися кластерами по правилу единственной связи:

$$r_{l(m,\,q)} = \rho(S_q,\ S(l, m)) = \frac{1}{2}\rho_{13} + \frac{1}{2}\rho_{23} - \frac{1}{2}|\rho_{13} - \rho_{23}| = \frac{50}{2} + \frac{41}{2} - \frac{1}{2}|50 - 41| = 41.$$

Новая матрица расстояний равна $D' = \begin{pmatrix} 0 & 41 \\ 41 & 0 \end{pmatrix}$.

14.3. Алгоритмы классификации методом кластерного анализа

Наиболее известные и хорошо зарекомендовавшие себя при решении прикладных задач алгоритмы кластерного анализа описаны в [17].

Далее будут рассмотрены два наиболее часто используемых метода кластеризации: 1) агломеративную процедуру иерархической классификации объектов или признаков с графическим представлением результатов классификации; 2) метод *k*-средних.

Агломеративный иерархический алгоритм кластер-анализа можно представить последовательностью следующих шагов. В начале процесса обучения каждый объект (признак) рассматривается как отдельный кластер. Далее:

1. Вычисляем матрицу расстояний *D*.
2. Находим два наиболее близких кластера.
3. Объединяем эти два кластера в один.
4. Вычисляем расстояния между новыми кластерами, формируя матрицу расстояний меньшей размерности.

Повторяем шаги 2 – 4 до тех пор, пока все кластеры не объединятся в один. Последовательность объединения можно представить графически в виде древовидной диаграммы, называемой дендрограммой (деревом классификации).

На рис. 14.6 представлены подлежащие классификации 8 объектов.

Используем для объединения правило единственной связи, а расстояние между кластерами будем определять по принципу «ближайшего соседа». Результат обучения в виде дерева классификации приведен на рис. 14.7. На этом рисунке можно выделить три кластера: в один из них включены объекты с номерами 6 – 8, в другой – объекты с номерами 3 – 5 и в третий – 1-й и 2-й объекты.

Метод *k*-средних [1] работает непосредственно с объектами, а не матрицей расстояний. В методе *k*-средних объект относится к тому классу, расстояние до которого минимально. В общем виде алгоритм работы метода *k*-средних состоит из следующих шагов.

1. Задается некоторое первоначальное разбиение на кластеры. Для этого устанавливается число классов K, и для итерации с номером $s = 0$ по данным неклассифицированной выборки вычисляются центры тяжести («средние» точки) кластеров $Z = \{ Z_1^s, \ Z_2^s, \ldots, Z_K^s \}$. Каждая такая точка является вектором, состоящим из p компонент.

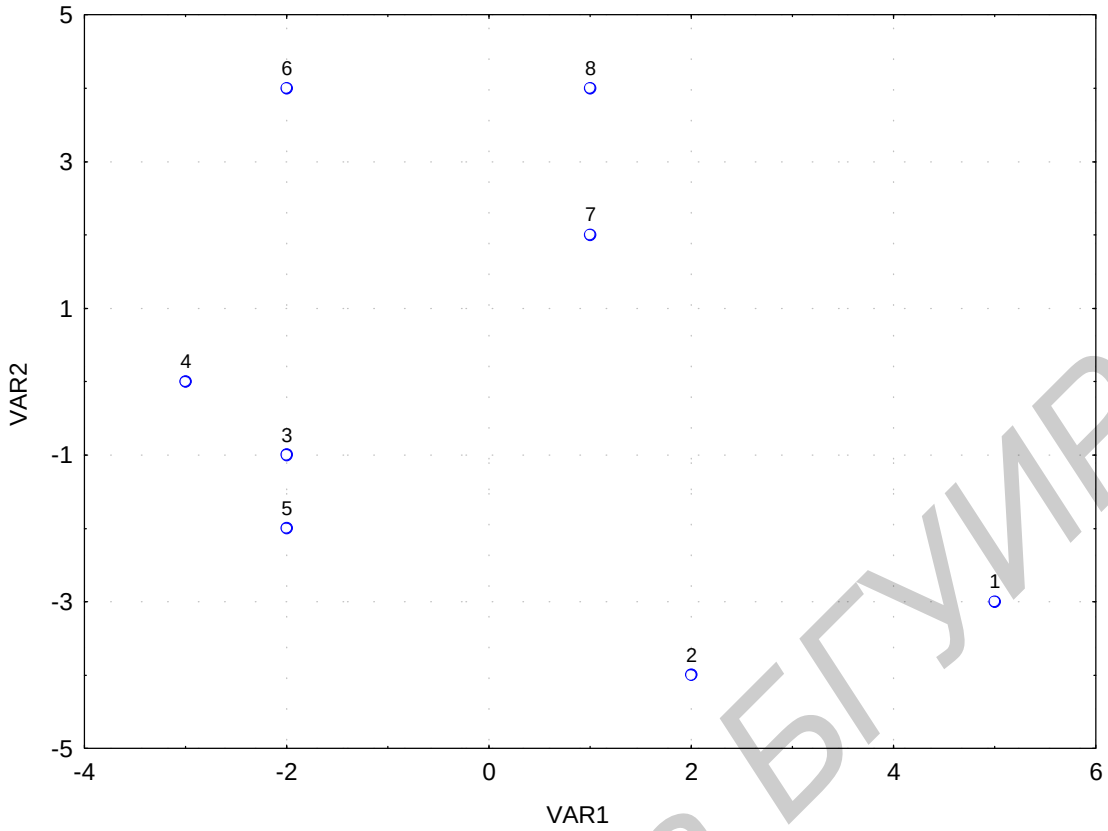


Рис. 14.6. Расположение подлежащих классификации объектов в виде точек на плоскости

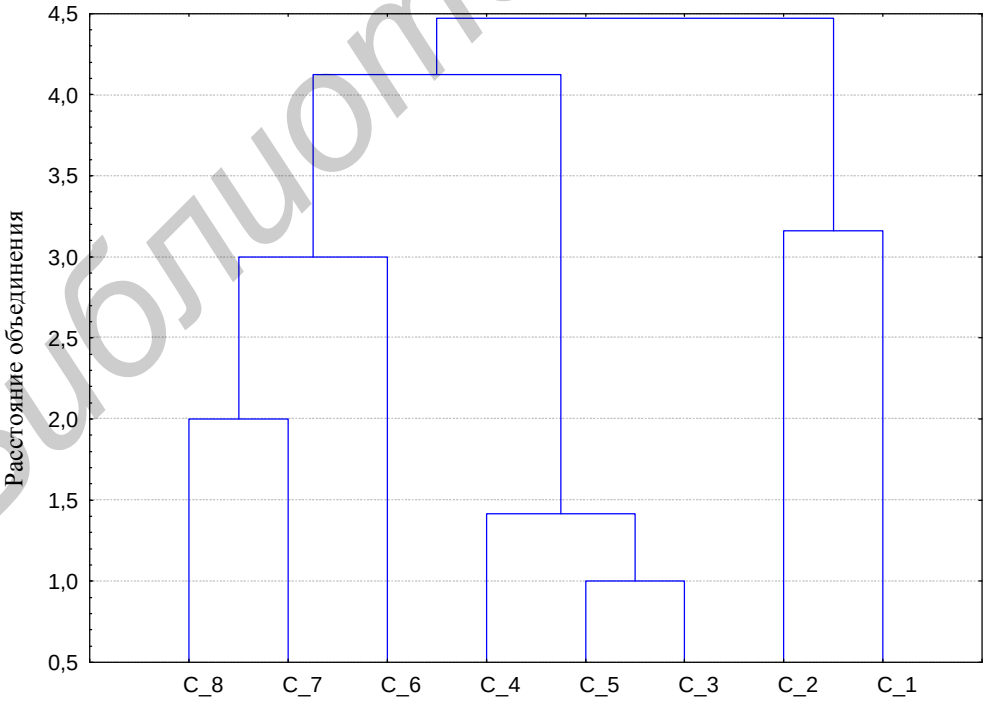


Рис. 14.7. Результаты обучения «без учителя» (иерархическая классификация 8 объектов по правилу единственной связи на основе евклидова расстояния)

2. Выполняется итерация $s = s + 1$.

3. Вычисляется матрица расстояний. Каждый объект, характеризуемый вектором признаков $X_j = (x_{1j}, x_{2j}, \ldots, x_{pj})$, $j = 1, 2, \ldots, n$, относится к одному из классов 1, 2, …, K по принципу ближайшего расстояния до центра класса.

4. Происходит перемещение точек: каждая точка перемещается в ближайший к ней кластер. Для каждого из сформированных классов определяются новые центры кластеров.

5. Шаги 2 – 4 повторяются, пока не будет найдена стабильная конфигурация (кластеры перестанут изменяться) или же число итераций не превысит заданное пользователем.

Метод *k*-средних формирует кластеры таким образом, чтобы минимизировать рассеяние внутри кластеров и максимизировать различия между ними.

15. ДЕРЕВЬЯ РЕШЕНИЙ

Ключевые понятия: дерево решений, структура дерева классификации, оценка качества разбиения, правила прекращения разбиения, построение дерева решений

15.1. Характеристики дерева решений

Дерево решений – это дерево, внутренние узлы которого представляют собой проверки для входных наблюдений (примеров, объектов) из обучающей выборки (обучающего множества), а вершины-листы являются категориями, классами (примеров, объектов). Пример дерева решений приведен на рис. 15.1.

Охарактеризуем данные, с которыми работает алгоритм построения дерева решений.

- **Описание признаков.** Вся информация об объектах (примерах) из предметной области должна описываться конечным набором признаков. Каждый признак должен иметь качественное или количественное (числовое) значение. Количество признаков должно быть фиксированным для всех примеров.

Рис. 15.1. Дерево решений

- **Принадлежность классу.** Каждый пример (объект) в обучающей выборке должен относиться к конкретному классу, т. е. один из признаков должен быть выбран в качестве имени или номера класса.
- **Объем классов.** Классы должны иметь конечное число примеров. Количество классов должно быть значительно меньше количества примеров.

Множество примеров может разбиваться на два или более подмножества. Если осуществляется разбиение на два подмножества, то мы получаем двоичное (бинарное) дерево решений. Дерево решений может классифицировать объекты в два или более классов.

Дерево решений – это способ представления правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение.

Под решающим правилом понимается логическая конструкция в виде продукции. Продукционные правила представляются в виде

если <посылка> **то** <заключение >.

В системах извлечения знаний в качестве посылки выступает описание объекта через его признаки, а заключением будет вывод о принадлежности объекта к определенному классу. В экспертных системах часто используются правила, в которых посылкой является описание ситуации, а заключением – действия, которые необходимо выполнить в данной ситуации.

Любое решающее дерево может быть преобразовано в набор продукционных правил: каждому пути от корня дерева до терминальной вершины соответствует одно продукционное правило. Его посылкой является конъюнкция условий «признак – значение», соответствующих пройденным вершинам и ребрам дерева, а заключением – имя или номер класса, соответствующего терминальной вершине.

- **Область применения деревьев решений.** Область применения деревьев решений в настоящее время широка, но все задачи, решаемые этим аппаратом, могут быть объединены в следующих три класса:
- *Описание данных*. Деревья решений позволяют хранить информацию о данных в компактной форме, вместо них мы можем хранить дерево решений, которое содержит точное описание объектов.
- *Классификация (распознававание).* Деревья решений отлично справляются с задачами классификации, т. е. отнесения объектов к одному из заранее известных классов. Целевая переменная должна иметь дискретные значения.
- *Регрессия*. Если целевая переменная имеет непрерывные значения, деревья решений позволяют установить зависимость целевой переменной от независимых (входных) переменных. Например, к этому классу относятся задачи численного прогнозирования (предсказания значений целевой переменной).

15.2. Структура дерева классификации

Рассмотрим структуру дерева классификации, предназначенного для принятия решения о типе урагана. Данные представляют собой значения двух координат: долготы (Longitude) и широты (Latitude), при которых 37 штормов достигли ураганной силы. Штормы классифицируются как ураганы двух типов, возникающие в Северной Атлантике: BARO (бароклинические) и TROP (тропические).

Ниже приводится дерево классификации (рис. 15.2).

Заглавие графа дает итоговую информацию о том, что дерево классификации имеет два разбиения и три терминальных узла. ***Терминальные узлы***, или, как их иногда называют, ***листья***, представляют собой узлы, в которых не принимается решение о дальнейшем разбиении. На графе терминальные узлы отмечаются пунктирной линией, в то время как остальные узлы решения отмечаются сплошной линией.

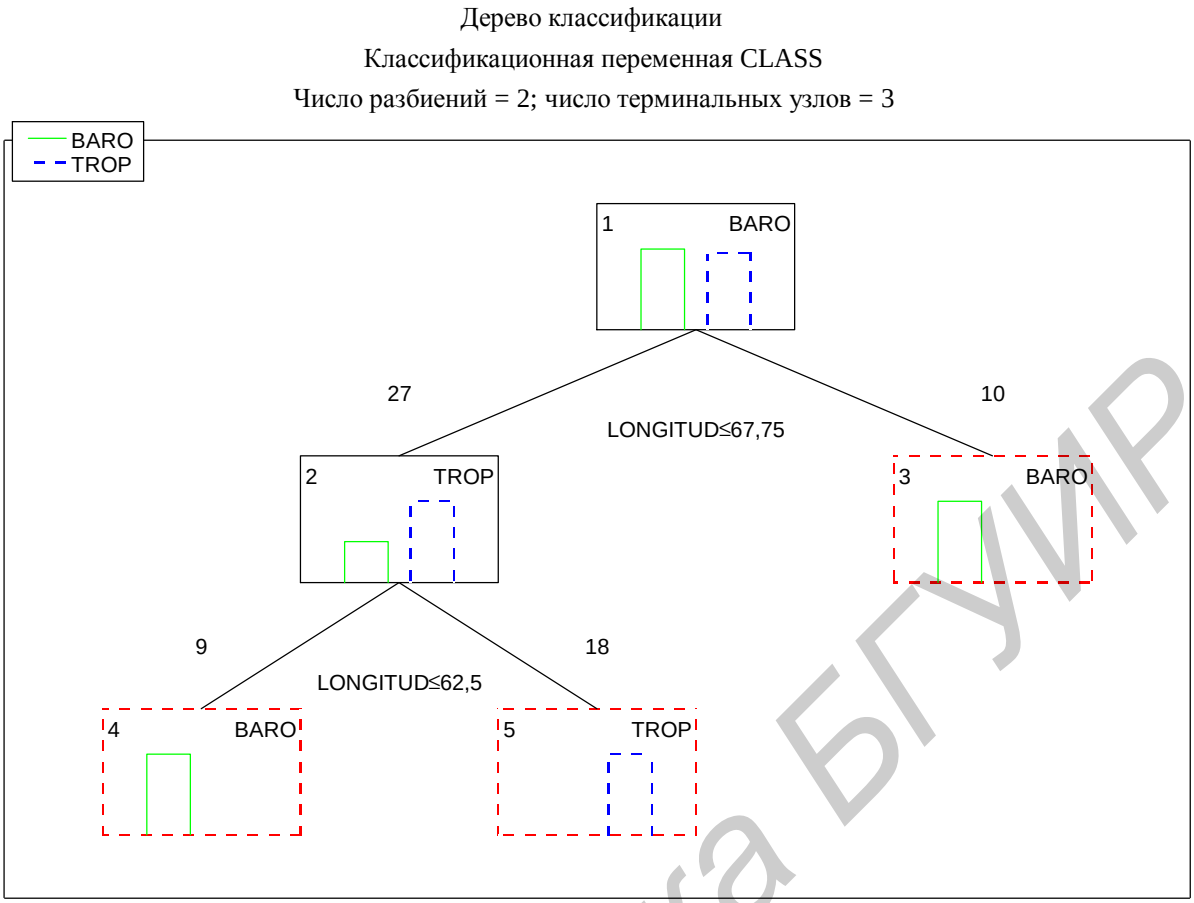


Рис. 15.2. Дерево классификации

Построение дерева начинается с узла, называемого ***корнем***. На графе он помечен цифрой 1. Первоначально все 37 ураганов приписываются корневому узлу и временно классифицируются как ураганы типа BARO (метка BARO в правом верхнем углу корневого узла). Тип BARO выбран в качестве начальной классификации, так как ураганов этого типа несколько больше, чем ураганов типа TROP. Это можно увидеть на гистограмме, расположенной внутри корневого узла. Легенда, идентифицирующая, какие столбцы гистограммы соответствуют ураганам типов BARO и TROP, расположена в верхнем левом углу графа.

