В. А. ДЮК

ЛОГИЧЕСКИЙ АНАЛИЗ ДАННЫХ

УЧЕБНОЕ ПОСОБИЕ

ЛАНЬ®
САНКТ-ПЕТЕРБУРГ • МОСКВА • КРАСНОДАР
2020

УДК 519.254
ББК 32.81я73

Д 95 **Дюк В. А.** Логический анализ данных : учебное пособие / В. А. Дюк. — Санкт-Петербург : Лань, 2020. — 80 с. : ил. — (Учебники для вузов. Специальная литература). — Текст : непосредственный.

ISBN 978-5-8114-4180-8

В учебном пособии приводится описание двух систем логического анализа данных, предназначенных для выявления в данных if-then правил. Первая система относится к наиболее представительному и популярному направлению, связанному с построением деревьев решений. Вторая система реализует метод ограниченного перебора. Описание систем сопровождается подробно разобранными примерами из различных предметных областей.

Книга предназначена для студентов старших курсов, магистров, аспирантов, научных работников и других специалистов, изучающих современные методы анализа данных.

УДК 519.254
ББК 32.81я73

Обложка
П. И. ПОЛЯКОВА

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Рынок программных продуктов в области интеллектуального анализа данных (ИАД) бурно развивается. Фактически каждый месяц в Интернете появляются анонсы новых инструментов для обнаружения знаний в базах. Помочь разобраться в том, какой из продуктов является наиболее подходящим, призваны специальные обзорные страницы Интернета (например, http://www.kdnuggest.com), на которых приводятся каталоги разработок, рассказывается о фирмах-разработчиках, ведутся дискуссии, сравниваются характеристики различных программ и т. д.

В данном учебном пособии приводится описание двух систем логического анализа данных. Первая система – See5 – относится к наиболее представительному и популярному направлению, связанному с построением деревьев решений. Вторая система – WizWhy – интересна тем, что ее разработчики утверждают, будто она способна обнаружить **ВСЕ** if-then правила в данных. Это утверждение подкрепляется сообщением о весьма большом количестве коммерческих структур, использующих WizWhy (более 30 000). Давайте сами убедимся в полезных свойствах предлагаемых подходов к логическому анализу данных.

1. Построение деревьев решений – система See5/C5.0

Система See5/C5.0 (Windows 95/98/NT) (http://www.rulequest.com) предназначена для анализа больших баз данных, содержащих до сотни тысяч записей и до сотни числовых или номинальных полей. Результат работы See5 выражается в виде деревьев решений и множества if-then правил. Система проста в обращении и не требует от пользователя специфических знаний в области прикладной статистики. Стоимость See5 – 740 долл., некоммерческая версия для обучения ограничена количеством анализируемых записей (до 200).

Проиллюстрируем процесс работы See5 на реальном примере из области медицинской диагностики. Исходные данные в рассматриваемом случае относятся к задаче дифференциальной диагностики заболеваний почек. Данные были получены в Российской медицинской академии [1].

Фрагмент исходных данных приведен в таблице 1.1. Это как раз тот вид данных, для обработки которых более всего подходит See5. Каждый объект (пациент) здесь принадлежит к одному из небольшого числа классов (здоров, множественные кисты, гидронефроз) и описывается 11 разнотипными признаками.

Таблица 1.1. Фрагмент исходных данных по дифференциальной диагностике заболеваний почек

Признак	Объект 1	Объект 2	...
Состояние почки *diagnosis*	Множественные кисты	Гидронефроз	...
Возраст пациента (число полных лет) *Age*	46	52	...
Пол пациента *Sex*	Женщина (F)	Мужчина (M)	...
Левая или правая почка *LR*	Правая почка (R)	Левая почка (L)	...
Длина почки (мм) *Length*	112	136	...
Ширина почки (мм) *Width*	68	69	...
Толщина почки (мм) *Thickness*	88	72	...
Толщина паренхимы (мм) *Thickpar*	18	18	...
Средняя скорость кровотока (см/с) *Speed*	2,3	12	...
Индекс резистентности *Index*	0,584	0,614	...
Ускорение артериального потока в систолу (см/с^2) *Accel*	459	291	...

Задача See5 состоит в предсказании диагностического класса какого-либо объекта по значениям его признаков. При этом, как мы увидим, See5 конструирует классификатор в виде дерева решений, которому, в свою очередь, может быть поставлено в соответствие некоторое множество логических правил.

1.1. Подготовка данных для See5

Каждой задаче, решаемой в системе See5, требуется присвоить свое собственное имя. Пусть в нашем случае это имя будет USR (Ultra Sonic Research). В процессе решения See5 использует и формирует несколько файлов с одинаковым именем и различными расширениями. Важно

точно соблюдать правила записи имен и расширений (система различает строчные и прописные буквы). Кроме того, отметим, что See5 поддерживает только латинские шрифты.

Файл имен переменных

Для работы See5 самыми необходимыми и существенными являются два файла – имен переменных и данных. В файле имен переменных с расширением *.names даются названия используемых признаков и классов.

Среди признаков различают две важные подгруппы:

- номинальные признаки (discrete attribute), количественные признаки (continuous attribute) и метки;
- явно определенные признаки, значения которых берутся непосредственно из файла данных, и неявно определенные признаки, задаваемые формулами (чаще всего употребляются явно определенные признаки).

Файл имен переменных **USR.names** в нашей задаче выглядит следующим образом:

```
diagnosis.          | the target attribute
diagnosis:            1, 2, 3
Age:                  continuous
Sex:                  F, M
LR:                   L, R
Length:               continuous
Width:                continuous
Thickness:            continuous
Thickpar:             continuous
Speed:                continuous
Index:                continuous
Accel:                continuous
```

Целевой признак **diagnosis** принимает три значения: 1 – в классе «здоровая почка»; 2 – в классе «множественные кисты» и 3 – в классе «гидронефроз». Признаки **Age** (возраст), **Length** (длина почки), **Width** (ширина почки), **Thickness** (толщина почки), **Thickpar** (толщина паренхимы), **Speed** (средняя скорость кровотока), **Index** (индекс резистентности) и **Accel** (ускорение артериального потока в систолу) являются количественными. Признак **Sex** (пол пациента) может иметь два значения F (female) и M (male), а признак **LR** (левая или правая почка) принимает значения L или R. Порядок записи имен переменных должен соответствовать их порядку в файле данных.

При подготовке файла имен переменных следует иметь в виду, что пробелы, пустые строки и знаки табуляции игнорируются системой (кроме, конечно, случаев, когда они применяются в именах переменных). Вертикальная черта «|» предназначена для записи напоминаний или комментариев.

После имени каждой явно определенной переменной вставляется двоеточие «:», а затем следует характеристика этой переменной. Возможны следующие характеристики:

- continuous – количественный признак;
- список значений переменной, разделенных запятой (для дискретной, номинальной переменной);
- максимальное значение N для дискретной переменной (эту характеристику рекомендуется применять очень осторожно, так как здесь исключается дополнительная проверка данных при их вводе в анализ);

- ignore – для признака, исключаемого из анализа;
- label – метка для идентификации отдельного объекта.

После имени каждой неявно определенной переменной также следует двоеточие и далее записывается формула. В формуле, где необходимо, используются скобки, а дискретные признаки ограничиваются кавычками. Ниже приведены доступные операторы.

- +, -, *, /, % (mod), ^ (возведение в степень);
- >, >=, <, <=, =, <> или != (не равно);
- and, or;
- sin(...), cos(...), tan(...), log(...), exp(...), int(...) (целая часть от).

В зависимости от применяемой формулы конечный результат может быть как количественным, так и давать логическое значение **true/false**.

Файл данных

Вторым файлом, необходимым для работы See5, является файл данных. Он имеет расширение *.data. В нашем случае это файл **USR.data**.

Каждому объекту в файле данных соответствует собственная строка. Если значение целевой переменной находится вверху файла имен переменных, строка начинается со значения этой целевой переменной. Затем через запятую следуют значения всех остальных признаков. Неизвестные значения переменных кодируются вопросительным знаком «?», после вертикальной черты «|» можно писать невоспринимаемые системой комментарии.

Ниже полностью приводится весь файл данных USR.data, который мы будем использовать для демонстрации возможностей See5.

1, 62, F, R, 127, 52, 43, 14, 13.3, 0.698, 140	2, 48, F, R, 100, 56, 44, 16, 18, 0.667, 114
1, 43, M, L, 103, 44, 49, 15, 16.3, 0.634, 291	2, 74, M, L, 104, 56, 56, 16, 18.2, 0.643, 88
1, 58, M, R, 103, 58, 46, 17, 16.5, 0.704, 143	2, 62, F, L, 118, 56, 41, 14, 18.3, 0.611, 347
1, 37, M, L, 112, 53, 51, 18, 18.2, 0.562, 189	2, 60, F, L, 108, 56, 50, 14, 18.6, 0.546, 216
1, 21, M, L, 126, 62, 45, 14, 18.5, 0.613, 116	2, 63, F, L, 110, 61, 56, 14, 19, 0.645, 216
1, 74, M, R, 115, 57, 49, 16, 11.1, 0.69, 85	2, 64, M, R, 123, 61, 57, 20, 11.1, 0.632, 173
1, 62, M, R, 103, 66, 45, 16, 11.2, 0.657, 65	2, 54, M, R, 105, 58, 43, 16, 21, 0.62, 230
1, 43, M, R, 104, 54, 46, 14, 11.3, 0.574, 629	2, 47, F, L, 116, 73, 56, 17, 21.4, 0.636, 178
1, 34, F, L, 110, 52, 42, 19, 11.3, 0.686, 152	2, 47, F, L, 109, 58, 48, 11, 21.5, 0.544, 266
1, 68, F, R, 112, 52, 42, 17, 11.4, 0.593, 258	2, 66, F, R, 98, 64, 56, 19, 22.3, 0.655, 111
1, 37, M, R, 119, 51, 41, 14, 11.8, 0.65, 101	2, 67, F, L, 103, 44, 42, 14, 22.6, 0.682, 656
1, 38, M, L, 107, 59, 56, 12, 20.9, 0.572, 279	2, 54, M, L, 119, 50, 48, 20, 23.5, 0.65, 188
1, 67, F, R, 113, 57, 47, 17, 20.9, 0.681, 115	2, 70, F, R, 105, 56, 41, 13, 23.5, 0.663, 242
1, 46, F, L, 107, 60, 36, 15, 21.8, 0.678, 352	2, 66, F, L, 111, 61, 50, 16, 24.7, 0.689, 189
1, 67, F, R, 135, 70, 67, 27, 24, 0.583, 51	2, 56, F, L, 99, 58, 47, 16, 25, 0.686, 196
1, 52, M, L, 111, 59, 50, 21, 26.6, 0.644, 379	2, 47, F, R, 99, 62, 50, 12, 26.2, 0.544, 235
1, 42, M, R, 82, 47, 42, 14, 26.8, 0.651, 538	2, 69, F, L, 125, 66, 51, 18, 26.2, 0.667, 275
1, 47, F, L, 114, 63, 51, 17, 27.8, 0.645, 589	2, 37, M, L, 120, 89, 61, 16, 27.9, 0.566, 218
1, 23, F, L, 128, 39, 56, 21, 30.6, 0.62, 255	2, 48, F, L, 113, 77, 49, 22, 30.5, 0.686, 210
1, 35, F, L, 114, 50, 41, 14, 31.2, 0.622, 445	2, 45, F, L, 115, 56, 52, 17, 34.2, 0.587, 410
1, 64, F, R, 97, 57, 39, 16, 34.7, 0.68, 368	2, 70, F, L, 108, 48, 42, 19, 36.1, 0.69, 219
1, 56, M, L, 125, 74, 72, 2, 46.6, 0.638, 685	2, 67, F, L, 122, 72, 64, 20, 43.3, 0.674, 229
2, 46, F, R, 112, 68, 88, 18, 2.3, 0.584, 459	3, 42, M, L, 104, 69, 51, 24, 14.5, 0.729, 211
2, 58, M, L, 129, 67, 58, 18, 12.5, 0.686, 151	3, 21, M, R, 144, 101, 49, 6, 16.3, 0.707, 194
2, 69, F, R, 115, 69, 44, 14, 13.2, 0.657, 231	3, 67, F, L, 99, 52, 52, 8, 16.3, 0.744, 332
2, 69, M, L, 126, 59, 49, 13, 14.1, 0.652, 282	3, 34, F, R, 105, 54, 46, 19, 11.4, 0.704, 98
2, 54, M, R, 98, 87, 41, 24, 14.3, 0.637, 352	3, 47, F, R, 116, 85, 64, 24, 11.5, 0.804, 416
2, 67, M, R, 111, 59, 47, 18, 14.6, 0.742, 242	3, 38, M, R, 118, 45, 60, 12, 11.9, 0.701, 354
2, 70, F, R, 108, 58, 37, 11, 14.6, 0.663, 139	3, 56, M, R, 118, 55, 43, 22, 21, 0.735, 165

2, 67, M, L, 129, 64, 58, 17, 15.2, 0.693, 382	3, 37, M, R, 114, 64, 52, 18, 21.8, 0.717, 225
2, 55, M, L, 125, 59, 48, 18, 15.4, 0.674, 330	3, 62, M, L, 134, 67, 53, 16, 23, 0.76, 321
2, 65, F, R, 111, 54, 51, 13, 15.4, 0.727, 257	3, 68, F, L, 106, 48, 46, 18, 24.5, 0.693, 224
2, 70, F, L, 108, 65, 51, 14, 16.3, 0.724, 250	3, 43, M, L, 136, 65, 57, 22, 25.6, 0.731, 351
2, 63, F, R, 121, 67, 55, 16, 16.4, 0.653, 148	3, 35, F, R, 124, 41, 67, 20, 26.3, 0.692, 286
2, 56, F, R, 99, 55, 47, 16, 16.8, 0.693, 103	3, 23, F, R, 127, 78, 62, 17, 31.2, 0.714, 349
2, 60, F, R, 105, 56, 46, 14, 17.7, 0.526, 219	3, 52, M, R, 125, 59, 44, 27, 31.3, 0.703, 545
2, 54, M, L, 107, 57, 43, 14, 17.9, 0.651, 254	3, 67, F, R, 114, 55, 36, 9, 41.5, 0.73, 310

Файлы тестовых данных (необязательные)

Для проверки качества построенного дерева решений и соответствующего множества логических правил в системе See5 предусмотрена возможность работы со специальными файлами, в которых содержатся дополнительные тестовые данные.

Третий вид файла, используемый системой See5, содержит новые тестовые объекты. Это то, что еще принято называть контрольной выборкой. Данный файл **USR.test** является необязательным и, если используется, имеет формат уже описанного файла **USR.data**.

Следующий вспомогательный файл **USR.cases** также является необязательным. Он содержит объекты с неизвестной классификацией.

Файл стоимости

Последний вид файла, обозначаемый **USR.costs**, содержит информацию о стоимости различных ошибок классификации. Заполнение этого файла является необязательным. Вместе с тем назначение штрафов за ошибки может оказаться весьма полезным при разработке некоторых приложений.

1.2. Интерфейс пользователя

В главном окне See5 располагается пять кнопок (рис. 1.1). Перечислим их слева направо.

С помощью кнопки **Locate Data** (местонахождение данных) вызывается окно для просмотра доступных файлов данных и их загрузки в систему.

Нажатием кнопки **Construct Classifier** (построение классификатора) производится обращение к окну диалога для выбора типа классификатора и установки его опций.

Кнопка **Stop** предназначена для останова процесса построения дерева решений.

Кнопка **Use Classifier** (использование классификатора) запускает процесс интерактивной классификации одного или более объектов.

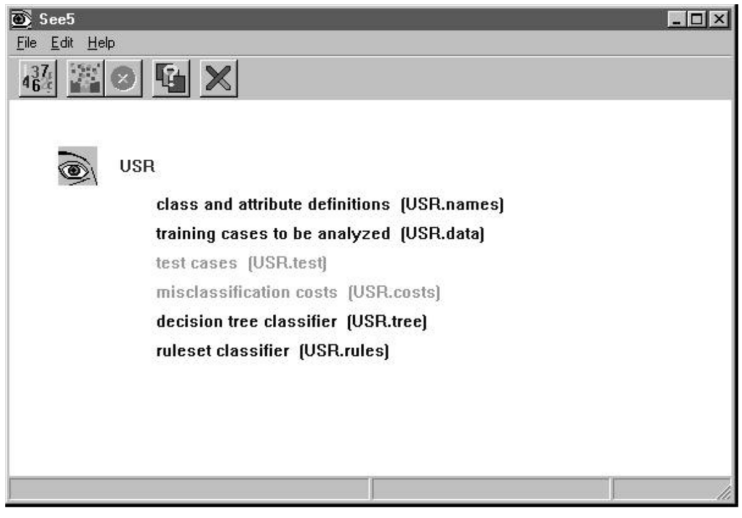


Рис. 1.1. Главное окно системы See5

С помощью кнопки **Cross-Reference** (перекрестная ссылка) вызывается окно, в котором наглядно раскрываются связи между объектами обучающей выборки и найденными правилами их классификации.

Все перечисленные функции доступны также из меню **File**. В свою очередь, в меню **Edit** предоставляется возможность редактирования файла имен данных и файла стоимости ошибок классификации.

Построение дерева решений

На первом этапе обработки данных обычно используются опции системы, установленные по умолчанию. Нажимаем кнопку **Construct Classifier** и затем в появившемся окне диалога (рис. 1.2) сразу нажимаем **OK** (предполагается, что файл данных **USR.data** уже загружен). Система выдает окно результатов, которые выглядят следующим образом (рис. 1.3).

В первой строке отчета о результатах дается информация об используемой версии системы See5 и текущее время. Затем в следующих двух строках говорится о том, что классифицирующей переменной служит **diagnosis** и прочтенный файл данных **USR.data** содержит 74 объекта, каждый из которых описан 11 признаками.

В следующих строках отчета отображено построенное дерево решений. Его можно проинтерпретировать следующим образом:

ЕСЛИ **Index** больше 0.69 и **Speed** больше 18, ТО класс № 3, иначе

ЕСЛИ **Index** больше 0.69 и **Speed** не больше 18 и **Thickness** не больше 46, ТО класс № 1 и т. д.

Каждая ветка дерева заканчивается указанием номера класса, к которому она приводит. Сразу за номером следует запись вида (n) или (n/m). Например, самая первая ветка заканчивается записью (12.0). Это означает, что данной ветке соответствует 12 объектов из определенного (третьего) класса. Последняя ветка заканчивается записью 1 (6.0/1.0), из чего следует, что эта ветка описывает класс № 1 и сюда попадает 6 объектов, из которых 1 попадает ошибочно. Величины n или m могут оказаться дробными в случае, когда на какую-либо ветку придется некоторое число объектов с неизвестными значениями признаков.

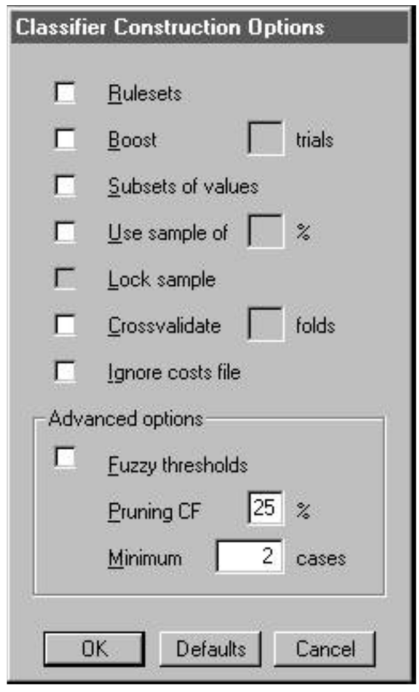
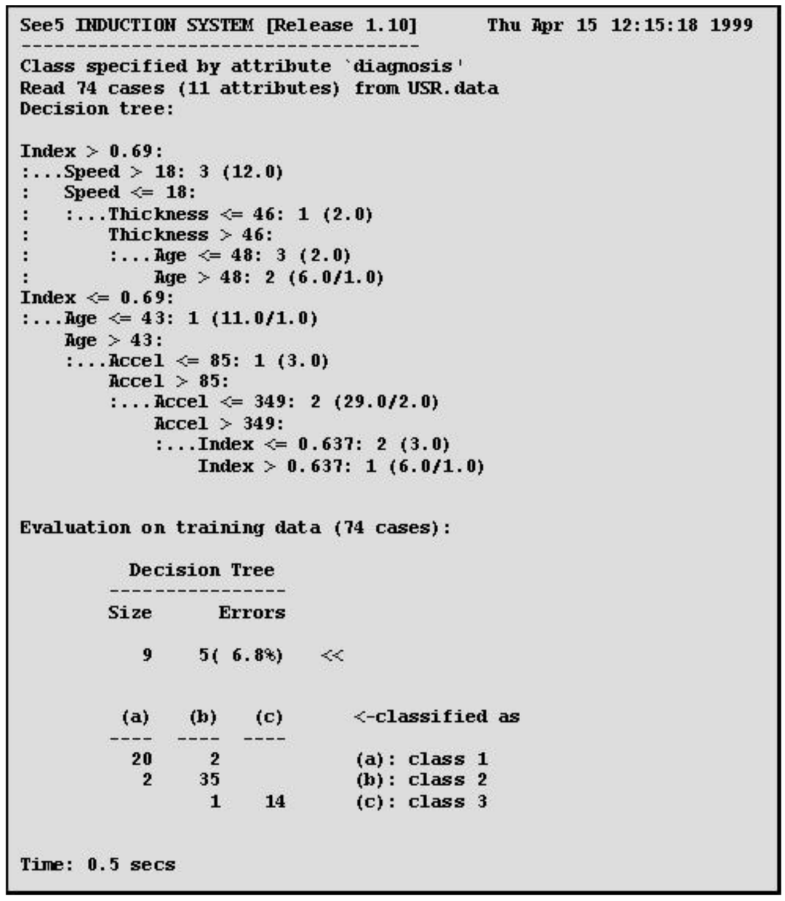


Рис. 1.2. Окно диалога для задания опций алгоритма конструирования классификатора



```
See5 INDUCTION SYSTEM [Release 1.10]      Thu Apr 15 12:15:18 1999
-----------------------------------
Class specified by attribute `diagnosis'
Read 74 cases (11 attributes) from USR.data
Decision tree:

Index > 0.69:
:...Speed > 18: 3 (12.0)
:   Speed <= 18:
:   :...Thickness <= 46: 1 (2.0)
:       Thickness > 46:
:       :...Age <= 48: 3 (2.0)
:           Age > 48: 2 (6.0/1.0)
Index <= 0.69:
:...Age <= 43: 1 (11.0/1.0)
    Age > 43:
    :...Accel <= 85: 1 (3.0)
        Accel > 85:
        :...Accel <= 349: 2 (29.0/2.0)
            Accel > 349:
            :...Index <= 0.637: 2 (3.0)
                Index > 0.637: 1 (6.0/1.0)


Evaluation on training data (74 cases):

         Decision Tree
       ----------------
       Size      Errors

          9    5( 6.8%)   <<


        (a)   (b)   (c)      <-classified as
       ----  ----  ----
         20     2           (a): class 1
          2    35           (b): class 2
                1    14     (c): class 3


Time: 0.5 secs
```

Рис. 1.3. Результаты построения начального дерева решений

В следующем разделе отчета приводятся характеристики сконструированного классификатора, оцениваемые на обучающей выборке. Здесь мы видим, что построенное дерево решений имеет 9 веток (size = 9), а ошибка классификации наблюдается на 5 объектах, что составляет 6,8%.

В завершающей части отчета дается таблица с детальным разбором результатов классификации. Исходя из данных этой таблицы, можно сказать, что из 1-го класса («здоровая почка») правильно классифицируется 20 объектов, а 2 объекта ошибочно относятся к классу 2; среди объектов 2-го класса («множественные кисты») 35 диагностируются правильно и 2 ошибочно признаются здоровыми; все объекты 3-го класса («гидронефроз») классифицируются правильно за исключением одного объекта, попадающего в класс № 2.

В заключение система See5 выдает сообщение о затраченном на решение времени. В нашем случае оно составило 0,5 с. Здесь надо отметить очень высокую скорость работы алгоритма See5, позволяющую оперативно обрабатывать высокоразмерные массивы информации, содержащие тысячи и десятки тысяч записей.

Можно еще более подробно разобрать результаты нашей классификации. Для этого нажмем в главном окне See5 кнопку **Cross-Reference** (перекрестная ссылка). Система выдаст окно, в левой половине которого нарисовано построенное дерево решений, а в правой половине перечисляются объекты, попавшие на ту или иную ветвь дерева. Чтобы выделить интересующую ветвь, нужно щелкнуть по ней левой кнопкой мыши (справа от ветви появится темный круг – на рис. 1.4 на него указывает стрелка). Кроме того, если щелкнуть мышью по номеру какого-либо объекта из правого поля, то система выдаст еще одно окно с именем **Case**, в котором приводятся значения признаков и выделенного объекта. В случае, показанном на рисунке 1.3, нас заинтересовала ветвь (**Index** <= 0.69 и **Age** <= 43), на которой находится 10 объектов из 1-го класса и 1 объект из 2-го класса.

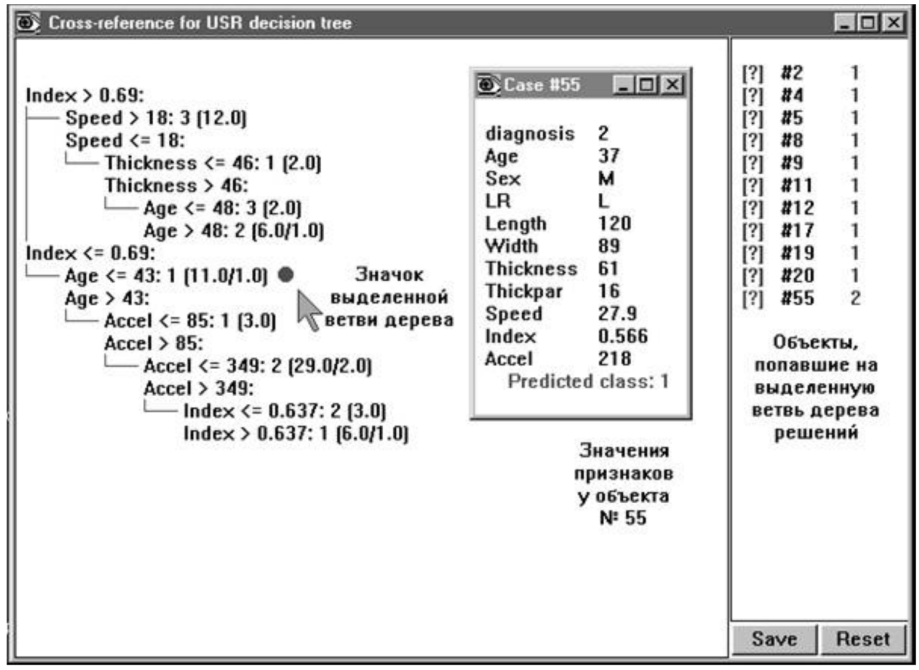


Рис. 1.4. Отображение результатов классификации в окне перекрестных ссылок

Преобразование дерева решений в набор правил

В ряде случаев полученное дерево решений может оказаться слишком сложным для восприятия. Например, при решении задач высокой размерности для неоднородных данных дерево нередко получается кустистым и довольно запутанным. Вместо того чтобы «ползать» по каждой полученной ветке, в системе See5 предусмотрена возможность преобразования дерева решений в набор правил if-then. Для этого требуется вызвать окно диалога для заданий опций конструируемого алгоритма (рис. 1.2) и поставить флажок в поле **Rulesets** (набор правил). После проведения такой операции система добавляет в окно отчета список правил, соответствующих рассчитанному дереву решений. Применительно к рассматриваемым данным по ультразвуковой диагностике это будет следующий список (табл. 1.2).

Таблица 1.2. Список выделенных правил

Rule 1: (cover 11) Age <= 43 Index <= 0.69 -> class 1 [0.846] Rule 2: (cover 10) Speed > 19 Index <= 0.69 Accel > 310 -> class 1 [0.750] Rule 3: (cover 14) LR = R Speed > 19 Index <= 0.69 -> class 1 [0.625]	Rule 4: (cover 5) Age <= 63 Speed <= 18 Index > 0.69 -> class 1 [0.429] Rule 5: (cover 15) Age > 43 LR = L Index <= 0.69 Accel <= 310 -> class 2 [0.941] Rule 6: (cover 15) Age > 43 Speed <= 19 Index <= 0.69 -> class 2 [0.941]	Rule 7: (cover 8) Age > 63 Speed <= 18 -> class 2 [0.800] Rule 8: (cover 17) Age > 43 Length <= 108 Index <= 0.69 -> class 2 [0.789] Rule 9: (cover 12) Speed > 18 Index > 0.69 -> class 3 [0.929]

Каждое правило состоит из следующих фрагментов:

- номер правила;
- количество объектов обучающей выборки, попадающих под действие правила (cover «n»);
- одно или несколько элементарных логических событий, входящих в состав правила (сложного логического высказывания);
- номер класса, которому соответствует данное правило;
- величина, принимающая значение от 0 до 1, которая выражает степень доверия к правилу (характеристика точности правила).