Корневой узел разбивается, образуя два новых узла. Текст, находящийся ниже корневого узла, описывает условие разделения. В нем указано, что ураганы со значением координаты долгота (Longitude), меньшим или равным 67,75, относятся к узлу 2 и предварительно классифицируются как ураганы типа TROP, а ураганы со значением координаты долгота (Longitude), большим 67,75, приписываются к узлу 3 и классифицируются как ураганы типа BARO. Значения 27 и 10, размещенные над узлами 2 и 3 соответственно, указывают число наблюдений (объектов), отнесенных к каждому из этих двух дочерних узлов из

их родительского (корневого) узла. Аналогичным образом производится разбиение в узле 2. В результате 9 ураганов со значением координаты долгота (Longitude), меньшим или равным 62,5, относятся к узлу 4 и классифицируются как ураганы типа BARO, а остальные 18 ураганов со значением координаты долгота (Longitude), большим 62,5, приписываются к узлу 4 и классифицируются как ураганы типа TROP.

В каждом узле строится гистограмма распределения объектов по классам.

Информация, связанная с построением дерева, представлена в табл. 15.1.

Таблица 15.1

Структура дерева классификации

Дерево	Дочерние узлы, число наблюдений в классе (*n*), предсказываемый класс, условие разбиения в узле						
Узел	Левая ветвь	Правая ветвь	*n* в классе BARO	*n* в классе TROP	Предсказанный класс	Константа разбиения	Разделяющая переменная
1	2	3	19	18	BARO	– 67,75	LONGITUDE
2	4	5	9	18	TROP	– 62,50	LONGITUDE
3			10	0	BARO		
4			9	0	BARO		
5			0	18	TROP		

Отметим, что в приведенной таблице узлы с номерами 3 и 5 идентифицируются как терминальные, так как в них не выполняется разбиение. Следует отметить также знак константы разбиения, например – 67,75 для разбиения в узле 1. На древовидном графе условие разбиения записывается как LONGITUDE $\leq$ 67,75, а не в эквивалентной форме – 67,75 + LONGITUDE ≤ 0. Это делается для экономии места на графике.

Решающее правило, представленное деревом классификации на рис. 15.2, формулируется следующим образом:

> Если значение координаты LONGITUDE $\geq$ 67,75 или значение координаты LONGITUDE $\leq$ 62,5, то возникший ураган относится к классу BARO, иначе ураган относится к классу TROP.

15.3. Вычислительные задачи древообразных классификаторов

В процессе построения дерева классификации необходимо решить следующие четыре основные задачи:

1) определение качества предсказания;
2) выбор разбиений;
3) определение правила прекращения разбиения;
4) нахождение дерева «правильного размера».

15.3.1. Определение качества предсказания

Пусть c_j – штраф за ошибочную классификацию, когда верна гипотеза H_j (j = 1, 2) о том, что объект принадлежит классу с номером j. Функция потерь Q определяется следующим образом:

$$Q = c_1 \pi_1 \alpha + c_2 \pi_2 \beta,$$

где π_i – априорная вероятность принадлежности i-му классу; α и β – вероятности ошибок первого и второго рода соответственно.

Вероятность δ ошибочной классификации может быть записана как

$$\delta = \mathrm{E}\,(y_k - \hat{y}\,(x_k))^2 / n.$$

Здесь E(·) – математическое ожидание; $y = j$, когда верна гипотеза j = 1, 2; $\hat{y}\ (X) = j$, когда принимаем гипотезу H_j.

В алгоритме разбиения CART оценки качества разбиения используется индекс *Gini*, являющийся показателем неопределенности в узле. Если набор данных T содержит данные n классов, тогда индекс *Gini* определяется как

$$Gini\ (T) = 1 - \sum_{i=1}^{n} p_i^2,$$

где p_i – вероятность (относительная частота) класса i в T.

Если набор T разбивается на две части T_1 и T_2 с числом объектов в каждом N_1 и N_2 соответственно, тогда показатель качества разбиения будет равен

$$Gini_{split}\ (T) = \frac{N_1}{N}\ Gini\ (T_1) + \frac{N_2}{N}\ Gini\ (T_2).$$

Наилучшим считается то разбиение, для которого $Gini_{split}(T)$ минимально.

Обозначим N – число объектов в узле-предке, L, R – число объектов в левом и правом потомке соответственно, l_i и r_i – число экземпляров i-го клас-

са в левом/правом потомке. Тогда качество разбиения оценивается по следующей формуле:

$$\tilde{G}_{split} = \frac{1}{L}\sum_{i=1}^{n} l_i^2 + \frac{1}{R}\sum_{i=1}^{n} r_i^2 \to \max.$$

В итоге, лучшим будет то разбиение, для которого величина $\tilde{G}_{split}$ максимальна.

15.3.2. Выбор разбиений

Следующим шагом в построении дерева классификации является выбор предикторных переменных, которые используются для разбиения объектов в узле на два множества объектов, соответствующих левой и правой ветви (левому и правому дочерним узлам).

Разбиение начинается с корневого узла, далее образуются дочерние узлы до тех пор, пока не будет завершен процесс разбиения. Вектор предикторных (предсказывающих) переменных, подаваемый на вход дерева, может содержать как числовые (порядковые), так и категоризованные номинальные переменные. В каждом узле разбиение может производиться двумя способами: по одной переменной или по значению линейной функции предсказывающих переменных.

Если переменная числового типа, то в узле формируется правило вида $x_i \le c$, где c – некоторый порог, который чаще всего выбирается как среднее арифметическое двух соседних упорядоченных значений переменной x_i обучающей выборки. Если переменная категориального типа, то в узле формируется правило $x_i \in V(x_i)$, где $V(x_i)$ – некоторое непустое подмножество множества значений переменной x_i в обучающей выборке. Следовательно, для n значений числовой переменной необходимо сравнить $n-1$ разбиение, а для категориального $(2^{n-1}-1)$. На каждом шаге построения дерева последовательно сравниваются все возможные разбиения для всех переменных и выбирается переменная, обеспечивающая наилучшее разбиение.

Для каждого узла вычисляются P-значения для проверки значимости связи номера класса с уровнями каждой предикторной переменной. Для категоризованных предикторов P-значения вычисляются с помощью χ^2-теста независимости классов и уровней категоризованной переменной в рассматриваемом узле.

Для порядковых предикторов *P*-значения вычисляются с помощью однофакторного дисперсионного анализа. В этом случае устанавливается зависимость между номером класса и значениями порядкового предиктора в рассматриваемом узле. Если наименьшее вычисленное *P*-значение меньше, чем заданный по умолчанию уровень значимости 0,05 для множественного сравнения Бонферрони, или меньше пороговых значений, заданных пользователем, то в качестве переменной для разбиения выбирается переменная с наименьшим *P*-значением. Если *P*-значения меньше порогового не найдены, то проверяется гипотеза о равенстве дисперсий с помощью робастного критерия Левена.

15.3.3. Определение правила прекращения разбиения

Используются три способа прекращения разбиения.

1. **«Чистая» классификация.** В древовидных классификаторах не устанавливаются пределы для числа разбиений. При «чистой» классификации разбиение прекращается, когда в каждом терминальном узле будут содержаться объекты только из одного класса. Такой способ на практике применяется редко, так как переменные, характеризующие реальные объекты, могут быть измерены с ошибками или «зашумлены».

2. **Задание минимального числа объектов *n***. Разбиение прекращается, когда в каждом терминальном узле будут содержаться объекты только из одного класса или число объектов из других классов не будет превышать заданного значения *n*.

3. **Задание доли объектов.** Разбиение прекращается, когда в каждом терминальном узле будут содержаться объекты только из одного класса или число объектов из других классов не будет превышать заданной доли от объема класса.

15.3.4. Нахождение дерева «правильного размера»

Размер дерева определяется числом его узлов. Если дерево очень большое, записать решающее правило оказывается чрезвычайно сложно, оно становится громоздким и теряет выразительность и компактность. Поэтому возникает необходимость использовать для принятия решения дерево не слишком сложное, но и не слишком проигрывающее в точности классификации. Для выбора дерева «правильного размера» из всех возможных можно использовать две стратегии.

Первая стратегия состоит в формировании дерева, размер которого определяется пользователем на основании знаний предыдущих исследований, ранее

полученной диагностической информации, опыта, интуиции и т. д. Другая стратегия заключается в использовании специальных автоматических процедур отсечения дерева.

В первой стратегии размер формируемого дерева определяется пользователем с помощью правила прекращения разбиений. Для этой цели задается доля объектов в терминальном узле, которая позволит дереву расти до нужных размеров.

Вторая стратегия строит дерево на основе достижения компромисса между размером дерева (сложностью) и потерями от ошибочной классификации (стоимостью ошибки).

Такое дерево строится путем отсечения поддеревьев. Основная проблема отсечения – большое количество всех возможных отсеченных поддеревьев для одного дерева. Если бинарное дерево имеет T листьев, тогда существует $\sim 1{,}5028369^{T}$ отсечённых поддеревьев. И если дерево имеет хотя бы 1000 листьев, тогда число отсечённых поддеревьев становится просто огромным.

Базовая идея метода – не рассматривать все возможные поддеревья, ограничившись только лучшими представителями согласно приведенному ниже методу.

Обозначим T – число листьев дерева, $R(T)$ – ошибка классификации дерева, равная отношению числа неправильно классифицированных объектов к числу объектов в обучающей выборке. Определим $C_\alpha(T)$ – полную стоимость (оценку показателя затраты – сложность) дерева как

$$C_\alpha(T) = R(T) + \alpha \cdot T,$$

где T – число листьев (терминальных узлов) дерева; α – некоторый параметр, изменяющийся от 0 до $+\infty$. Полная стоимость дерева состоит из двух компонент – ошибки классификации дерева и штрафа за его сложность. Если ошибка классификации дерева неизменна, тогда с увеличением α полная стоимость дерева будет увеличиваться. Тогда в зависимости от α менее ветвистое дерево, дающее бóльшую ошибку классификации, может стоить меньше, чем дающее меньшую ошибку, но более ветвистое.

Определим $T_{\max}$ – максимальное по размеру дерево, которое предстоит обрезать. Если мы зафиксируем значение α, тогда существует наименьшее минимизируемое поддерево α, которое выполняет следующие условия:

$$C_a(T(a)) = \min_{T \le T_{\max}} C_a(T),$$

$$\textit{if } C_\alpha(T) = C_\alpha(T(\alpha)) \textit{ then } T(\alpha) \le T.$$

Первое условие говорит, что не существует такого поддерева дерева T_{max}, которое имело бы меньшую стоимость, чем $T(\alpha)$ при этом значении α. Второе условие говорит, что если существует более одного поддерева, имеющего данную полную стоимость, тогда мы выбираем наименьшее дерево.

Алгоритм вычисления T_1 из T_{max} прост. Нужно найти любую пару листьев с общим предком, которые могут быть объединены, т. е. отсечены в родительский узел без увеличения ошибки классификации. $R(t) = R(l) + R(r)$, где r и l – листья узла t. Продолжать поиск, пока таких пар листьев больше не останется. Так мы получим дерево, имеющее такую же стоимость, как T_{max} при $\alpha = 0$, но менее ветвистое, чем T_{max}.

15.4. Построение дерева решений

Идею построения деревьев решений на основе примеров рассмотрим по Р. Куинлану. Им разработан известный алгоритм ID3 (Induction of Decision trees) [2]. Пусть задано некоторое обучающее множество T, содержащее объекты (примеры), каждый из которых характеризуется n атрибутами (признаками), причем один из них указывает на принадлежность объекта к определенному классу. Назовем признаки, которые задают свойства каждого примера обучающей выборки, предсказывающими (предикторными) атрибутами. Такие признаки могут быть бинарными, количественными или качественными. Признак, который для каждого примера задает принадлежность к формируемому понятию, называется предсказываемым. Этот признак также входит в обучающую выборку.

Пусть через $i = 1, \ldots, m$ обозначены классы. Тогда существуют три ситуации:

1) множество T содержит один или более примеров, относящихся к одному классу C_k. Тогда дерево решений для T – это лист (терминальный узел), определяющий класс C_k;

2) множество T не содержит ни одного примера, т. е. это пустое множество. Тогда это снова лист, и класс, ассоциированный с листом, выбирается из другого множества, отличного от T, скажем, из множества, ассоциированного с родителем;

3) множество T содержит примеры, относящиеся к разным классам. В этом случае следует разбить множество T на некоторые подмножества. Для этого выбирается один из признаков, имеющий два и более отличных друг от друга значения $O_1, O_2, \ldots, O_s$. Множество T разбивается на подмножества $T_1, T_2, \ldots, T_s$,

где каждое подмножество T_i содержит все примеры, имеющие значение O_i для выбранного признака. Эта процедура будет рекурсивно продолжаться до тех пор, пока конечное множество не будет состоять из примеров, относящихся к одному и тому же классу.

Вышеописанная процедура лежит в основе многих современных алгоритмов построения деревьев решений. Очевидно, что при использовании данной методики построение дерева решений будет происходить сверху вниз.

Для построения дерева на каждом внутреннем узле необходимо найти такое условие (проверку), которое бы разбивало множество, ассоциированное с этим узлом, на подмножества. В условие должен быть включен один из атрибутов (признаков). Общее правило для выбора атрибута можно сформулировать следующим образом: выбранный атрибут должен разбить множество так, чтобы получаемые в итоге подмножества состояли из объектов, принадлежащих к одному классу, или количество объектов из других классов в каждом из этих подмножеств было бы как можно меньше. Такой атрибут считается наиболее информативным среди всех атрибутов, еще не рассмотренных на пути от корня дерева. В качестве меры информативности обычно используется теоретико-информационное понятие энтропии. Возможны и другие критерии. Например, при построении древовидных классификаторов применяются статистические критерии, на основании которых производится выбор признака для разбиения множества объектов в узле.

Рассмотрим информационный критерий выбора [2]. Если имеется n равновероятных значений признака, то вероятность p каждого из них равна $1/n$, и информация, связанная со значением признака, равна $-\log p = \log n$ (log обозначает логарифм по основанию два). В общем случае, если мы имеем дискретное распределение

$$P = (p_1, p_2, \ldots, p_n), \tag{15.1}$$

то передаваемая информация вычисляется по формуле

$$I(P) = -\sum_{i=1}^{n} p_i \log p_i . \tag{15.2}$$

Значение $I(P)$ дает оценку среднего количества информации, необходимого для определения класса примера из множества S.

Чем ближе распределение к равномерному, тем больше его энтропия.

Если множество S примеров (объектов) разбито на попарно непересекающиеся классы $C_1, C_2, \ldots, C_k$, то информация, необходимая для того, чтобы уста-

новить класс примера, равна $Info(S) = I(P)$, где P – дискретное распределение вероятностей появления соответствующего примера при условии его принадлежности к классу $C_1, C_2, \ldots, C_k$. Каждая из оценок вероятностей p_i (15.1) того, что случайно выбранный пример из множества S будет принадлежать к классу C_i, вычисляется как

$$p_i = \frac{|C_i|}{|S|},$$

где $|C_i|$, $|S|$ – мощности как отдельных классов, так и всей обучающей выборки соответственно.

Разбив множество примеров на основе значений некоторого признака X на подмножества $S_1, S_2, \ldots, S_n$, мы можем вычислить $Info\ (S)$ как взвешенное среднее информации, необходимой для установления принадлежности примера к определенному классу в каждом подмножестве:

$$Info\ (X, S) = \sum_{i=1}^{n} \frac{|S_i|}{|S|} Info(S_i). \qquad (15.3)$$

Величина

$$Gain\ (X, S) = Info\ (S) - Info\ (X, S) \qquad (15.4)$$

показывает количество информации, которое мы получаем благодаря признаку X. Алгоритм ID3 использует эту величину для оценки информативности признака при построении решающих деревьев, это позволяет получать деревья минимального размера. Критерий (15.4) вычисляется для всех признаков. Выбирается признак, максимизирующий данное выражение. Этот признак будет являться условием разбиения в текущем узле дерева.

Если в процессе работы алгоритма получен узел, ассоциированный с пустым множеством (т. е. ни один пример не попал в данный узел), то он помечается как лист (терминальный узел), и в качестве решения для листа выбирается наиболее часто встречающийся класс у непосредственного предка данного листа.

Поясним, почему критерий (15.4) должен максимизироваться. Из свойств энтропии известно, что максимально возможное значение энтропии достигается в том случае, когда все сообщения равновероятны. В нашем случае, энтропия (15.3) достигает своего максимума, когда частота появления классов в примерах множества S равновероятна. Нам же необходимо выбрать такой признак, чтобы при разбиении

по нему один из классов имел наибольшую вероятность появления. Это возможно в том случае, когда энтропия (15.3) будет иметь минимальное значение и, соответственно, критерий (15.4) достигнет своего максимума.

Если признаки являются количественными (числовыми), то следует выбрать некий порог, с которым должны сравниваться все значения признака. Пусть количественный признак имеет конечное число значений. Обозначим их $\{v_1, v_2, \ldots, v_n\}$. Предварительно отсортируем все значения. Тогда любое значение, лежащее между v_i и v_{i+1}, делит все примеры на два множества: те, которые лежат слева от этого значения $\{v_1, v_2, \ldots, v_i\}$, и те, что справа $\{v_{i+1}, v_{i+2}, \ldots, v_n\}$. В качестве порога можно выбрать среднее между значениями v_i и v_{i+1}:

$$TH_i = \frac{v_i + v_{i+1}}{2}.$$

Таким образом, мы существенно упростили задачу нахождения порога и привели к рассмотрению всего $n-1$ потенциальных пороговых значений $TH_1, TH_2, \ldots, TH_{n-1}$. Формулы (15.2) – (15.4) последовательно применяются ко всем потенциальным пороговым значениям. Затем среди них выбирается то, которое дает максимальное значение по критерию (15.4). Далее это значение сравнивается со значениями критерия (15.4), подсчитанными для остальных признаков. Если выяснится, что среди всех признаков данный числовой признак имеет максимальное значение по критерию (15.4), то в качестве условия разбиения выбирается именно он.

Алгоритм ID3 основан на следующей рекурсивной процедуре.

1. Выбирается признак для корневого узла дерева и формируются ветви для каждого из возможных значений этого признака.

2. Дерево используется для классификации обучающего множества. Если все примеры на некотором листе принадлежат одному классу, то этот лист помечается именем этого класса.

Если все листья помечены именами классов, алгоритм заканчивает работу. В противном случае узел помечается именем очередного признака, и создаются ветви для каждого из возможных значений этого признака, после чего алгоритм снова выполняет шаг 2.

16. СУЩНОСТЬ ЗАДАЧ РАСПОЗНАВАНИЯ. КЛАССИФИКАЦИЯ ПОСРЕДСТВОМ ЗАДАНИЯ ГРАНИЦЫ РАЗДЕЛЕНИЯ

Ключевые понятия: классификация и распознавание, решающее правило, алгоритм распознавания

16.1. Понятия классификации и распознавания

Разделение рассматриваемой совокупности объектов или явлений на однородные (в определенном смысле) группы называется ***классификацией*** [17]. При этом термин «классификация» используют в зависимости от контекста для обозначения как самого процесса разделения, так и его результата. Это понятие тесно связано с такими терминами, как группировка, типологизация, систематизация, дискриминация, кластеризация, распознавание и является одним из основополагающих в практической и научной деятельности человека.

В работах, посвященных созданию интеллектуальных систем, основанных на индуктивных методах, используются термины ***классификация и распознавание***. Эти понятия семантически близки. Под ***задачей распознавания***, как правило, понимается в первую очередь задача формирования ***решающего правила***, на основании которого объекты, принадлежащие классу, могут быть отделены от остальных. Когда мы говорим о получении решающего правила, речь идет о создании алгоритма, способного эффективно отличить объекты, принадлежащие классу, от прочих.

Решающее правило может быть задано по-разному. Прежде всего любое решающее правило – это набор условий, которым должны удовлетворять примеры формируемого понятия. Способ задания решающего правила может быть различным, например, логическое выражение, дерево решений, набор продукционных правил. Правило должно обеспечивать максимально возможную точность распознавания, отражать особенности примеров класса и быть компактным.

Задачу классификации часто понимают как задачу отнесения вновь предъявленных объектов, отличных от объектов обучающей выборки, к тому или иному классу на основе построенного решающего правила.

Проверка точности полученного решающего правила осуществляется на множестве контрольных примеров или путем переклассификации обучающей выборки.

В многомерном анализе данных указанные две задачи решаются в рамках методов классификации.

В табл. 16.1 предложена одна из возможных классификаций методов распознавания.

Таблица 16.1

Классификация методов распознавания		Условия применимости	Ограничения (недостатки)
Методы, основанные на операциях с признаками	Методы, основанные на оценках плотностей распределения значений признаков	Задачи с известным распределением (как правило, нормальным)	Необходимость перебора всей обучающей выборки при распознавании, высокая чувствительность к репрезентативности обучающей выборки
	Методы, основанные на предположениях о классе решающих функций	Классы должны быть хорошо разделяемыми, система признаков – ортонормированная	Должен быть заранее известен вид решающей функции. Невозможность учета новых знаний о корреляциях между признаками
	Логические методы	Задачи небольшой размерности пространства признаков	При отборе логических решающих правил (конъюнкций) необходим полный перебор. Высокая вычислительная трудоемкость
	Лингвистические (структурные) методы	Задачи небольшой размерности пространства признаков	Задача восстановления (определения) грамматики по некоторому множеству высказываний (описаний объектов), является трудноформализуемой
Методы, основанные на операциях с объектами	Метод сравнения с прототипом	Задачи небольшой размерности пространства признаков	Высокая зависимость результатов классификации от меры расстояния (метрики)
	Метод k-ближайших соседей	Задачи небольшой размерности по количеству классов и признаков	Высокая зависимость результатов классификации от меры расстояния (метрики). Необходимость полного перебора обучающей выборки при распознавании. Вычислительная трудоемкость
	Алгоритмы вычисления оценок (голосования) АВО	Задачи небольшой размерности по количеству классов и признаков	Зависимость результатов классификации от меры расстояния (метрики). Необходимость полного перебора обучающей выборки при распознавании. Высокая техническая сложность метода
	Коллективы решающих правил	Задачи небольшой размерности по количеству классов и признаков	Очень высокая техническая сложность метода, теоретические проблемы, как при определении областей компетенции частных методов, так и в самих частных методах

16.2. Математическая постановка задачи распознавания

Для определения принадлежности объекта к одному из заранее известных классов строится правило, в соответствии с которым устанавливается, к какому из классов может быть отнесен классифицируемый (распознаваемый) объект. Алгоритмы классификации основаны на сравнении меры близости или меры сходства классифицируемого объекта с каждым классом. При этом если выбранная мера близости данного объекта с каким-либо классом Ω_i, $i = 1, \ldots, m$, превышает меру его близости с другими классами, то принимается решение о принадлежности этого объекта классу Ω_i.

Пусть результатом наблюдения над объектом ω является реализация p-мерного вектора $X = (x_1, \ldots, x_p)^{\mathrm{T}}$. **Задача классификации** формулируется следующим образом: **требуется установить правило $d(x)$, согласно которому по наблюденному значению вектора X объект относится к одному их возможных классов Ω_i, $i = 1, \ldots, m$.**

Обозначим R^p пространство возможных значений наблюдаемых векторов x. Каждая процедура классификации эквивалентна разбиению пространства R^p на области $R_1, \ldots, R_m$ и установлению принадлежности наблюдения x области R_i: если $x \in R_i$, то объект относят к классу Ω_i. ***Решением*** $d(x) = d_0$ ***является номер класса***, к которому относят наблюдение ($d = \{1, \ldots, m\}$).

Исходная информация, необходимая для решения задачи распознавания (классификации), состоит из двух частей: 1) из априорных сведений об исследуемых классах; 2) из выборочной, т. е. из так называемых «обучающих выборок», представляющих собой распределенные по классам анализируемых объектов-векторов признаков.

Априорные сведения об исследуемых классах представляют собой обычно информацию о способах их описания. Эти сведения получаются из теоретических умозаключений о природе исследуемых объектов или как результат предварительных исследований. Выборочная исходная информация получается в виде экспертных оценок или с помощью специально организованного экспериментального исследования. Математические методы решения задачи распознавания зависят от наличия априорной и предварительной выборочной информации.

Результатом решения задачи распознавания (классификации) должны быть:

- *набор* наиболее *информативных* объясняющих *переменных* (переменные отбираются по определенному правилу из числа исходных описательных признаков либо строятся как некоторые их комбинации);
- правило отнесения (*классификатор*, *функция классификации*, *дискриминантная функция*) каждого нового объекта, заданного значениями своих описательных признаков, к одному из заданных классов или образов.

16.3. Основные задачи классификации (распознавания)

Охарактеризуем основные задачи классификации (распознавания) [6, 17, 23].

Задача 1 заключается в определении полного перечня признаков, характеризующих объекты или явления, для классификации которых разрабатывается данная система. ***Признаки*** могут быть подразделены на ***детерминированные***, ***вероятностные***, ***логические*** и ***структурные***.

Все перечисленные виды признаков, описывающих объекты классификации, могут появляться при задании исходных данных в одной из следующих форм либо при их сочетаниях: 1) экспертные данные, численная и символьная информация общего вида; 2) полученные в различных частях спектра излучений изображения (оптические, инфракрасные, ультразвуковые и т. д.) и затем преобразованные в цифровую форму; 3) сигналы (длинные числовые последовательности).

Задача 2 состоит в формировании априорной совокупности признаков. С учетом результатов решения **задачи 1** в эту совокупность включаются только те признаки, относительно которых может быть получена априорная информация, необходимая для описания классов с помощью этих признаков.

Задача 3 состоит в описании всех классов на языке признаков.

Если признаки классифицируемых объектов детерминированные, то описанием каждого класса объектов на языке этих признаков является его эталон, т. е. точка, сумма расстояний которой от точек, описывающих объекты, принадлежащие данному классу, минимальна.

Если признаки классифицируемых объектов *логические* и имеют *количественные выражения*, то для описания классов объектов на языке признаков необходимо определить диапазоны значений признаков, соответствующие классам Ω_i, $i = 1, \ldots, m$. При этом каждый из отрезков может рассматриваться как *элемен-*

тарное логическое высказывание A, B, C, … . Если признаки распознаваемых объектов суть *суждения качественного характера*, то каждый из них также рассматривается как элементарное логическое высказывание A', B', C', … . Для описания классов на языке этих признаков необходимо выяснить, какими из них характеризуется каждый класс, после этого установить зависимости в форме булевых соотношений между признаками $A, B, C, \ldots, A', B', C'$ и классами Ω_i, $i = 1, \ldots, m$.

Если распределение объектов по областям R_i ($i = 1, \ldots, m$) p-мерного пространства признаков является *вероятностным*, то для описания классов необходимо определить характеристики этих распределений: *функции плотности вероятности* $f_i(x_1, x_2, ..., x_p)$ значений признаков $x_1, x_2, ..., x_p$ при следующем условии: априорные вероятности того, что случайным образом выбранный из общей совокупности объект окажется принадлежащим Ω_i, равны $P(\Omega_i)$.

Если признаки классифицируемых объектов *структурные*, то описаниями классов являются *языки*, состоящие из предложений, каждое из которых характеризует структурные особенности объектов, принадлежащих исключительно одному из классов.

Во многих задачах распознавания со сложной структурой образов число признаков, требующееся для построения разделяющих поверхностей, очень велико. Целесообразно рассмотреть такие образы, как композицию менее сложных структурных компонентов. Эти структурные компоненты могут быть построены из простых элементов с помощью грамматик. Основные элементы (примитивы) представляют собой простейшие образы, по которым осуществляется распознавание, а сложный образ описывается в терминах взаимосвязи этих примитивов. На рис. 16.1 приведен пример*, иллюстрирующий этот подход.

Правило, разделяющее два типа хромосом (субмедианный и телецентрический), может быть построено с использованием пяти кривых-примитивов, или терминальных символов. Правила продукции определяют порядок размещения или конкатенации (сцепления) примитивов при формировании образа. Каждый класс образов описывается единственным множеством продукционных правил. Нетерминальные символы используются для частичных или промежуточных описаний образов на основе примитивов.

---

* Dubes, R. Algorithms for clustering data / R. Dubes, A. Jain. – New York : Prentice-Hall, 1988

Основным преимуществом структурного подхода является то, что он определяет, как данный образ может быть сконструирован из терминальных элементов. Такое описание классов используется в тех случаях, когда образы имеют определенную структуру, которая может быть зафиксирована множеством правил, как, например, текстурированные изображения, форма контуров, форма электрокрдиограммы (ЭКГ) и т. п. Грамматика нормальной и ненормальной форм ЭКГ определяется на основе комплекса зубцов $P - Q - R - S$. Конкретная кривая ЭКГ классифицируется как нормальная или ненормальная в зависимости от того, какая из двух грамматик корректно выполнит ее синтаксический разбор.

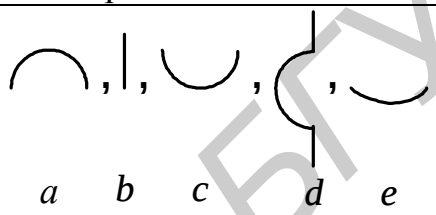
<table>
<tr><td colspan="3">Хромосомная грамматика</td></tr>
<tr><td>Терминальные символы</td><td colspan="2"> 
a b c d e</td></tr>
<tr><td>Нетерминальные символы</td><td>S: стартовый символ
A: плечевая пара
B: низ
C: сторона</td><td>D: правая часть
E: левая часть
F: плечо</td></tr>
<tr><td colspan="3">Образ класса
  </td></tr>
<tr><td colspan="2">Субмедиальный</td><td>Телецентрический</td></tr>
<tr><td colspan="3">Продукционные правила</td></tr>
<tr><td colspan="3">$S \rightarrow A\ A$ (субмедиальный)
$S \rightarrow B\ A$ (телецентрический)
$A \rightarrow C\ A | A\ C | F\ D | E\ F$ $\quad B \rightarrow b\ B | B\ b | e$
$E \rightarrow F\ c$ $\quad C \rightarrow bc | C\ b | d$
$D \rightarrow c\ F$ $\quad F \rightarrow bF | Fb | a$</td></tr>
</table>

Рис. 16.1. Пример структурного распознавания образов

Задача 4 заключается в разбиении априорного пространства признаков на области, соответствующие классам. Подобное разбиение должно быть выполнено так, чтобы при этом обеспечивалось минимальное значение ошибок, возникающих при классификации неизвестных объектов или явлений.

Пусть объект (наблюдение) описывается вектором $x = (x_1, x_2, ..., x_p)$, состоящим из p компонент-признаков. Тогда в пространстве признаков каждый объект будет представлять p-мерную точку.

Положим, произведено разбиение объектов на классы Ω_i ($i = 1, \ldots, m$). Требуется выделить в пространстве признаков области R_i, эквивалентные классам: если объект, имеющий признаки x_j^0, $j = 1, \ldots, p$, относится к классу Ω_i, то представляющая его точка в пространстве признаков принадлежит области R_i.

Помимо геометрической, существует и алгебраическая трактовка задачи: требуется построить разделяющие функции $F_i(x_1, x_2, ..., x_p)$, $i = 1, \ldots, m$, обладающие следующим свойством: если объект, имеющий признаки x_j^0, $j = 1, \ldots, p$, относится к классу Ω_i, то величина $F_i(x_1^0, x_2^0, ..., x_p^0)$ должна быть наибольшей. Если x_k обозначает вектор признаков объектов, относящихся к классу Ω_k, то $F_k(x_k) > F_l(x_k)$, $k, l = 1, \ldots, m$, $k \neq l$.

Таким образом, в пространстве признаков **граница разбиений, называемая решающей границей между областями R_i, соответствующими классам Ω_i, выражается уравнением** $F_k(x) - F_l(x) = 0$.

Задача 5 состоит в выборе алгоритмов классификации (распознавания), обеспечивающих отнесение классифицируемого объекта или явления к тому или другому классу.