Для более детального рассмотрения множества правил, подобно тому как это делалось с деревом решений, можно обратиться к окну перекрестных ссылок (**Cross-Reference**).

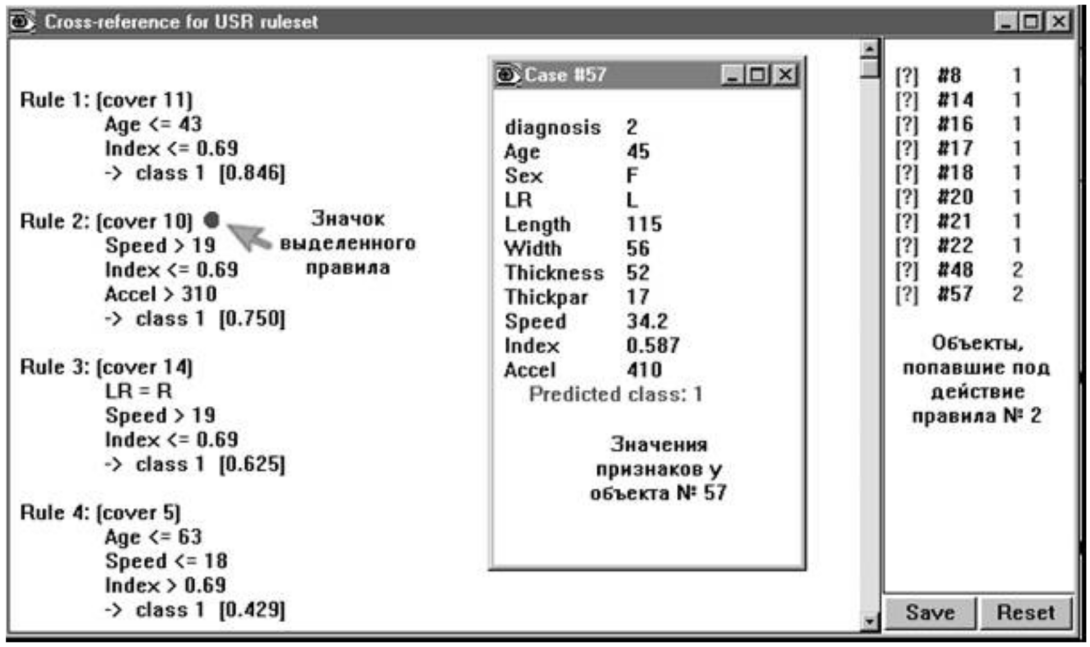


Рис. 1.5. Детальный разбор полученных правил в окне перекрестных ссылок

В целом, как уже говорилось, результаты в виде набора правил являются более простыми и понятными, чем в виде деревьев решений. Каждое правило ясным образом описывает связь между набором значений признаков и идентификатором класса. Более того, количество правил, сгенерированных из дерева решений, нередко оказывается несколько меньшим, чем число веток на дереве, а результат может оказаться более точным (в нашем случае этого, правда, не произошло). Вместе с тем для больших баз данных генерация множества правил требует ощутимых временных затрат.

1.3. Усиление решения (Boosting)

Идея усиления решения заключается в конструировании не одного, как в рассмотренном выше случае, а сразу нескольких деревьев решений. При этом главное требование к таким деревьям решений заключается в том, чтобы они как можно меньше дублировали друг друга. В системе See5 данная идея реализуется следующим образом.

На первом шаге конструируется начальное дерево решений (такое дерево применительно к данным по ультразвуковой диагностике почек было рассмотрено выше). Как следует из представленных результатов, классификатор, построенный на основе начального дерева, дает ошибки на некоторых объектах. Так, в нашем случае наблюдается 5 ошибок на 74 объектах обучающей выборки (рис. 1.6).

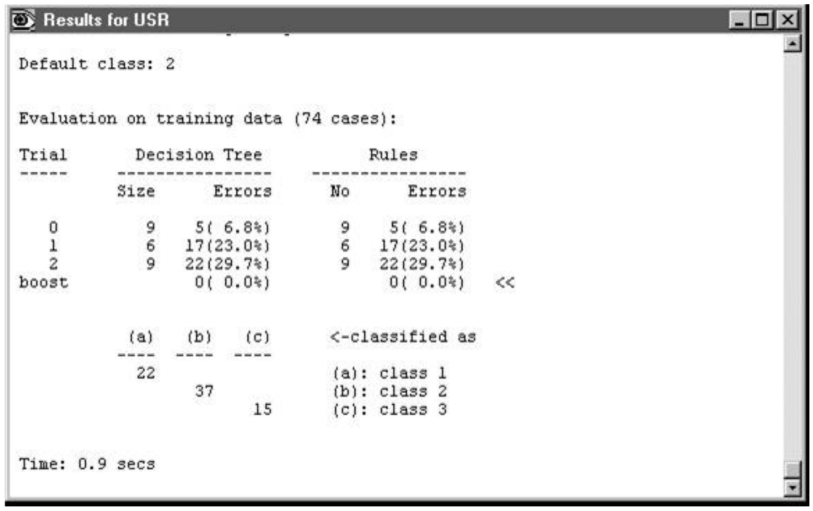


Рис. 1.6. Окно отчета о результатах построения трех деревьев решений

На втором шаге при конструировании следующего дерева делается попытка избежать ранее сделанных ошибок. Следствием такой попытки считается существенное отличие второго дерева от начального. Полученное дерево также будет приводить к ошибочным решениям, но уже на других объектах. На следующем шаге работы алгоритма очередное дерево строится с учетом ошибок всех предыдущих деревьев решений.

Для запуска процесса усиления решения требуется установить флажок **Boost** в диалоговом окне для задания опций работы алгоритма (рис. 1.2). Кроме того, в этом же окне нужно задать общее число строящихся деревьев решений. Это число проставляется в поле **trials**.

Понятно, что построение множества деревьев решений требует дополнительного времени. Но временные издержки способны вполне окупиться – точность классификации, как правило, значительно повышается.

Разработчики See5 утверждают, что при использовании 10 деревьев решений ошибки классификации снижаются в среднем на 25%. Посмотрим, как это будет выглядеть на наших числовых данных. Установим флажок **Boost** и в поле **trials** запишем цифру 3 (попробуем построить 3 дерева решений). Нажимаем **OK** и получаем окно отчета с информацией о результатах решения (рис. 1.6).

Как следует из отчета, второе дерево решений классифицирует данные с ошибкой 23%, а для третьего дерева эта ошибка составляет 29,7% (вообще говоря, нумерация деревьев начинается с цифры 0). Но все три дерева решений вместе классифицируют данные без ошибок (запись в строке boost). Для достижения такого безошибочного результата, как видно из отчета, потребовалось использование 9 + 6 + 9 = 24 правил.

1.4. Использование правил для принятия решений

Построенное множество правил применяется для принятия решения о принадлежности того или иного объекта какому-либо классу. При этом бывают ситуации, когда один и тот же объект попадает под действие сразу нескольких правил, в том числе правил, описывающих разные классы. Подобные внутренние конфликты могут быть разрешены двумя способами. В первом способе предпочтение отдается одному правилу, имеющему более высокую степень доверия (более высокую точность). Второй способ связан с обобщением результатов разных правил для принятия окончательного решения.

В системе See5 принят второй способ – каждое сработавшее правило подает голос для отнесения какого-либо объекта к изучаемым классам. Голоса суммируются с весами, равными вычисленным степеням доверия, и объект считается принадлежащим к классу, для которого набирается наибольшая взвешенная сумма голосов.

1.5. Смягчение порогов

В системе See5 предусмотрена еще одна возможность улучшения качества классификации. Но на сей раз эта возможность касается не столько точности результатов, сколько повышения их устойчивости к возможным флуктуациям значений признаков. Она связана с введением нечетких (мягких) порогов, на сравнении с которыми основывается выбор той или иной ветви дерева решений.

В диалоговом окне для задания параметров алгоритма See5 (рис. 1.2) имеется специальная опция для смягчения порогов. Это опция **fuzzy thresholds** (размытые пороги). При обращении к ней вместо одного порога задаются три значения – нижняя граница **LB**, верхняя граница **UB** и центральное значение **T**. Если значение переменной лежит ниже **LB** или выше **UB**, то исследуются соответствующие единственные ветви дерева. Если же значение переменной попадает между **LB** и **UB**, то исследуются одновременно две ветви дерева и выбирается наиболее правдоподобный результат классификации. Значения **LB**, **UB** и **T** система определяет автоматически.

Пример дерева решений с размытыми порогами приведен на рисунке 1.7. Здесь каждый порог представлен в виде <= **LB** (**T**) или >= **UB** (**T**)**.**

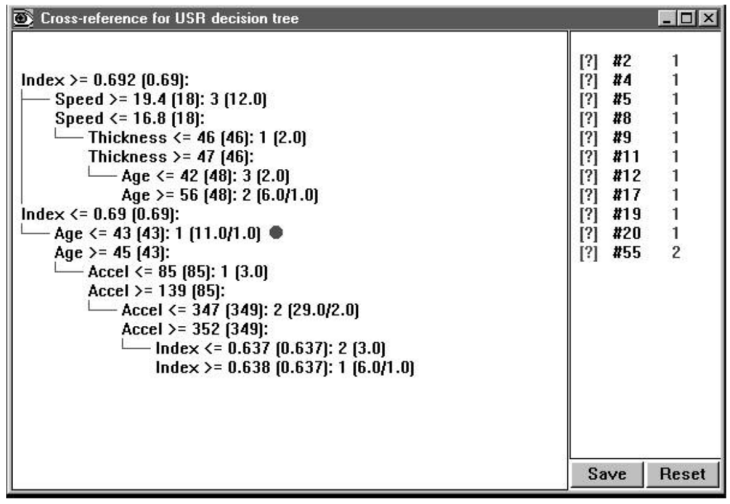


Рис. 1.7. Дерево решений с размытыми порогами

1.6. Дополнительные настройки алгоритма

В системе See5 предусмотрены опции для дополнительной настройки алгоритма построения деревьев решений. Они предназначены для пользователя, желающего поэкспериментировать и, возможно, попытаться улучшить найденный результат.

Во-первых, сюда относится опция **pruning CF** (ConFidence level – уровень доверия), предназначенная для отсечения статистически несостоятельных ветвей дерева решений. По умолчанию система выставляет значение уровня доверия 25% (рис. 1.2). Изменение этого значения приводит к соответствующему изменению размера дерева решений.

Во-вторых, на точность классификации может существенно влиять опция **Minimum ... cases**. В поле этой опции выставляется число, ограничивающее минимальное количество объектов на ветке дерева решений. Чем меньше будет это число, тем более «кустистым» станет дерево и тем точнее производится «подгонка» дерева под требуемую классификацию.

1.7. Перекрестная проверка

Для получения надежных оценок качества построенных классификаторов в системе See5 используется так называемая перекрестная проверка. Она осуществляется следующим образом.

Вся выборка объектов разбивается на «m» блоков примерно одного размера и с одинаковым распределением классов. Затем последовательно каждый блок используется как контрольный

набор объектов для тестирования классификатора, построенного на основе внешних для данного блока объектов. Число блоков вводится в поле **crossvalidate** диалогового окна для задания опций алгоритма конструирования классификатора. Результат перекрестной проверки отображается в нижней части окна отчета.

1.8. Выборка из больших наборов данных

Несмотря на высокое быстродействие системы See5, конструирование классификаторов на полном наборе исходных данных при их большом количестве может занимать довольно много времени. Это становится особенно заметно при использовании дополнительных опций алгоритма, например опции для усиления решения (**boosting**).

See5 имеет возможность работы не с полным набором данных, а с некоторой выборкой из исходного набора. Для этого предусмотрена специальная опция **Use sample of X%** (рис. 1.2). При использовании указанной опции осуществляются две операции. Во-первых, из исходного набора случайным образом извлекается X% объектов и на их основе конструируется классификатор. И, во-вторых, производится тестирование построенного классификатора на другой непересекающейся выборке объема X% (если $X < 50\%$) либо на всех оставшихся объектах (если $X \geq 50\%$).

При очередном обращении к опции **Use sample of X%** будет сделана новая случайная выборка из исходных данных, построен и протестирован новый классификатор. Но в системе See5 имеется также возможность зафиксировать выборку. Для этого необходимо поставить флажок в поле **Lock sample**.

На рисунке 1.8 приведен результат построения дерева решений на выборке половинного объема от исходных данных. На обучающей выборке достигнут неплохой эффект классификации – ошибка составляет всего 5,4%. Вместе с тем на контрольной выборке, объем которой равен половине объема исходных данных, процент правильной классификации резко падает до 35,1%. Это заставляет задуматься о том, насколько построенное дерево решений и соответствующие if-then правила отражают объективную реальность, и, скорее всего, продолжить поиск более устойчивого варианта решения.

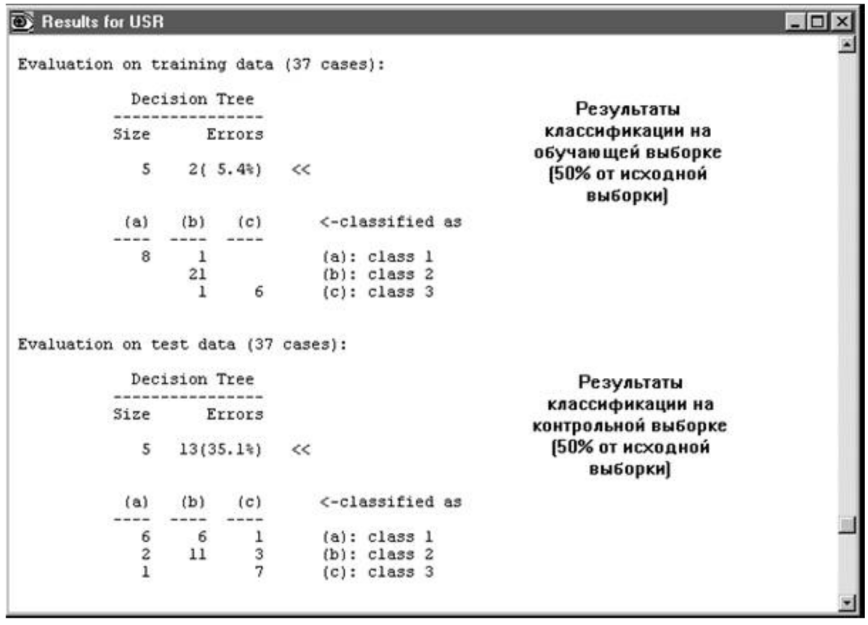


Рис. 1.8. Результаты классификации данных ультразвуковой диагностики на обучающей и контрольной выборках

1.9. Учет стоимости различных ошибок классификации

До сего момента, анализируя данные по ультразвуковой диагностике заболеваний, мы считали все виды ошибок классификации эквивалентными. Мы давали оценку качества построенного классификатора, просто подсчитывая общее число ошибок. Но в реальной жизни стоимость различных ошибок может быть разной. Например, если мы ошибочно сочтем здорового человека больным и направим его на дополнительное обследование, это будет не так страшно, как в случае ошибочного отнесения больного к группе здоровых. Соответственно, при оценке качества построенного дерева решений часто бывает необходимо вводить в анализ веса различных ошибок.

В системе See5 для учета стоимости различных ошибок классификации создается специальный файл *.costs. Он содержит строки следующего вида:

предсказанный класс, истинный класс: стоимость ошибки,

где «стоимость ошибки» – неотрицательное действительное число.

Число строк, характеризующих комбинации «предсказанный класс – истинный класс», в этом файле может быть любым. Если стоимость какой-либо ошибки не определена явно, то система назначает эту стоимость равной 1.

Предположим, что стоимость ошибочного отнесения больных почек к классу здоровых в нашем случае будет равна 10, а стоимость всех остальных видов ошибок равна 5. Тогда файл для учета различной стоимости ошибок **USR.costs** может выглядеть следующим образом:

```
| costs file for USR
1, 2: 10  | стоимость ошибочного отнесения класса 2 к классу 1
1, 3: 10  | стоимость ошибочного отнесения класса 3 к классу 1
2, 1: 5   | стоимость ошибочного отнесения класса 1 к классу 2
2, 3: 5   | стоимость ошибочного отнесения класса 3 к классу 2
3, 1: 5   | стоимость ошибочного отнесения класса 1 к классу 3
3, 2: 5   | стоимость ошибочного отнесения класса 2 к классу 3
```

Результаты обработки данных с разделением на обучающую и контрольную выборки (по 50%) и с учетом стоимости различных ошибок приведены на рисунке 1.9.

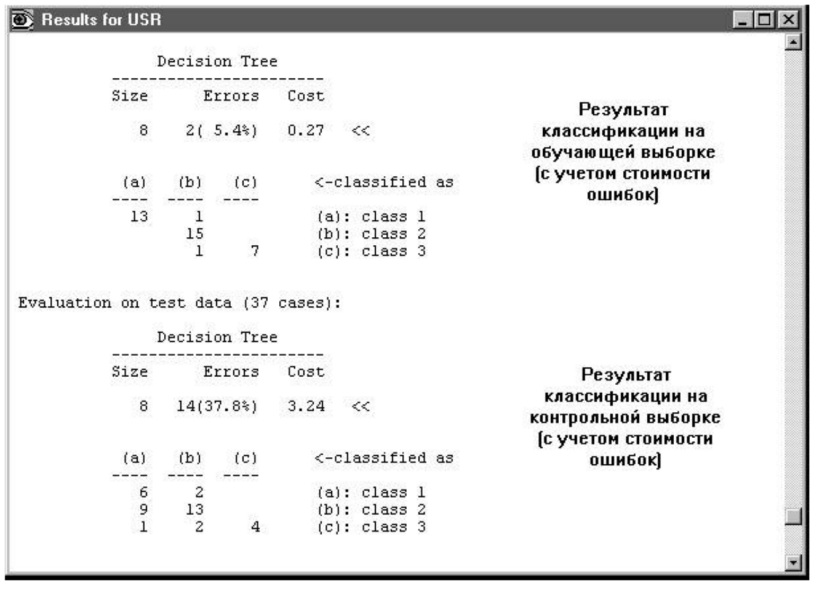


Рис. 1.9. Результаты классификации с учетом стоимости различных ошибок

Для редактирования файла стоимости различных ошибок классификации следует его вызвать из меню **Edit | costs file** и внести необходимые изменения в автоматически инициализированном редакторе WordPad. Можно исключить учет стоимости ошибок, если поставить флажок **Ignore costs file** в окне диалога для задания параметров алгоритма построения деревьев решений (рис. 1.2).

1.10. Использование классификаторов

После того как пользователь признает какой-либо вариант дерева решений удовлетворительным, ему предоставляется возможность испытать этот вариант в интерактивном режиме на новых данных. Нажимаем в главном окне See5 кнопку **Use Classifier** и тем самым активизируем алгоритм классификации данных, соответствующий самому последнему варианту дерева решений. На экран выводится специальное окно для ввода значений поочередно предъявляемых признаков (рис. 1.10). Количество таких признаков может быть разным, ведь в зависимости от ответов реализуется та или иная ветвь дерева решений. После ввода всех затребованных значений на экран выводится окно, в котором указываются предсказанный класс и уровень доверия к результату классификации (рис. 1.11).

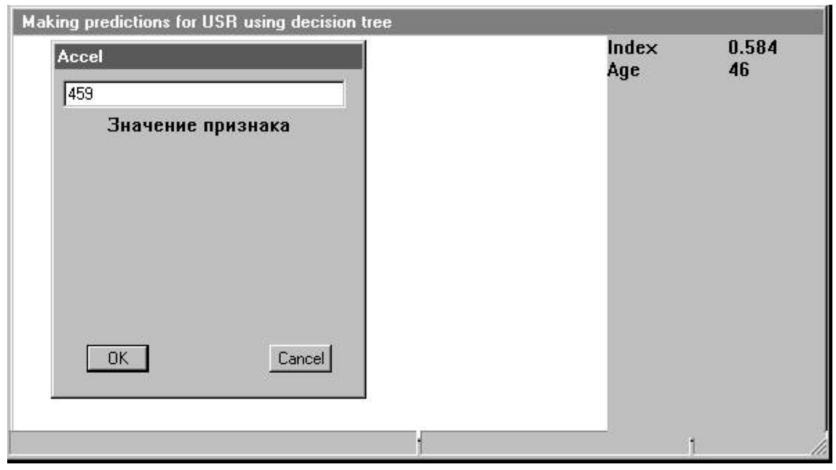


Рис. 1.10. Интерактивный режим классификации данных

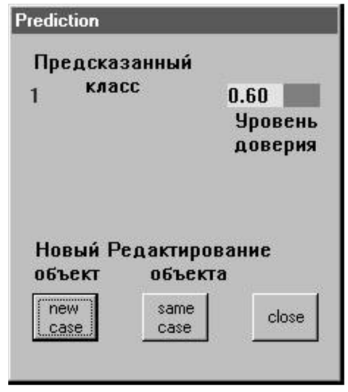


Рис. 1.11. Результат интерактивной классификации

1.11. Детальная проверка и сохранение результатов

Завершающая стадия работы с See5 обычно заключается в детальном просмотре результатов работы построенного классификатора в окне перекрестных ссылок. После нажатия соответствующей кнопки (**Cross-Reference**) на экране появляется диалоговое окно, в котором предлагается выбрать файл с данными для классификации (рис. 1.12). Это может быть исходный файл данных (в нашем случае **USR.data**), файл с тестовыми данными (**USR.test**) или файл, содержащий объекты с неизвестной классификацией (**USR.cases**).

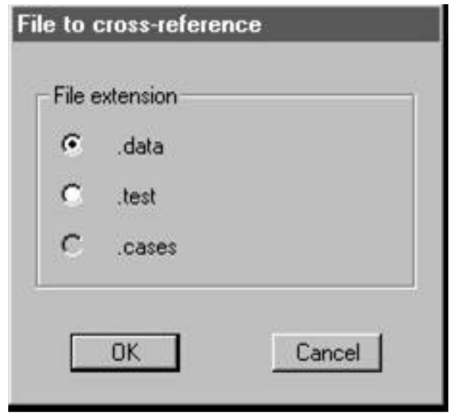


Рис. 1.12. Выбор файла данных для классификации

Выбрав требуемый файл, нажимаем **OK**. На экране появляется окно перекрестных ссылок, в левой половине которого сначала изображено полное дерево решений, а в правой представлен список объектов, подвергнутых классификации. Некоторые возможности работы с окном перекрестных ссылок обсуждались выше. Здесь остановимся еще на двух возможностях.

Первая заключается в возможности поэлементного просмотра для выбранного объекта ветки построенного дерева решения. Для этого нужно щелкнуть левой кнопкой мыши в правом поле окна перекрестных ссылок на требуемом объекте – в левом поле автоматически отобразится соответствующая ветка. Так, в случае, показанном на рисунке 1.13, для изучения был выбран объект № 4 (около него появился темный кружок). Как видим, с этим объектом соотносится достаточно короткая ветка решения (**Index ≤ 0,69 & Age > 43 & Accel ≤ 85**). Аналогичным образом можно разобрать результаты классификации всех других доступных объектов (нажатием кнопки **Reset** возвращается исходное изображение полного дерева решений).

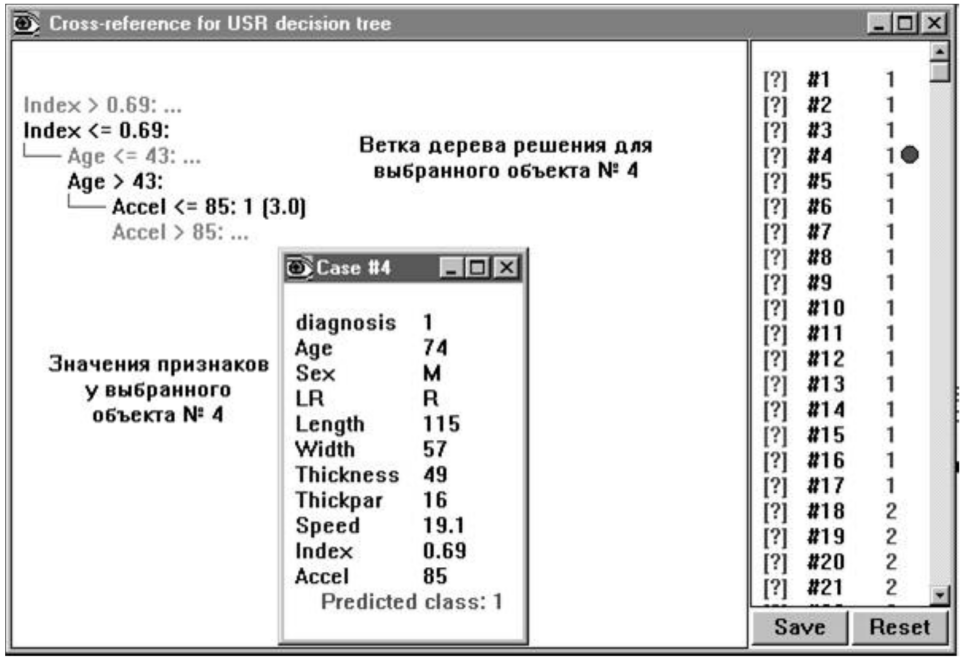


Рис. 1.13. Просмотр результатов классификации в окне перекрестных ссылок

Вторая возможность заключается в сохранении полученных результатов. Причем здесь существенным является *выборочное* сохранение. А именно после нажатия кнопки **Save**, расположенной в правом нижнем углу окна перекрестных ссылок, сохраняться в текстовом формате будут только результаты, относящиеся к текущему отображению дерева решений (целиком или его части).

2. WizWhy – система поиска логических правил в данных

Система WizWhy предприятия WizSoft (http://www.wizsoft.com) является современным представителем подхода, реализующего ограниченный перебор. Хотя авторы системы не раскрывают специфику алгоритма, положенного в основу работы WizWhy, вывод о наличии здесь ограниченного перебора был сделан по результатам тщательного тестирования системы (изучались результаты, зависимости времени их получения от числа анализируемых параметров и др.).

Алгоритмы ограниченного перебора были предложены в середине 1960-х гг. М. М. Бонгардом для поиска логических закономерностей в данных. С тех пор они продемонстрировали свою эффективность при решении множества задач из самых различных областей.

Эти алгоритмы вычисляют частоты комбинаций простых логических событий в подгруппах данных. Примеры простых логических событий: $X = a$; $X < a$; $X > a$; $a < X \le b$ и др., где X – какой-либо параметр, a и b – константы. Ограничением служит длина комбинации простых логических событий (у М. М. Бонгарда она была равна 3). На основании анализа вычисленных частот делается заключение о полезности той или иной комбинации для установления ассоциации в данных, для классификации, прогнозирования и пр.

Авторы системы WizWhy утверждают, что она автоматически извлекает из данных ***ВСЕ*** if-then правила. На самом деле это, конечно, не так. Во-первых, максимальная длина комбинации в правиле if-then в системе WizWhy равна 6, и, во-вторых, с самого начала работы алгоритма производится эвристический поиск простых логических событий, на которых потом строится весь дальнейший анализ. Тем не менее система WizWhy является на сегодняшний день одним из лидеров на рынке продуктов Data Mining. Это не лишено оснований. Система демонстрирует более высокие показатели при решении ряда практических задач, чем все остальные алгоритмы. Стоимость системы составляет около 4 тыс. долл., количество пользователей – приблизительно 30 тыс. Демонстрационная версия WizWhy ограничена только количеством анализируемых записей – 1000 объектов.

2.1. Общие свойства системы WizWhy

Авторы WizWhy акцентируют внимание на следующих общих свойствах системы:

- выявление ВСЕХ if-then правил;
- вычисление вероятности ошибки для каждого правила;
- определение наилучшей сегментации числовых переменных;
- вычисление прогностической силы каждого признака;
- обобщение полученных правил и зависимостей;
- выявление необычных феноменов в данных;
- использование обнаруженных правил для прогнозирования;
- выражение прогноза в виде списка релевантных правил;
- вычисление ошибки прогноза;
- прогноз с учетом стоимости ошибок.

В качестве достоинств WizWhy дополнительно отмечают такие:

- на прогнозы системы не влияют субъективные причины;
- пользователям системы не требуется специальных знаний в прикладной статистике;
- более точные и быстрые вычисления, чем у других методов Data Mining.

Для большей убедительности авторы WizWhy противопоставляют свою систему нейросетевому подходу и алгоритмам построения деревьев решений и утверждают, что WizWhy, обладая более высокими характеристиками, вытесняет другие программные продукты с рынка Data Mining.

2.2. Загрузка и управление данными

Первое, что нужно сделать при работе с WizWhy, – это загрузить анализируемый файл данных. Здесь имеется несколько возможностей:

- вы можете подготавливать и читать файлы ASCII;
- вы можете напрямую работать с файлами dBase (*.dbf), MS Access (*.mdb), Oracle и таблицами MS SQL;
- вы можете воспринимать наборы данных посредством ODBC (Open Database Connectivity);
- для начала работы с процедурой загрузки прежде всего следует обратиться к закладке **Basic Data** в окне диалога с именем текущего проекта (рис. 2.1). Здесь в поле **Open Data of Type** нужно указать тип загружаемых данных;
- для примера возьмем таблицу с данными по ультразвуковой диагностике почек в текстовом формате ASCII (разделителем колонок является знак табуляции, в первой строке таблицы данных записаны имена переменных). Укажем требуемый тип данных и в появившемся окне диалога выберем файл USR.txt – на экран выводится окно диалога системы WizWhy для редактирования и преобразования текстовых файлов (рис. 2.2).

В поле **Record Type** (тип записи) устанавливаем переключатель в положение **Delimited** (данные с разделителем) и проставляем флажок в позиции **First record for fields names**, говорящий о том, что имена переменных располагаются в первой строке таблицы данных. В поле **Field Delimiters** (разделитель) ставим флажок в позиции **Tab** (знак табуляции). Нажимаем кнопку **Parse**, после чего система производит автоматический грамматический разбор наших данных. Просматриваем результаты этого разбора и при необходимости вносим коррективы – в поле **Column (field)** предоставляются возможности для изменения имен и типов переменных, а также для отказа от импорта каких-либо колонок. Нажимаем **OK**. Система импортирует данные для дальнейшей обработки, что отражается в диалоговом окне для управления данными **Basic Data** (рис. 2.3).

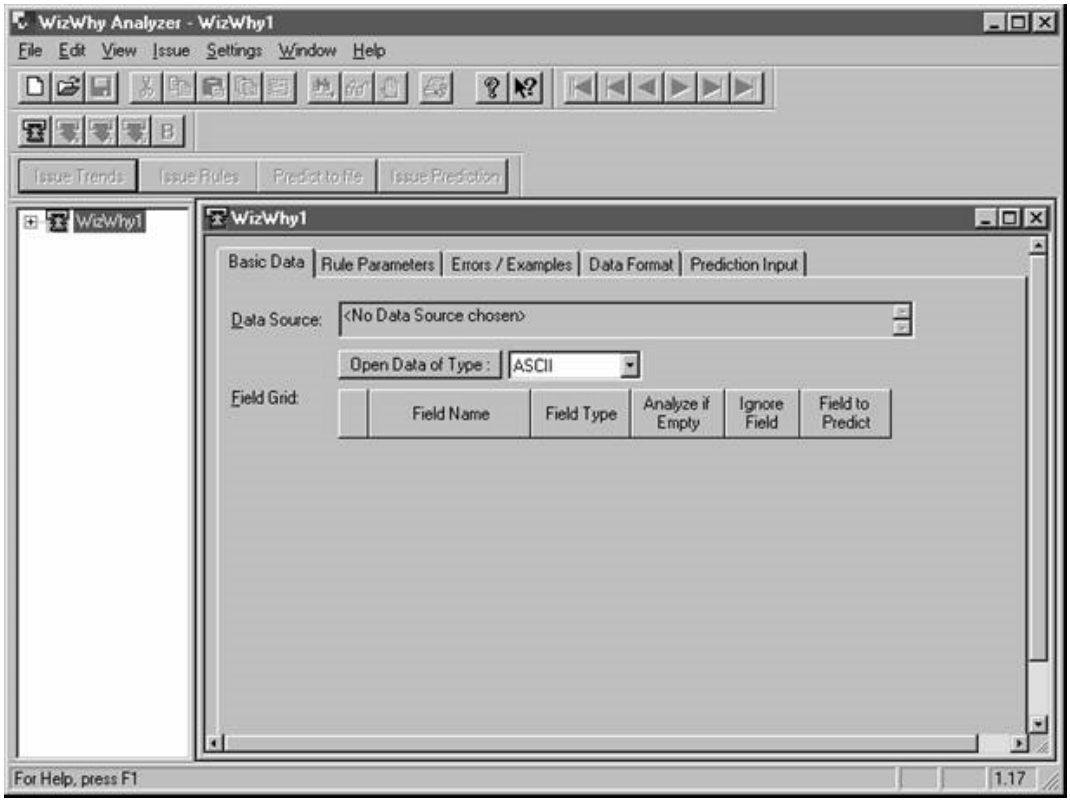


Рис. 2.1. Начало работы с системой WizWhy

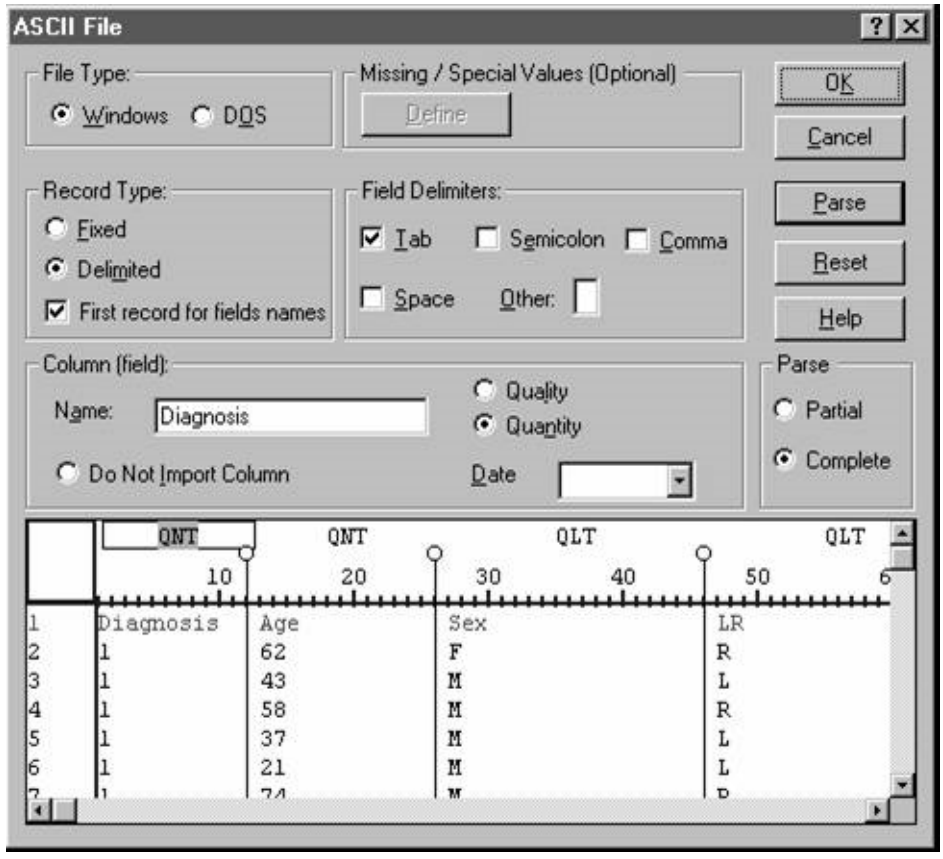


Рис. 2.2. Диалоговое окно для чтения данных в текстовом формате

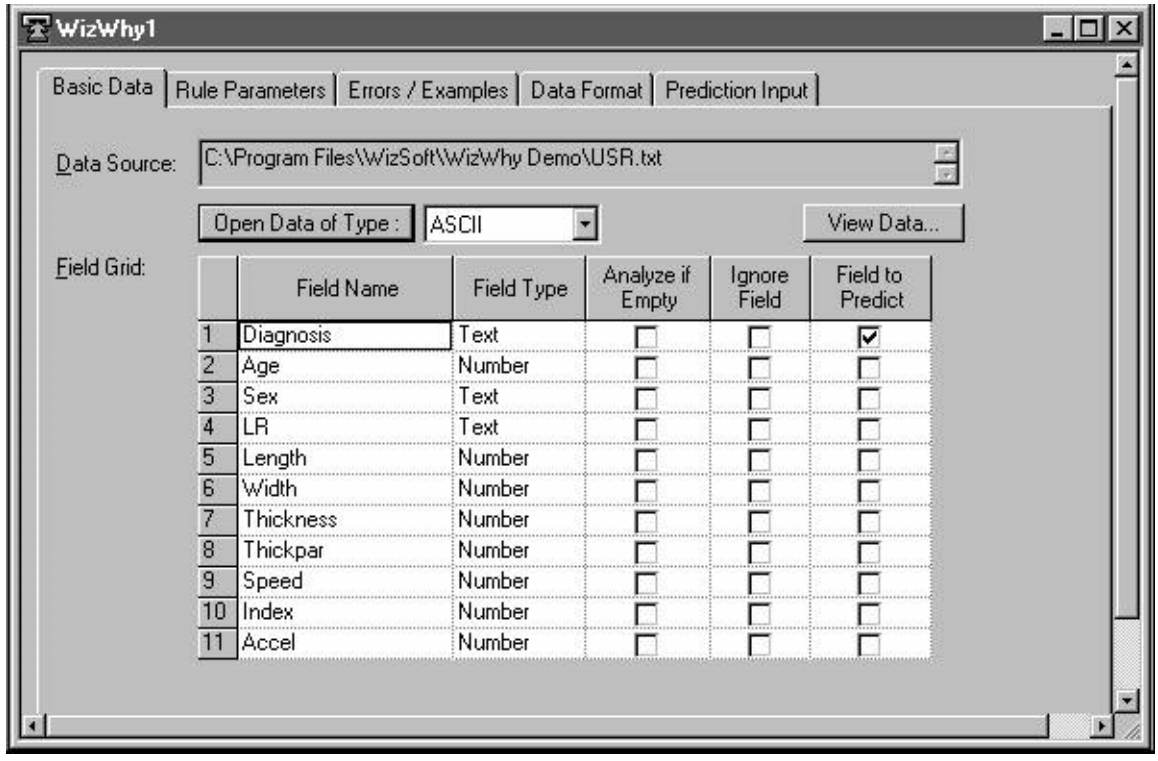


Рис. 2.3. Окно диалога для управления данными

В поле **Data Source** указываются местоположение и имя файла, из которого были импортированы данные. Кнопка **View Data** предназначена для вызова окна для просмотра загруженных

данных (в нем демонстрируется 100 первых строк таблицы данных). В поле **Field Grid** отображаются имена и типы введенных переменных и предоставляются возможности проведения следующих операций:

- назначение целевой или так называемой зависимой (dependent) переменной. Это переменная, значения которой будут связываться с помощью if-then правил со значениями так называемых независимых (independent) переменных. В нашем случае такой целевой переменной является **Diagnosis** – выставляем флажок в соответствующей позиции колонки **Field to Predict**;
- модификация переменных. В колонке **Field Name** можно редактировать имена переменных. Для этого нужно щелкнуть на соответствующей позиции и ввести новое имя. Кроме того, в позициях колонки **Field Type** можно изменять тип переменной. Например, заменить тип **Text** (текстовый, номинальный) на **Number** (количественный) или **Date** (дата) в формате День-Месяц-Год (Год-Месяц-День) и т. п. Здесь заметим, что в зависимости от выбранного типа данных в дальнейшем к переменной применяются различные процедуры обработки.

В системе WizWhy предусмотрен также случай, когда пропуски в таблице данных (пустые ячейки) представляют собой самостоятельные информативные события. Для учета подобных пропусков в значениях какой-либо переменной ставится флажок против нее в колонке **Analyze if Empty**. В свою очередь, если имеется необходимость исключить переменную из анализа, нужно выставить флажок в колонке **Ignore Field**.

В нашем примере текстовый формат имеют три переменные – целевой признак **Diagnosis**, признак **Sex** (пол пациента) и признак **LR** (левая или правая почка). Остальные переменные **Age** (возраст), **Length** (длина почки), **Width** (ширина почки), **Thickness** (толщина почки), **Thickpar** (толщина паренхимы), **Speed** (средняя скорость кровотока), **Index** (индекс резистентности) и **Accel** (ускорение артериального потока в систолу) являются количественными.

2.3. Задание параметров процедуры поиска правил

В системе WizWhy целевой признак разделяет все множество объектов на две части. Это делается следующим образом.

Если целевая переменная является текстовой (номинальной), WizWhy просматривает все объекты (записи) и отбирает те из них, для которых целевая переменная имеет выбранное значение. Отобранные таким образом объекты составляют первую группу. Правила, характерные для данной группы, называются if-then правилами. Оставшиеся объекты составляют вторую группу, и для этой группы характерные правила обозначаются как if-then-NOT правила.

Если целевой признак является количественным, пользователь должен указать область значений этого признака. Правила if-then будут определяться для этой указанной области. В свою очередь, if-then-NOT правила будут описывать объекты, не попавшие в выделенную область.

В рассматриваемом нами практическом примере целевой признак **Diagnosis** номинальный. Он принимает три значения: 1 – в классе «здоровая почка»; 2 – в классе «множественные кисты» и 3 – в классе «гидронефроз». Будем искать в данных if-then правила для объектов с диагнозом «множественные кисты». Для этого с помощью закладки **Rule Parameters** (параметры правил) войдем в соответствующее окно диалога и в поле **Predicted Value** проставим значение «2» (рис. 2.4).

После задания области значений целевой переменной или, как в нашем случае, ее одного значения система WizWhy читает данные и вычисляет простые статистики, которые могут быть использованы в дальнейшем анализе. Так, например, справа от поля **Predicted Value** система выводит значение частоты, с которой в анализируемых данных встречается значение Diagnosis = 2. Как указывают авторы, чтение больших наборов данных способно занимать

много времени. Пользователь может прекратить процесс чтения, нажав кнопку **Cancel** на специальной панели. В этом случае дальнейшему исследованию подвергается только та информация, которая была прочитана. Но при желании процесс чтения данных можно повторить.

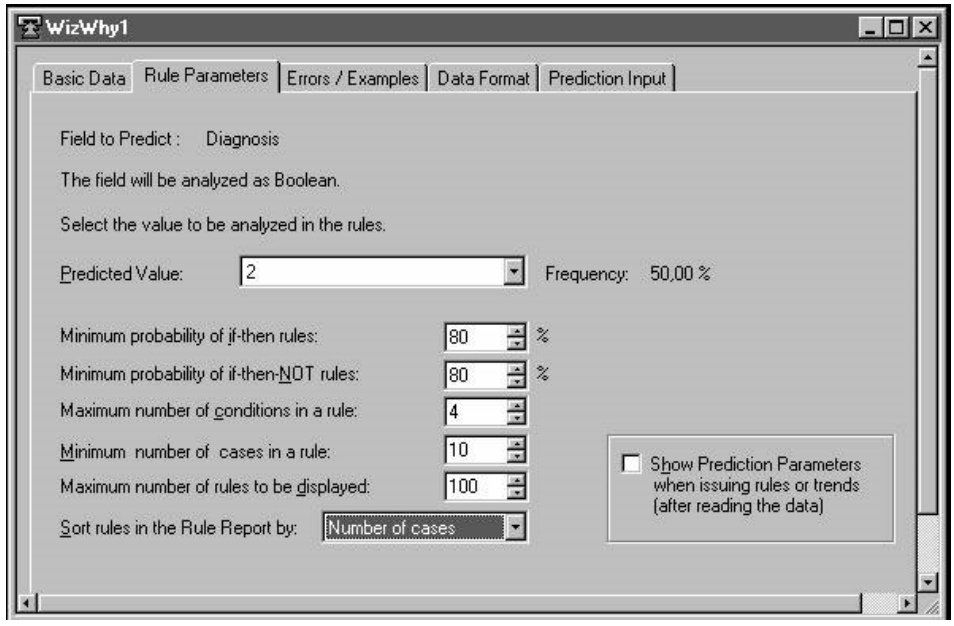


Рис. 2.4. Окно диалога для задания параметров процедуры поиска логических правил

Следующим шагом является задание собственно параметров правил, поиск которых будет осуществляться в прочитанных данных. Сюда прежде всего относятся **Minimum probability of if-then rules** (минимальная вероятность if-then правил) и **Minimum probability of if-then-NOT rules** (минимальная вероятность if-then-NOT правил). Эти параметры есть не что иное, как **точность** правила, охарактеризованная в предудыщей главе. Проставим в соответствующих полях окна диалога одинаковые значения указанных вероятностей 80%. Это означает, что в системе WizWhy формулируется требование обнаруживать правила, которые будут ошибаться не более чем в 20% случаев (имеются в виду ошибки на анализируемой выборке).

В принципе можно задавать любые значения минимальных вероятностей от 0 до 100%. Но следует хорошо представлять, что, задав слишком низкий уровень точности, мы получим большое количество правил, среди которых будет много малоинформативных компонентов. В свою очередь, выставив требование 100%, мы, скорее всего, ничего не получим вообще.

Еще одним важным параметром служит **Maximum number of conditions in a rule** (максимальное число условий в правиле). Это максимальное количество элементарных логических событий в одном правиле. Хотя авторы системы ничего не говорят о предельном значении данного параметра, установлено, что оно равно 6.

Следущим параметром, который необходимо задать для работы процедуры поиска правил, является **Minimum number of cases in a rule** (минимальное число объектов в правиле). Выставим здесь значение 10, обозначив тем самым наше желание обнаружить в данных правила, «покрывающие» не менее 10 объектов. Нижний предел составляет 4 объекта.

Последние операции в работе с рассматриваемым окном диалога касаются способов выдачи результатов. Во-первых, нужно ввести параметр **Maximum number of rules to be displayed** (максимальное количество отображаемых правил). Этот параметр не влияет на работу процедуры поиска правил. Он предназначен только для ограничения количества правил, выдаваемых в отчет (**Rule Report**). Далее следует указать способ сортировки правил в отчете (по уровню значимости **Significance Level**, по точности **% Probability**, по количеству объектов **No. of Cases in a Rule**).

Наконец, в правом нижнем углу окна диалога для задания параметров процедуры поиска правил можно поставить флажок, если имеется желание перед стартом процедуры дополнительно просматривать и корректировать ее параметры. Полностью подготовленное окно диалога для нашего примера по ультразвуковой диагностике почек приведено на рисунке 2.4.

2.4. Работа с окном диалога Ошибки/Примеры (Errors/Examples)

Окно диалога Ошибки/Примеры показано на рисунке 2.5. Оно разделено на два поля: «Стоимость ошибок прогноза» (**Prediction Error Costs**) и «Представить примеры» (**Present examples where**).

В поле «Стоимость ошибок прогноза» требуется ввести соответствующие значения по отдельности для двух видов ошибок: пропуска объекта (**Cost of a miss**) и ложной тревоги (**Cost of a false alarm**). По умолчанию эти значения равны 1. Но, как обсуждалось в предыдущем разделе, учет различной стоимости указанных ошибок может оказаться весьма ценным при решении практических задач.

В поле «Представить примеры» можно выразить желание просмотреть примеры работы выявляемых правил. Если поставить флажок в позиции **Rule is in effect**, то система будет формировать в отчете для каждого правила список номеров объектов, для которых правило не ошибается. Длина списка ограничивается заданным числом. Соответственно, флажок в позиции **Rule is not in effect** запрашивает у системы выдачу списка номеров объектов, на которых какое-либо правило работает с ошибкой.

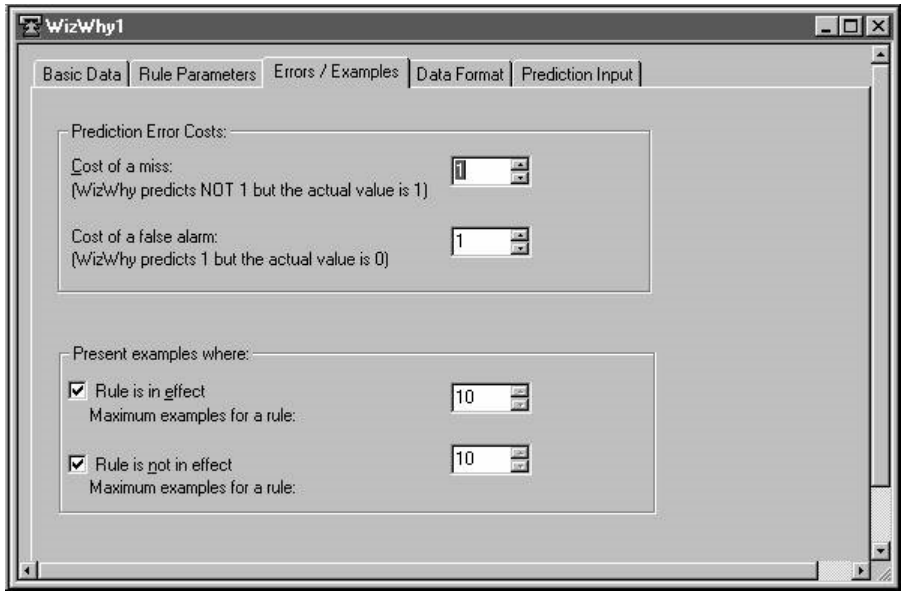


Рис. 2.5. Окно диалога Ошибки/Примеры

2.5. Работа с другими окнами диалога

Окно диалога **Data Format** предназначено для задания и корректировки формата информации, с которой работает WizWhy (рис. 2.6). Прежде всего сюда относится формат данных. В поле **Number and Currency Format** имеется возможность задавать количество цифр и виды разделителей в числовых и денежных данных, а в поле **Data Format** – выбирать формат для записи дат.

Кроме того, в нижней части окна диалога предусмотрены опции, выбор которых определяет место выдачи отчета о результатах работы системы (на принтер, на экран, в текстовый файл, в RTF-файл). В поле **Subheading** заносится подзаголовок отчета. Нажатием кнопки **Font** в правом нижнем углу вызывается окно диалога для выбора используемых шрифтов.

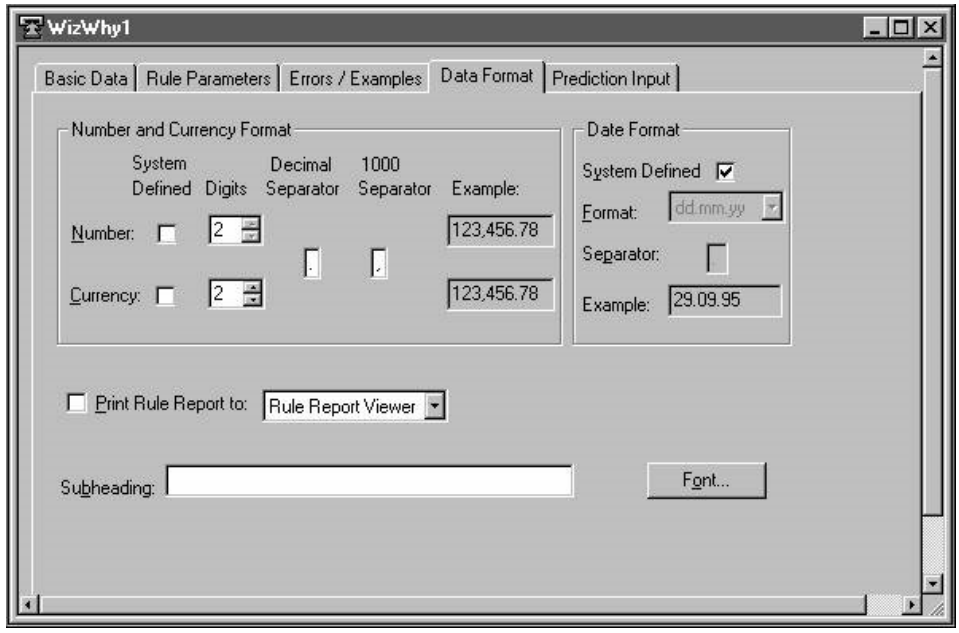


Рис. 2.6. Окно диалога для изменения формата информации

Последнее окно диалога – **Prediction Input** предназначено для ввода, просмотра и коррекции внешних данных, на которых требуется проверить действие найденных правил. Оно изображено на рисунке 2.7. Работа с этим окном аналогична работе с уже рассмотренным окном диалога **Basic Data**.

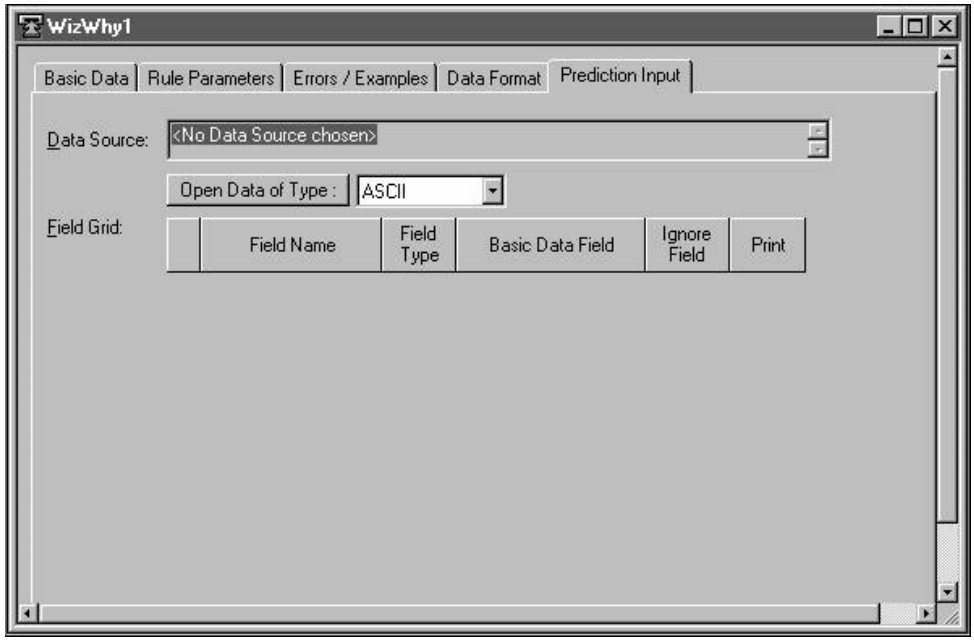


Рис. 2.7. Окно диалога для ввода внешних данных

2.6. Результаты работы системы

После внесения необходимой информации в рассмотренные выше окна диалога можно приступить к поиску правил в загруженных данных. Для этого нужно нажать кнопку **Issue Rules** (выдача правил) – система WizWhy выдает три отчета.

1. Отчет о правилах (**Rule Report**), в котором перечисляются обнаруженные правила с указанием их характеристик.

2. Отчет о трендах (**Trend Report**), в котором представляются результаты сегментации отдельных признаков.
3. Отчет о неожиданных правилах (**Unexpected Rule Report**).

Рассмотрим указанные отчеты более подробно.

Отчет о правилах

Отчет о правилах размещен в трех окнах (рис. 2.8).

1. Левое окно – Список правил (**Rule List**).

2. Правое верхнее окно – Содержание записи в деталях (**Record Details Grid**).

3. Правое нижнее окно – Индекс признака (**Field Index**).

Список правил

Список правил предваряется информацией о заданных параметрах поиска. Здесь, на примере данных по ультразвуковой диагностике почек, как видим, говорится, что общее число обработанных записей (объектов) составляет 74, минимальные вероятности правил if-then и if-then-NOT равны по 0,8, минимальное количество объектов для правил – 2. Затем подтверждается, что правила находятся для переменной **Diagnosis**, конкретно для значения этой переменной, равного 2. Также указывается, что стоимости ошибок в виде пропусков и ложных тревог составляют 1, а средняя вероятность (априорная вероятность) прогнозируемого значения переменной равна 0,5.