Алгоритмы классификации основаны на сравнении меры близости или меры сходства классифицируемого объекта с каждым классом. При этом если выбранная мера близости данного объекта с каким-либо классом Ω_k, $k = 1, \ldots, m$ превышает меру его близости с другими классами, то принимается решение о принадлежности этого объекта классу Ω_k.

В алгоритмах классификации, базирующихся на использовании как детерминированных, так и вероятностных признаков, в качестве меры близости объектов, групп объектов, представляющих собой классы, объекты и группы объектов наиболее часто применяются метрики в виде расстояний.

В алгоритмах распознавания, базирующихся на использовании логических признаков, не используется понятие «мера близости». Когда построено описание классов на языке логических признаков в виде соответствующих булевых

соотношений (эквивалентности или импликаций), при подстановке в эти соотношения значений признаков, характеризующих распознаваемый объект, получаем решение, к какому классу или каким классам объект может быть отнесен и к каким он не относится.

В алгоритмах распознавания, использующих структурные (лингвистические) признаки, понятие меры близости также может не использоваться. Когда построены языки для описания классов в виде совокупности предложений, характеризующих структурные особенности объектов каждого класса, то распознавание неизвестного объекта осуществляется идентификацией предложения, описывающего этот объект, с одним из предложений языка – элемента описания соответствующего класса.

Виды систем классификации (распознавания). В зависимости от того, с какого рода информацией работает алгоритм распознавания, системы распознавания (классификации) могут быть разделены на детерминированные, вероятностные, логические, структурные и комбинированные. Каждая из этих систем использует определенные математические методы классификации, реализованные в виде алгоритмов.

16.4. Детерминированные системы

В детерминированных системах для построения алгоритмов распознавания используются геометрические меры близости, основанные на измерении расстояний между распознаваемым объектом и эталонами классов. В общем случае применение детерминированных методов распознавания предусматривает наличие координат эталонов классов в пространстве признаков или координат объектов, принадлежащих соответствующим классам.

Методы распознавания, применяемые в детерминированных системах, работают с детерминированными признаками.

Детерминированные признаки – это признаки, принимающие конкретные числовые значения. При рассмотрении детерминированных признаков ошибками измерений пренебрегают.

16.4.1. Алгоритм распознавания, основанный на принципе разделения

Суть принципа разделения состоит в следующем. Во многих задачах описания объектов задаются наборами значений числовых признаков (*p*-мерными

векторами). Тогда объекты можно интерпретировать как точки p-мерного пространства. Их описания, принадлежащие разным классам, могут быть разделены поверхностями достаточно простого вида.

Воспользуемся классом разделяющих поверхностей в виде ***гиперплоскостей***:

$$\sum_{i=1}^{p} a_i x_i + a_{p+1} = 0.$$

Пусть множество допустимых объектов разделено на два класса: K_1, K_2, $K_1 \cap K_2 = \varnothing$. Пусть также известно, что объекты S_1, …, S_m принадлежат K_1, объекты S_{m+1}, …, S_q – K_2. Эти объекты неравнозначны. Поэтому введем их числовые характеристики: $\gamma(S_i) = \gamma_i$ – вес объекта S_i, $i = 1, 2, \ldots, m, m+1, \ldots, q$. Таким образом, алгоритм распознавания может быть охарактеризован параметрами $a_1, \ldots, a_{p+1}$ – коэффициентами в уравнении гиперплоскости и $\gamma_1, \ldots, \gamma_q$ – весами объектов, классификация которых была приведена ранее. Распознавание объекта S_i с описанием $I(S_i)$ производится следующим образом.

Пусть $f(x_1, \ldots, x_p) = \sum_{i=1}^{p} a_i x_i + a_{p+1}$. Разделим объекты S_1, …, S_m на множества K_1^+, K_1^-; $S_i \in K_1^+$, если $f(I(S_i)) > 0$; $S_i \in K_1^-$, если $f(I(S_i)) < 0$.

Аналогично объекты S_{m+1}, …, S_q разделим на множества K_2^+, K_2^-. Рассмотрим величины

$$\gamma(K_1^+) = \sum_{S_i \in K_1^+} \gamma(S_i), \quad \gamma(K_1^-) = \sum_{S_i \in K_1^-} \gamma(S_i)$$

и аналогичные им величины $\gamma(K_2^+)$ и $\gamma(K_2^-)$. Вычислим $f(I(S))$. Сопоставим S два числа: $\Gamma_1(S)$, $\Gamma_2(S)$ – значение функции принадлежности S классам K_1 и K_2 соответственно. Если $f(I(S)) > 0$, то

$$\Gamma_1(S) = (\gamma(K_1^+) + \gamma(K_2^-)) / (\gamma(K_1^-) + \gamma(K_2^+)),$$

$$\Gamma_2(S) = (\gamma(K_2^+) + \gamma(K_1^-)) / (\gamma(K_1^+) + \gamma(K_2^-)).$$

При $f(I(S)) < 0$ $\Gamma_1(S) = (\gamma(K_1^-) + \gamma(K_2^+)) / (\gamma(K_1^+) + \gamma(K_2^-))$, аналогично вычисляется $\Gamma_2(S)$. По значениям $\Gamma_1(S)$ и $\Gamma_2(S)$ принимается решение об отнесении S к K_1 или K_2. Эта процедура задается **решающим правилом**, которое может быть записано следующим образом:

если $\Gamma_1(S) - \Gamma_2(S) > \delta$, то $S \in K_1$,

если $\Gamma_2(S) - \Gamma_1(S) > \delta$, то $S \in K_2$,

если $|\Gamma_1(S) - \Gamma_2(S)| \le \delta$, то решение не принимается, алгоритм отказывается от классификации S. Здесь δ – параметр решающего правила.

Построенный на принципе разделения алгоритм классификации основан на следующих предположениях: 1) элементы классов K_1 и K_2 разделяются гиперплоскостью; 2) элементы классов не равнозначны по важности, меру этой важности можно выразить числом.

Пример 16.1. Предположим, что области признаков, характеризующие два диагноза заболевания, можно разделить плоскостью, так что точки, соответствующие диагнозам D_1 и D_2, лежат по разные стороны от разделяющей плоскости. Пусть классы (диагнозы) описываются с помощью двух признаков. В этом случае уравнение разделяющей плоскости (прямой) имеет вид $f(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2$, где α_0, α_1 и α_2 – числовые коэффициенты, характеризующие положение разделяющей плоскости; x_1 и x_2 – числовые значения признаков (координаты пространства признаков).

Решающее правило. Если точка, описывающая признаки больного, находится выше плоскости, то $f(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 > 0$, и устанавливается диагноз D_1. При $f(x) = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 < 0$ устанавливается диагноз D_2. В общем случае для решающего правила составляется линейная функция

$$d(x) = \begin{cases} D_1, \text{ если } \sum_k a_k x_k \ge h; \\ D_2, \text{ если } \sum_k a_k x_k \le -h; \end{cases}$$

неопределенный ответ, если $-h < \sum_k a_k x_k < h$.

Величины δ_1 и δ_2 характеризуют порог распознавания. Чем больше значения δ_1 и δ_2, тем выше надежность распознавания, но одновременно тем больше число отказов от установления диагноза.

Классификация посредством задания границы между классами применяется и для вероятностных признаков (случайных величин). Известен ***метод классификации при помощи линейной гиперплоскости*:**

$h(x) = v_0 + v_1 x_1 + v_2 x_2 + \ldots + v_p x_p$.

В дискриминантном анализе для числа классов $m = 2$ используется следующее решающее правило: классифицируемый объект относится к первому классу, если $h(x) < 0$; в противном случае объект относится ко второму классу. Гиперплоскость $h(x) = 0$ строится таким образом, чтобы минимизировать ошибку классификации второго рода β при заданной ошибке классификации первого рода α.

Можно получить также ***кусочно-линейный классификатор***. Разделяющая поверхность в этом случае является кусочно-линейной, т. е. состоящей из кусков гиперплоскостей. Вид разделяющей поверхности может быть разнообразным и зависит от взаимного расположения классифицируемых совокупностей.

17. СТАТИСТИЧЕСКИЕ АЛГОРИТМЫ РАСПОЗНАВАНИЯ

Ключевые понятия: вероятностные системы распознавания, правила классификации при известных плотностях распределения, классификация при наличии обучающих выборок

В зависимости от того, с какого рода информацией работает алгоритм распознавания, системы распознавания (классификации) могут быть разделены на детерминированные, вероятностные, логические, структурные и комбинированные. Каждая из этих систем использует определенные математические методы классификации, реализованные в виде алгоритмов.

17.1. Вероятностные системы распознавания

В вероятностных системах для построения алгоритмов распознавания используются вероятностные методы, основанные на теории статистических решений. В общем случае применение вероятностных методов распознавания предусматривает наличие вероятностных зависимостей между признаками распознаваемых объектов и классами, к которым эти объекты принадлежат.

Априорная информация об исследуемых классах представляется в виде распределений. При этом возможны два случая. В первом случае распределения векторов описательных признаков X внутри классов предполагаются известными. Они задаются аналитически или с помощью перечисления всех возможных значений X. Во втором случае распределения X внутри классов опре-

деляются лишь частично. Тогда необходимо использовать два вида информации: предположения о свойствах распределений и обучающую выборку.

Для статистических методов распознавания применяется классификация, приведенная на рис. 17.1.

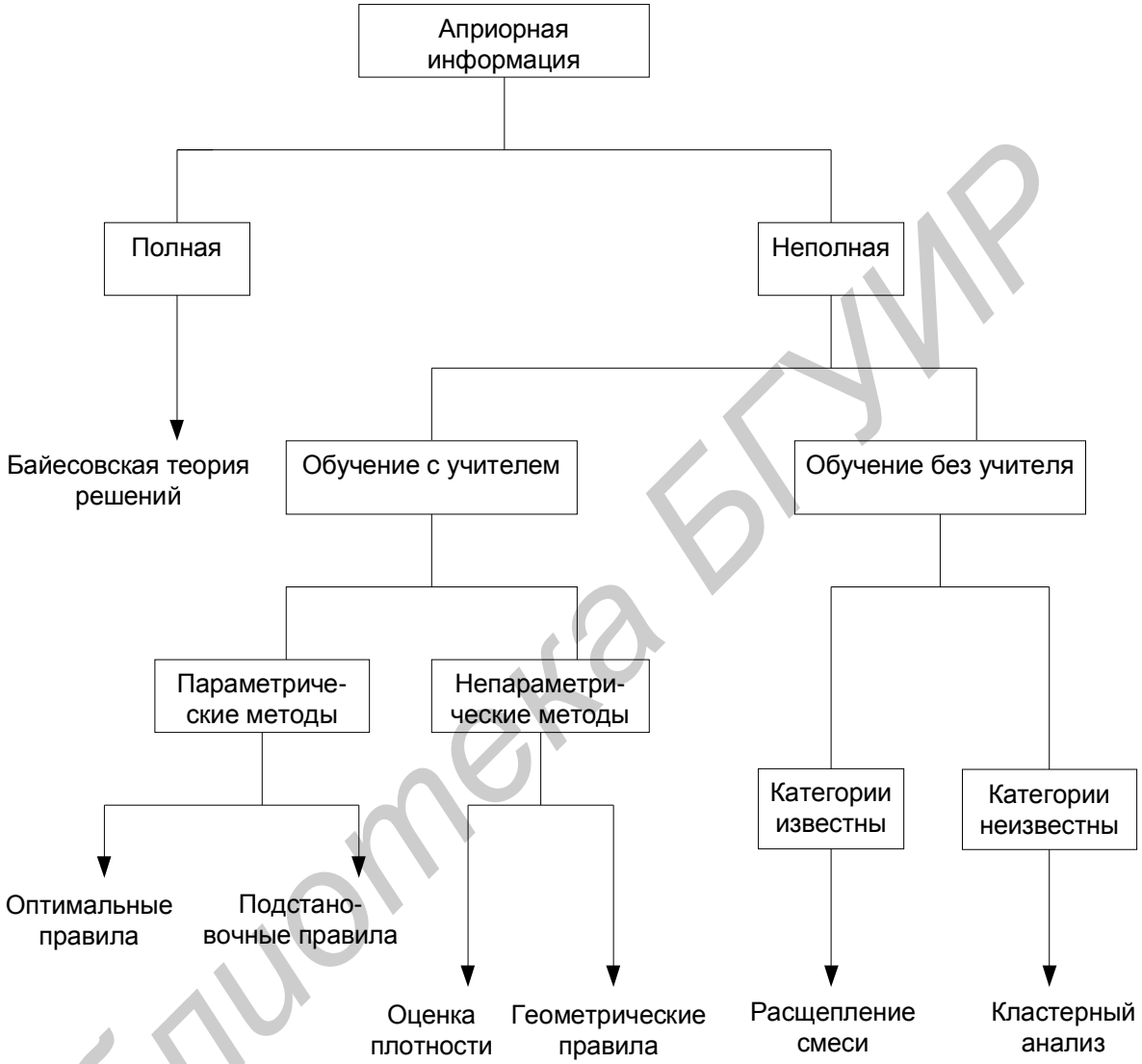


Рис. 17.1. Классификация статистических методов распознавания

Вероятностные признаки, используемые для описания классов, – это признаки, случайные значения которых распределены по всем классам объектов. При этом решение о принадлежности распознаваемого объекта к тому или иному классу может приниматься только на основании конкретных значений признаков данного объекта, полученных в результате проведения соответствующих экспериментов. Признаки следует рассматривать как вероятностные и в том случае, когда измерение их числовых значений производится с такими

ошибками, что по результатам измерений невозможно определенно сказать, какое числовое значение данная величина приняла.

17.2. Правила классификации при известных плотностях распределения

17.2.1. Правило классификации максимального правдоподобия

Обозначим $f_j(x)$ плотность распределения для класса Ω_j, $j = 1, \ldots, m$. Правило максимального правдоподобия (МП-правило) относит x к тому классу Ω_j, для которого функция правдоподобия L_j максимальна:

$$L_j = f_j(x) = \max_i f_i(x).$$

Области пространства признаков R_j (см. подразд. 16.2), задаваемые МП-правилом классификации, определятся как

$$R_j = \{x: L_j(x) > L_i(x)\},\ i = 1, \ldots, m,\ i \neq j.$$

Классифицируя наблюдения, мы можем столкнуться с ошибкой классификации. Для $m = 2$ вероятность отнесения наблюдения во второй класс при условии, что это наблюдение принадлежит первому классу, может быть вычислена как

$$p_{21} = P(X \in R_2 | \Omega_1) = \int_{R_2} f_1(x)dx.$$

Аналогично вероятность классифицировать наблюдение в первый класс при условии, что это наблюдение принадлежит второму классу, равна

$$p_{12} = P(X \in R_1 | \Omega_2) = \int_{R_1} f_2(x)dx.$$

Обозначим $C(i|j)$ потери от ошибочной классификации наблюдения из класса Ω_j. Потери могут быть представлены матрицей потерь:

Действительный номер класса	Результаты классификации (принимаемое решение)	
	Ω_1	Ω_2
Ω_1	0	$C(2\|1)$
Ω_2	$C(1\|2)$	0

Пусть p_j – априорная вероятность класса Ω_j. Ожидаемые *потери от ошибочной классификации*

$$EC = C(2|1)\, p_{21}\, p_1 + C(1|2)\, p_{12}\, p_2. \qquad (17.1)$$

Мы заинтересованы в построении такого правила классификации, которое обеспечивало бы малые ожидаемые потери *EC* или минимизировало их по классу правил. Дискриминантное правило, минимизирующее *EC* (17.1) (ECM-правило), для двух классов задается как

$$R_1 = \left\{ x: \frac{f_1}{f_2} \geq \left(\frac{C(1|2)}{C(2|1)} \right) \frac{p_2}{p_1} \right\},$$
$$R_2 = \left\{ x: \frac{f_1}{f_2} < \left(\frac{C(1|2)}{C(2|1)} \right) \frac{p_2}{p_1} \right\}. \qquad (17.2)$$

Правило максимального правдоподобия является частным случаем правила (17.2) при $C(2|1) = C(1|2)$ и $p_1 = p$.

Пример 17.1. Определить области значений R_1 и R_2 и построить МП-правило классификации для случая, когда два класса описываются нормальными плотностями $N(\mu_1, \sigma_1^2)$ и $N(\mu_2, \sigma_2^2)$.