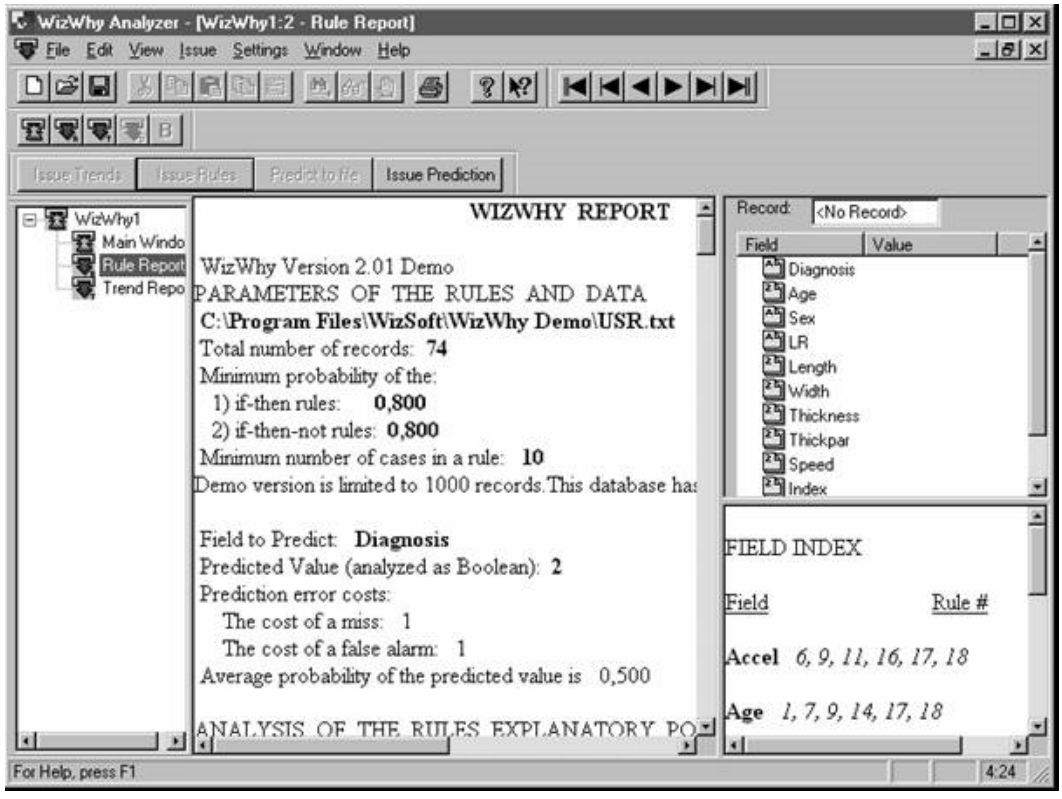


Рис. 2.8. Общий вид отчета о правилах в системе WizWhy

Далее система выдает следующий блок общей информации об обнаруженных правилах:

ANALYSIS OF THE RULES EXPLANATORY POWER

Decision point: Predict 2 when conclusive probability is more than **0,460**

Number of misses: 2

Number of false alarms: 5

Total number of errors: 7

Total cost of errors: 7

Success rate when predicting 2: 0,868

Success rate when predicting NOT2: 0,931

Number of records with no relevant rules: 7

Average cost (per record): 0,104

Expected average cost (per record): 0,500

Improvement Factor: 4,786

Из приведенного блока можно почерпнуть сведения о значениях некоторых служебных параметров – **Decision point** (точка решения), **Average cost** (средние потери на запись), **Expected average cost** (ожидаемые средние потери) и **Improvement Factor** (выигрыш), представляющий собой отношение ожидаемых средних потерь к реальным потерям на запись. Кроме того, в блоке содержатся сведения о прогнозирующей способности всей совокупности обнаруженных правил – количество пропусков при прогнозировании (**Number of misses**), число ложных тревог (**Number of false alarms**), общее количество ошибок (**Total number of errors**), общие потери (**Total cost of errors**), вероятность успешного прогнозирования для класса 2 (**Success rate when predicting 2**), вероятность успешного прогнозирования альтернативного класса (**Success rate when predicting NOT 2**) и количество объектов, не охваченных выделенными правилами (**Number of records with no relevant rules**).

Список правил состоит из правил, упорядоченных по заданному критерию (в нашем случае по числу объектов, описываемых правилом). В данных по ультразвуковой диагностике почек при установленных параметрах процедуры система WizWhy обнаружила 19 правил. Они приведены в таблице 2.1.

Таблица 2.1. Таблица if-then правил, обнаруженных WizWhy

1) *If* **Age** *is* **21,00 ... 43,00** (average = **34,89**) Then **Diagnosis** *is not* **2** Rule's probability: **0,947** The rule exists in **18** records. Significance Level: Error probability < 0,1 Positive Examples (records' serial numbers): 2, 4, 5, 8, 9, 11, 12, 17, 19, 20 Negative Examples (records' serial numbers): 55 2) *If* **Index** *is* **0,70 ... 0,80** (average = **0,73**) Then **Diagnosis** *is not* **2** Rule's probability: **0,833** The rule exists in **15** records. Significance Level: Error probability < 0,2 Positive Examples (records' serial numbers): 1, 3, 60, 61, 62, 63, 64, 65, 66, 67 Negative Examples (records' serial numbers): 28, 32, 33 3) *If* Width *is* 39,00 ... 53,00 (average = 48,00) Then **Diagnosis** *is not* **2** Rule's probability: **0,813**	11) *If* **Width** *is* **56,00 ... 89,00** (average = **65,23**) and Accel *is* 210,00 ... 242,00 (average = 224,92) Then Diagnosis *is* 2 Rule's probability: **0,846** The rule exists in **11** records. Significance Level: Error probability < 0,2 Positive Examples (records' serial numbers): 25, 28, 36, 41, 42, 44, 50, 53, 55, 56 Negative Examples (records' serial numbers): 60, 67 12) *If* **Sex** *is* **M** and **LR** *is* **R** and **Speed** *is* **16,30 ... 39,30** (average = **21,47**) Then **Diagnosis** *is not* **2** Rule's probability: **0,846** The rule exists in **11** records. Significance Level: Error probability < 0,2 Positive Examples (records' serial numbers): 3, 6, 7, 8, 11, 17, 61, 65, 66, 67 Negative Examples (records' serial numbers): 43, 44 13) *If* **Speed** *is* **2,30 ... 15,40** (average = **13,28**)

The rule exists in **13** records.
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers): 1, 2, 4, 9, 10, 11, 17, 19, 20, 62
Negative Examples (records' serial numbers): 48, 49, 58

4) *If* Speed *is* 16,30 ... 41,50 (average = 23,97)
and **Index** *is* **0,70 ... 0,80** (average = **0,73**)
Then
Diagnosis *is not* **2**
Rule's probability: **0,929**
The rule exists in **13** records.
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers): 3, 61, 62, 63, 64, 65, 66, 67, 68, 70
Negative Examples (records' serial numbers): 33

5) *If* **Index** *is* **0,65 ... 0,67** (average = **0,66**)
Then
Diagnosis *is* 2
Rule's probability: **0,800**
The rule exists in **12** records.
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers): 25, 26, 29, 31, 34, 37, 38, 47, 49, 50
Negative Examples (records' serial numbers): 7, 11, 17

6) *If* Accel *is* 210,00 ... 242,00 (average = 224,47)
Then
Diagnosis *is* 2
Rule's probability: **0,800**
The rule exists in **12** records.
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers): 25, 28, 36, 41, 42, 44, 50, 53, 55, 56
Negative Examples (records' serial numbers): 60, 67, 69

7) *If* **Age** *is* **48,00 ... 70,00** (average = **62,77**)
and **Index** *is* **0,65 ... 0,67** (average = **0,66**)
Then
Diagnosis *is* 2
Rule's probability: **0,923**
The rule exists in **12** records.
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers): 25, 26, 29, 31, 34, 37, 38, 47, 49, 50
Negative Examples (records' serial numbers): 7

8) *If* **Sex** *is* **F**
and **LR** *is* **L**
and **Width** *is* **56,00 ... 77,00** (average = **63,00**)
Then
Diagnosis *is* **2**
Rule's probability: **0,833**
The rule exists in **10** records.
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers): 23, 24, 25, 26, 27, 28, 29, 30, 31, 32
Negative Examples (records' serial numbers): 1, 60

14) *If* **Age** *is* **46,00 ... 70,00** (average = **62,00**)
and **Speed** *is* **2,30 ... 15,40** (average = **13,17**)
Then
Diagnosis *is* **2**
Rule's probability: **0,909**
The rule exists in **10** records.
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers): 23, 24, 25, 26, 27, 28, 29, 30, 31, 32
Negative Examples (records' serial numbers): 1

15) *If* **Width** *is* **54,00 ... 87,00** (average = **64,82**)
and **Speed** *is* **2,30 ... 15,40** (average = **13,28**)
Then
Diagnosis *is* **2**
Rule's probability: **0,909**
The rule exists in **10** records.
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers): 23, 24, 25, 26, 27, 28, 29, 30, 31, 32
Negative Examples (records' serial numbers): 60

16) *If* **Speed** *is* **17,70 ... 43,30** (average = **25,84**)
and **Accel** *is* **210,00 ... 242,00** (average = **223,58**)
Then
Diagnosis *is* **2**
Rule's probability: **0,833**
The rule exists in **10** records.
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers): 36, 41, 42, 44, 50, 53, 55, 56, 58, 59
Negative Examples (records' serial numbers): 67, 69

17) *If* **Age** *is* **23,00 ... 43,00** (average = **36,30**)
and **Accel** *is* **255,00 ... 629,00** (average = **377,70**)
Then
Diagnosis *is not* **2**
Rule's probability: **1,000**
The rule exists in **10** records.
Significance Level: Error probability < 0,1

Then
Diagnosis *is* 2
Rule's probability: **0,857**
The rule exists in **12** records.
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
33, 40, 41, 42, 45, 46, 51, 52, 54, 56
Negative Examples (records' serial numbers):
14, 18

9) *If* **Age** *is* **47,00 ... 70,00** (average = **61,92**)
and Accel *is* 210,00 ... 242,00 (average = 226,08)
Then
Diagnosis *is* 2
Rule's probability: **0,917**
The rule exists in **11** records.
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
25, 28, 36, 41, 42, 44, 50, 53, 56, 58
Negative Examples (records' serial numbers):
69

10) *If* **Width** *is* **56,00 ... 72,00** (average = **62,42**)
and **Index** *is* **0,65 ... 0,67** (average = **0,66**)
Then
Diagnosis *is* 2
Rule's probability: **0,917**
The rule exists in **11** records.
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
25, 26, 29, 31, 34, 37, 38, 47, 50, 54
Negative Examples (records' serial numbers):
7
Positive Examples (records' serial numbers):
2, 8, 12, 17, 19, 20, 65, 70, 71, 72

18) *If* **Age** *is* **47,00 ... 70,00** (average = **60,50**)
and **Width** *is* **56,00 ... 77,00** (average = **62,60**)
and **Accel** *is* **210,00 ... 242,00** (average = **227,00**)
Then
Diagnosis *is* **2**
Rule's probability: **1,000**
The rule exists in **10** records.
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
25, 28, 36, 41, 42, 44, 50, 53, 56, 59

19) *If* **LR** *is* **R**
and **Speed** *is* **16,30 ... 41,50** (average = **25,44**)
and **Index** *is* **0,70 ... 0,80** (average = **0,72**)
Then
Diagnosis *is not* **2**
Rule's probability: **1,000**
The rule exists in **10** records.
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
3, 61, 63, 64, 65, 66, 67, 72, 73, 74

Для пояснения полученных результатов рассмотрим более подробно, например, правило № 19:

19) *If* **LR** *is* **R**
and **Speed** *is* **16,30 ... 41,50** (average = **25,44**)
and **Index** *is* **0,70 ... 0,80** (average = **0,72**)
Then
Diagnosis *is not* **2**
Rule's probability: **1,000**
The rule exists in **10** records.
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
3, 61, 63, 64, 65, 66, 67, 72, 73, 74

Это правило представляет собой конъюнкцию трех элементарных высказываний. Первое – **LR** *is* **R** – говорит о том, что правило относится только к правой почке. Второе – **Speed** *is* **16,30 ... 41,50** – определяет диапазон значений для средней скорости кровотока, а третье – **Index** *is* **0,70 ... 0,80** – описывает интервал значений индекса резистентности. Высказывание – **Diagnosis** *is not* **2** – означает, что правило характерно для объектов, не имеющих диагноза «множественные кисты».

Запись – *Rule's probability*: **1,000** – означает, что точность правила в данном случае равна 1. Следующая запись – *The rule exists in* **10** *records* – характеризует объем множества объектов, для которых

справедливо рассматриваемое правило, а другая запись – *Significance Level: Error probability* < 0,1 – касается статистической оценки уровня значимости полученного правила (как видим, доверие к правилу превышает 90%). Последняя запись – *Positive Examples (records' serial numbers)* – означает «положительные примеры», которые затем представлены как номера записей (объектов) в наборе данных.

Система WizWhy предоставляет возможность визуализации полученного правила. Для этого нужно щелкнуть на правиле левой кнопкой мыши и затем с помощью правой кнопки вызвать контекстное меню, в котором выбрать диаграмму правила **Rule Chart** (рис. 2.9). Эта диаграмма иллюстрирует отдельные компоненты правила и дает графическое отображение совокупного взаимодействия переменных.

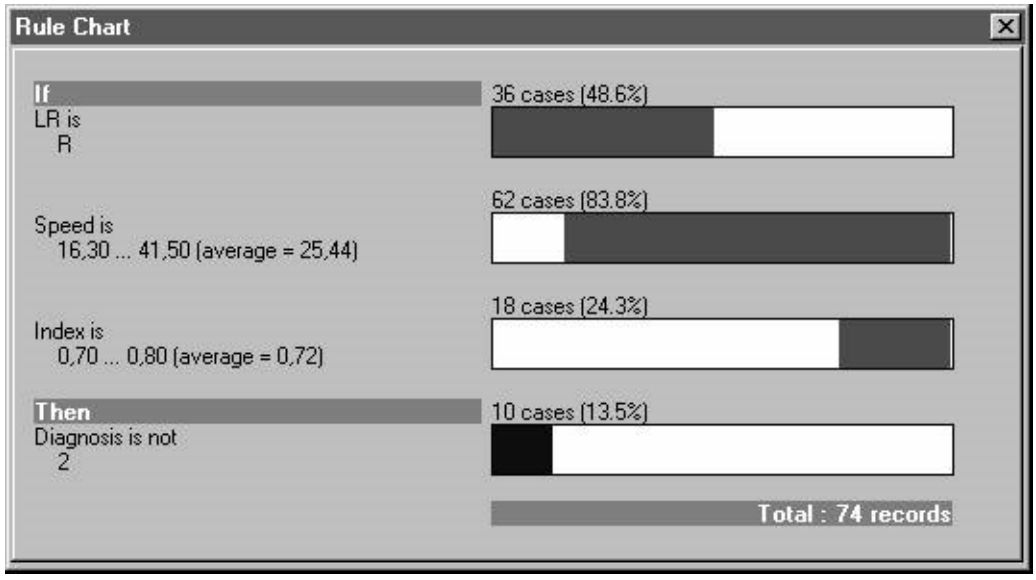


Рис. 2.9. Диаграмма выделенного правила № 19

Содержание записи в деталях

Окно «Содержание записи в деталях» позволяет просмотреть значения признаков для каждого объекта. Для этого требуется ввести номер объекта в поле **Record** и нажать клавишу **Enter**. Пример для объекта № 25 приведен на рисунке 2.10.

Рис. 2.10. Содержание записи в деталях

Другая возможность состоит в том, что если дважды щелкнуть левой кнопкой мыши на номере объекта в списке правил, который там приведен в качестве положительного или отрицательного примера, соответствующие значения признаков отобразятся в рассматриваемом окне. При этом целевая переменная будет отмечена специальным значком красного цвета, а все остальные – значками зеленого цвета. Кроме того, на значках, расположенных сразу слева от названия признаков, указываются типы данных признаков.

Индекс признака

В окне «Индекс признака», расположенном в правом нижнем углу, отображаются порядковые номера правил, в которых появляются те или иные признаки (рис. 2.11). Можно просмотреть все окно, используя прокрутку. Также в системе предусмотрена другая возможность – если в списке правил дважды щелкнуть мышью на каком-либо признаке в любом из правил, то этот признак будет автоматически выделен в окне «Индекс признака». По представляемой информации удобно выносить суждения о полезности признаков (о коэффициенте использования признаков) для классификации данных и прогнозирования. В свою очередь, если дважды щелкнуть в окне «Индекс признака» по любому номеру правила, то это правило моментально будет выделено в списке правил.

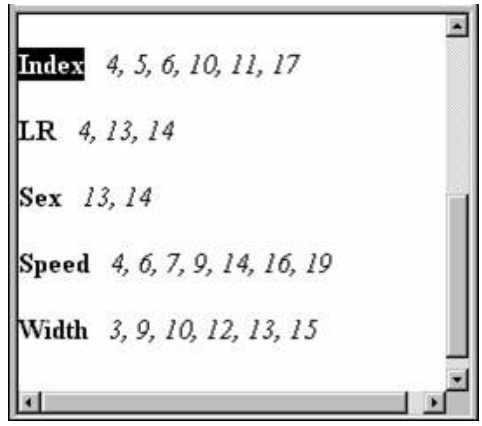


Рис. 2.11. Индекс признака

Распечатка и экспорт правил

Для распечатки правил или их экспорта в другой файл требуется нажать соответствующую кнопку печати в главном окне WizWhy – на экране появится специальное окно диалога **Print Rules** (рис. 2.12).

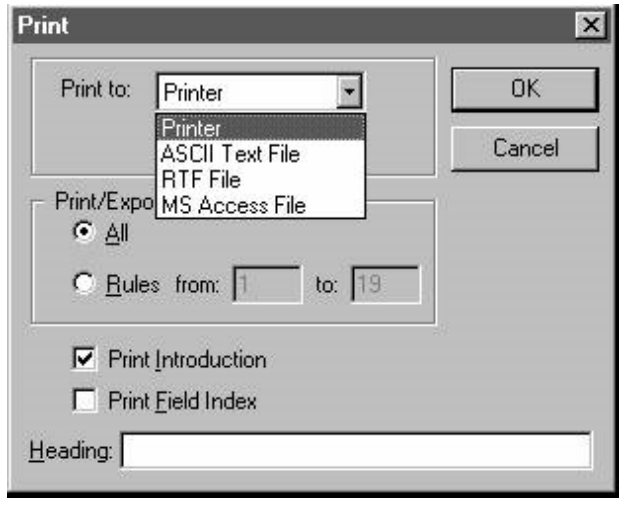


Рис. 2.12. Окно диалога для распечатки и экспорта выделенных правил

В поле **Print to** указывается адрес, по которому направляется результирующая информация. Это может быть принтер, файлы ASCII или RTF, а также файл MS Access.

В поле **Print/Export Range** указывается диапазон порядковых номеров правил, которые должны быть распечатаны или экспортированы. В нижней части окна диалога проставляются по необходимости флажки для распечатки или экспорта введения к списку правил **Print Introduction** и содержимого окна «Индекс признака». Кроме того, в поле **Heading** можно ввести заголовок для результирующей информации.

В системе WizWhy предусмотрена также возможность экспорта и сохранения полученных правил в виде операторов SQL. Для этого необходимо войти в меню **Issue | SQL Statement...** – система выдает окно диалога, показанное на рисунке 2.13. В этом окне можно редактировать совокупность операторов и с помощью переключателей адресовать данную совокупность в текстовый файл либо в буфер обмена.

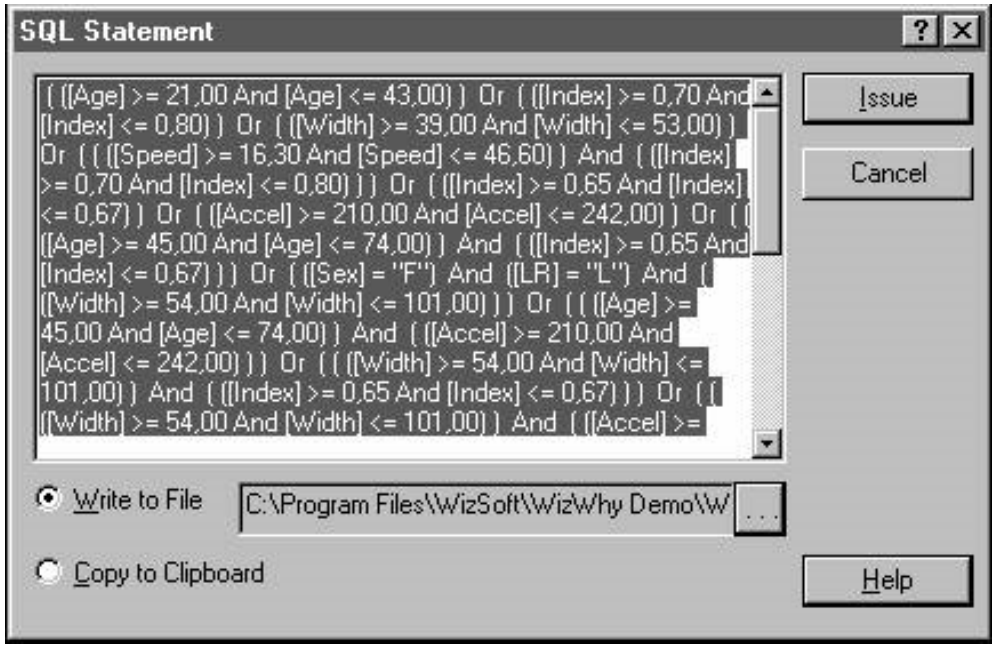


Рис. 2.13. Результаты работы WizWhy в виде операторов SQL

Отчет о трендах

Отчет о трендах представляет результаты сегментации отдельных признаков. Окно данного отчета разделено на три области (рис. 2.14).

В области, расположенной в левом верхнем углу, мы задаем анализируемый признак (**Field to be analyzed**). Здесь можно не только выбирать требуемый признак, но и сортировать признаки по какому-либо критерию (в алфавитном порядке, по номеру поля, по информативности).

Другие две области предназначены для отражения отношений между значениями признака и зависимой переменной. В верхней правой области окна отчета приводятся статистические характеристики сегментов выделенного признака. И наконец, в нижней области отчета приводится графическая иллюстрация информативности каждого сегмента.

На графике по горизонтальной оси располагаются сегменты, на которые выбранный признак автоматически разбивается системой WizWhy. По вертикальной оси откладывается отношение количества объектов класса if-then к общему количеству объектов, попадающих в сегмент. Таким образом, высота столбиков на графике отражает информативность сегментов. Если столбик выше горизонтальной черты, то, значит, в данный сегмент чаще попадают объекты класса if-then, а если ниже горизонтальной черты – класса if-then-NOT. В свою очередь, ширина столбиков пропорциональна количеству объектов, относящихся к данному сегменту.

Рис. 2.14. Отчет о трендах

Отчет о неожиданных правилах

В системе WizWhy введено представление о так называемых неожиданных правилах (unexpected rules). Под неожиданными понимаются правила в виде конъюнкции двух и более простых высказываний, комбинация которых дает точность и полноту прогноза выше, чем это можно было бы ожидать при независимости простых высказываний. Это представление, по-видимому, имеет цель – дополнительно заинтриговать конечного пользователя возможностью открывать в данных нетривиальные закономерности.

В нашем случае система не обнаружила таких неожиданных правил. Однако можно попытаться это сделать, если мы изменим задание на поиск правил. Например, уменьшим минимальную вероятность if-then и if-then-NOT правил с 80 до 70% в окне **Rule Parameters** (рис. 2.14). Проделаем указанную операцию и нажнем кнопку **Issue Rules** – система обнаружит теперь в данных по ультразвуковой диагностике 38 правил, и среди них будут четыре неожиданных, отчет о которых выдается в специальном окне (рис. 2.15).

Окно отчета о неожиданных правилах разделено на три секции. В левой верхней секции отображается в стандартной форме найденное неожиданное правило. Правая верхняя секция содержит информацию об элементах, из которых составлено неожиданное правило. И наконец, нижняя секция предназначена для сортировки неожиданных правил и графического представления результатов.

Так, в нашем случае первое неожиданное правило, изображенное на рисунке 2.15, расшифровывается следующим образом: если (пол женский) и (ширина почки в интервале от 61 до 77) и (ускорение кровотока от 148 до 275), то диагноз «множественные кисты». Данное правило вместе с рассчитанными характеристиками приведено ниже. Здесь по сравнению с ранее рассмотренными характеристиками выдаются две новые – уровень неожиданности (**Level of Unlikelihood**) и ожидавшаяся вероятность правила (**Expected rule probability**). Как видим, за

счет взаимосвязи элементов правила точность целого правила составила 1 и оказалась значительно выше ожидавшейся (0,81).

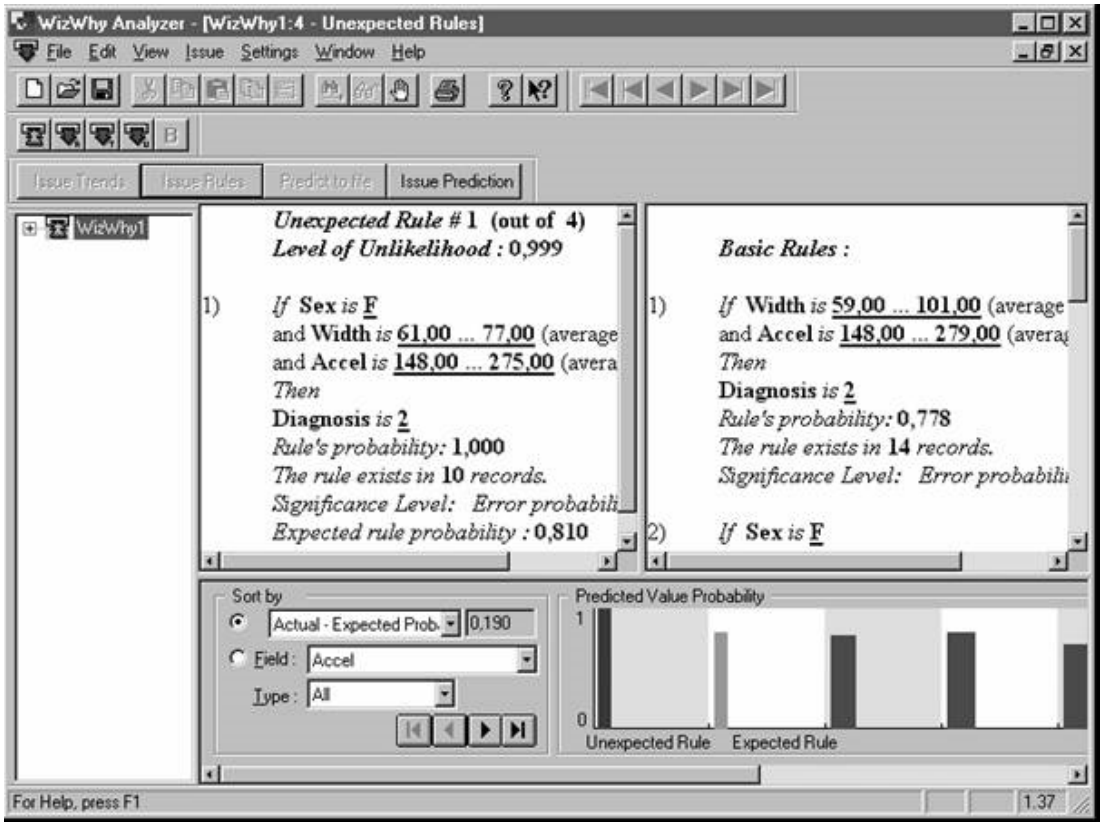


Рис. 2.15. Отчет о неожиданных правилах

Unexpected Rule # 1 (out of 4)

Level of Unlikelihood: 0,999

1) If **Sex** is **F**

and Width *is* 61,00 ... 77,00 (average = 67,30)

and Accel *is* 148,00 ... 275,00 (average = 216,10)

Then

Diagnosis *is* 2

Rule's probability: **1,000**

The rule exists in **10** records.

Significance Level: Error probability < 0,1

Expected rule probability: **0,810**

Actual minus Expected probability: **0,190**

В правой верхней секции приводится статистический разбор компонентов, из которых состоит неожиданное правило. Он состоит из двух частей (табл. 2.2).

Базисные правила (***Basic Rules***) представляют собой комбинации простых событий, входящих в неожиданное правило. В нашем случае, так как неожиданное правило состоит из трех простых событий, число таких комбинаций также составит 3.

Базисные тренды (***Basic Trends***) – это статистический разбор сегментов анализируемых переменных, составляющих собственно простые логические события.

Таблица 2.2. Разбор компонент неожиданного правила

Basic Rules	Basic Trends
1) *If* Width *is* 59,00 ... 101,00 (average = 69,00) and Accel *is* 148,00 ... 279,00 (average = 214,11) Then Diagnosis *is* 2 Rule's probability: **0,778** The rule exists in **14** records. Significance Level: Error probability < 0,2 2) If **Sex** is **F** and Accel *is* 148,00 ... 275,00 (average = 222,14) Then Diagnosis *is* 2 Rule's probability: **0,810** The rule exists in **17** records. Significance Level: Error probability < 0,2 3) If **Sex** is **F** and Width *is* 60,00 ... 85,00 (average = 68,29) Then Diagnosis *is* 2 Rule's probability: **0,706** The rule exists in **12** records. Significance Level: Error probability < 0,3	4) *If* Accel *is* 148,00 ... 279,00 (average = 217,18) Then Diagnosis *is* 2 Rule's probability: **0,706** The rule exists in **24** records. Significance Level: Error probability < 0,3 5) If **Sex** is **F** Then Diagnosis *is* 2 Trend's probability: **0,595** The trend exists in **25** records. 6) *If* Width *is* 59,00 ... 101,00 Then Diagnosis *is* 2 Trend's probability: **0,556** The trend exists in **20** records.

Как видим из таблицы, все компоненты неожиданного правила по отдельности имеют точность существенно ниже 1 – самое высокое значение точности наблюдается у базисного правила № 2, представляющего собой комбинацию двух простых событий (**Sex** *is* **F**) и (**Accel** *is* **148,00 ... 275,00**).

Нижняя секция отчета о неожиданных правилах разделена на две части (рис. 2.15). В левой части располагаются элементы управления для сортировки этих правил. По умолчанию правила проранжированы по величине разности между реальной и ожидавшейся точностями правила. Если установить переключатель в поле **Field** и выбрать из списка какой-либо признак, то будут отображаться только те неожиданные правила, в которых встречается указанный признак. В свою очередь, в поле **Type** можно выбрать один из трех типов фильтров правил – **All** (все правила), **if-then** правила и **if-then-NOT.**

В правой части нижней секции отчета о неожиданных правилах дается графическое представление характеристик правил и их составляющих. Первый слева столбик относится к найденному неожиданному правилу – его высота равна точности, а ширина пропорциональна количеству покрываемых объектов. Следующий столбик отображает ожидавшиеся характеристики правила, а остальные столбики соответствуют описанным выше базисным правилам и трендам. Если щелкнуть левой кнопкой мыши по какому-либо столбику, то система WizWhy автоматически изменит содержание верхних окон отчета о неожиданных правилах. Можно также щелнуть на столбике правой кнопкой мыши – появляется контекстное меню, в котором можно заказать иллюстрацию в виде диаграммы правила (Rule chat).

2.7. Предсказание на основе полученных правил

В системе WizWhy предусмотрены две возможности использования обнаруженных правил для предсказания значений целевого показателя на новом материале.

Первая возможность заключается в ручном вводе значений признаков и обработке нового одиночного объекта (записи). Она реализуется следующим образом.

Нажимаем кнопку **Issue Prediction** – на экран выдается окно диалога для ручного ввода значений признаков (рис. 2.16). После заполнения окошек предложенной таблицы (здесь возможны пропуски) нажимаем кнопку **Issue Report** – система находит релевантные правила и создает отчет, в котором подробно описываются как конечный результат предсказания, так и характеристики каждого отдельного правила, сработавшего для данного объекта. Этот отчет приводится ниже.

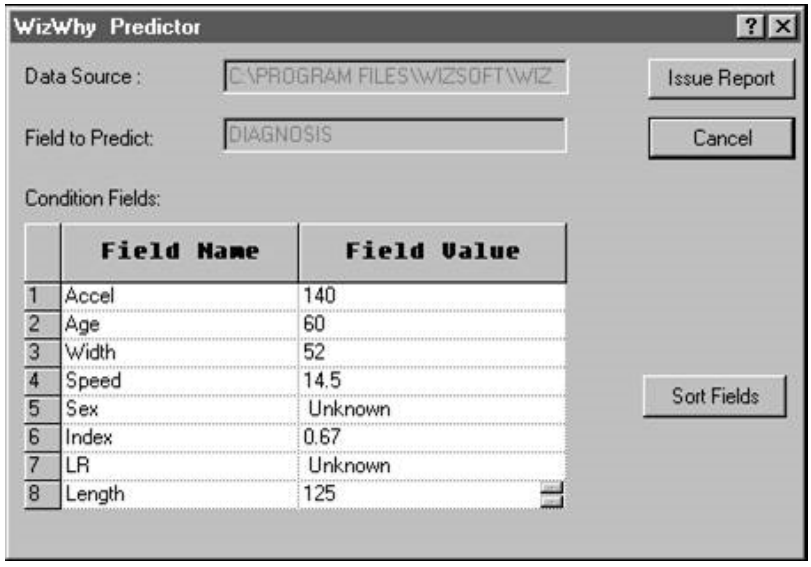


Рис. 2.16. Окно диалога для ручного ввода значений признаков

WizWhy PREDICTION REPORT

Condition Fields:
Age = 60,00
Length = 125,00
Width = 52,00
Speed = 14,50
Index = 0,67
Accel = 140,00
Field to Predict: **Diagnosis**
Subject for Prediction: **Diagnosis** is **2**
Prediction's significance level: Error probability = 0,259
Primary Prediction's probability: 0,500
Conclusive Prediction's probability: 0,549
Prediction: No **2**
Relevant rules:

1) *If* **Age** *is* **56,00 ... 74,00** (average = **64,81**)
and Speed *is* 12,50 ... 18,30 (average = 15,59)
Then
Diagnosis *is* 2
Rule's probability: **0,813**
The rule exists in **13** records.
Significance Level: Error probability < 0,2

2) *If* **Age** *is* **56,00 ... 74,00** (average = **66,31**)
and **Index** *is* **0,63 ... 0,67** (average = **0,65**)
Then
Diagnosis *is* 2
Rule's probability: **0,846**
Significance Level: Error probability < 0,2

3) *If* Length *is* 112,00 ... 128,00 (average = 118,64)
and Width *is* 39,00 ... 55,00 (average = 49,36)
Then
Diagnosis *is not* 2
Rule's probability: **0,909**
The rule exists in **10** records.
Significance Level: Error probability < 0,2

Как видим, в нашем случае система выдала предсказание, что рассматриваемый объект не относится к классу № 2 (Prediction: No **2**). Это решение система приняла на основании трех релевантных правил. Хотя два первых правила говорят, что объект является представителем класса № 2 (диагноз «множественные кисты»), но их «побеждает» третье правило, имеющие более высокую точность (0,909).

Вторая возможность использования множества правил заключается в обработке сразу большого массива новой информации. Для этого сначала, перейдя к закладке **Prediction Input** в окне диалога для ввода данных (рис. 2.17), нужно указать файл, в котором записана новая информация. Пусть в нашем случае это будет тот же самый файл с обучающей выборкой **USR.txt**. Затем требуется задать имя файла, в который будут записываться результаты предсказания. Данная операция осуществляется с помощью кнопки **Print result to…** И наконец, нажимается кнопка **Predict to file** – система производит необходимые расчеты и сообщает, что результаты успешно записаны в указанный файл результатов, который приведен в таблице 2.3.

Таблица 2.3. Содержание файла результатов предсказания

№ п/п	"Diagnosis"	"Sign_Level"	"Concl_Prob"	"Prediction"
1	"1"	0.492	0.481	No 2
2	"1"	0.559	0.089	No 2
3	"1"	0.406	0.526	No 2
4	"1"	0.689	0.088	No 2
5	"1"	0.020.	0.030	No 2
6	"1"	0.198	0.072	No 2
7	"1"	0.374	0.532	No 2
8	"1"	0.000.	0.032	No 2
9	"1"	0.511	0.459	No 2
10	"1"	0.186	0.068	No 2
11	"1"	0.173	0.051	No 2
12	"1"	0.433	0.450.	No 2
13	"1"	0.272	0.077	No 2
14	"1"	0.466	0.514	No 2
15	"1"	0.330	0.525	No 2
16	"1"	0.200	0.528	No 2
17	"1"	0.243	0.059	No 2
18	"1"	0.152	0.610	No 2
19	"1"	0.588	0.460	No 2
20	"1"	0.002	0.016	No 2
21	"1"	0.489	0.454	No 2
22	"1"	0.145	0.624	2
23	"2"	0.361	0.743	2
24	"2"	0.097	0.952	2
25	"2"	0.009	0.995	2
26	"2"	0.000	0.960	2
27	"2"	0.384	0.737	2

Продолжение табл. 2.3

№ п/п	"Diagnosis"	"Sign_Level"	"Concl_Prob"	"Prediction"
28	"2"	0.097	0.952	2
29	"2"	0.000	0.963	2
30	"2"	0.210	0.788	2
31	"2"	0.275	0.759	2
32	"2"	0.285	0.745	2
33	"2"	0.009	0.995	2
34	"2"	0.009	0.995	2
35	"2"	0.191	0.613	2
36	"2"	0.000	0.962	2
37	"2"	0.000	0.961	2
38	"2"	0.000	0.961	2
39	"2"	0.001	0.854	2
40	"2"	0.228	0.789	2
41	"2"	0.205	0.933	2
42	"2"	0.009	0.995	2
43	"2"	0.097	0.952	2
44	"2"	0.705	0.497	No 2
45	"2"	0.097	0.952	2
46	"2"	0.181	0.770	2
47	"2"	0.000	0.857	2
48	"2"	0.230	0.422	No 2
49	"2"	0.194	0.626	2
50	"2"	0.000	0.975	2
51	"2"	0.009	0.995	2
52	"2"	0.001	0.965	2
53	"2"	0.097	0.952	2
54	"2"	0.009	0.995	2
55	"2"	0.377	0.503	No 2
56	"2"	0.097	0.952	2
57	"2"	0.715	0.498	No 2
58	"2"	0.240	0.930	2
59	"2"	0.009	0.995	2
60	"3"	0.446	0.458	No 2
61	"3"	0.097	0.048	No 2
62	"3"	0.610	0.512	No 2
63	"3"	0.000	0.024	No 2
64	"3"	0.097	0.048	No 2
65	"3"	0.097	0.048	No 2
66	"3"	0.097	0.048	No 2
67	"3"	0.097	0.048	No 2
68	"3"	0.097	0.048	No 2
69	"3"	0.297	0.774	2
70	"3"	0.097	0.048	No 2
71	"3"	0.000	0.022	No 2
72	"3"	0.097	0.048	No 2
73	"3"	0.097	0.048	No 2
74	"3"	0.097	0.048	No 2

3. Практические примеры

3.1. Сравнение структуры интеллекта физиков и лириков

В рассматриваемом примере исследуются экспериментальные данные, представляющие собой результаты психологического тестирования учащихся специализированных школ Санкт-Петербурга с физико-математическим и гуманитарным уклоном. Сначала эти данные будут обработаны традиционными методами статистического анализа. Затем мы применим к экспериментальному материалу технологии поиска логических закономерностей, описанные в соответствующих главах.

Нам предстоит на практике убедиться, что к исследованиям людей далеко не всегда следует подходить с позиций классической математической статистики. Это приводит к малосодержательным и нередко бесполезным выводам. В то же время применение аппарата поиска в данных логических высказываний способно привести к раскрытию ценных многоаспектных знаний о людях.

Общая характеристика данных

Объектами исследования являются 76 учащихся специализированных школ:

- 38 человек – учащиеся 10-х классов физико-математической школы № 30. В дальнейшем они будут называться «**физики**»;
- 38 человек – учащиеся 10-х классов гуманитарных школ (21 учащийся литературной школы № 27 и 17 учащихся художественной школы № 363). Назовем их «**лирики**».

Целью исследования являлось определение различий в структуре интеллекта физиков и лириков. Естественно было заранее предположить, что у физиков более развит «левополушарный» вербальный, а у лириков – «правополушарный» невербальный интеллект. Исходя из этого, выбирался инструмент психологического исследования, позволяющий оценивать как вербальную, так и невербальную стороны мышления. Конкретно использовались субтесты IN, AN, GE, AR и PL популярного теста Р. Амтхауэра. Ниже этим субтестам дается краткая характеристика.

Тест № 1 (IN) – «дополнение предложений». Он предназначен для оценки способности к рассуждению, здравого смысла, сложившейся самостоятельности мышления (для юношеского возраста). По отношению к этому тесту будем использовать термин «здравомыслие».

Тест № 3 (AN) – «аналогии». Этот тест отражает способность комбинировать, подвижность и непостоянство мышления.

Тест № 4 (GE) – «обобщение». Тест предназначен для определения способности к абстрактному мышлению, образованию понятий, умению словесно выражать мысль.

Тест № 5 (AR) – «арифметические задачи». Данный тест отражает развитость практического численного мышления.

Тест № 7 (PL) – «выбор геометрического образца». Тест направлен на оценку воображения, богатства представлений, наглядного целостного мышления. Сопоставим этому тесту термин «воображение».

Результаты психологического тестирования физиков и лириков представлены в таблице 3.1. В этой же таблице указан пол испытуемых. Среди физиков 12 девушек и 26 юношей, среди лириков 21 девушка и 17 юношей.

Таблица 3.1. Результаты психологического тестирования физиков и лириков

№ п/п	Тест 1 (здравомыслие)	Тест 3 (аналогии)	Тест 4 (обобщение)	Тест 5 (численное мышление)	Тест 7 (воображение)	Пол
Физики						
1	15	16	19	13	8	Мужской
2	16	11	16	20	8	Мужской
3	11	16	14	15	12	Мужской
4	15	10	9	18	10	Мужской
5	8	11	10	16	10	Женский
6	11	13	13	7	11	Женский
7	7	9	13	11	13	Женский
8	11	15	14	12	8	Женский
9	12	15	18	10	10	Женский
10	14	10	16	12	10	Мужской
11	8	4	16	14	8	Мужской
12	9	13	12	15	6	Мужской
13	10	11	18	12	10	Мужской
14	10	11	16	14	9	Мужской
15	11	10	14	13	15	Мужской
16	10	15	17	19	11	Мужской
17	12	13	13	11	8	Мужской
18	11	12	16	17	10	Мужской
19	9	12	16	14	11	Женский
20	11	14	11	17	13	Мужской
21	11	14	15	10	7	Мужской
22	15	17	16	12	12	Мужской
23	11	12	16	12	6	Мужской
24	13	7	17	6	8	Мужской
25	13	15	13	7	10	Мужской
26	18	15	17	7	11	Женский
27	18	17	15	9	8	Мужской
28	16	18	23	6	10	Мужской
29	15	18	23	15	11	Мужской
30	14	10	19	10	13	Женский
31	13	10	10	6	7	Мужской
32	14	15	19	8	7	Женский
33	18	15	18	9	10	Женский
34	10	12	18	10	11	Мужской
35	14	12	16	7	13	Мужской
36	17	17	24	13	9	Женский
37	10	12	22	9	14	Мужской
38	11	15	19	11	8	Женский
Лирики						
39	6	7	10	4	10	Мужской
40	5	7	15	2	7	Женский
41	9	10	15	4	9	Женский
42	8	14	15	5	14	Женский
43	10	7	11	8	6	Мужской
44	12	8	9	8	7	Мужской
45	6	8	8	7	9	Мужской
46	10	12	20	14	11	Мужской
47	14	4	14	7	11	Мужской
48	6	6	12	11	9	Мужской

Продолжение табл. 3.1

№ п/п	Тест 1 (здравомыслие)	Тест 3 (аналогии)	Тест 4 (обобщение)	Тест 5 (численное мышление)	Тест 7 (воображение)	Пол
49	8	10	19	3	5	Женский
50	8	9	20	3	5	Женский
51	4	7	12	5	10	Женский
52	11	12	16	7	6	Женский
53	7	6	5	3	4	Мужской
54	10	3	9	5	6	Женский
55	15	14	19	7	6	Женский
56	10	11	12	8	15	Мужской
57	6	10	12	9	15	Мужской
58	10	9	12	5	10	Женский
59	10	9	12	5	10	Мужской
60	11	12	11	5	11	Мужской
61	10	11	13	4	11	Мужской
62	7	10	12	7	10	Женский
63	10	12	12	6	5	Мужской
64	7	11	10	5	8	Женский
65	9	10	16	4	12	Женский
66	10	8	18	4	7	Женский
67	12	10	19	6	13	Женский
68	15	4	14	5	9	Женский
69	12	9	18	4	9	Женский
70	11	6	18	5	8	Мужской
71	12	15	20	6	8	Мужской
72	13	14	25	15	13	Женский
73	15	8	14	14	4	Женский
74	13	15	19	7	8	Женский
75	10	12	18	6	11	Мужской
76	13	9	14	8	10	Женский

Сравнение средних значений результатов тестирования в группах физиков и лириков

Сравнение средних значений каких-либо показателей с помощью статистических Т-критериев считается важнейшим видом анализа экспериментально-психологических данных. Применим этот метод к нашим данным. В таблице 3.2 приведены средние значения и стандартные отклонения результатов субтестов по группам физиков и лириков.

Таблица 3.2. Средние значения и стандартные отклонения результатов субтестов

	Тест 1		Тест 3		Тест 4		Тест 5		Тест 7	
	Физики	Лирики	Физики	Лирики	Физики	Лирики	Физики	Лирики	Физики	Лирики
Среднее	12,4	9,8	12,9	9,4	16,1	14,4	11,8	6,3	9,9	9,0
Стандарт. отклонение	2,9	2,9	3,1	3,0	3,5	4,2	3,8	9,1	2,2	2,9

На основании статистических критериев по данным таблицы можно сделать следующие выводы.

1. Среднее «здравомыслие» в группе физиков выше среднего «здравомыслия» в группе лириков с достоверностью 99%.

2. Средняя «способность к аналогии» в группе физиков выше средней «способности к аналогии» в группе лириков с достоверностью 99%.

3. Средняя «способность к обобщению» в группе физиков выше средней «способности к обобщению» в группе лириков с достоверностью 95%.

4. Средняя «способность к численному мышлению» в группе физиков выше средней «способности к численному мышлению» в группе лириков более чем с 99%-ной достоверностью.

На рисунке 3.1, где приведены гистограммы распределения тестовых оценок, невооруженным глазом видно, что особенно сильное различие между физиками и лириками наблюдается в способности к численному мышлению (тест 5).

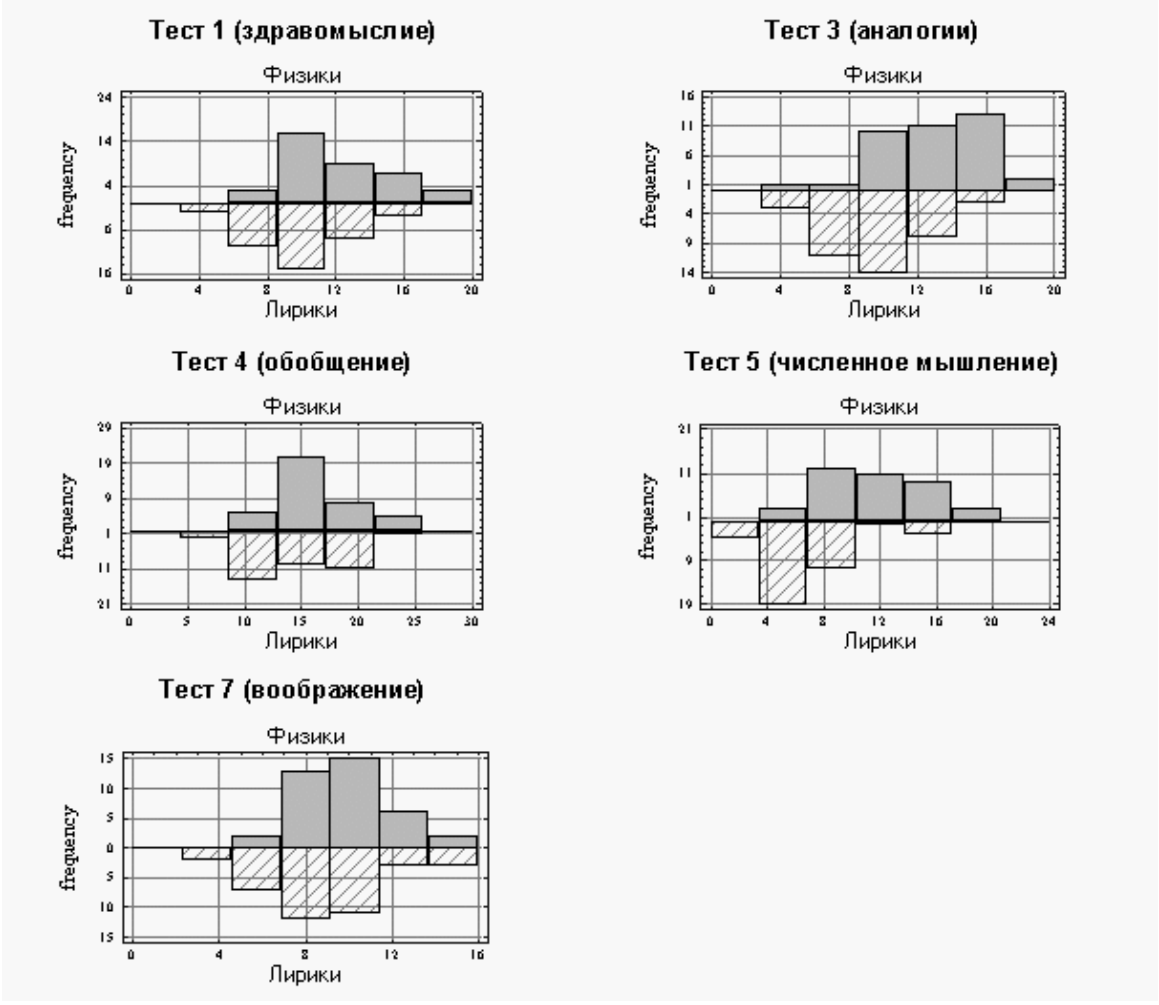


Рис. 3.1. Сравнительные гистограммы распределения тестовых оценок

Полученные результаты классического статистического анализа экспериментальных данных достаточно тривиальны. Можно было заранее без проведения специального исследования предположить, что у учащихся физико-математических школ окажутся в среднем более высокие способности к установлению аналогий, обобщению (абстрактному мышлению) и умению решать арифметические задачи. Хотя слегка удивляют их более высокая способность к само-

стоятельному мышлению (результаты анализа по тесту 1), а также низкие способности гуманитариев решать простые арифметические задачи (тест 5).

Вместе с тем представленный экспериментальный материал дает основание для попыток установления более глубоких и содержательных отличий гуманитариев от естественников, чем это выявляется с помощью классического статистического аппарата, оперирующего усредненными характеристиками совокупностей объектов. Вообще усреднение характеристик объектов с высоким уровнем сложности системной организации является во многом фиктивной операцией, не имеющей смысла (по выражению Б. С. Ястремского, не имеет смысла средняя высота дома на улице, состоящей из дворцов и лачуг). Совокупности таких объектов являются принципиально неоднородными, полиморфными и для описания закономерностей таких совокупностей должен использоваться аппарат, умеющий выявлять и учитывать указанные неоднородности.

Такими свойствами обладают методы индуктивного логического вывода, позволяющие находить в ограниченном наборе экспериментальных данных логические закономерности. Мы будем искать логические закономерности вида (событие 1) и (событие 2) и … и (событие N), характерные для каждой из исследуемых групп учащихся. Под событием здесь понимается, что значение какого-либо определенного теста попадает в некоторый заданный интервал значений. Таким образом, логическая закономерность в данном случае представляет собой «комплекс психологических характеристик, часто встречающийся у одной группы испытуемых и редко у другой».

Поиск логических закономерностей системой WizWhy

Ниже применяются следующие обозначения:

T1 – тест 1; T3 – тест 3; T4 – тест 4; T5 – тест 5; T7 – тест 7; Sex – пол (1 – мужской, 0 – женский); значок «&» – логическая связка «и».

Сводный отчет системы WizWhy

PARAMETERS OF THE RULES AND DATA
D:\Old\Knowledge\Школа
Total number of records: **76**
Minimum probability of the:
1) if-then rules: **0,700**
2) if-then-not rules: **0,700**
Minimum number of cases in a rule: **10**
Field to Predict: **Class**
Predicted Value (analyzed as Boolean): **Physics**
Prediction error costs:
The cost of a miss: 1
The cost of a false alarm: 1
Average probability of the predicted value is 0,500

ANALYSIS OF THE RULES EXPLANATORY POWER
Decision point: Predict Physics when conclusive probability is more than **0,454**
Number of misses: 2
Number of false alarms: 9
Total number of errors: 11
Total cost of errors: 11
Success rate when predicting Physics: 0,800
Success rate when predicting NOT Physics: 0,935
Number of records with no relevant rules: 0

Average cost (per record): 0,145
Expected average cost (per record): 0,500
Improvement Factor: 3,455

Правила, обнаруженные в данных системой WizWhy

1) *If* **Здравомыслие** *is* **4,00 ... 10,00** (average = **8,29**)
and **Арифметич. задачи** *is* **2,00 ... 6,00** (average = **4,29**)
Then
Class *is not* **Physics**
Rule's probability: **1,000**
The rule exists in **17** *records.*
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
39, 40, 41, 42, 49, 50, 51, 53, 54, 58

2) *If* **Арифметич. задачи** *is* **2,00 ... 6,00** (average = **4,69**)
Then
Class *is not* **Physics**
Rule's probability: **0,885**
The rule exists in **23** *records.*
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
39, 40, 41, 42, 49, 50, 51, 53, 54, 58
Negative Examples (records' serial numbers):
24, 28, 31

3) *If* **Здравомыслие** *is* **4,00 ... 10,00** (average = **7,68**)
and **Аналогии** *is* **3,00 ... 10,00** (average = **7,84**)
Then
Class *is not* **Physics**
Rule's probability: **0,895**
The rule exists in **17** *records.*
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
39, 40, 41, 43, 45, 48, 49, 50, 51, 53
Negative Examples (records' serial numbers):
7, 11

4) *If* **Аналогии** *is* **12,00 ... 18,00** (average = **14,46**)
and **Арифметич. задачи** *is* **7,00 ... 19,00** (average = **11,67**)
and **Воображение** *is* **8,00 ... 14,00** (average = **10,46**)
Then
Class *is* **Physics**
Rule's probability: **0,875**
The rule exists in **21** *records.*
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
1, 3, 6, 8, 9, 16, 17, 18, 19, 20
Negative Examples (records' serial numbers):
46, 72, 74

5) *If* **Обобщение** *is* **15,00 ... 17,00** (average = **16,09**)
and **Арифметич. задачи** *is* **7,00 ... 20,00** (average = **13,18**)
and **Воображение** *is* **8,00 ... 13,00** (average = **10,09**)
Then
Class *is* **Physics**
Rule's probability: **1,000**
The rule exists in **11** *records.*
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
2, 10, 11, 14, 16, 18, 19, 22, 26, 27

6) *If* **Здравомыслие** *is* **4,00 ... 10,00** (average = **7,45**)
and **Аналогии** *is* **3,00 ... 10,00** (average = **7,45**)
and **Обобщение** *is* **5,00 ... 12,00** (average = **10,45**)
Then
Class *is not* **Physics**
Rule's probability: **1,000**
The rule exists in **11** *records.*
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
39, 43, 45, 48, 51, 53, 54, 57, 58, 59

7) *If* **Аналогии** *is* **12,00 ... 18,00** (average = **14,23**)
and **Арифметич. задачи** *is* **7,00 ... 19,00** (average = **11,30**)
Then

15) *If* **Арифметич. задачи** *is* **7,00 ... 20,00** (average = **11,63**)
and **Воображение** *is* **8,00 ... 15,00** (average = **10,63**)
Then
Class *is* **Physics**
Rule's probability: **0,756**
The rule exists in **31** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Negative Examples (records' serial numbers):
45, 46, 47, 48, 56, 57, 62, 72, 74, 76

16) *If* **Аналогии** *is* **3,00 ... 10,00** (average = **7,86**)
and **Обобщение** *is* **5,00 ... 12,00** (average = **10,21**)
Then
Class *is not* **Physics**
Rule's probability: **0,857**
The rule exists in **12** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
39, 43, 44, 45, 48, 51, 53, 54, 57, 58
Negative Examples (records' serial numbers):
4, 31

17) *If* **Здравомыслие** *is* **14,00 ... 18,00** (average = **15,44**)
Then
Class *is* **Physics**
Rule's probability: **0,778**
The rule exists in **14** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
1, 2, 4, 10, 22, 26, 27, 28, 29, 30
Negative Examples (records' serial numbers):
47, 55, 68, 73

18) *If* **Аналогии** *is* **3,00 ... 10,00** (average = **8,00**)
Then
Class *is not* **Physics**
Rule's probability: **0,758**
The rule exists in **25** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
39, 40, 41, 43, 44, 45, 47, 48, 49, 50
Negative Examples (records' serial numbers):
4, 7, 10, 11, 15, 24, 30, 31

19) *If* **Здравомыслие** *is* **14,00 ... 18,00** (average = **15,44**)
and **Арифметич. задачи** *is* **7,00 ... 20,00** (average = **11,31**)
Then
Class *is* **Physics**
Rule's probability: **0,813**
The rule exists in **13** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
1, 2, 4, 10, 22, 26, 27, 29, 30, 32
Negative Examples (records' serial numbers):
47, 55, 73

20) *If* **Аналогии** *is* **12,00 ... 18,00** (average = **14,41**)
and **Воображение** *is* **8,00 ... 14,00** (average = **10,52**)
Then
Class *is* **Physics**
Rule's probability: **0,759**
The rule exists in **22** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
1, 3, 6, 8, 9, 16, 17, 18, 19, 20
Negative Examples (records' serial numbers):
42, 46, 60, 71, 72, 74, 75

21) *If* **Обобщение** *is* **15,00 ... 17,00** (average = **16,00**)

Class *is* **Physics**
Rule's probability: **0,833**
The rule exists in **25** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
1, 3, 6, 8, 9, 12, 16, 17, 18, 19
Negative Examples (records' serial numbers):
46, 52, 55, 72, 74

8) *If* **Обобщение** *is* **15,00 ... 17,00** (average = **16,00**)
and **Арифметич. задачи** *is* **7,00 ... 20,00** (average = **12,43**)
Then
Class *is* **Physics**
Rule's probability: **0,929**
The rule exists in **13** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
2, 10, 11, 14, 16, 18, 19, 21, 22, 23
Negative Examples (records' serial numbers):
52