Решение. Функция правдоподобия L_i определится выражением

$$L_i(x) = (2\pi\sigma_i^2)^{-1/2} \exp\left(-0{,}5 \left(\frac{x-\mu_i}{\sigma_i} \right)^2 \right), \; i = 1, 2.$$

Наблюдение относится к классу Ω_1 ($x \in \Omega_1$), если $L_1(x) \geq L_2(x)$. Это условие эквивалентно соотношению

$$\frac{\sigma_1}{\sigma_2} \exp\left(-0{,}5 \left(\frac{x-\mu_1}{\sigma_1} \right)^2 - 0{,}5 \left(\frac{x-\mu_2}{\sigma_2} \right)^2 \right) \geq 1$$

или

$$x^2 \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right) - 2x \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} \right) + \left(\frac{\mu_1^2}{\sigma_1^2} - \frac{\mu_2^2}{\sigma_2^2} \right) \leq 2 \ln \frac{\sigma_2}{\sigma_1}. \qquad (17.3)$$

Пусть $\mu_1 = 0$; $\sigma_1 = 1$; $\mu_2 = 1$; $\sigma_2 = 0{,}5$. Формула (17.3) дает

$$R_1 = \{x: \; x \leq \frac{1}{3}(4 - \sqrt{4 + 6\ln 2}) \text{ или } x \geq \frac{1}{3}(4 + \sqrt{4 + 6\ln 2})\},$$

$R_2 = R \setminus R_1$.

В случае равных дисперсий получаем **правило классификации** (при $\mu_1 < \mu_2$):

$$x \text{ относится к классу } \begin{cases} \Omega_1, \text{ если } x \in R_1 = \left\{x: x \le \frac{1}{2}(\mu_1+\mu_2)\right\}, \\ \Omega_2, \text{ если } x \in R_2 = \left\{x: x > \frac{1}{2}(\mu_1+\mu_2)\right\}. \end{cases}$$

На рис. 17.2 изображены две плотности распределения, на оси абсцисс расположены области значений признака, соответствующие двум классам.

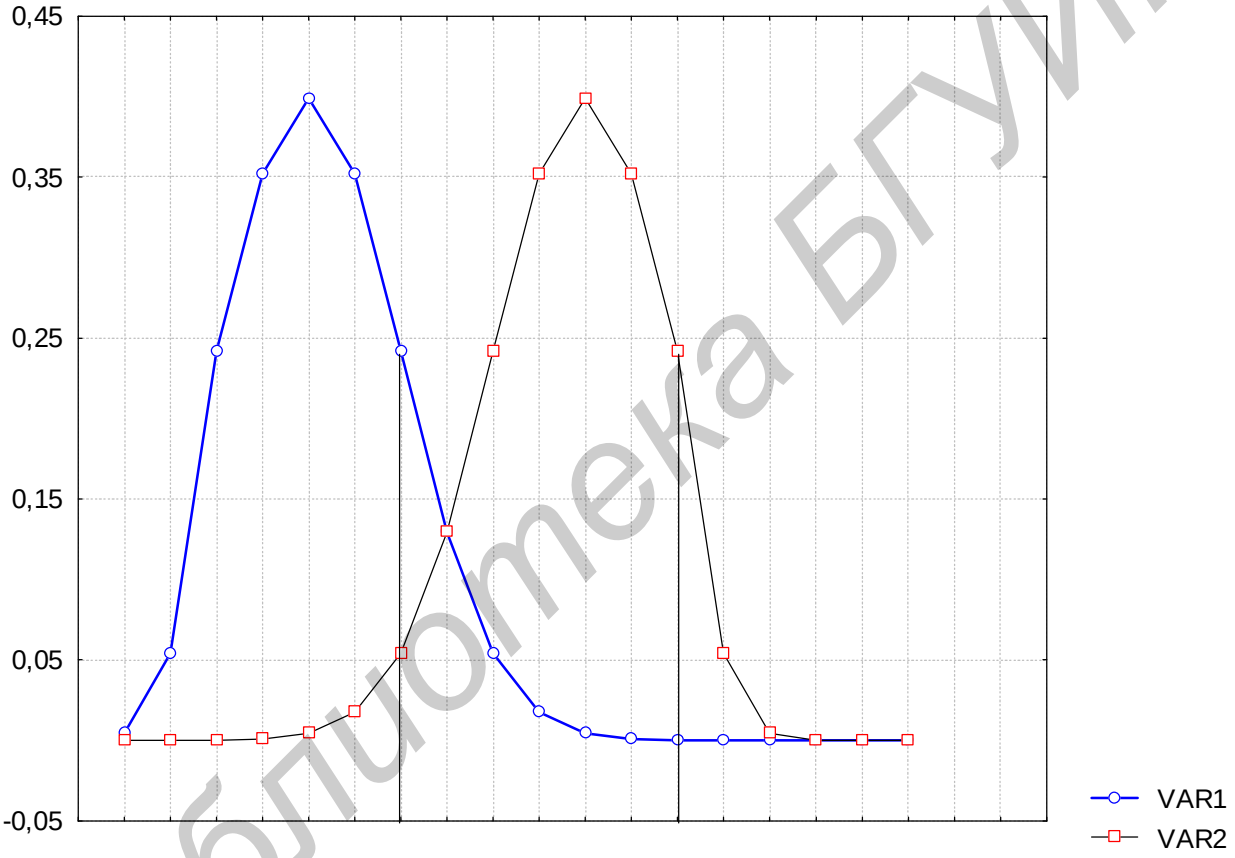


Рис. 17.2. Области значений признака, соответствующие двум классам

Пример 17.2. Определить области значений R_1 и R_2 из условия минимизации ожидаемых потерь от ошибочной классификации EC (17.1). Пусть Ω_1 представляет класс «неудачных» клиентов банка (соответственно Ω_2 – класс «хороших» клиентов).

Решение. Если «неудачные» клиенты классифицированы как «хорошие» клиенты, потери составляют $C(2|1)$. Аналогично определим $C(1|2)$ как потери от ошибочной классификации «хороших» клиентов как «неудачных».

Обозначим γ выигрыш банка от правильной классификации «хорошего» клиента. Общий выигрыш банка составит:

$$G(R_2) = -C(2|1)\,p_1 \int I(x \in R_2)\, f_1(x)dx - C(1|2)\,p_2 \int (1 - I(x \in R_2))\, f_2(x)dx +$$

$$+ \gamma\, p_2 \int I(x \in R_2)\, f_2(x)dx =$$

$$= C(2|1)\,p_1 \int I(x \in R_2)(-C(2|1)p_1 f_1(x) + (C(1|2) + \gamma)p_2 f_2(x))dx.$$

Здесь $I(x)$ – индикаторная функция множества A: $I(x) = 1$, если $x \in A$, и $I(x) = 0$ – в противном случае.

Так как первый член в этом уравнении постоянный, максимум достигается для $R_2 = \{ x: -C(2|1)\,p_1\, f_1(x) + (C(1|2) + \gamma)\, p_2\, f_2(x) \geq 0\}$.

Это эквивалентно

$$R_2 = \{ x: \ \ f_1(x) / f_2(x) \geq \frac{C(2|1)p_1}{(C(1|2+\gamma)p_2} \}.$$

Последнее соотношение при $\gamma = 0$ соответствует множеству R_2, задаваемому (17.2).

Для многомерного нормального распределения правило классификации может быть связано с **расстоянием Махаланобиса**. Пусть классы описываются многомерными нормальными плотностями $f_i(x, \mu_i, \Sigma)$ с равными ковариационными матрицами Σ.

МП-правило на основании расстояния Махаланобиса формулируется следующим образом:

наблюдение x относится к классу Ω_j, где j I {1, …, m} есть величина, минимизирующая квадрат расстояния Махаланобиса между x и центром класса μ_i:

$$\delta^2 (x, \mu_i) = (x - \mu_i)^{\mathrm{T}} \Sigma^{-1} (x - \mu_i),\ i = 1, \ldots, m. \qquad (17.4)$$

В случае $m = 2$

$$x \text{ I } R_1 \text{ Ù } \alpha^{\mathrm{T}}(x - \mu) \geq 0, \text{ где } \alpha = \Sigma^{-1}(\mu_1 - \mu_2) \text{ и } \mu = 0{,}5\,(\mu_1 + \mu_2). \qquad (17.5)$$

17.2.2. Байесовское правило классификации

Обозначим p_j априорные вероятности классов Ω_j и $\sum_{j=1}^{m} p_j = 1$. Правило классификации имеет вид:

> x относится к классу Ω_j, для которого величина $p_i f_i(x)$ максимальна:
> $$p_j f_j(x) = \max_i p_i f_i(x).$$

Таким образом, правило классификации определяется областью значений признаков $R_j = \{x: p_j f_j(x) \geq p_i f_i(x)\}$, $i = 1, \ldots, m$, $i \neq j$. Очевидно, байесовское правило идентично МП-правилу при $p_j = 1/m$. Дальнейшая модификация правила состоит в том, что наблюдение x относится к классу Ω_j с определенной вероятностью $\varphi_j(x)$, такой, что $\sum_{j=1}^{m} \varphi_j(x) = 1$ для всех x. Такое правило называется *рандомизированным дискретным правилом*.

Рандомизированное дискретное правило является обобщением детерминированного правила, так как для него

$$\varphi_j(x) = \begin{cases} 1, & \text{если } p_j f_j(x) = \max_i p_i f_i(x), \\ 0, & \text{в противном случае.} \end{cases}$$

Для того чтобы узнать, какое правило лучше, введем меру сравнения. Обозначим

$$p_{ij} = \int \varphi_i(x) f_i(x) dx \quad - \qquad (17.6)$$

вероятность отнесения наблюдения x в класс Ω_i, если оно принадлежит классу Ω_j. Решающее правило с вероятностью p_{ij} лучше любого другого дискриминантного правила с вероятностью p'_{ij}, если $p_{ii} \geq p'_{ii}$, $i = 1, \ldots, m$.

17.2.3. Вероятность ошибочной классификации для правила максимального правдоподобия

Рассмотрим случай $m = 2$ класса. Пусть класс описывается нормальной плотностью $f_i(x) = n_p(\mu_i, \Sigma)$. Вычислим вероятность ошибочной классификации для МП-правила. Рассмотрим, например, $p_{12} = P(X \text{ I } R_1|\Omega_2)$. На основании (17.5) имеем

$$p_{12} = P\{\alpha^{\mathrm{T}}(x - \mu) > 0 \mid \Omega_2\}.$$

Если X I R_2, $\alpha^{\mathrm{T}}(x - \mu)$ имеет нормальное распределение $N(-\frac{1}{2}\delta^2, \delta^2)$, где $\delta^2 = (\mu_1 - \mu_2)^{\mathrm{T}} \Sigma^{-1} (\mu_1 - \mu_2)$ – квадрат расстояния Махаланобиса между центрами двух классов, тогда мы получим $p_{12} = \Phi(-\frac{1}{2}\delta)$. Аналогично, вероятность классификации в класс Ω_2, если x принадлежит Ω_1, будет равна $p_{21} = \Phi(-\frac{1}{2}\delta)$.

17.2.4. Классификация при различных ковариационных матрицах

Минимум ожидаемых потерь от ошибочной классификации EC зависит от отношения плотностей $f_1(x) / f_2(x)$ или в эквивалентной форме от $\ln(f_1(x))$ – $-\ln(f_2(x))$. Если ковариационные матрицы для обеих плотностей распределения различны, правило будет иметь вид:

$$R_1 = \left\{ x: \frac{1}{2} x^{\mathrm{T}} (\Sigma_1^{-1} - \Sigma_2^{-1}) x + (m_1^{\mathrm{T}} \Sigma_1^{-1} - m_2^{\mathrm{T}} \Sigma_2^{-1}) x - k \geq \ln \left(\frac{C(1|2)}{C(2|1)} \right) \frac{p_2}{p_1} \right\},$$

$$R_2 = \left\{ x: \frac{1}{2} x^{\mathrm{T}} (\Sigma_1^{-1} - \Sigma_2^{-1}) x + (m_1^{\mathrm{T}} \Sigma_1^{-1} - m_2^{\mathrm{T}} \Sigma_2^{-1}) x - k < \ln \left(\frac{C(1|2)}{C(2|1)} \right) \frac{p_2}{p_1} \right\},$$

где $k = \frac{1}{2} \ln \frac{|\Sigma_1^{-1}|}{|\Sigma_2^{-1}|} + \frac{1}{2} (\mu_1^{\mathrm{T}} \Sigma_1^{-1} \mu_1 - \mu_2^{\mathrm{T}} \Sigma_2^{-1} \mu_2)$.

Области, соответствующие каждому классу, определяются квадратической функцией. Квадратическое правило классификации совпадает с правилами, использующими предположение $\Sigma_1 = \Sigma_2$, так как в этом случае

$$\frac{1}{2} x^{\mathrm{T}} (\Sigma_1^{-1} - \Sigma_2^{-1}) x = 0.$$

17.2.5. Линейные дискриминантные функции Фишера

Идея Р. А. Фишера состоит в том, чтобы построить решающее правило на основе таких линейных функций, которые максимизируют квадрат расстояния между средними значениями классов. Если $Y = Xa$ линейная комбинация наблюдений, тогда общая сумма квадратов равна $\sum_{i=1}^{n} (y_i - \bar{y})^2$. Обозначим внутригрупповую матрицу ковариаций W; межгрупповую матрицу ковариаций – B.

Тогда внутригрупповая сумма квадратов вычислится как

$$\sum_{j=1}^{m} Y_j^{\mathrm{T}} H_j Y_j = a^{\mathrm{T}} W a,$$

Где H_j – ($n_j \times n_j$) – центрирующая матрица; Y_j – j-я подматрица Y, соответствующая наблюдениям из класса j. Внутригрупповая сумма квадратов является мерой рассеяния (вариацией) внутри групп.

Межгрупповая сумма квадратов равна $\sum_{j=1}^{m} (\bar{y}_j - \bar{y})^2 = a^{\mathrm{T}} B a$ является мерой рассеяния средних значений в группах относительно общего среднего.

Необходимо найти такие функции a, которые бы максимизировали

$$\gamma = \frac{a^{\mathrm{T}} B a}{a^{\mathrm{T}} W a}.$$

Решением этой задачи является собственный вектор матрицы $W^{-1}B$, соответствующий максимальному собственному значению.

Теперь легко получить решающе правило: классифицировать наблюдение X в класс j, для которого $a^{\mathrm{T}} \bar{X}_j$ ближе к $a^{\mathrm{T}} X$, т. е.

классифицировать X в класс Π_j, где $j = \arg\min_j \lvert a^{\mathrm{T}}(x - \bar{X}_j)\rvert$.

Для $m = 2$ получаем простое правило классификации:

отнести наблюдение X в класс Π_1, если $a^{\mathrm{T}}(x - 0{,}5(\bar{X}_1 + \bar{X}_2)) > 0$,
отнести наблюдение X в класс Π_2, если $a^{\mathrm{T}}(x - 0{,}5(\bar{X}_1 + \bar{X}_2)) \le 0$,

17.3. Классификация при наличии обучающих выборок

17.3.1. Подстановочное правило классификации

При построении рассмотренных выше правил предполагалось, что классы описаны полностью с помощью плотностей распределений. Если распределения неизвестны, то можно выдвинуть предположение о виде распределения, а недостающая информация об их параметрах может быть получена на основе обучающей выборки.