9) *If* **Здравомыслие** *is* **14,00 ... 18,00** (average = **15,62**)
and **Арифметич. задачи** *is* **7,00 ... 20,00** (average = **11,69**)
and **Воображение** *is* **8,00 ... 13,00** (average = **10,31**)
Then
Class *is* **Physics**
Rule's probability: **0,923**
The rule exists in **12** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
1, 2, 4, 10, 22, 26, 27, 29, 30, 33
Negative Examples (records' serial numbers):
47

10) *If* **Здравомыслие** *is* **14,00 ... 18,00** (average = **15,91**)
and **Аналогии** *is* **12,00 ... 18,00** (average = **15,82**)
Then
Class *is* **Physics**
Rule's probability: **0,909**
The rule exists in **10** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
1, 22, 26, 27, 28, 29, 32, 33, 35, 36
Negative Examples (records' serial numbers):
55

11) *If* **Здравомыслие** *is* **14,00 ... 18,00** (average = **15,60**)
and **Воображение** *is* **8,00 ... 13,00** (average = **10,20**)
Then
Class *is* **Physics**
Rule's probability: **0,867**
The rule exists in **13** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
1, 2, 4, 10, 22, 26, 27, 28, 29, 30
Negative Examples (records' serial numbers):
47, 68

12) *If* **Здравомыслие** *is* **4,00 ... 10,00** (average = **7,88**)
and **Обобщение** *is* **5,00 ... 12,00** (average = **10,69**)
Then
Class *is not* **Physics**
Rule's probability: **0,875**
The rule exists in **14** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
39, 43, 45, 48, 51, 53, 54, 56, 57, 58
Negative Examples (records' serial numbers):
5, 12

13) *If* **Здравомыслие** *is* **4,00 ... 10,00** (average = **7,27**)
and **Обобщение** *is* **8,00 ... 12,00** (average = **11,09**)
and **Воображение** *is* **8,00 ... 15,00** (average = **10,55**)
Then
Class *is not* **Physics**
Rule's probability: **0,909**
The rule exists in **10** *records.*
Significance Level: Error probability < 0,2

and **Воображение** *is* **8,00 ... 14,00** (average = **10,27**)
Then
Class *is* **Physics**
Rule's probability: **0,800**
The rule exists in **12** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
2, 10, 11, 14, 16, 18, 19, 22, 24, 26
Negative Examples (records' serial numbers):
41, 42, 65

22) *If* **Аналогии** *is* **12,00 ... 18,00** (average = **14,17**)
Then
Class *is* **Physics**
Rule's probability: **0,722**
The rule exists in **26** *records.*
Significance Level: Error probability < 0,3
Positive Examples (records' serial numbers):
1, 3, 6, 8, 9, 12, 16, 17, 18, 19
Negative Examples (records' serial numbers):
42, 46, 52, 55, 60, 63, 71, 72, 74, 75

23) *If* **Обобщение** *is* **5,00 ... 12,00** (average = **10,52**)
Then
Class *is not* **Physics**
Rule's probability: **0,762**
The rule exists in **16** *records.*
Significance Level: Error probability < 0,3
Positive Examples (records' serial numbers):
39, 43, 44, 45, 48, 51, 53, 54, 56, 57
Negative Examples (records' serial numbers):
4, 5, 12, 20, 31

24) *If* **Арифметич. задачи** *is* **7,00 ... 20,00** (average = **11,32**)
Then
Class *is* **Physics**
Rule's probability: **0,700**
The rule exists in **35** *records.*
Significance Level: Error probability < 0,3
Positive Examples (records' serial numbers):
1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Negative Examples (records' serial numbers):
43, 44, 45, 46, 47, 48, 52, 55, 56, 57

25) *If* **Обобщение** *is* **8,00 ... 12,00** (average = **10,93**)
and **Воображение** *is* **8,00 ... 15,00** (average = **10,71**)
Then
Class *is not* **Physics**
Rule's probability: **0,786**
The rule exists in **11** *records.*
Significance Level: Error probability < 0,3
Positive Examples (records' serial numbers):
39, 45, 48, 51, 56, 57, 58, 59, 60, 62
Negative Examples (records' serial numbers):
4, 5, 20

26) *If* **Здравомыслие** *is* **4,00 ... 10,00** (average = **8,44**)
Then
Class *is not* **Physics**
Rule's probability: **0,706**
The rule exists in **24** *records.*
Significance Level: Error probability < 0,3
Positive Examples (records' serial numbers):
39, 40, 41, 42, 43, 45, 46, 48, 49, 50
Negative Examples (records' serial numbers):
5, 7, 11, 12, 13, 14, 16, 19, 34, 37

27) *If* **Обобщение** *is* **15,00 ... 17,00** (average = **15,89**)
Then
Class *is* **Physics**
Rule's probability: **0,737**
The rule exists in **14** *records.*
Significance Level: Error probability < 0,3
Positive Examples (records' serial numbers):
2, 10, 11, 14, 16, 18, 19, 21, 22, 23
Negative Examples (records' serial numbers):
40, 41, 42, 52, 65

28) *If* **Обобщение** *is* **18,00 ... 25,00** (average = **20,00**)

Positive Examples (records' serial numbers): ***39, 45, 48, 51, 56, 57, 58, 59, 62, 64*** *Negative Examples (records' serial numbers):* ***5*** 14) *If* **Аналогии** *is* **4,00 ... 10,00** (average = **8,00**) and **Арифметич. задачи** *is* **4,00 ... 6,00** (average = **4,82**) and **Воображение** *is* **8,00 ... 13,00** (average = **9,82**) *Then* **Class** *is not* **Physics** *Rule's probability:* **0,909** *The rule exists in* **10** *records.* *Significance Level: Error probability* < 0,2 *Positive Examples (records' serial numbers):* ***39, 41, 51, 58, 59, 65, 67, 68, 69, 70*** *Negative Examples (records' serial numbers):* ***24***	and **Арифметич. задачи** *is* **7,00 ... 15,00** (average = **10,87**) *Then* **Class** *is* **Physics** *Rule's probability:* **0,733** *The rule exists in* **11** *records.* *Significance Level: Error probability* < 0,3 *Positive Examples (records' serial numbers):* ***1, 9, 13, 29, 30, 32, 33, 34, 36, 37*** *Negative Examples (records' serial numbers):* ***46, 55, 72, 74*** 29) *If* **Воображение** *is* **4,00 ... 7,00** (average = **5,94**) *Then* **Class** *is not* **Physics** *Rule's probability:* **0,706** *The rule exists in* **12** *records.* *Significance Level: Error probability* < 0,3 *Positive Examples (records' serial numbers):* ***40, 43, 44, 49, 50, 52, 53, 54, 55, 63*** *Negative Examples (records' serial numbers):* ***12, 21, 23, 31, 32***

Приведенные логические закономерности демонстрируют значительную неоднородность объектов исследования – людей. Как мы видим, потребовалось 29 (система WizWhy) разнообразных высказываний, чтобы с разных сторон осветить особенности многоплановой структуры интеллекта физиков и лириков. При этом достаточно выпукло проявились определенные слабости мыслительной деятельности у учащихся гуманитарных школ. Буквально все высказывания, относящиеся к лирикам, говорят об их средних и низких тестовых оценках, в то время как у физиков оценки всегда средние и высокие (наиболее яркие высказывания выделены в таблице). Возможно, такая картина связана с ограниченностью данного психологического эксперимента. Но в любом случае полученные результаты могут служить весомыми аргументами в известном споре о физиках и лириках.

3.2. Влияние возраста и стажа работников на производительность труда

Встречаясь с объявлениями о приеме на работу, нередко можно заметить, что в этих объявлениях имеются ограничения на возраст сотрудников и пожелания относительно их стажа работы по специальности. В рассматриваемом примере нам предстоит разобраться, насколько бывают обоснованы подобные претензии работодателей.

Исходные данные заимствованы из книги [2]. Это результаты обследования 60 работников производства, у которых фиксировалась средняя часовая выработка в натуральных единицах продукции. Данные обследования отражены в таблице 3.3.

Таблица 3.3. Данные обследования работников производства

Стаж	**Возраст**		
	от 25 до 35 лет	**от 35 до 45 лет**	**от 45 до 55 лет**
от 1 до 4 лет	19 20 20 20 22	19 20 20 23 25	18 19 20 21 21
от 4 до 7 лет	30 31 32 32 34	20 29 30 31 31	19 25 25 26 26
от 7 до 10 лет	35 35 39 40 41	36 40 41 42 45	24 24 24 25 25
свыше 10 лет	40 40 41 41 42	28 31 35 36 40	20 24 25 31 32

Наш анализ будет состоять из двух частей. В первой части мы его проведем по классической статистической схеме с использованием аппарата многофакторного дисперсионного анализа. Во второй части к анализируемым данным будет применена технология обнаружения логических закономерностей.

Дисперсионный анализ

Дисперсионный анализ применяется для обнаружения влияния выделенного (контролируемого) набора факторов на результативный признак. Факторы обычно измеряются в неколичественной шкале, а результативный признак выражается числом или вектором с числовыми компонентами.

Идея дисперсионного анализа состоит в разложении общей дисперсии результативного признака на части, обусловленные влиянием контролируемых факторов, и остаточную дисперсию, объясняемую неконтролируемым влиянием или случайными обстоятельствами. Выводы о существенности влияния контролируемых факторов на результат производятся путем сравнения частей общей дисперсии при выполнении требования нормальности распределения результативного признака.

Известно много моделей дисперсионного анализа. Они классифицируются, с одной стороны, по математической природе факторов (детерминированные, случайные и смешанные) и, с другой стороны, по числу контролируемых факторов (однофакторные и многофакторные модели). Модели с более чем одним фактором дают возможность исследовать влияние на результат не только отдельных контролируемых факторов (главные влияния), но и их наложение (взаимодействия). По способу организации исходных данных выделяют полные и неполные m-факторные планы, полные и неполные блочные планы и рандомизированные (случайные) блочные планы.

Для проведения дисперсионного анализа будем использовать пакет STATGRAPHICS *Plus* for Windows, в котором реализованы все перечисленные выше модели дисперсионного анализа.

Раскроем электронную таблицу STATGRAPHICS и введем в нее значения результативного признака **output** (производительность) и закодированные значения градаций контролируемых факторов **age** (возраст) и **record** (стаж), как это показано на рисунке 3.2.

Выберем **Compare | Analysis of Variance | Multifactor ANOVA**. Заполним окно многофакторного дисперсионного анализа (рис. 3.3).

Нажмем **OK**. На экране появится сводка множественного дисперсионного анализа, в которой подтверждается, что к обработке принято 60 наблюдений, для которых зафиксированы значения двух факторов. Внизу под этими сообщениями включено сообщение **StatAdvisor** (Стат-Консультанта) с рекомендациями по проведению дальнейшего анализа.

	output	age	record	Col_4	Col_5	Col_6	Col_7
1	19	1	1				
2	20	1	1				
3	20	1	1				
4	20	1	1				
5	22	1	1				
6	19	2	1				
7	20	2	1				
8	20	2	1				
9	23	2	1				
10	25	2	1				
11	18	3	1				
12	19	3	1				
13	20	3	1				
14	21	3	1				
15	23	3	1				
16	30	1	2				
17	31	1	2				
18	32	1	2				
19	32	1	2				
20	34	1	2				
21	20	2	2				
22	29	2	2				
23	30	2	2				
24	31	2	2				
25	31	2	2				
26	19	3	2				

Рис. 3.2. Результаты обследования работников производства

Вызовем окно табличных опций, нажав вторую слева кнопку в нижнем ряду кнопок (рис. 3.4). Установим флажок **ANOVA Table** (таблица дисперсионного анализа) и нажмем **OK**. Щелкнув дважды на окне с этой таблицей, раскроем его на все рабочее поле (рис. 3.5).

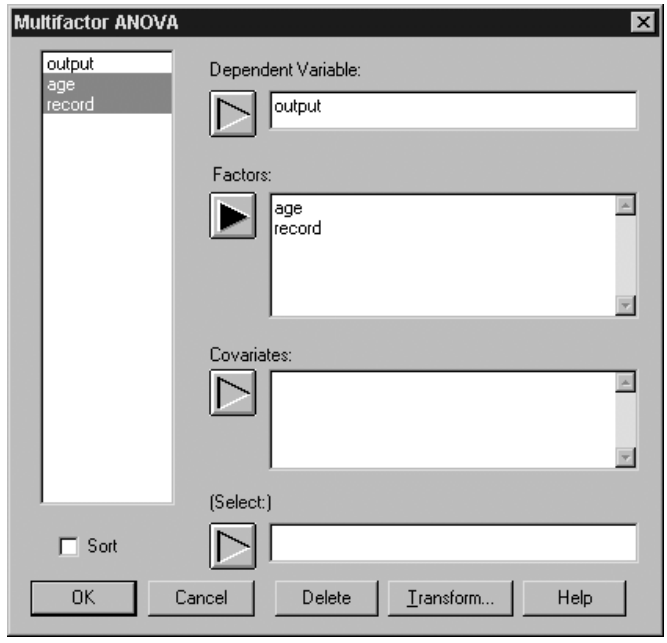


Рис. 3.3. Окно диалога многофакторного дисперсионного анализа

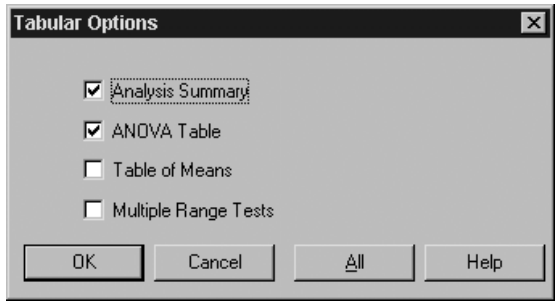


Рис. 3.4. Табличные окна дисперсионного анализа

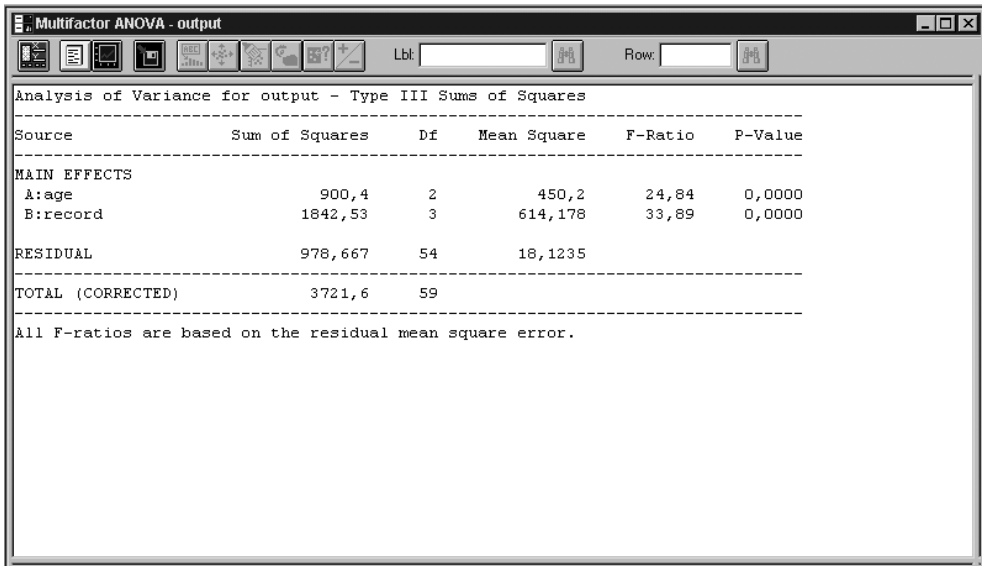


Рис. 3.5. Исходная таблица дисперсионного анализа

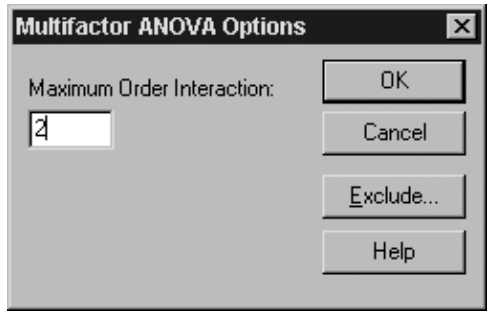


Рис. 3.6. Окно диалога для задания порядка взаимодействия факторов

Multifactor ANOVA - output

Lbl: Row:

Analysis of Variance for output - Type III Sums of Squares

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
MAIN EFFECTS					
A:age	900,4	2	450,2	48,98	0,0000
B:record	1842,53	3	614,178	66,82	0,0000
INTERACTIONS					
AB	537,467	6	89,5778	9,75	0,0000
RESIDUAL	441,2	48	9,19167		
TOTAL (CORRECTED)	3721,6	59			

All F-ratios are based on the residual mean square error.

Рис. 3.7. Таблица дисперсионного анализа с оценкой значимости совокупного влияния возраста и стажа работников на производительность труда

На основании табличных чисел (а также по сообщению **StatAdvisor**) делаем заключение, что на производительность труда оказывают влияние оба фактора по отдельности – и возраст работника, и его трудовой стаж. Доверие к такому выводу – более 99%. Можно, кроме того, оценить и совместное влияние двух факторов.

Щелкнем правой кнопкой мыши на табличном окне и выберем **Analysis Options**. Появится окно диалога для ввода различных взаимодействий факторов и задания их порядка (рис. 3.6).

Введем порядок взаимодействия, равный 2, и нажмем **OK**. В таблицу многофакторного дисперсионного анализа будут добавлены оценки статистической значимости совместного влияния возраста и стажа работников на их производительность труда (рис. 3.7).

Как следует из полученных цифр, на производительность труда изучаемой совокупности работников существенно влияют совместно действующие возраст и стаж. Уровень доверия к такому выводу составляет выше 99%. Можно еще более углубить проводимое исследование, воспользовавшись многосторонними оценками различных компонент факторного взаимодействия и дополнительными статистическими тестами, реализованными в процедуре дисперсионного анализа STATGRAPHICS. Но, как говорится, лучше один раз увидеть, чем сто раз услышать. Поэтому воспользуемся графическими возможностями отображения результатов анализа.

Нажмем кнопку графических опций (третья слева в нижнем ряду кнопок) и установим флажки **Means Plot** (график средних) и **Interactions Plot** (график взаимодействий). Нажмем **OK** (рис. 3.8).

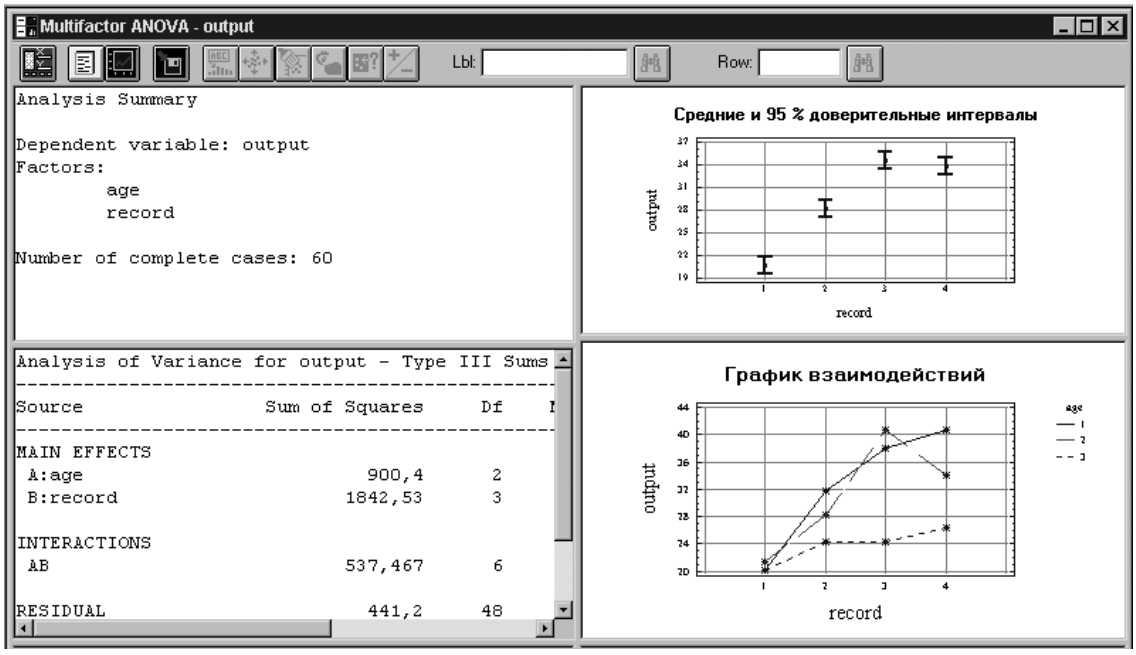


Рис. 3.8. Табличные и графические отображения результатов

В верхнем графическом окне показан график зависимости средних значений производительности труда от стажа и очерчены доверительные интервалы для этих средних. Хорошо видно, что стаж несомненно влияет на результативный признак. Вместе с тем, похоже, производительность достигает своего пика у работников со стажем от 7 до 10 лет, а затем начинает снижаться.

Полученная картина проясняется, если взглянуть на нижнее графическое окно, где приведена картинка, иллюстрирующая взаимодействие возраста и стажа. Из нее следует, что производительность труда постоянно увеличивается с ростом стажа у молодых работников (25–35 лет). Для второй возрастной группы (35–40 лет) такой рост наблюдается только для тех работников, стаж которых не превышает 10 лет. Затем производительность у них резко падает. Для третьей возрастной группы (45–55 лет) характерна вообще самая низкая производительность труда, значение которой остается почти на одном и том же уровне независимо от стажа работы.

Отобразим результаты дисперсионного анализа в ином ракурсе. Для этого будем щелкать правой кнопкой на каждом графическом окне, выбирая из контекстного меню пункт **Pane Options**, и заменять в соответствующих окнах диалога фактор **record** (стаж) на фактор **age** (возраст). Теперь на всех графиках по оси абсцисс будут отображаться возрастные категории. Пример одного из окон диалога приведен на рисунке 3.9.

Раскроем полученные графические окна двумя щелчками левой кнопки мыши. Получим следующие картинки (рис. 3.10 и 3.11).

Первый график показывает уменьшение производительности труда с возрастом. Из второго следует, что максимальная производительность труда наблюдается у первой возрастной группы (25–35 лет) со стажем свыше 4–7 лет и у второй возрастной группы (35–45 лет) со стажем 7–10 лет. Также видно, что при незначительном стаже, независимо от возраста, производительность труда всегда является самой низкой. Кроме того, можно еще раз заметить, что в третьей возрастной группе (45–55 лет) производительность труда существенно снижается.

На этих выводах, как правило, анализ данных завершается. Мы вроде бы удовлетворили свое любопытство и можем теперь аргументированно со ссылкой на статистические критерии обосновать желание работодателей иметь дело с молодыми специалистами, имеющими, однако, достаточно большой стаж.

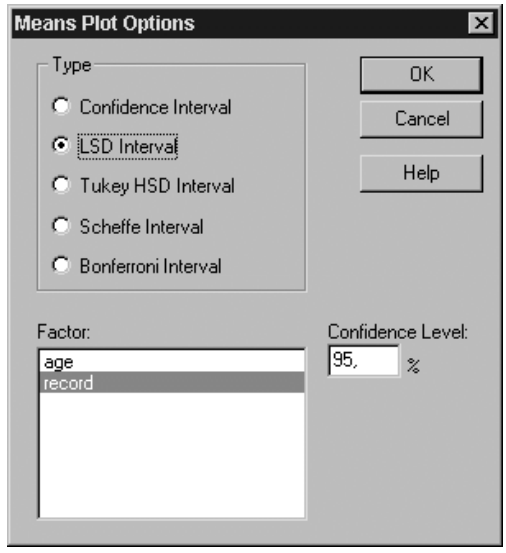


Рис. 3.9. Пример окна диалога для задания параметров графических отображений результатов дисперсионного анализа

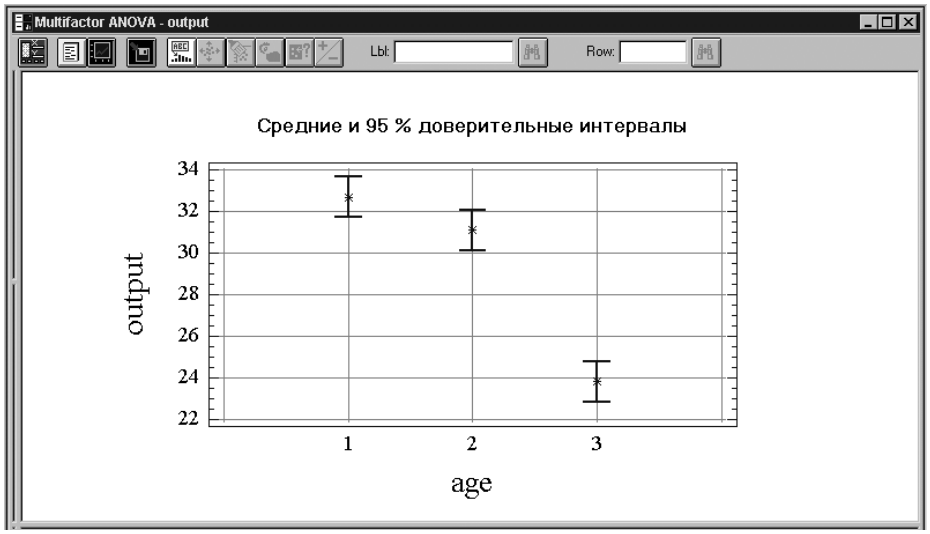


Рис. 3.10. Влияние возраста работников на производительность труда

Рис. 3.11. Влияние взаимодействия возраста и стажа на производительность труда

Вместе с тем наша аргументация основана на анализе усредненных характеристик и высвечивает лишь общие тенденции в рассматриваемом вопросе. А это вряд ли уместно, когда речь идет о конкретных людях, по отношению к которым требуется принимать административное решение. Как мы убедимся ниже, с помощью технологии обнаружения логических закономерностей в данных можно сделать гораздо более ответственные выводы по анализируемой ситуации.

Обработка данных системой WizWhy

Система WizWhy после обработки данных о производительности труда выдала следующий отчет.

Total number of records: **60**
Minimum probability of the:
1) if-then rules: **0.730**
2) if-then-not rules: **0.630**
Minimum number of cases in a rule: **8**
Field to Predict: **Производительность**
Predicted Value (analyzed as Boolean): ***more than* 25**
Prediction error costs:
The cost of a miss: 1
The cost of a false alarm: 1
Average probability of the predicted value is 0.550

ANALYSIS OF THE RULES EXPLANATORY POWER
Decision point: Predict more than 25 when conclusive probability is more than **0.706**
Number of misses: 4
Number of false alarms: 1
Total number of errors: 5
Total cost of errors: 5
Success rate when predicting more than 25: 0.967
Success rate when predicting NOT more than 25: 0.867
Number of records with no relevant rules: 0
Average cost (per record): 0.083
Expected average cost (per record): 0.450
Improvement Factor: 5.400

IF-THEN RULES:

1) *If* **Стаж** *is* **1**
Then
Производительность *is not* ***more than* 25**
Rule's probability: **1.000**
The rule exists in **15** *records.*
Significance Level: Error probability < 0.1
Positive Examples (records' serial numbers):
1, 2, 3, 4, 6, 7, 8, 9, 10, 11

2) *If* **Возраст** *is* **1**
and **Стаж** *is* **2 ... 4** (average = **3**)
Then
Производительность *is* ***more than* 25**
Rule's probability: **1.000**
The rule exists in **15** *records.*
Significance Level: Error probability < 0.1
Positive Examples (records' serial numbers):
32, 34, 39, 40, 42, 43, 44, 48, 49, 51

3) *If* **Возраст** *is* **2**
and **Стаж** *is* **2 ... 4** (average = **3**)
Then
Производительность *is* ***more than* 25**
Rule's probability: **0.933**
The rule exists in **14** *records.*
Significance Level: Error probability < 0.1
Positive Examples (records' serial numbers):
30, 31, 33, 35, 36, 37, 45, 46, 47, 50

Negative Examples (records' serial numbers):
12

4) *If* **Возраст** *is* **3**
Then
Производительность *is not* ***more than*** **25**
Rule's probability: **0.800**
The rule exists in **16** *records.*
Significance Level: Error probability < 0.2
Positive Examples (records' serial numbers):
1, 4, 5, 11, 13, 14, 17, 18, 19, 20
Negative Examples (records' serial numbers):
28, 29, 38, 41

5) *If* **Стаж** *is* **2 ... 4** (average = **3**)
Then
Производительность *is* ***more than*** **25**
Rule's probability: **0.733**
The rule exists in **33** *records.*
Significance Level: Error probability < 0.3
Positive Examples (records' serial numbers):
28, 29, 30, 31, 32, 33, 34, 35, 36, 37
Negative Examples (records' serial numbers):
5, 12, 13, 18, 19, 20, 21, 23, 24, 25

6) *If* **Возраст** *is* **1**
Then
Производительность *is* ***more than*** **25**
Rule's probability: **0.750**
The rule exists in **15** *records.*
Significance Level: Error probability < 0.3
Positive Examples (records' serial numbers):
32, 34, 39, 40, 42, 43, 44, 48, 49, 51
Negative Examples (records' serial numbers):
2, 6, 7, 8, 15

Из найденых 6 правил 4 описывают группу работников со средней и высокой производительностью труда, а 2 относятся к низкопроизводительным работникам. Рассмотрим эти правила раздельно по группам.

Группа с высокой производительностью

В первую очередь обращает на себя внимание правило № 2. Оно расшифровывается следующим образом: если (возраст от 25 до 35 лет) и (стаж больше 4 лет), то высокая производительность труда. Это правило безошибочно и описывает 15 работников.

Другое правило № 3 также имеет достаточно высокую точность, апеллирует к лицам от 35 до 45 лет, имеющим стаж также более 4 лет. Данное правило описывает 14 работников и делает всего одну ошибку.

Оставшиеся два правила обладают меньшей точностью и представляют собой высказывания отдельно по возрасту и стажу. Так правило № 5 с точностью 0,73 говорит о том, средняя и высокая производительность наблюдается у 33 работников, имеющих стаж более 4 лет. Правило № 6 с точностью 0,75 утверждает, что 15 высокопроизводительных работников отличаются сравнительно молодым возрастом (от 25 до 35 лет).

Группа с низкой производительностью

Правило № 1 здесь говорит о том, что при небольшом стаже никогда не следует ожидать хорошей производительности от работников. Это правило описывает 15 случаев низкой производительности со 100%-ной точностью.

И наконец, правило № 4 состоит в том, что 80% людей в возрасте от 45 лет и выше не показывают удовлетворительных результатов работы в условиях рассмотренного производства.

3.3. Выяснение причин неурожайности сельскохозяйственных участков

Исходные данные

Исходные данные заимствованы из книги [3].

На 43 опытных участках по возделыванию риса был получен различный урожай. Агротехника возделывания культуры характеризовалась следующими признаками:

x_1 – предшественник (в баллах);
x_2 – количество удобрений (ц на 1 га);
x_3 – прополка (раз);
x_4 – число дней от залива до сброса воды;
x_5 – число дней от косовицы до обмолота.

Экспериментальные данные представлены в таблице 3.4.

Таблица 3.4. Значения признаков для участков с различной урожайностью риса

№ п/п	Урожайность, ц с 1 га	Предшественник, баллы	Кол-во удобрений, ц на 1 га	Про-полка, раз	Число дней от залива до сброса воды	Число дней от косовицы до обмолота
	y	x_1	x_2	x_3	x_4	x_5
Группа 1						
1	36	2,8	1,47	1,2	115	8
2	36,1	3	1,23	1,3	117	7
3	36,1	2,7	1,31	1,4	114	9
4	36,2	3	1,5	1,5	119	10
5	36,4	3,2	1,14	1,6	120	7
6	36,9	2,8	1,22	1,6	121	11
7	37,5	2,7	1,3	1,3	122	8
8	37,8	3,3	1,24	1,3	118	10
9	38,2	2,8	1,16	1,9	119	7
10	38,6	2,7	1,22	1,6	117	9
11	38,9	2,8	1,35	1,2	119	10
12	39	2,9	1,4	1,4	115	8
13	39	3,1	1,36	1,3	120	11
14	39,2	2,8	1,23	1,6	114	10
15	39,4	2,7	1,3	1,4	118	9
16	39,5	3	1,41	1,3	117	8
17	39,7	2,9	1,28	1,4	120	12
18	39,7	3,1	1,36	1,2	121	9
19	39,8	2,8	1,32	1,4	118	7
20	40	2,9	1,4	1,5	118	10
Группа 2						
21	41,2	3,2	1,05	1,5	109	9
22	41,4	2,8	1,1	1,2	108	10
23	41,6	2,9	1,2	1,6	118	10
24	41,8	3	1,12	1,3	110	14
25	41,9	3,3	1,08	1,4	112	12
26	42,2	2,7	1,13	1,5	111	15
27	42,5	3	1,18	1,7	112	12
28	42,8	3,1	1,22	1,3	113	14
29	43,1	3,3	1,25	1,8	112	13
30	43,1	2,9	1,1	1,7	113	10
31	43,2	2,8	1,2	1,8	112	15
32	43,6	3,2	1,26	1,6	113	9
33	43,7	3,4	1,28	1,8	110	12
34	43,8	3,5	1,22	1,9	114	13
35	43,8	3	1,19	1,7	108	16
36	43,9	2,8	1,29	1,7	108	12
37	43,9	2,9	1,24	1,6	112	10
38	44,2	3	1,17	1,8	114	9

Продолжение табл. 3.4

№ п/п	Урожайность, ц с 1 га	Предшественник, баллы	Кол-во удобрений, ц на 1 га	Прополка, раз	Число дней от залива до сброса воды	Число дней от косовицы до обмолота
	y	x_1	x_2	x_3	x_4	x_5
39	44,6	3,3	1,25	1,3	115	11
40	44,8	3,4	1,27	1,7	112	12
41	44,9	3,5	1,26	1,5	111	14
42	44,9	3,1	1,3	1,5	119	11
43	45	3,2	1,24	1,6	110	13

В первоисточнике на основании специальных критериев математической статистики делается вывод, что рассматриваемая совокупность из 43 опытных участков по возделыванию риса не может считаться однородной. На этом обсуждение экспериментального материала фактически заканчивается. Главный вопрос о взаимосвязях между агротехническими мероприятиями и урожайностью рассматриваемой сельскохозяйственной культуры остался нерешенным.

Ниже представлены результаты комплексного исследования экспериментальных данных из таблицы 3.4. На первом этапе эти данные были обработаны рядом традиционных методов прикладной статистики с помощью пакета программ STATGRAPHICS *Plus* for Windows. А на втором этапе поиска ответа на сформулированный вопрос применялись методы поиска логических закономерностей в данных.

Используемые обозначения: **yield** – урожайность; **predec** – предшественник, **fertil** – количество удобрений; **weeding** – прополка; **water** – число дней от залива до сброса воды; **trashing** – число дней от косовицы до обмолота.

Комплексная обработка данных традиционными методами

Сравнение средних значений признаков

На рисунке 3.12 приведены совмещенные гистограммы распределений значений всех анализируемых признаков (**class** = 0 – объекты с низкой урожайностью, **class** = 1 – с высокой).

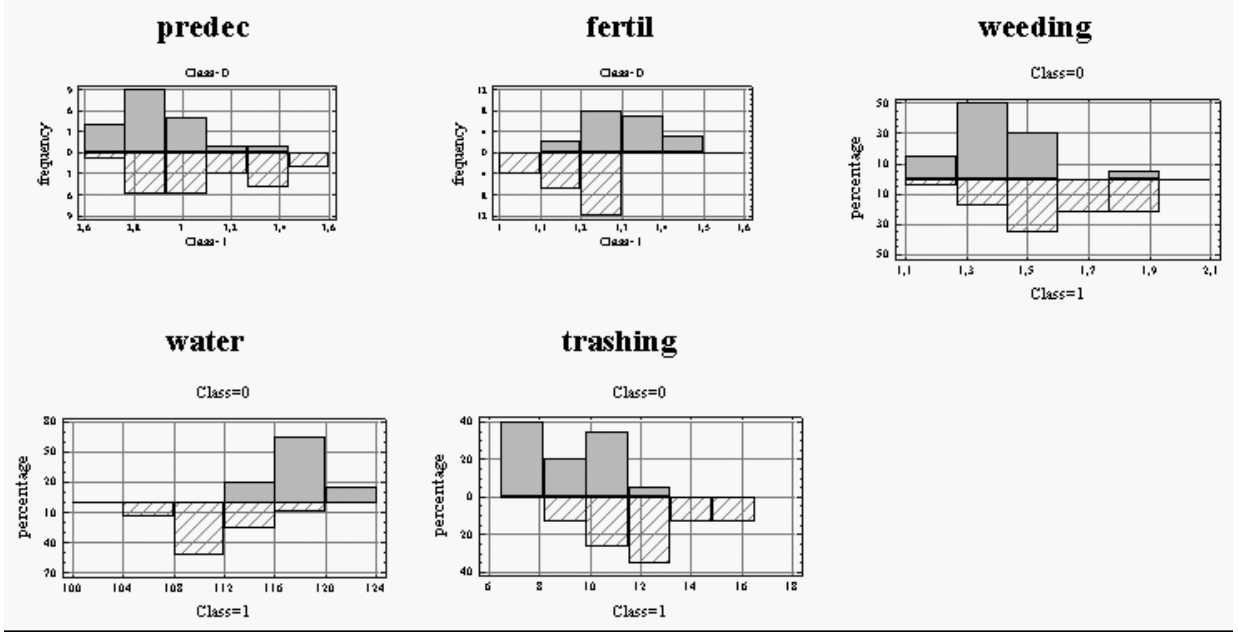


Рис. 3.12. Гистограммы распределения значений признаков

Из гистограмм следует, что более высокая урожайность чаще встречается на участках с более высокой балльной оценкой предшественника, со средними и ниже количествами внесенных удобрений, с более частой прополкой, с более ранним сбросом воды и более поздним обмолотом после косовицы. Эти выводы подтверждаются результатами Т-тестов для сравнения средних значений на уровне значимости 0,001.

Вместе с тем более глубокие выводы делаются на основании многомерного анализа экспериментальных данных, в котором учитывается совокупное взаимодействие признаков.

Метод главных компонент

В результате применения метода главных компонент (МГК) оказалось, что разделение двух классов сельскохозяйственных участков имеет проекции объектов на 1-ю главную компоненту, на которую приходится 45% дисперсии анализируемой выборки. Весовые коэффициенты для этой главной компоненты приведены в таблице 3.5.

Таблица 3.5. Веса признаков в первой главной компоненте

Признак	**predec**	**fertil**	**weeding**	**water**	**trashing**
Вес	0,32	–0,43	0,40	–0,54	0,50

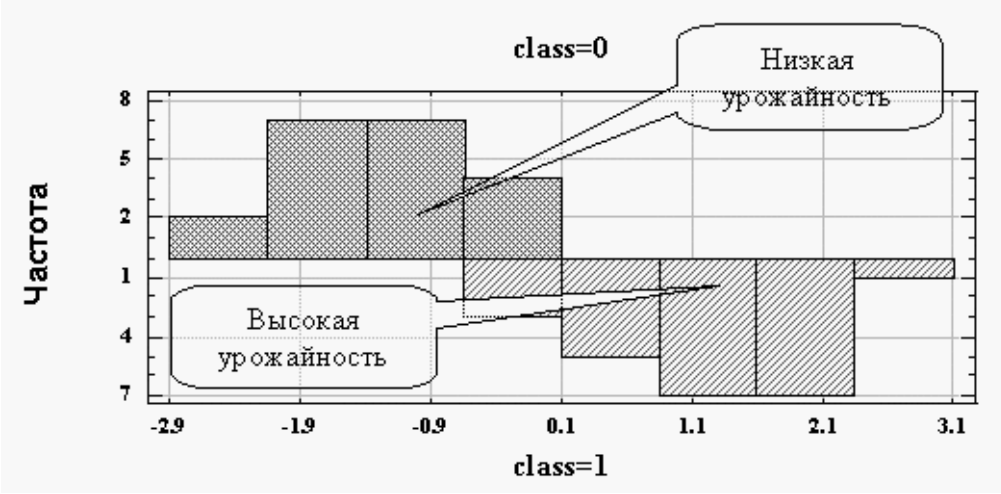


Рис. 3.13. Проекция объектов на 1-ю главную компоненту

Представление о разделении проекций классов на 1-ю главную компоненту дает рисунок 3.13.

Из рисунка следует, что на проекции имеется сравнительно небольшая область неопределенности, в которую попадают 3 участка с высокой урожайностью и 4 с низкой. В целом смысл, который имеет 1-я главная компонента, определяемый по весам входящих в нее признаков, совпадает с интерпретацией одномерного анализа – повышение урожайности риса положительно связано с балльной оценкой предшественника, с частотой прополки и с временным интервалом от косовицы до молотьбы и отрицательно связано с количеством внесенных удобрений и числом дней от залива до сброса воды.

Множественный регрессионный анализ

Сводка множественного регрессионного анализа, в котором независимой переменной выступает урожайность участков **yield**, а предикторами служат вышеописанные признаки $x_1, \ldots, x_5$, приведена ниже. При построении регрессионной модели применялся алгоритм последовательного уменьшения группы признаков.

Таблица 3.6. Сводка множественного регрессионного анализа

Dependent variable: yield

Parameter	Estimate	Standard Error	T Statistic	P-Value
CONSTANT	55,4292	12,6433	4,38407	0,0001
predec	3,71752	1,40207	2,65144	0,0115
water	-0,262156	0,09313	-2,81495	0,0076
trashing	0,424119	0,161582	2,6248	0,0123

Analysis of Variance

Source	Sum of Squares	Df	Mean Square	F-Ratio	P-Value
Model	208,339	3	69,4465	17,40	0,0000
Residual	155,661	39	3,9913		
Total (Corr.)	364,0	42			

R-squared = 57,2361 percent
R-squared (adjusted for d.f.) = 53,9466 percent
Standard Error of Est. = 1,99782
Mean absolute error = 1,50813
Durbin-Watson statistic = 1,04477

Как следует из полученной сводки, регрессионная модель заслуживает доверия более 99%. Однако коэффициент детерминации сравнительно невысок и составляет 57%, а средняя абсолютная ошибка слишком велика, чтобы модель могла претендовать на точный прогноз урожайности, и равна 1,5. Фактически эта модель пригодна для ориентировочного осмысления статистической связи независимой переменной и предикторов. Из модели следует, что рост урожайности риса положительно связан с балльной оценкой предшественника и числом дней между косовицей и обмолотом и отрицательно связан с числом дней от залива до сброса воды.

Дискриминантный анализ

Применялся классический вариант дискриминантного анализа, основанный на определении канонических направлений в исходном пространстве признаков. Обучающая информация задавалась переменной **class**, которая принимает значение 0 для объектов с низкой урожайностью и значение 1 в группе объектов с высокой урожайностью. Дискриминантный анализ проводился с применением алгоритма последовательного уменьшения группы признаков. Результаты приведены в таблицах 3.7 и 3.8. Гистограмма распределения значений дискриминантной функции показана на рисунке 3.14.

Таблица 3.7. Стандартизированные коэффициенты дискриминантной функции

Predec	**Fertil**	**Water**	**Trashing**
– 0,459109	0,476217	0,680748	–0,446469

Таблица 3.8. Таблица классификаций

Номер класса	**Объем группы**	**Предсказанный класс**	
		0	**1**
0	20	20 (100%)	0 (0%)
1	23	2 (8,7%)	21 (91,3%)
Процент правильной классификации – 95,35%			

По формальному эффекту дискриминантная функции обеспечивает несколько лучшее разделение групп участков с различной урожайностью риса по сравнению с проекциями объектов на 1-ю главную компоненту. Здесь только два участка с высокой урожайностью ошибочно относятся к группе с низкой урожайностью, а коэффициент канонической корреляции дискриминантной функции с переменной **class** составляет 0,87.

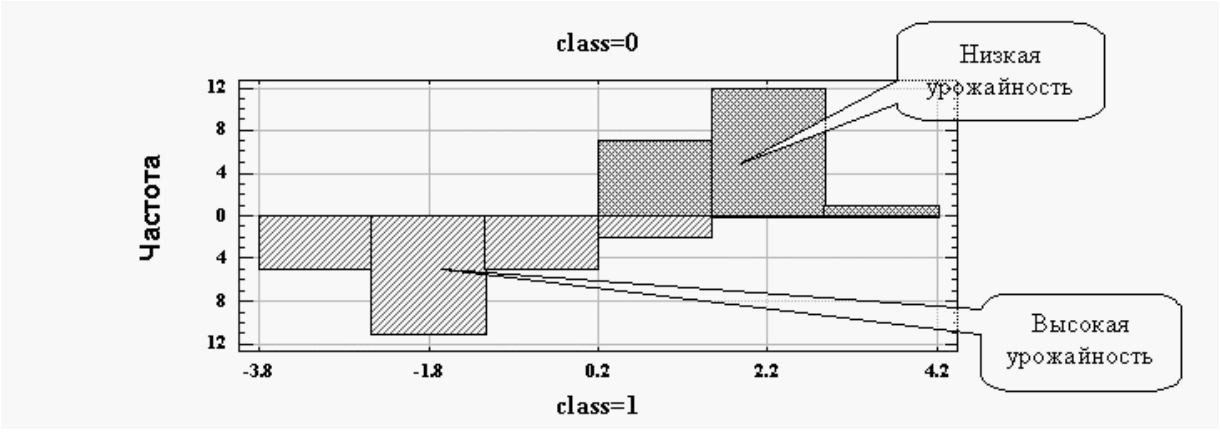


Рис. 3.14. Гистограмма распределения значений дискриминантной функции

Вместе с тем содержательная сторона проведенного дискриминантного анализа, как и в предыдущих видах анализа, носит качественный характер. А именно – можно лишь утверждать, что повышение урожайности положительно связано с балльной оценкой предшественника **predec** и числом дней от косовицы до обмолота **trashing** и отрицательно связано с количеством внесенных удобрений **fertil** и временным интервалом от залива до сброса воды **water**.

В целом результаты комплексной обработки агротехнических характеристик участков с различной урожайностью риса с применением аппарата одномерного и многомерного статистического анализа дают основание лишь для выявления общих тенденций. Кроме того, наблюдаются некоторые расхождения результатов различных методов. Так, метод главных компонент говорит о важности учета всех пяти агротехнических признаков, модель дискриминантного анализа не испытывает необходимости учета прополки, а регрессионный анализ, кроме всего этого, показывает, что не требуется привлекать для объяснений агротехнических закономерностей количество удобрений. Это происходит главным образом из-за внутригрупповой неоднородности экспериментальных данных.

Агротехническая система имеет высокий уровень сложности и для более глубокого и полного проникновения в ее суть должен применяться аппарат, умеющий выявлять и учитывать структурные неоднородности данных об этой системе. Такими свойствами обладают методы индуктивного логического вывода, позволяющие находить в ограниченном наборе экспериментальных фактов логические закономерности.

Результаты обработки данных системой See5

Decision tree:

```
(дней от залива до сброса воды) <= 113: 0 (18.0)
(дней от залива до сброса воды) > 113:
:...(дней от косовицы до обмолота) <= 10: 1 (11.0/2.0)
    (дней от косовицы до обмолота) > 10:
    :...(дней от залива до сброса воды) <= 119: 0 (3.0)
        (дней от залива до сброса воды) > 119: 1 (3.0)
```

Extracted rules:

```
Rule 1: (cover 18)
        (дней от залива до сброса воды) <= 113
        ->  высокая урожайность  [0.950]
Rule 2: (cover 16)
        (дней от залива до сброса воды) <= 119
        (дней от косовицы до обмолота) > 10
        ->  высокая урожайность  [0.944]
```

```
Rule 3: (cover 6)
        (дней от залива до сброса воды) > 119
        ->  низкая урожайность  [0.875]
Rule 4: (cover 19)
        (дней от залива до сброса воды) > 113
        (дней от косовицы до обмолота) <= 10
        ->  низкая урожайность  [0.857]
```

Ошибки классификации на обучающей выборке:

```
           Decision Tree           Rules
         ----------------    ----------------
         Size      Errors      No      Errors
            4    2( 4.7%)       4    2( 4.7%)   <<

          (a)   (b)   <-classified as
         ----  ----
           21     2   (a): высокая урожайность
                 20   (b): низкая урожайность
```

Результаты обработки данных системой WizWhy

Ниже представлен отчет системы WizWhy для следующих установочных параметров поиска логических правил в экспериментальных данных.

- Целевая переменная – **урожай.**
- Минимальная вероятность if-then правила – 0,7.
- Минимальная вероятность if-then-NOT правила – 0,7.
- Минимальное количество объектов, покрываемых правилом, – 10.

PARAMETERS OF THE RULES AND DATA
Total number of records: **43**
Minimum probability of the:
1) if-then rules: **0,700**
2) if-then-not rules: **0,700**
Minimum number of cases in a rule: **10**

Field to Predict: **Урожай**
Predicted Value (analyzed as Boolean): ***more than*** **40,93**
Prediction error costs:
The cost of a miss: 1
The cost of a false alarm: 1
Average probability of the predicted value is 0,535

ANALYSIS OF THE RULES EXPLANATORY POWER
Decision point: Predict more than 40,93 when conclusive probability is more than **0,484**
Number of misses: 1
Number of false alarms: 2
Total number of errors: 3
Total cost of errors: 3
Success rate when predicting more than 40,93: 0,917
Success rate when predicting NOT more than 40,93: 0,947
Number of records with no relevant rules: 0
Average cost (per record): 0,070
Expected average cost (per record): 0,488
Improvement Factor: 7,000

IF-THEN RULES:

1) *If* **Дней от залива до сброса воды** *is* **108,00 ... 113,00** (average = **110,89**) *Then* **Урожай** *is* ***more than*** **40,93** *Rule's probability:* **1,000**	10) *If* **К-во удобрений** *is* **1,08 ... 1,27** (average = **1,20**) and **Дней от косовицы до обмолота** *is* **10,00 ... 16,00** (average = **12,25**) *Then* **Урожай** *is* ***more than*** **40,93**

The rule exists in **18** *records.*
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
21, 22, 24, 25, 26, 27, 28, 29, 30, 31
2) *If* **Предшественник** *is* **2,70 ... 2,80** (average = **2,76**) and **Дней от залива до сброса воды** *is* **114,00 ... 122,00** (average = **117,70**)
Then
Урожай *is not* ***more than*** **40,93**
Rule's probability: **1,000**
The rule exists in **10** *records.*
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
1, 3, 6, 7, 9, 10, 11, 14, 15, 19
3) *If* **К-во удобрений** *is* **1,28 ... 1,47** (average = **1,35**) and **Прополка** *is* **1,20 ... 1,40** (average = **1,32**)
Then
Урожай *is not* ***more than*** **40,93**
Rule's probability: **1,000**
The rule exists in **11** *records.*
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
1, 3, 7, 11, 12, 13, 15, 16, 17, 18
4) *If* **К-во удобрений** *is* **1,28 ... 1,50** (average = **1,36**) and **Дней от залива до сброса воды** *is* **114,00 ... 122,00** (average = **118,21**)
Then
Урожай *is not* ***more than*** **40,93**
Rule's probability: **0,929**
The rule exists in **13** *records.*
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
1, 3, 4, 7, 11, 12, 13, 15, 16, 17
Negative Examples (records' serial numbers):
42
5) *If* **Прополка** *is* **1,20 ... 1,40** (average = **1,31**) and **Дней от залива до сброса воды** *is* **114,00 ... 122,00** (average = **117,79**)
Then
Урожай *is not* ***more than*** **40,93**
Rule's probability: **0,929**
The rule exists in **13** *records.*
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
1, 2, 3, 7, 8, 11, 12, 13, 15, 16
Negative Examples (records' serial numbers):
39
6) *If* **Дней от залива до сброса воды** *is* **114,00 ... 122,00** (average = **117,46**) and **Дней от косовицы до обмолота** *is* **7,00 ... 9,00** (average = **8,08**)
Then
Урожай *is not* ***more than*** **40,93**
Rule's probability: **0,923**
The rule exists in **12** *records.*
Significance Level: Error probability < 0,1
Positive Examples (records' serial numbers):
1, 2, 3, 5, 7, 9, 10, 12, 15, 16
Negative Examples (records' serial numbers):
38
7) *If* **Дней от залива до сброса воды** *is* **114,00 ...**
Rule's probability: **0,850**
The rule exists in **17** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
22, 23, 24, 25, 26, 27, 28, 29, 30, 31
Negative Examples (records' serial numbers):
6, 8, 14
11) *If* **Предшественник** *is* **3,20 ... 3,50** (average = **3,32**)
Then
Урожай *is* ***more than*** **40,93**
Rule's probability: **0,833**
The rule exists in **10** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
21, 25, 29, 32, 33, 34, 39, 40, 41, 43
Negative Examples (records' serial numbers):
5, 8
12) *If* **Прополка** *is* **1,50 ... 1,90** (average = **1,65**) and **Дней от косовицы до обмолота** *is* **10,00 ... 16,00** (average = **12,05**)
Then
Урожай *is* ***more than*** **40,93**
Rule's probability: **0,789**
The rule exists in **15** *records.*
Significance Level: Error probability < 0,2
Positive Examples (records' serial numbers):
23, 26, 27, 29, 30, 31, 33, 34, 35, 36
Negative Examples (records' serial numbers):
4, 6, 14, 20
13) *If* **Предшественник** *is* **2,70 ... 2,80** (average = **2,76**)
Then
Урожай *is not* ***more than*** **40,93**
Rule's probability: **0,714**
The rule exists in **10** *records.*
Significance Level: Error probability < 0,3
Positive Examples (records' serial numbers):
1, 3, 6, 7, 9, 10, 11, 14, 15, 19
Negative Examples (records' serial numbers):
22, 26, 31, 36
14) *If* **Прополка** *is* **1,20 ... 1,40** (average = **1,31**)
Then
Урожай *is not* ***more than*** **40,93**
Rule's probability: **0,722**
The rule exists in **13** *records.*
Significance Level: Error probability < 0,3
Positive Examples (records' serial numbers):
1, 2, 3, 7, 8, 11, 12, 13, 15, 16
Negative Examples (records' serial numbers):
22, 24, 25, 28, 39
15) *If* **К-во удобрений** *is* **1,05 ... 1,27** (average = **1,19**)
Then
Урожай *is* ***more than*** **40,93**
Rule's probability: **0,741**
The rule exists in **20** *records.*
Significance Level: Error probability < 0,3
Positive Examples (records' serial numbers):
21, 22, 23, 24, 25, 26, 27, 28, 29, 30
Negative Examples (records' serial numbers):
2, 5, 6, 8, 9, 10, 14
16) *If* **Прополка** *is* **1,50 ... 1,90** (average = **1,65**)
Then

122,00 (average = **117,68**) *Then* **Урожай** *is not* ***more than*** **40,93** *Rule's probability:* **0,800** *The rule exists in* **20** *records.* *Significance Level: Error probability* < 0,2 *Positive Examples (records' serial numbers):* ***1, 2, 3, 4, 5, 6, 7, 8, 9, 10*** *Negative Examples (records' serial numbers):* ***23, 34, 38, 39, 42*** 8) *If* **К-во удобрений** *is* **1,28 ... 1,50** (average = **1,35**) *Then* **Урожай** *is not* ***more than*** **40,93** *Rule's probability:* **0,813** *The rule exists in* **13** *records.* *Significance Level: Error probability* < 0,2 *Positive Examples (records' serial numbers):* ***1, 3, 4, 7, 11, 12, 13, 15, 16, 17*** *Negative Examples (records' serial numbers):* ***33, 36, 42*** 9) *If* **Дней от косовицы до обмолота** *is* **7,00 ... 9,00** (average = **8,20**) *Then* **Урожай** *is not* ***more than*** **40,93** *Rule's probability:* **0,800** *The rule exists in* **12** *records.* *Significance Level: Error probability* < 0,2 *Positive Examples (records' serial numbers):* ***1, 2, 3, 5, 7, 9, 10, 12, 15, 16*** *Negative Examples (records' serial numbers):* ***21, 32, 38***	**Урожай** *is* ***more than*** **40,93** *Rule's probability:* **0,720** *The rule exists in* **18** *records.* *Significance Level: Error probability* < 0,3 *Positive Examples (records' serial numbers):* ***21, 23, 26, 27, 29, 30, 31, 32, 33, 34*** *Negative Examples (records' serial numbers):* ***4, 5, 6, 9, 10, 14, 20*** 17) *If* **Дней от косовицы до обмолота** *is* **10,00 ... 16,00** (average = **11,89**) *Then* **Урожай** *is* ***more than*** **40,93** *Rule's probability:* **0,714** *The rule exists in* **20** *records.* *Significance Level: Error probability* < 0,3 *Positive Examples (records' serial numbers):* ***22, 23, 24, 25, 26, 27, 28, 29, 30, 31*** *Negative Examples (records' serial numbers):* ***4, 6, 8, 11, 13, 14, 17, 20***

Полученная система 17 логических правил обладает следующими характеристиками:

- вероятность правильного предсказания высокого урожая (больше 40,93 ц/га) – 0,917;
- вероятность правильного предсказания низкого урожая (меньше 40,93 ц/га) – 0,947.

Фактически полученные правила представляют собой инструкцию по агротехнике с указанием конкретных значений факторов и их комбинаций, влияющих на урожайность рисовых участков. Это довольно обстоятельная и длинная инструкция, но, по-видимому, иначе вряд ли возможно описать сложную систему многофакторного взаимодействия с удовлетворительной точностью.