Предположим, что данные j-го класса подчиняются p-мерному нормальному распределению $f_j(x) = n_p(\mu_j, \Sigma)$. В обучающей выборке представлено m классов, каждый объемом n_j. Вычислим оценки неизвестных параметров: среднее $\overline{X}_j$ – для математического ожидания μ_j и выборочную ковариационную матрицу S_j – для Σ:

$$\overline{X}_q^{(j)} = \frac{1}{n_j}\sum_{k=1}^{n_j} x_{kq}^{(j)},\ q = 1, \ldots, p;$$

оценки $(S_j)_{rs}$ элементов ковариационной матрицы Σ вычисляются как

$$(S_j)_{rs} = \frac{1}{n_j - 1}\sum_{k=1}^{n_j}(x_{kr}^{(j)} - \overline{X}_r^{(j)})(x_{ks}^{(j)} - \overline{X}_s^{(j)});\ r, s = 1, \ldots, p. \qquad (17.7)$$

Оценка общей ковариационной матрицы S_T вычисляется по формуле

$$S_T = \sum_{j=1}^{m}\frac{n_j S_j}{n - m},\ n = \sum_{j=1}^{m} n_j.$$

Затем полученные оценки подставляются в решающее правило. Такой способ называется ***подстановочным алгоритмом*** построения правила классификации. Например, подстановочный вариант МП-правила, задаваемого соотношениями (17.4) и (17.5), может быть получен следующим образом. Вычислим по обучающей выборке оценку d^2 квадрата расстояния Махаланобиса между x и $\overline{X}_i$:

$$d^2(x, \mu_i) = (x - \overline{X}_i)^{\mathrm{T}} S^{-1} (x - \overline{X}_i),\ i = 1, \ldots, m.$$

Тогда решающее правило примет вид:

наблюдение x относится к классу Ω_j, для которого оценка $\overline{d}^2$ квадрата расстояния Махаланобиса между x и $\overline{X}_i$ минимальна.

Для $m = 3$ области значений x, соответствующие каждому классу, будут задаваться с помощью следующих функций:

$$h_{12}(x) = (\overline{X}_1 - \overline{X}_2)^{\mathrm{T}} S^{-1} (x - 1/2(\overline{X}_1 + \overline{X}_2)),$$

$$h_{13}(x) = (\overline{X}_1 - \overline{X}_3)^{\mathrm{T}} S^{-1} (x - 1/2(\overline{X}_1 + \overline{X}_3)),$$

$$h_{23}(x) = (\overline{X}_2 - \overline{X}_3)^{\mathrm{T}} S^{-1} (x - 1/2(\overline{X}_2 + \overline{X}_3)).$$

Правило классификации в этом случае будет иметь вид:

наблюдение x относится к классу $\begin{cases} \Omega_1, \text{ если } h_{12}(x) \geq 0 \text{ и } h_{13}(x) \geq 0, \\ \Omega_2, \text{ если } h_{12}(x) < 0 \text{ и } h_{23}(x) \geq 0, \\ \Omega_3, \text{ если } h_{13}(x) < 0 \text{ и } h_{23}(x) < 0. \end{cases}$

17.3.2. Оценка вероятности ошибочной классификации

Вероятность ошибочной классификации, задаваемая выражением (17.6), может быть вычислена путем замены неизвестных значений параметров распределения их оценками, полученными по обучающей выборке.

Для МП-правила (m = 2) в случае нормальных распределений получим

$$\hat{p}_{12} = \hat{p}_{21} = \Phi(-\frac{1}{2}d),$$

где $d^2 = (\overline{X}_1 - \overline{X}_2)^{\mathrm{T}} S_T^{-1} (\overline{X}_1 - \overline{X}_2)$ – оценка квадрата расстояния Махаланобиса между центрами двух классов.

Вероятность ошибочной классификации может быть также оценена либо по новой, экзаменационной выборке, либо путем реклассификации обучающей выборки. Мы классифицируем каждое наблюдение x_i, i = 1, …, n, в m классов согласно выбранному правилу. Обозначим n_{ij} число объектов из класса Ω_j, которые были распознаны как относящиеся к классу Ω_i. Оценкой вероятности ошибочной классификации p_{ij} является $\hat{p}_{ij} = \frac{n_{ij}}{n_j}$. Результаты вычислений представляются в виде матрицы классификации, строками которой являются номера классов в экзаменационной выборке (наблюдаемый класс), а столбцами – решение, полученное на основании решающего правила (предсказанный класс).

При вычислении частоты ошибочных решений обучающая выборка использовалась дважды: первый раз – для построения правила классификации, второй раз – для оценки качества этого правила. Такой подход дает в среднем завышенную оценку качества классификации по сравнению с той же оценкой качества по не зависимым от обучения данным. Это означает, что полученные по обучающей выборке значения ошибок будут меньше ожидаемых, а значения расстояний – больше. Предложен ряд способов, позволяющих корректировать это завышение [17]: 1) *разбиение имеющихся данных* на экзаменационную и

обучающую выборки; 2) применение к параметру качества *поправки на смещение*; 3) использование *метода скользящего экзамена*.

Для двух классов метод скользящего может быть представлен следующими шагами.

Шаг 1. Начинаем процедуру с класса Ω_1. Отделяем одно наблюдение от выборки и строим РП по оставшимся $n - 1$ наблюдениям.

Шаг 2. Применяем РП к выделенному на шаге 1 наблюдению.

Шаг 3. Повторяем шаги 1 и 2 до тех пор, пока не будут классифицированы все наблюдения из Ω_1. Подсчитываем число n'_{21} ошибочно классифицированных наблюдений из первого класса.

Шаг 4. Повторяем шаги 1 – 3 для класса Ω_2. Подсчитываем число n'_{12} ошибочно классифицированных наблюдений из второго класса.

Шаг 5. Оцениваем вероятности ошибочной классификации $\hat{p}'_{12} = n'_{12}/n_2$ и $\hat{p}'_{21} = n'_{21}/n_1$ и фактическую долю $\hat{d}$ ошибочно классифицированных наблюдений:

$$\hat{d} = \frac{n'_{21} + n'_{12}}{n_1 + n_2}.$$

17.3.3. Основные этапы решения задачи классификации

Рассмотрим основные этапы решения задачи классификации.

1. Формирование априорной совокупности признаков.
2. Описание классов на языке признаков.
3. Разбиение априорного пространства признаков на области, соответствующие классам.
4. Выбор алгоритмов классификации.
5. Построение функций классификации.
6. Формирование решающего правила и выполнение классификации объектов.
7. Определение точности построенного правила.

Пример 17.3. Используя метод дискриминантного анализа, вычислить значения коэффициентов функции классификации, построить решающее правило и оценить его точность.

Исходные данные представляют собой классифицированную выборку (m = 2) значений шести переменных, характеризующих подлинные (TYPE = 1) и поддельные (TYPE = 2) 1000-франковые швейцарские банкноты*. Ниже в таблице приведено описание переменных.

Имя признака	Характеристика
1. Length	Длина банкноты
2. Leftheight	Высота банкноты, измеренная по левому краю
3. Rightheight	Высота банкноты, измеренная по правому краю
4. Lowermargin	Расстояние от внутренней рамки до нижнего края
5. Uppermargin	Расстояние от внутренней рамки до верхнего края
6. Diagonal	Длина диагонали

Объем выборки n = 105 (n_1 = 51, n_2 = 54), размерность вектора наблюдения p = 6. Для проведения расчетов исходные значения признаков были стандартизованы. Стандартизованные данные (переменные: длина банкноты и длина диагонали) представлены на рис. 17.3.

1. Формирование априорной совокупности признаков

По условию задачи априорная совокупность признаков включает p = 6 признаков.

Определим числовые характеристики исходных данных. Средние значения компонент вектора наблюдений для двух типов банкнот приведены ниже.

Векторы средних значений $\overline{X}^{(1)}$ для класса подлинных банкнот и $\overline{X}^{(2)}$ для класса поддельных банкнот, являющиеся оценками векторов математических ожиданий, равны соответственно $\overline{X}^{(1)}$ = (0,3126; –0,5484; –0,6445; –0,7356; –0,6807; 0,9485), $\overline{X}^{(2)}$ = (–0,1664; 0,5047; 0,6851; 0,7058; 0,6290; –0,8953).

* Веб-страница MD*base. – Режим доступа : http://www.mdtech.de

Класс	Среднее значение						
	Length	Leftheight	Right-height	Lower-margin	Upper-margin	Diagonal	n_i
Подлинные банкноты	0,3126	–0,5484	–0,6445	–0,7356	–0,6807	0,9485	51
Поддельные банкноты	–0,1664	0,5047	0,6851	0,7058	0,6290	–0,8953	54
Общее среднее	0,0663	–0,0068	0,0393	0,0057	–0,0071	0,0003	105

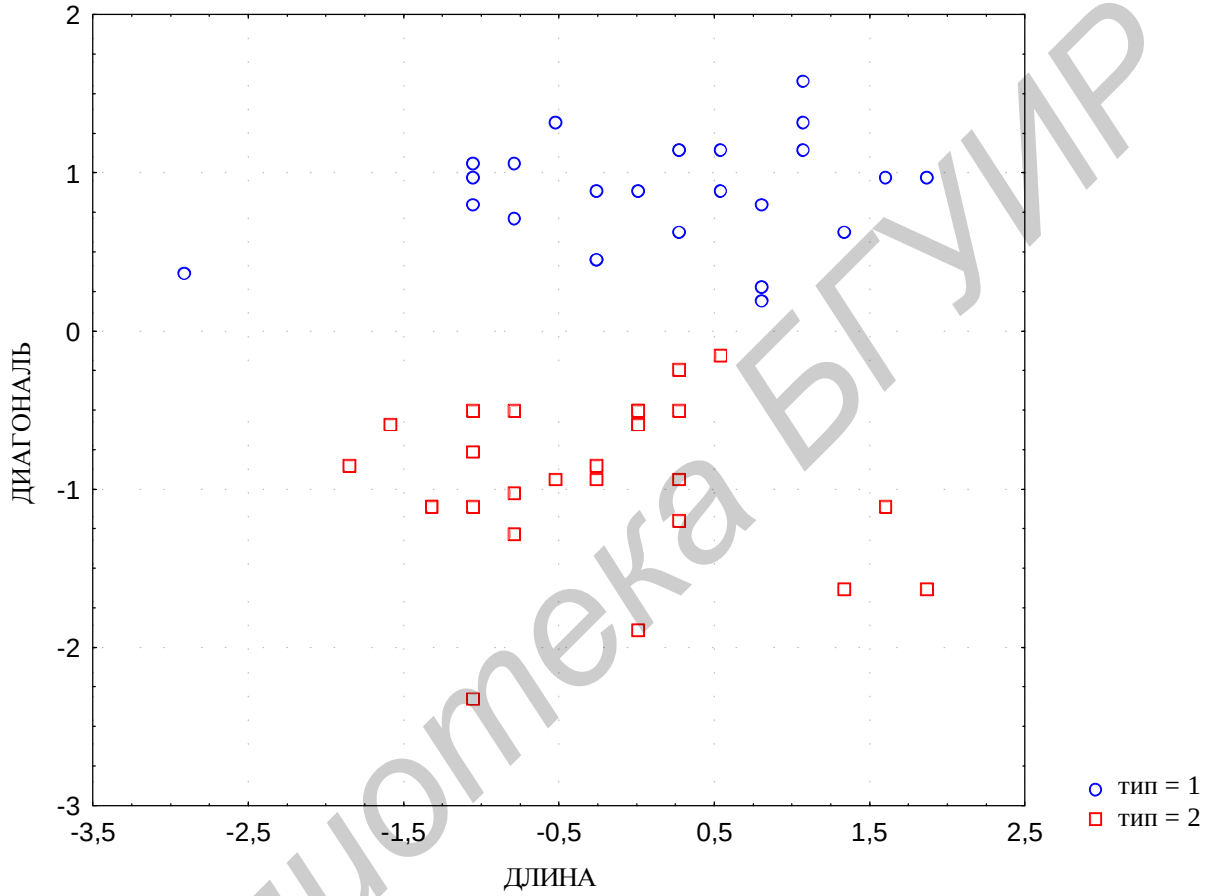


Рис. 17.3. Длина диагонали и длина банкноты двух типов банкнот (подлинных и поддельных)

Выборочная объединенная внутригрупповая матрица ковариаций равна:

	Length	Leftheight	Rightheight	Lowermargin	Uppermargin	Diagonal
Length	1,15	0,37	0,29	–0,05	0,12	–0,01
Leftheight	0,37	0,68	0,40	0,03	–0,01	–0,05
Rightheight	0,29	0,40	0,58	0,04	0,01	0,03
Lowermargin	–0,05	0,03	0,04	0,43	–0,34	0,07
Uppermargin	0,12	–0,01	0,01	–0,34	0,61	–0,05
Diagonal	–0,01	–0,05	0,03	0,07	–0,05	0,18

2. Описание классов на языке признаков

Полагая, что исходные признаки являются вероятностными, для описания классов необходимо определить функции плотности вероятности $f_i(x_1, x_2, ..., x_p)$ значений признаков $x_1, x_2, ..., x_p$, $i = 1, 2$ при следующем условии: априорные вероятности того, что случайным образом выбранный из общей совокупности объект окажется принадлежащим классу Ω_i,равны p_i: $p_1 = 0{,}49$, $p_2 = 0{,}51$ (априорные вероятности классов пропорциональны объемам классов n_i).

3. Разбиение априорного пространства признаков на области, соответствующие классам

Для анализируемых данных разбиение априорного пространства признаков на области, соответствующие классам, задано с помощью классификационной переменной, определяющей номер или имя класса. В нашей задаче это переменная TYPE (тип банкноты), имеющая два значения 1 (подлинные банкноты) и 2 (поддельные банкноты).

4. Выбор алгоритмов классификации

Для решения задачи используем подстановочный алгоритм в модели Фишера, так как у нас отсутствует информация о виде плотности распределения и ее параметрах. Сформулируем предположения модели.

Предположения: теоретические распределения – многомерные нормальные $F_i = N(\mu_i, \mathbf{\Sigma})$, $|\mathbf{\Sigma}| > 0$, $i = 1, 2$; параметры распределения: μ_i (вектор математического ожидания), $\mathbf{\Sigma}$ (ковариационная матрица) – неизвестны. Параметры m_i и $\mathbf{\Sigma}$ оцениваются по выборочным данным. Значения оценок приведены в п. 1.

5. Построение функций классификации

Вычисленные с помощью подстановочного алгоритма функции классификации имеют вид:

$h_1(x) = -7{,}74 + 0{,}76x_1 + 0{,}47x_2 - 1{,}70x_3 - 5{,}51x_4 - 3{,}80x_5 + 6{,}66x_6$ (подлинные банкноты),

$h_2(x) = -6{,}98 - 0{,}59x_1 - 0{,}68x_2 + 1{,}84x_3 + 5{,}21x_4 + 3{,}52x_5 - 6{,}40x_6$ (поддельные банкноты).

6. Формирование решающего правила и выполнение классификации объектов

Правило классификации, основанное на значении функции классификации, имеет вид:

распознаваемый объект x относится к классу $i = 1, \ldots, m$, для которого значение линейной дискриминантной функции $h_i(x)$ является максимальным.

Определим, к какому типу банкнот будет относиться банкнота, которая характеризуется следующими показателями: $x_1 = -0{,}26$, $x_2 = -1{,}17$, $x_3 = -0{,}64$, $x_4 = 0{,}5$, $x_5 = -1{,}3$, $x_6 = 1{,}49$. Для этого подставим значения x в функции классификации для первого и второго классов. В результате получим $h_1(x) = 10{,}22$; $h_2(x) = -24{,}00$. Принимаем решение: банкнота будет относиться к классу подлинных, так как *для этого класса значение линейной дискриминантной функции* $h_1(x)$ *является максимальным*.

Классифицируем это же наблюдение с помощью правила, использующего расстояние Махаланобиса. По формуле (12.6) вычислим оценки расстояния между заданным вектором x и центрами первого и второго классов: $D_1^2 = 2{,}77$, $D_2^2 = 71{,}21$. *Наблюдение будет относиться к тому классу, расстояние Махаланобиса до которого минимально*. Следовательно, распознаваемая банкнота классифицируется как подлинная, так как $D_1^2 < D_2^2$.

Оценим апостериорные вероятности принадлежности рассматриваемого наблюдения x первому (TYPE = 1) и второму (TYPE = 2) классам: $\hat{P}(\Omega_1 | x) = 1{,}00$; $\hat{P}(\Omega_2 | x) = 0{,}00$. Воспользуемся решающим правилом: *классифицируемый объект x относится к классу* $i = 1, \ldots, m$, *для которого значение апостериорной вероятности* $P(\Omega_i | x)$ *является максимальным*. На основании этого правила следует сделать заключение о том, что идентифицируемая банкнота относится к классу подлинных банкнот.

7. Определение точности построенного правила

Определим ошибку решающего правила. Для этого классифицируем экзаменационную выборку и вычислим частоту ошибочных решений. Результаты вычислений представлены в виде матрицы классификации, строками которой

являются номера классов в экзаменационной выборке (наблюдаемый класс), а столбцами – решение, полученное на основании решающего правила (предсказанный класс).

Результаты классификации

Наблюдаемый класс	% правильных решений	Число решений	
		Предсказанный класс	
		подлинные	Поддельные
Подлинные банкноты	99	99	1
Поддельные банкноты	100	0	100
Всего	99,5	99	101

Из матрицы классификации следует, что общая частота ошибочных решений составила 0,5 %, а для класса подлинных банкнот – 1 %.

18. ЛОГИЧЕСКИЕ И СТРУКТУРНЫЕ МЕТОДЫ РАСПОЗНАВАНИЯ

Ключевые понятия: логические модели распознавания, структурные модели распознавания, процедура распознавания

В зависимости от того, с какого рода информацией работает алгоритм распознавания, системы распознавания (классификации) могут быть разделены на детерминированные, вероятностные, логические, структурные и комбинированные. Каждая из этих систем использует определенные математические методы классификации, реализованные в виде алгоритмов.

18.1. Логические методы распознавания

В ***логических системах*** для построения алгоритмов распознавания используются логические методы распознавания, основанные на дискретном анализе и базирующемся на нем исчислении высказываний. В общем случае применение логических методов распознавания предусматривает наличие логических связей, выраженных через систему булевых уравнений, в которой переменные – логические признаки распознаваемых объектов, а неизвестные величины – классы, к которым эти объекты относятся.

Логические признаки. ***Логические признаки*** объектов можно рассматривать как элементарные высказывания, принимающие с полной определенностью два

значения истинности: «да», «нет» или «истина», «ложь». К логическим признакам относятся прежде всего признаки, не имеющие количественного выражения. Эти признаки представляют собой суждения качественного характера: наличие (отсутствие) некоторых свойств или некоторых элементов у распознаваемых объектов или явлений. В качестве логических признаков можно рассматривать, например, качество изделий массовой продукции («годен – брак»), результат выполнения некоторого задания («выполнено – не выполнено»), наличие в прошлом определенных заболеваний и т. д. К логическим признакам можно отнести такие, у которых важна не сама величина, а лишь факт попадания или непопадания ее в заданный интервал.

Логические модели распознавания. В процессе решения прикладных задач обработки данных сформировались семейства (модели) алгоритмов для решения задач распознавания. Укажем ряд алгоритмов, получивших практическое применение.

В ***моделях, основанных на исчислении высказываний,*** в частности на аппарате алгебры логики, классы и признаки объектов рассматриваются как логические переменные, а описание классов на языке признаков представляется в форме булевых соотношений.

Область применения логических методов. Применение методов алгебры логики необходимо тогда, когда существенны не только количественные соотношения между величинами, характеризующими рассматриваемые процессы, но и связывающие их логические зависимости. При распознавании эти методы используют в случаях, когда отсутствуют сведения о количественном распределении объектов по пространственным, временным или каким-то другим интервалам в соответствующем пространстве признаков, а имеются лишь детерминированные логические связи между анализируемыми объектами и их признаками.

Приведем **примеры задач**, для решения которых требуется применение методов алгебры логики: диагностика заболевания пациента на основе данных наблюдения и известных априорных зависимостей между видами заболеваний и соответствующими признаками; диагностика технических систем на основе данных наблюдений и известных априорных зависимостей между видами неисправностей и функциональными характеристиками компонент системы.

Для решения задач распознавания используется математический аппарат булевой алгебры. Он применяется для исчисления высказываний, установления

зависимости и независимости высказываний, нахождения явного вида логической зависимости, а также для решения булевых алгебраических уравнений с одним (или более) неизвестным.

В логических системах распознавания классы и признаки объектов рассматриваются как логические переменные. Введем следующие обозначения для классов и признаков.

Пусть множество объектов разделено на классы Ω_i, $i = 1, \ldots, m$, а для описания объектов используются признаки $A_1, A_2, ..., A_n$.

Вся априорная информация о классах объектов выражает:

1) связь между высказываниями $\Omega_1, ..., \Omega_m$ и $A_1, ..., A_n$;

2) зависимости между признаками $A_1, ..., A_n$;

3) зависимости между классами $\Omega_1, ..., \Omega_m$.

Предположим, что вся априорная информация представлена в форме булевых соотношений:

$$\begin{cases} \Omega_i = f_i(A_1, ..., A_n);\ \Omega_j = h_j(A_1, ..., A_n); \\ L_i(A_1, ..., A_n) \to \Omega_i; \\ F_i(A_1, ..., A_n; \Omega_1, ..., \Omega_m) = H_l(A_1, ..., A_n; \Omega_1, ..., \Omega_m); \\ \Phi_k(A_1, ..., A_n; \Omega_1, ..., \Omega_m) = 1; \\ y_r(A_1, ..., A_n) = 1;\ y_s(A_1, ..., A_n) = 1, ... \end{cases} \quad (18.1)$$

Предположим также, что наряду с (18.1) в результате эксперимента получены данные, касающиеся части признаков $A_1, ..., A_n$, характеризующих объекты классов $\Omega_1, ..., \Omega_m$, и что эти данные выражены как булева функция $G(A_1, ..., A_n) = 1$.

Прямая задача распознавания [6] состоит в том, чтобы определить, какие выводы можно сделать относительно классов $\Omega_1, ..., \Omega_m$ на основе априорной информации (18.1) и апостериорной информации $G(A_1, ..., A_n) = 1$, т. е. требуется определить неизвестную функцию $F(\Omega_1, ..., \Omega_m)$, удовлетворяющую уравнению

$$\overline{G}(A_1, ..., A_n) + F(\Omega_1, ..., \Omega_m) = 1 \quad (18.2)$$

при ограничениях (18.1).

Сопряженная задача заключается в том, чтобы установить, какие совокупности признаков $A_1, ..., A_n$ должны иметь место, если известны некоторые

сведения о классах $\Omega_1, ..., \Omega_m$, т. е. требуется определить неизвестную функцию $G_1(A_1, ..., A_n)$, удовлетворяющую уравнению

$$\overline{F_1}(\Omega_1, ..., \Omega_m) + G_1(A_1, ..., A_n) = 1 \qquad (18.3)$$

при заданной функции $F(\Omega_1, ..., \Omega_m)$ и связях (18.2).

Обратная задача распознавания [6] заключается в том, чтобы определить множество априорно неизвестных посылок $G(A_1, ..., A_n)$, из которых следуют некоторые выводы $F(\Omega_1, ..., \Omega_m)$ при условии, что признаки $A_1, ..., A_n$ и классы $\Omega_1, ..., \Omega_m$ связаны зависимостями (18.3).

Пример 18.1. Рассмотрим пример представления медицинских знаний с помощью операций математической логики. Допустим, что из медицинских наблюдений известны следующие связи: 1) признак A_1 появляется при диагнозе D_2; 2) если имеется диагноз D_1 и отсутствует диагноз D_2, то должен появляться признак A_2; 3) если появляются оба признака A_1 и A_2 вместе, то могут быть диагнозы D_1 и D_2 одновременно.

Первое условие записывается в виде $D_2 \rightarrow A_1$, второе – $D_1 \wedge D_2 \rightarrow A_2$, третье – $A_1 \vee A_2 \rightarrow D_1 \vee D_2$.

Так как все эти условия справедливы одновременно, то они могут быть записаны в виде булевой функции событий:

$E = [D_2 \rightarrow A_1] \wedge [D_1 \wedge D_2 \rightarrow A_2] \wedge [A_1 \vee A_2 \rightarrow D_1 \vee D_2]$.

18.2. Структурные методы распознавания

18.2.1. Структурные (лингвистические) системы

В ***структурных (лингвистических) системах*** для построения алгоритмов распознавания используются специальные грамматики, порождающие языки. Языки состоят из предложений, каждое из которых описывает объекты, принадлежащие конкретному классу. Применение структурных методов требует наличия совокупностей предложений для описания множества объектов, принадлежащих всем классам. При этом множество предложений должно быть подразделено на подмножества по числу классов. Элементами подмножеств являются предложения, описывающие объекты, принадлежащие данному подмножеству (классу). Таким образом, априорными описаниями классов являются

совокупности предложений, каждое их которых соответствует конкретному объекту, принадлежащему данному классу.

Структурные признаки. ***Структурные (лингвистические, синтаксические) признаки*** представляют собой непроизводные элементы (символы) структуры объекта. Иначе эти элементы (константы) называют терминалами. Каждый объект может рассматриваться как цепочка терминалов или как предложение. При этом если предложение, описывающее неизвестный распознаваемый объект, относится к языку данного класса, то выносится решение о принадлежности объекта этому классу.

18.2.2. Структурные (лингвистические) модели распознавания

В структурных (лингвистических) моделях распознавания описание объектов производится языковыми средствами. Правила языка, определяющие способы построения объектов из непроизводных элементов, называют грамматикой. В соответствии с грамматикой объект представляется предложением на этом языке. Априорными описаниями классов являются *структурные описания* – формальные конструкции. Они используются при анализе иерархической структуры объекта и отношений, существующих между отдельными элементами этой структуры.

Этапы процедуры распознавания. ***Процедура распознавания*** на основе использования структурных методов состоит из следующих этапов:

1) предварительной обработки, описания или представления объекта;
2) синтаксического анализа.

Предварительная обработка. На этапе предварительной обработки предъявленный для распознавания объект подвергается кодированию и фильтрации, восстановлению и улучшению качества.

Объект после предварительной обработки представляется некоторой структурой языкового типа (например цепочкой или графом). Процесс получения представления объекта включает в себя процедуры:

1) разбиения (сегментации) объекта;
2) выделения признаков – непроизводных элементов.

В результате каждый объект получает свое представление с помощью некоторого набора непроизводных элементов и ряда фиксированных синтаксических операций. Например, при использовании операции соединения объект получает представление в виде некоторой цепочки соединенных непроизводных элементов.

Синтаксический анализ. На этапе ***синтаксического анализа*** система распознавания должна обнаруживать синтаксические связи, существующие в объекте. Решение о синтаксической правильности представления объекта, т. е. о принадлежности его к некоторому классу, задаваемому определенной синтаксической системой или грамматикой, вырабатывается ***синтаксическим анализатором*** **(*блоком грамматического разбора*)**. При выполнении синтаксического анализа (грамматического разбора) анализатор воспроизводит полное синтаксическое описание объекта в виде дерева грамматического разбора, если соответствующий объект является синтаксически правильным. В противном случае объект либо отклоняется, либо подвергается анализу с помощью других заданных грамматик, которыми могут описываться другие классы изучаемых объектов.

Процедура распознавания – это сравнение с эталоном. Цепочка непроизводных элементов, представляющая анализируемый объект, сопоставляется с цепочками непроизводных элементов, описывающих классы. Распознаваемый объект с помощью выбранного *критерия согласия* (*подобия*) относится к тому классу, с которым обнаруживается наибольшая близость.

Область применения. Структурные методы распознавания применяются для распознавания изображений в следующих областях: техническое зрение роботов, медицинские системы обследования и диагностики (рентгенография, компьютерная томография, ангиография), аэрофотосъемка и космическая съемка со спутников, неразрушающий контроль в промышленности и т. д.

Структурные методы применяются также в тех случаях, когда информация о распознаваемых объектах или явлениях содержится в записях соответствующих сигналов (электрокардиограмм, инфракрасных сигналов, порождаемых исследуемыми объектами, акустических сигналов функционирующих систем и т. д.). Для определения признаков используется разложение в ряды по ортогональным функциям. При этом в качестве признаков берутся коэффициенты разложения. Возможно использование в качестве признаков и некоторых характерных элементов экспериментальных кривых (точки минимума, максимума и др.), например, электрокардиограммы, содержащей типичные структурные элементы – зубцы (экстремальные точки) *P, Q, R, S, T*, связанные с циклами деятельности сердца.

Использование в качестве признаков характерных элементов экспериментальных кривых, их структуры базируется на том факте, что структура сигналов,

отраженных распознаваемыми объектами или порожденных ими, однозначно определяется структурой наблюдаемого объекта.

18.2.3. Реализация процесса распознавания на основе структурных методов

Задачи распознавания. Для распознавания неизвестного объекта на основе структурных методов [6] необходимо решить следующие задачи:

- найти его непроизводные элементы и отношения между ними;
- с помощью синтаксического анализа (грамматического разбора) установить, согласуется ли описание объекта с грамматикой, которая по предположению могла его породить.

Для формирования соответствующей грамматики можно воспользоваться либо априорными сведениями о распознаваемых объектах, либо результатами изучения конечного выборочного множества «типичных» объектов. В первом случае говорят о ***задании грамматики на основе эвристических соображений***, во втором – о ***выводе грамматики***.

Описание классов. Допустим, что мы имеем дело с двумя классами объектов (Ω_1 и Ω_2) и объекты, входящие в эти классы, можно описать с помощью признаков, принадлежащих некоторому конечному множеству. Эти ***признаки можно считать основными символами*** (***элементы основного словаря***) и обозначать через V (их называют также ***непроизводными символами***, ***непроизводными элементами***). Каждый объект может рассматриваться как цепочка или предложение, поскольку он составлен из элементов основного словаря.

Пусть также существует ***грамматика*** **Г** такая, что порождаемый ею язык состоит из предложений (объектов), принадлежащих исключительно одному из классов, например классу Ω_1. Эта грамматика может быть использована для классификации объектов.

Решающее правило задается следующим образом: ***предъявленный неизвестный объект можно отнести к классу*** Ω_1***, если он является предложением языка L*****(Г)**. ***В противном случае объект приписывается классу*** Ω_2. В данном случае объект попадает в класс Ω_2 исключительно потому, что он не входит в класс Ω_1: если оказывается, что объект не является грамматически правильным предложением в смысле грамматики Г, то предполагается, что он должен принадлежать классу Ω_2.

На самом же деле объект может не относиться и к классу Ω_2. Поэтому необходимо располагать грамматиками Γ_1 и Γ_2, порождающими языки $L(\Gamma_1)$ и $L(\Gamma_2)$ соответственно. ***Объект попадает в тот класс, в языке которого он оказывается грамматически правильным предложением***. Если последнее не выполняется, то очевидно, что данный объект не принадлежит ни к одному из двух заданных классов, и, следовательно, требуется еще одна грамматика Γ_3 и т. д. В случае m классов рассматривается m грамматик и связанных с ними языков $L(\mathbf{\Gamma}_i)$, $i = 1, \ldots, m$. ***Распознаваемый объект относится к классу Ω_i, если он является грамматически правильным предложением языка*** $L(\Gamma_i)$. Если объект оказывается грамматически правильным предложением более чем одного языка, решение относительно его принадлежности принимается точно таким же образом, как это делается при использовании других методов распознавания, когда оказывается, что результаты реализации процедуры распознавания не позволяют определенно отнести объект к одному из заданных классов.

Практическое использование структурного метода обычно требует решения следующих основных проблем:

1) построения адекватного описания распознаваемых объектов;
2) выбора грамматики;
3) реализации процесса распознавания с помощью процедур синтаксического анализа;
4) использования процедур обучения для вывода грамматик;
5) применения в рамках структурного подхода процедур из других методов распознавания, например, статистических (для учета искажений случайного характера), кластерного анализа и т. д.

Основные приемы применения структурных методов распознавания рассмотрены в [9].

ЛИТЕРАТУРА

1. Анализ геоинформационных данных. Компьютерный практикум : учеб. пособие / М. Д. Степанова, С. А. Самодумкин, А. Н. Крючков, Н. А. Гулякина ; под науч. ред. В. В. Голенкова. − Минск : БГУИР, 2004.

2. Достоверный и правдоподобный вывод в интеллектуальных системах / В. Н. Вагин [и др.]. − М. : Физматлит, 2004.

3. Волковец, А. И. Теория вероятностей и математическая статистика : конспект лекций / А. И. Волковец, А. Б. Гуринович. − Минск : БГУИР, 2003.

4. Гаврилова, Т. А. Базы знаний интеллектуальных систем : учебник / Т. А. Гаврилова, В. Ф. Хорошевский. − СПб. : Питер, 2000.

5. Гаек, П. Автоматическое образование гипотез / П. Гаек, Т. Гавранек. – М. : Наука, 1984.

6. Горелик, А. Л. Методы распознавания : учеб. пособие / А. Л. Горелик, В. А. Скрипкин. − М. : Высш. шк., 1986.

7. Дубров, А. М. Многомерные статистические методы : учеб. пособие / А. М. Дубров, В. С. Мхитарян, Л. И. Трошин. − М. : Финансы и статистика, 1998.

8. Жевняк, Р. М. Высшая математика : учеб. пособие. В 5 ч. Ч. 5 / Р. М. Жевняк, А. А. Карпук. − Минск : Выш. шк., 1988.

9. Журавлев, Ю. И. Распознавание образов и распознавание изображений / Ю. И. Журавлев, И. Б. Гуревич // Распознавание, классификация и прогноз. − М. : Наука, 1989. – Вып. 2.

10. Интеллектуальные технологии в геоинформационных системах : учеб. пособие / А. Н. Крючков, С. А. Самодумкин, М. Д. Степанова, Н. А. Гулякина ; под науч. ред. В. В. Голенкова. − Минск : БГУИР, 2006.

11. Искусственный интеллект : справоч. В 3 кн. / под ред. Д. А. Поспелова. − М. : Радио и связь, 1990.

12. Кузнецов, С. О. Об одной модели обучения и классификации, основанной на операции сходства / С. О. Кузнецов, В. К. Финн // Обозрение прикладной и промышленной математики. – Т. 3. – Вып. 1. – 1996.

13. Ларичев, О. И. Теория и методы принятия решений, а также Хроника событий в волшебных странах : учебник / О. И. Ларичев. − М. : Логос, 2002.

14. Малыхина, Г. И. Логика : учеб. пособие / Г. И. Малыхина. − Минск : Выш. шк., 2002.

15. Минто, В. Дедуктивная и индуктивная логика / В. Минто. − М. : Комета, 1995.

16. Мюллер, П. Таблицы по математической статистике / П. Мюллер, П. Нойман, Р. Шторм. − М. : Финансы и статистика, 1982.

17. Прикладная статистика. Классификация и снижение размерности / С. А. Айвазян [и др.]. − М. : Финансы и статистика, 1989.

18. Представление и обработка знаний в графодинамических ассоциативных машинах / В. В. Голенков [и др]. − Минск : БГУИР, 2001.

19. Прикладная статистика. Основы моделирования и первичная обработка данных / С. А. Айвазян, И. С. Енюков, Л. Д. Мешалкин. − М. : Финансы и статистика, 1983.

20. Статические и динамические экспертные системы : учеб. пособие / Э. В. Попов [и др]. − М. : Финансы и статистика, 1996.

21. Степанова, М. Д. Статистические основы индуктивного вывода : метод. пособие / М. Д. Степанова. − Минск : БГУИР, 1999.

22. Степанова, М. Д. Проверка статистических гипотез : учеб.-метод. пособие / М. Д. Степанова. − Минск : БГУИР, 2000.

23. Степанова, М. Д. Математические методы диагностики в медицинских интеллектуальных системах : учеб.-метод. пособие / М. Д. Степанова, С. А. Самодумкин, Т. Л. Лемешева. − Минск : БГУИР, 2000.

24. Харин, Ю. С. Математическая и прикладная статистика : учеб. пособие / Ю. С. Харин, Е. Е. Жук. – Минск : БГУ, 2005.

25. Харин, Ю. С. Практикум на ЭВМ по математической статистике : учеб. пособие / Ю. С. Харин, М. Д. Степанова. – Минск : Университетское, 1987.

ПРИЛОЖЕНИЕ 1

КРИТЕРИИ ПРОВЕРКИ ГИПОТЕЗ

Таблица П1.1

Гипотезы однородности

Нулевая гипотеза	Допущения	Альтер-нативная гипотеза	Статистика критерия	Область отклонения
$\mu_1 = \mu_2$	Нормальность и независимость данных, дисперсии σ_1^2 и σ_2^2 неизвестны, $\sigma_1^2 = \sigma_2^2$	$\mu_1 - \mu_2 > 0$ $\mu_1 - \mu_2 < 0$ $\mu_1 - \mu_2 \neq 0$	$t = \frac{(\bar{X}_1 - \bar{X}_2)}{s\sqrt{1/n_1 + 1/n_2}}$ $s^2 = \frac{\sum_{i=1}^{2}(n_i - 1)s_i^2}{n_1 + n_2 - 2}$ $s_i^2 = \frac{\sum_{j=1}^{n_i}(x_{ij} - \bar{X}_i)^2}{n_i - 1}$	$t > t_{1-\alpha, f}$ $t < t_{\alpha, f}$ $\lvert t\rvert > t_{1-\alpha/2, f}$ $t_{\varepsilon, f}$ – квантиль уровня ε t-распределения с $f = n_1 + n_2 - 2$ степенями свободы
$\mu_1 - \mu_2 = \delta$	Нормальность и независимость данных, дисперсии σ_1^2 и σ_2^2 неизвестны, $\sigma_1^2 \neq \sigma_2^2$	$\mu_1 - \mu_2 > \delta$ $\mu_1 - \mu_2 < \delta$ $\mu_1 - \mu_2 \neq \delta$	$t = \frac{(\bar{X}_1 - \bar{X}_2) - \delta}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$ $n = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{\frac{(s_1^2/n_1)^2}{n_1 - 1} + \frac{(s_2^2/n_2)^2}{n_2 - 1}}$	$t > t_{1-\alpha, \nu}$ $t < t_{\alpha, \nu}$ $\lvert t\rvert > t_{1-\alpha/2, \nu}$ $t_{\varepsilon, \nu}$ – квантиль уровня ε t-распределения с ν степенями свободы
$\sigma_1^2 = \sigma_2^2$	Нормальность и независимость данных	$\sigma_1^2 > \sigma_2^2$ $\sigma_1^2 < \sigma_2^2$ $\sigma_1^2 \neq \sigma_2^2$	$F = s_1^2 / s_2^2$	$F > F_{1-\alpha, f_1, f_2}$ $F < F_{\alpha, f_1, f_2}$ $F < F_{\alpha/2, f_1, f_2}$ или $F > F_{1-\alpha/2, f_1, f_2}$ F_ε – квантиль уровня ε F-распределения с $f_1 = n_1 - 1$ и $f_2 = n_2 - 1$ степенями свободы

Окончание табл. П1.1

Нулевая гипотеза	Допущения	Альтернативная гипотеза	Статистика критерия	Область отклонения
$p_1 - p_2 = 0$	Биномиальные испытания, n_1 и n_2 велико, независимость данных	$p_1 - p_2 > 0$ $p_1 - p_2 < 0$ $p_1 - p_2 \neq 0$	$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(1/n_1 + 1/n_2)}}$ $\hat{p} = \frac{X_1 + X_2}{n_1 + n_2}$ $\hat{p}_i = \frac{X_i}{n_i}$ X_i – число интересующих нас событий в выборке объемом n_i	$Z > z_{1-a}$ $Z < z_{\alpha}$ $\lvert Z \rvert > z_{1-a/2}$ z_{ε} – квантиль уровня ε стандартного нормального распределения

Таблица П1.2

Гипотезы о числовых значениях параметров

Нулевая гипотеза	Допущения	Альтернативная гипотеза	Статистика критерия	Область отклонения
$\mu = \mu_0$	Нормальность данных, дисперсия s_0^2 известна	$\mu > \mu_0$ $\mu < \mu_0$ $\mu \neq \mu_0$	$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$	$Z > z_{1-a}$ $Z < z_{\alpha}$ $\lvert Z \rvert > z_{1-a/2}$ z_{ε} – квантиль уровня ε стандартного нормального распределения
$\mu = \mu_0$	Нормальность данных, дисперсия s_0^2 неизвестна	$\mu > \mu_0$ $\mu < \mu_0$ $\mu \neq \mu_0$	$T = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$	$t > t_{1-a,n-1}$ $t < t_{a,n-1}$ $\lvert t \rvert > t_{1-a/2,n-1}$ $t_{\varepsilon,\nu}$ – квантиль уровня ε t-распределения с ν степенями свободы

Окончание табл. П1.2

Нулевая гипотеза	Допущения	Альтер-нативная гипотеза	Статистика критерия	Область отклонения
$p = p_0$	Биномиальные испытания, n большое	$p > p_0$ $p < p_0$ $p \neq p_0$	$Z = \frac{\hat{p} - np_0}{\sqrt{p_0(1-p_0)n}}$	$Z > z_{1-\alpha}$ $Z < z_{\alpha}$ $\lvert Z \rvert > z_{1-\alpha/2}$ z_{ε} – квантиль уровня ε стандартного нормального распределения
$\sigma^2 = \sigma_0^2$	Нормальность данных	$\sigma^2 > \sigma_0^2$ $\sigma^2 < \sigma_0^2$ $\sigma^2 \neq \sigma_0^2$	$\chi^2 = (n-1)\frac{s^2}{s_0^2}$	$\chi^2 > \chi^2_{1-\alpha,\, n-1}$ $\chi^2 < \chi^2_{\alpha,\, n-1}$ $\chi^2 < \chi^2_{\alpha/2,\, n-1}$ или $\chi^2 > \chi^2_{1-\alpha/2,\, n-1}$ $\chi^2_{\varepsilon,\, f}$ – квантиль уровня ε χ^2-распределения с $f = n - 1$ степенями свободы

ОСНОВНЫЕ ВЕРОЯТНОСТНЫЕ РАСПРЕДЕЛЕНИЯ

П.2.1. Непрерывные распределения

Нормальное распределение

Плотность распределения $f(x) = \frac{1}{\sqrt{2p}\,s} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$;

параметры распределения: μ – математическое ожидание, σ^2 – дисперсия.

Логарифмически нормальное (логнормальное) распределение

Плотность распределения $f_L(x) = \frac{1}{\sqrt{2p}\,xs} \exp\left(-\frac{\ln(x/m)^2}{2\sigma^2}\right)$;

параметры распределения: m – медиана, $\sigma > 0$ – параметр формы.

В статистических вычислениях используется другой параметр: μ – математическое ожидание случайной величины $\ln L$ (L – случайная величина с логнормальным распределением). Параметры μ и m связаны следующим соотношением: $m = \exp\mu$; $\mu = \ln m$. Обозначим $w = \exp(\sigma^2)$, тогда математическое ожидание логарифмически нормального распределения равно $m\exp(\sigma^2/2) = m\sqrt{w}$, дисперсия – $m^2\, w(w-1)$.

Оценки параметров:

$\hat{m} = \exp\hat{m}$, $\hat{m} = \frac{1}{n}\sum_{i=1}^{n} \ln x_i$;

$\hat{s}^2 = \frac{1}{n-1}\sum_{i=1}^{n} (\ln x_i - \hat{m})^2$.

Гамма-распределение

Плотность распределения $f_\gamma(x) = \lambda^a x^{a-1} \exp(-\lambda x) / \Gamma(a)$, $x \geq 0$,

параметр масштаба $\lambda > 0$, параметр формы $a > 0$,

математическое ожидание a/λ,

дисперсия a/λ^2,

$\Gamma(\cdot)$ – гамма-функция.

Оценки параметров: $\hat{l} = \overline{X} / s^2$, $\hat{a} = (\overline{X}/s)^2$.

Распределение Вейбулла – Гнеденко

Функция распределения

$F(x) = 1 - \exp(-a\,x^b)$, $x \geq 0$,

параметры распределения: параметр масштаба $a > 0$, параметр формы $b > 0$.

Экспоненциальное распределение

Функция распределения

$F(x) = 1 - \exp(-\lambda\,x)$, $x \geq 0$,

$\lambda > 0$ – параметр распределения,

математическое ожидание $1/\lambda$,

дисперсия $1/\lambda^2$.

Равномерное распределение

Функция распределения

$F(x) = (x - a) / b$,

параметры распределения: параметр расположения a – нижняя граница области значений, параметр масштаба b – размер области значений. Вместо b можно использовать параметр $h = a + b$ – верхнюю границу области значений.

Плотность вероятностей $1/b$.

Оценки параметров (метод моментов): $\hat{a} = \bar{x} - \sqrt{3}\,s$, $\hat{b} = \sqrt{12}\,s$.

Здесь $\bar{x}$ – выборочное среднее, s^2 – выборочная дисперсия.

П.2.2. Дискретные распределения

Биномиальное распределение

Функция распределения $F(x) = \sum_{i=0}^{x} C_n^i\, p^i (1-p)^{n-i}$,

параметры распределения: n – целое положительное число (число испытаний), p – параметр испытания Бернулли, $0 < p < 1$,

математическое ожидание np,

дисперсия $np(1-p)$,

оценка параметра $\hat{p} = x/n$.

Распределение Пуассона

Функция распределения $F(x) = \sum_{i=0}^{x} \frac{\lambda^i \exp(-\lambda)}{i!}$,

параметр распределения $\lambda > 0$, математическое ожидание и дисперсия равны λ,

оценка параметра $\hat{\imath} = \overline{X}$.

Геометрическое распределение

Функция распределения $F(x) = \sum_{x=0}^{n} p(1-p)^x = 1 - q^n$, $(q = 1 - p)$,

область значений $n \geq 1$, n – целое.

Значение n – число испытаний Бернулли с вероятностью успеха p вплоть до появления первого успеха (включая также первый успех),

p – параметр испытания Бернулли, $0 < p < 1$,

математическое ожидание $1/p$,

дисперсия $(1 - p)/ p$.

П.2.3. Законы распределения вероятностей, используемые в статистических вычислениях

χ^2-распределение

Плотность распределения

$$f_{\chi^2}(x) = \frac{1}{2^{m/2}\Gamma(m/2)} x^{m/2-1} e^{-x/2},\ x > 0,$$

m – положительное число, параметр распределения – число степеней свободы,

$\Gamma(\bullet)$ – гамма-функция Эйлера.

Плотность распределения при $m \leq 2$ постоянно убывает, а при $m > 2$ имеет единственный максимум в точке $x_{\text{mod}} = m - 2$,

математическое ожидание m,

дисперсия $2m$.

t-распределение Стьюдента

Плотность распределения

$$f_t(x) = \frac{\Gamma(\frac{m+1}{2})}{m\sqrt{p}\,\Gamma(m/2)}(1 + x^2/m)^{-(m+1)/2}, -\infty < x < \infty,$$

m – положительное число, параметр распределения – число степеней свободы, математическое ожидание, мода, медиана равны 0,

дисперсия $\frac{m}{m - 2}$ (существует только при $m > 2$).

F-распределение

Плотность распределения

$$f_F(x) = \frac{\Gamma(\frac{m_1 + m_2}{2}) m_1^{m_1/2} m_2^{m_2/2}}{\Gamma(m_1/2)\Gamma(m_2/2)} \frac{x^{m_1/2-1}}{(m_1 x + m_2)^{(m_1 + m_2)/2}}, x \geq 0,$$

m_1 и m_2 – положительные числа, параметры распределения – числа степеней свободы числителя и знаменателя,

математическое ожидание $\frac{m_2}{m_2 - 2}$ (существует при $m_2 > 2$),

дисперсия $\frac{2m_2^2(m_1 + m_2 - 2)}{(m_2 - 6)\sqrt{m_1 + m_2 - 2}}$ (существует при $m_2 > 6$).

Учебное издание

Голенков Владимир Васильевич
Степанова Маргарита Дмитриевна
Самодумкин Сергей Александрович
Гулякина Наталья Анатольевна

СТАТИСТИЧЕСКИЕ ОСНОВЫ ИНДУКТИВНОГО ВЫВОДА

УЧЕБНОЕ ПОСОБИЕ

Редактор *Т. Н. Крюкова*
Корректор *Л. А. Шичко*
Компьютерная верстка *Е. С. Чайковская*

Подписано в печать 08.04.2009. Формат 60x84 $^{1}/_{16}$. Бумага офсетная. Гарнитура «Таймс».
Печать ризографическая. Усл. печ. л. 11,97. Уч.-изд. л. 10,0. Тираж 150 экз. Заказ 196.

Издатель и полиграфическое исполнение: Учреждение образования
«Белорусский государственный университет информатики и радиоэлектроники»
ЛИ №02330/0494371 от 16.03.2009. ЛП №02330/0131666 от 30.04.2004.
220013, Минск, П. Бровки, 6