Сравнивая результаты двух систем See5 и WizWhy, бросается в глаза, что система See5 обнаружила всего 4 правила, но этого оказалось достаточно для описания рассмотренной агротехнической ситуации (только 2 ошибки на обучающей выборке). Вместе с тем одно из четырех обнаруженных правил в системе See5 обладает малой полнотой (охватывает 6 случаев). Это, по-видимому, несколько снижает его ценность.

3.4. Прогнозирование продолжительности ремиссий при алкоголизме

Исходным материалом для исследования служили исторические данные о 266 пациентах, проходивших лечение в отделении лечения больных алкоголизмом Психоневрологического института им. В. М. Бехтерева. Эти данные собирались на основе «Прогностической карты ремиссий при алкоголизме» [4], которая включает более 400 признаков, отражающих анамнестические сведения о больном и его социально-психологическую характеристику, а также клинические и социально-психологические данные о динамике ремиссий.

Приведенные ниже результаты по поиску правил для прогноза длительности ремиссий являются предварительными.

Общая характеристика данных

Обработке подвергались данные о клинико-психологических особенностях больных алкоголизмом на начальном этапе формирования ремиссии. Это признаки x130–x144 из прогностической карты за исключением признака x131 (данный признак не рассматривался, так как в его измерениях было довольно много пропущенных значений). Кроме того, в анализ был введен признак x125, характеризующий продолжительность спонтанных ремиссий в прошлом. Всего 15 признаков. Количество больных – 191. Все больные были разделены на две группы: 1-я группа – с продолжительностью ремиссии до 1 года (84 чел.); 2-я группа – с продолжительностью ремиссии больше 1 года (107 чел.). Ниже раскрывается содержание значений используемых признаков.

- x125 – **спонтанные ремиссии в прошлом**: 1 – нет, 2 – продолжительностью до 6 мес., 3 – до 1 года, 4 – более 2 лет;
- x130 – **влечение к алкоголю**: 1 – нет, 2 – эпизодическое, 3 – постоянное;
- x132 – **тревога**: 1 – нет, 2 – слабо/умеренно выраженная, 3 – выраженная;
- x133 – **внутреннее напряжение**: 1 – нет, 2 – слабо/умеренно выраженное, 3 – выраженное;
- x134 – **снижение настроения**: 1 – нет, 2 – слабо/умеренно выраженное, 3 – выраженное;
- x135 – **дисфория**: 1 – нет, 2 – слабо/умеренно выраженная, 3 – выраженная;
- x136 – **апатия**: 1 – нет, 2 – слабо/умеренно выраженная, 3 – выраженная;
- x137 – **эйфория**: 1 – нет, 2 – слабо/умеренно выраженная, 3 – выраженная;
- x138 – **дистимия**: 1 – нет, 2 – слабо/умеренно выраженная, 3 – выраженная;
- x139 – **астенические растройства**: 1 – нет, 2 – слабо/умеренно выраженные, 3 – выраженные;
- x140 – **неврозоподобные расстройства**: 1 – нет, 2 – слабо/умеренно выраженные, 3 – выраженные;
- x141 – **психопатоподобные расстройства**: 1 – нет, 2 – слабо/умеренно выраженные, 3 – выраженные;
- x142 – **психоорганические нарушения**: 1 – нет, 2 – слабо/умеренно выраженные, 3 – выраженные;
- x143 – **критика к болезни**: 1 – нет, 2 – слабо/умеренно выраженная, 3 – выраженная;
- x144 – **установка на трезвость**: 1 – нет, 2 – слабо/умеренно выраженная, 3 – выраженная.

Исходные данные приведены в таблице 3.9.

Таблица 3.9. Исходные данные

Group	x125	x130	x132	x133	x134	x135	x136	x137	x138	x139	x140	x141	x142	x143	x144
1	1	2	2	2	1	1	2	1	1	2	1	2	2	2	2
1	1	2	2	1	1	2	1	2	2	2	1	2	1	2	2
1	1	3	2	2	3	2	3	1	3	2	3	2	3	2	3
1	1	2	3	2	1	2	1	1	1	1	1	2	1	2	2
1	1	2	2	2	2	3	2	1	3	3	1	3	3	3	2
1	1	3	2	2	2	3	1	1	2	2	1	3	1	3	3
1	1	2	2	2	2	2	1	2	1	1	1	2	1	2	2
1	1	1	1	1	1	2	2	1	2	1	1	2	1	2	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2
1	1	1	1	1	1	1	1	2	1	2	1	1	1	2	1
1	1	1	1	1	1	1	1	2	1	1	1	1	1	2	2
1	1	2	2	2	1	2	1	1	1	1	1	1	1	2	2
1	2	1	2	2	1	1	1	1	1	1	2	1	1	1	2
1	1	2	1	1	1	1	1	1	1	1	1	1	1	2	2
1	1	2	2	1	1	2	2	1	1	2	2	1	1	1	1
1	1	1	2	1	1	1	1	1	1	2	1	1	1	2	2
1	1	2	1	2	1	2	1	1	2	1	1	1	1	2	1
1	2	1	2	2	2	2	1	2	2	2	2	1	1	1	1

Продолжение табл. 3.9

Group	x125	x130	x132	x133	x134	x135	x136	x137	x138	x139	x140	x141	x142	x143	x144
1	1	1	1	1	1	2	1	1	1	1	1	2	1	3	2
1	2	1	1	1	1	1	2	1	1	1	1	1	1	2	2
1	1	2	1	1	1	2	1	1	1	2	1	2	2	2	2
1	1	2	2	2	1	2	1	1	1	1	1	1	1	3	2
1	1	2	2	3	2	2	1	1	2	1	2	2	1	1	2
1	1	2	3	1	1	2	1	1	1	2	1	2	1	2	2
1	1	2	1	1	2	1	2	1	2	2	1	1	3	2	2
1	1	3	2	2	2	1	3	1	1	2	0	1	2	2	2
1	1	3	2	2	2	3	3	1	2	2	1	2	2	2	2
1	1	2	2	2	2	2	2	1	2	2	2	3	2	2	2
1	1	2	2	2	1	1	1	2	1	1	1	2	1	2	2
1	1	2	2	2	1	2	1	2	1	2	1	2	1	2	2
1	4	2	2	1	1	1	1	1	1	2	2	1	1	2	1
1	1	2	2	1	2	1	2	1	1	2	2	2	2	2	2
1	1	2	1	2	1	1	1	1	1	2	1	1	1	2	2
1	1	2	2	2	2	2	1	1	1	1	1	2	1	2	2
1	1	2	2	2	1	3	1	1	2	2	1	3	1	2	2
1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	2
1	1	2	1	1	1	3	1	1	1	2	1	2	2	2	2
1	1	1	2	1	1	2	1	1	1	1	1	2	1	2	2
1	1	2	1	1	1	2	1	2	1	1	1	1	1	2	1
1	1	2	1	2	1	2	1	2	2	1	1	2	2	3	2
1	1	2	1	1	2	2	2	1	1	2	1	2	2	3	3
1	1	2	1	2	2	1	2	1	1	2	1	3	1	2	2
1	1	1	2	1	1	1	1	2	1	1	1	1	1	2	1
1	2	1	1	1	1	1	1	2	1	1	1	1	1	2	2
1	1	2	2	1	1	2	1	2	1	1	1	2	2	2	2
1	2	2	2	1	1	2	1	2	1	2	1	2	2	2	1
1	1	2	2	2	2	2	1	1	1	2	1	2	1	2	2
1	1	2	1	1	1	2	1	2	1	1	1	2	1	2	2
1	1	1	1	1	1	1	1	2	1	1	1	1	1	2	1
1	1	2	2	2	1	1	1	1	1	2	2	2	2	2	2
1	1	2	2	1	1	1	1	1	1	1	1	1	2	2	2
1	2	1	2	2	2	1	1	1	1	1	1	1	1	2	1
1	1	1	2	2	1	2	1	1	1	1	1	2	2	2	2
1	1	2	2	1	1	1	1	1	1	1	1	2	2	2	2
1	1	2	2	2	2	2	2	1	1	2	1	2	1	2	2
1	1	2	2	2	1	1	1	2	1	1	1	1	1	2	2
1	1	2	2	2	1	2	1	2	1	2	1	1	1	2	1
1	1	2	1	2	2	2	1	1	1	1	2	3	1	2	2
1	1	2	1	1	1	1	2	1	1	2	1	1	2	2	2
1	1	2	1	2	1	2	1	2	2	1	1	3	2	2	2
1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
1	1	2	1	1	1	1	1	2	1	1	1	1	1	2	1
1	1	2	1	1	1	1	1	2	1	1	1	1	2	2	1
1	1	2	1	2	1	2	2	1	2	2	1	2	2	3	2
1	1	1	2	1	1	2	1	1	1	2	1	1	2	2	2
1	1	1	2	2	2	2	1	1	2	1	1	2	1	2	2
1	1	1	1	2	1	2	2	1	1	1	1	2	2	2	2
1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
1	1	1	1	1	1	1	1	1	1	2	1	1	1	2	2
1	1	2	1	2	1	1	1	1	1	2	2	1	1	2	1
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	2	1	1	2	1	1	1	2	1	2	2	2	2
1	2	1	1	1	1	1	1	1	1	2	1	1	2	2	2
1	1	1	1	1	1	2	1	2	1	2	1	3	2	2	2
1	1	1	1	1	1	2	1	1	1	1	1	2	1	2	2
1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1
1	1	2	2	2	1	2	1	1	2	1	2	2	1	2	2
1	1	1	1	2	2	1	1	1	2	2	2	1	1	2	1
1	1	2	2	1	1	2	1	2	1	2	1	1	2	2	2
1	2	1	1	2	1	2	1	1	2	1	1	2	1	2	2

Продолжение табл. 3.9

Group	x125	x130	x132	x133	x134	x135	x136	x137	x138	x139	x140	x141	x142	x143	x144
1	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
1	1	2	1	1	1	1	1	2	1	1	1	2	1	2	2
1	1	1	1	1	1	1	1	2	1	1	1	2	1	2	1
1	2	1	2	2	1	1	1	1	1	1	2	1	1	2	1
2	3	1	1	2	1	1	1	1	2	1	2	1	1	2	1
2	1	1	1	2	1	2	1	1	1	1	1	2	1	2	1
2	1	1	2	2	2	1	1	1	1	2	2	1	1	1	1
2	1	2	2	2	1	1	1	1	2	2	1	2	1	1	1
2	1	1	1	1	1	1	1	2	1	1	1	1	1	2	2
2	1	2	2	2	2	2	1	1	2	1	1	2	1	2	2
2	1	2	2	2	1	3	1	2	1	2	1	3	2	2	1
2	1	2	2	2	1	2	1	2	2	2	1	2	2	2	2
2	2	2	2	2	2	1	2	1	1	2	1	2	1	2	1
2	2	2	2	2	1	1	0	1	1	2	2	1	2	1	1
2	1	2	1	2	1	1	1	1	1	1	1	2	1	2	1
2	1	2	2	1	1	1	1	1	1	1	1	1	1	2	2
2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	1	2	2	1	1	1	1	1	1	1	2	1	1	1	1
2	2	1	1	1	1	2	1	2	1	1	1	1	1	2	1
2	2	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	1	2	2	2	2	2	1	1	1	1	1	1	1	2	2
2	1	1	2	2	1	1	1	1	1	1	1	1	1	1	1
2	1	3	2	2	2	2	2	1	2	2	1	2	1	2	2
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	1	1	1	1	1	1	1	1	1	2	1	1	1	2	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	1	2	1	1	1	1	1	2	1	1	1	1	1	2	2
2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	3	1	1	1	1	1	1	1	1	1	1	1	1	2	2
2	2	1	1	1	1	1	1	2	1	1	1	2	1	2	1
2	2	1	1	1	2	2	1	1	1	1	1	2	3	2	2
2	2	1	1	1	1	1	1	1	1	2	1	1	1	2	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	1	1	1	2	2	2	1	1	2	2	1	2	3	2	1
2	2	1	1	2	1	2	1	1	1	1	1	2	1	2	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	2	2	1	1	1	1	1	2	2	1	1	2	2
2	1	2	2	2	1	1	1	1	1	1	1	2	1	2	2
2	1	2	2	2	1	1	1	1	1	1	1	2	1	2	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	1	2	1	2	2	2	1	1	1	2	1	1	1	1	1
2	1	1	1	2	1	2	2	1	1	1	1	1	1	2	1
2	1	2	2	1	1	1	1	1	1	1	1	1	1	1	1
2	1	2	2	1	1	1	1	1	1	1	1	2	1	2	2
2	1	2	1	1	1	2	1	2	1	1	1	2	1	2	1
2	1	1	1	1	1	1	1	2	1	1	1	1	1	2	1
2	2	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	1	1	2	1	1	2	1	1	1	2	1	2	1	2	2
2	1	1	2	2	1	1	1	1	2	1	2	1	1	2	1
2	1	1	1	1	1	1	1	2	2	1	1	1	2	2	1
2	2	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	1	2	2	2	2	2	1	1	2	2	2	1	1	2	1
2	1	1	1	2	2	2	1	1	2	1	1	2	1	2	2
2	1	2	2	1	1	2	1	1	1	1	1	1	1	2	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	3	2	2	2	2	3	1	2	2	1	2	1	2	2
2	4	2	2	2	2	2	2	1	2	1	1	1	2	2	1
2	1	3	3	3	2	1	1	1	2	3	3	1	2	1	1

Продолжение табл. 3.9

Group	x125	x130	x132	x133	x134	x135	x136	x137	x138	x139	x140	x141	x142	x143	x144
2	3	2	2	2	2	2	1	1	1	2	2	1	1	1	1
2	1	2	2	3	1	3	1	1	2	1	1	2	2	1	1
2	1	1	1	1	1	1	1	2	1	1	1	1	1	2	1
2	1	1	2	2	1	1	1	1	1	1	1	1	1	2	1
2	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	2	2	2	2	2	1	1	2	1	1	2	1	3	2
2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	2	1	1	1	1	2	1	1	2	1	1	2	2	2
2	1	2	1	1	1	2	1	1	1	1	1	1	1	2	2
2	1	1	2	2	1	1	1	1	1	2	1	1	1	2	1
2	1	2	2	1	1	1	1	1	1	2	1	1	1	1	1
2	1	2	2	2	1	1	1	1	1	1	1	1	1	1	1
2	1	2	2	1	1	1	1	1	1	1	1	2	1	2	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	1	2	2	2	1	3	1	2	1	1	1	2	2	2	2
2	1	2	2	2	1	1	1	1	1	2	1	1	1	1	1
2	1	1	2	1	1	1	1	1	1	1	1	1	1	2	1
2	1	2	2	2	1	2	1	1	2	2	1	2	1	2	2
2	3	2	1	2	1	2	1	2	1	1	1	2	1	2	1
2	4	2	2	1	2	2	1	2	1	1	1	1	1	2	1
2	2	1	2	2	1	1	1	1	1	2	2	1	1	1	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	4	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	2	2	2	1	1	1	2	1	1	1	1	1	1	1
2	4	1	1	1	1	2	1	2	1	1	1	1	1	2	1
2	1	2	1	1	1	1	1	1	1	2	1	2	2	1	1
2	1	1	2	2	1	2	2	1	1	2	1	2	1	2	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	1	2	1	1	2	1	2	1	1	2	1	1	2	2	2
2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	2	2	2	2	1	1	1	1	2	2	1	2	2	2
2	1	1	1	2	1	1	2	1	1	2	1	1	2	2	1
2	2	1	1	1	1	2	1	2	1	2	1	1	1	2	2
2	2	1	1	1	1	1	1	1	1	2	1	1	2	2	1
2	2	1	1	1	1	1	2	1	1	1	1	1	1	1	1
2	1	1	1	2	1	2	2	1	1	2	1	1	2	2	1
2	1	2	1	2	2	1	2	1	1	2	1	1	1	2	2
2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	3	1	1	1	1	1	1	1	1	1	1	1	1	2	1
2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	2	1	1	1	1	1	2	2
2	2	2	2	1	1	1	1	1	1	2	1	1	1	2	1
2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	2	1	2	1	1	1	2	1	2	1
2	1	1	1	2	1	1	1	1	1	1	1	1	1	2	1

Частотный анализ признаков

На рисунке 3.15 приведены графические иллюстрации результатов частотного анализа. В эту таблицу включены признаки, которые могут считаться прогностически важными по критерию хи-квадрат ($p < 0,05$).

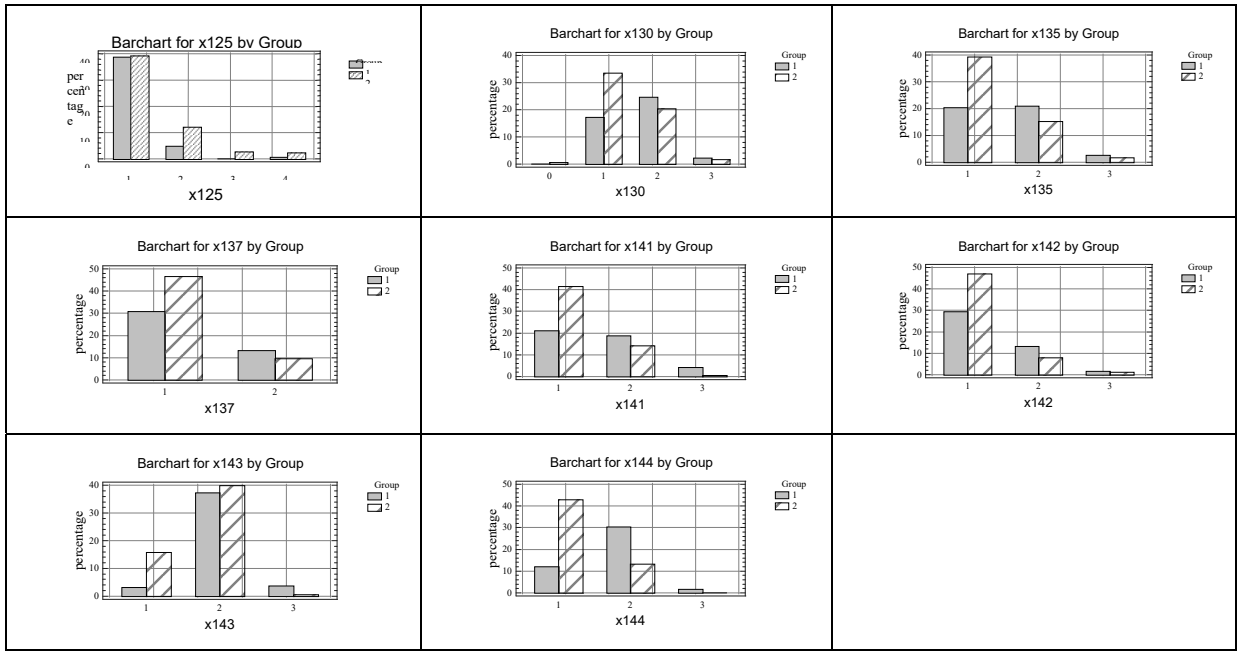


Рис. 3.15. Гистограммы распределения значений информативных признаков

Дискриминантный анализ

Приведенные выше результаты частотного анализа дали основание полагать, что исходное пространство признаков в значительной степени способно отражать значение целевого показателя – длительности ремиссии (большое число признаков по отдельности имеет статистически значимую связь с целевым показателем). Это подтвердили последующие результаты дискриминантного анализа, который проводился по классической схеме, дополненной алгоритмом последовательного уменьшения группы признаков. Получена следующая дискриминантная функция:

$$F = – 4{,}9 – 0{,}4 \cdot x125 + 0{,}9 \cdot x137 + 0{,}7 \cdot x140 + 0{,}6 \cdot x143 + 1{,}7 \cdot x144.$$

Дискриминантная функций обеспечивает 74,4% правильной классификации, что иллюстрируется на рисунке 3.16.

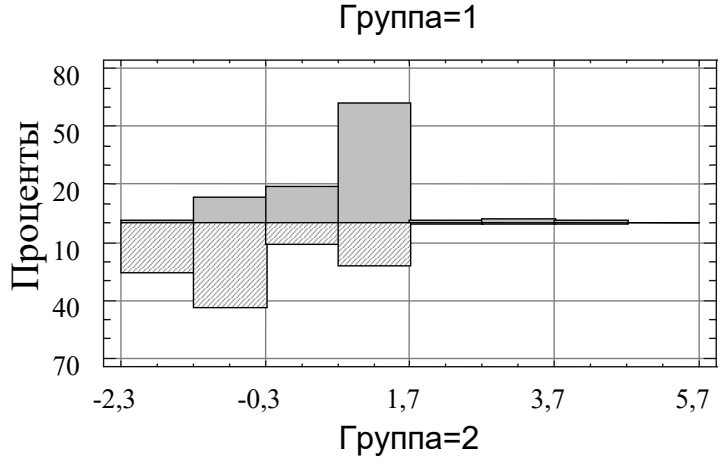


Рис. 3.16. Сравнительные гистограммы распределения значений дискриминантной функции

Как видим, гистограммы распределения значений дискриминантной функции в сравниваемых классах все же достаточно сильно пересекаются. «Хорошие» решения получаются только на краях распределения.

Подобная картина вообще характерна для медико-психологических данных. Линейная модель не способна описывать сложную и неоднородную структуру классов исследуемых объектов. Она отражает только самую общую «грубую» тенденцию.

Результаты обработки данных системой WizWhy

Те же самые данные были подвергнуты обработке системой WizWhy с целью обнаружения логических правил if-then и дальнейшего их использования для осуществления прогнозирования продолжительности воздержания больных алкоголизмом от употребления спиртных напитков.

Всего система WizWhy обнаружила 327 логических правила. Ниже приводится выдержка из ее отчета (50 первых логических правила).

WizWhy REPORT

Total number of records: **191**
Minimum probability of the:
1) if-then rules: **0.800**
2) if-then-not rules: **0.800**
Minimum number of cases in a rule: **15**

Field to Predict: **Group**
Predicted Value (analyzed as Boolean): **2**
Prediction error costs:
The cost of a miss: 1
The cost of a false alarm: 1
Average probability of the predicted value is 0.560

ANALYSIS OF THE RULES EXPLANATORY POWER
Decision point: Predict 2 when conclusive probability is more than **0.830**
Number of misses: 19
Number of false alarms: 18
Total number of errors: 37
Total cost of errors: 37
Success rate when predicting 2: 0.818
Success rate when predicting NOT 2: 0.768
Number of records with no relevant rules: 10
Average cost (per record): 0.204
Expected average cost (per record): 0.440
Improvement Factor: 2.151

IF-THEN RULES:

1)	*If* **x125** *is* **1** and **x135** *is* **2** and **x138** *is* **1** and **x141** *is* **2** and **x144** *is* **2** *Then* **Group** *is not* **2** *Rule's probability:* **0.941** *The rule exists in* **16** *records.* *Significance Level: Error probability* < 0.1
2)	*If* **x137** *is* **1** and **x138** *is* **1** and **x140** *is* **1** and **x144** *is* **1**
26)	*If* **x138** *is* **1** and **x141** *is* **2** and **x144** *is* **2** *Then* **Group** *is not* **2** *Rule's probability:* **0.821** *The rule exists in* **23** *records.* *Significance Level: Error probability* < 0.1
27)	*If* **x125** *is* **2** and **x134** *is* **1** and **x140** *is* **1** and **x141** *is* **1** and **x144** *is* **1** *Then*

Then
Group *is* **2**
Rule's probability: **0.919**
The rule exists in **57** *records.*
Significance Level: Error probability < 0.1

3) *If* **x134** *is* **1**
and **x135** *is* **1**
and **x137** *is* **1**
and **x140** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.926**
The rule exists in **50** *records.*
Significance Level: Error probability < 0.1

4) *If* **x125** *is* **1**
and **x135** *is* **2**
and **x138** *is* **1**
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.880**
The rule exists in **22** *records.*
Significance Level: Error probability < 0.1

5) *If* **x125** *is* **1**
and **x138** *is* **1**
and **x141** *is* **2**
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.852**
The rule exists in **23** *records.*
Significance Level: Error probability < 0.1

6) *If* **x134** *is* **1**
and **x138** *is* **1**
and **x142** *is* **2**
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.882**
The rule exists in **15** *records.*
Significance Level: Error probability < 0.1

7) *If* **x135** *is* **1**
and **x137** *is* **1**
and **x143** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.963**
The rule exists in **26** *records.*
Significance Level: Error probability < 0.1

8) *If* **x125** *is* **1**
and **x137** *is* **1**
and **x142** *is* **2 ... 3** (average = **2**)
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.857**
The rule exists in **18** *records.*
Significance Level: Error probability < 0.1

9) *If* **x137** *is* **1**
Group *is* **2**
Rule's probability: **1.000**
The rule exists in **16** *records.*
Significance Level: Error probability < 0.1

28) *If* **x125** *is* **2**
and **x132** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **1.000**
The rule exists in **17** *records.*
Significance Level: Error probability < 0.1

29) *If* **x134** *is* **1**
and **x135** *is* **2**
and **x138** *is* **1**
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.850**
The rule exists in **17** *records.*
Significance Level: Error probability < 0.1

30) *If* **x135** *is* **2**
and **x138** *is* **1**
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.815**
The rule exists in **22** *records.*
Significance Level: Error probability < 0.1

31) *If* **x137** *is* **1**
and **x143** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.935**
The rule exists in **29** *records.*
Significance Level: Error probability < 0.1

32) *If* **x125** *is* **1**
and **x142** *is* **2 ... 3** (average = **2**)
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.821**
The rule exists in **23** *records.*
Significance Level: Error probability < 0.1

33) *If* **x134** *is* **1**
and **x135** *is* **2**
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.821**
The rule exists in **23** *records.*
Significance Level: Error probability < 0.1

34) *If* **x125** *is* **2**
and **x134** *is* **1**
and **x137** *is* **1**
and **x140** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **1.000**

and **x140** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.897**
The rule exists in **61** *records.*
Significance Level: Error probability < 0.1

10) *If* **x136** *is* **1**
and **x137** *is* **1**
and **x143** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.964**
The rule exists in **27** *records.*
Significance Level: Error probability < 0.1

11) *If* **x135** *is* **2**
and **x138** *is* **1**
and **x141** *is* **2**
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.889**
The rule exists in **16** *records.*
Significance Level: Error probability < 0.1

12) *If* **x134** *is* **1**
and **x142** *is* **2**
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.857**
The rule exists in **18** *records.*
Significance Level: Error probability < 0.1

13) *If* **x125** *is* **1**
and **x134** *is* **1**
and **x135** *is* **2**
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.846**
The rule exists in **22** *records.*
Significance Level: Error probability < 0.1

14) *If* **x137** *is* **1**
and **x139** *is* **2**
and **x140** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **1.000**
The rule exists in **15** *records.*
Significance Level: Error probability < 0.1

15) *If* **x130** *is* **2**
and **x138** *is* **1**
and **x141** *is* **2**
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.850**
The rule exists in **17** *records.*
Significance Level: Error probability < 0.1

16) *If* **x133** *is* **1**
The rule exists in **16** *records.*
Significance Level: Error probability < 0.1

35) *If* **x133** *is* **1**
and **x135** *is* **1**
and **x137** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.891**
The rule exists in **41** *records.*
Significance Level: Error probability < 0.1

36) *If* **x137** *is* **1**
and **x142** *is* **2 ... 3** (average = **2**)
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.826**
The rule exists in **19** *records.*
Significance Level: Error probability < 0.1

37) *If* **x137** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.854**
The rule exists in **70** *records.*
Significance Level: Error probability < 0.1

38) *If* **x134** *is* **1**
and **x137** *is* **1**
and **x138** *is* **1**
and **x139** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.900**
The rule exists in **45** *records.*
Significance Level: Error probability < 0.1

39) *If* **x134** *is* **1**
and **x136** *is* **1**
and **x137** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.875**
The rule exists in **56** *records.*
Significance Level: Error probability < 0.1

40) *If* **x132** *is* **1**
and **x137** *is* **1**
and **x138** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.896**
The rule exists in **43** *records.*
Significance Level: Error probability < 0.1

41) *If* **x125** *is* **1**
and **x132** *is* **2**
and **x136** *is* **1**
and **x137** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**

and **x137** *is* **1**
and **x138** *is* **1**
and **x139** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.900**
The rule exists in **36** *records.*
Significance Level: Error probability < 0.1

17) *If* **x134** *is* **1**
and **x135** *is* **2**
and **x141** *is* **2**
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.850**
The rule exists in **17** *records.*
Significance Level: Error probability < 0.1

18) *If* **x137** *is* **1**
and **x140** *is* **1**
and **x143** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.958**
The rule exists in **23** *records.*
Significance Level: Error probability < 0.1

19) *If* **x125** *is* **2**
and **x134** *is* **1**
and **x140** *is* **1**
and **x142** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **1.000**
The rule exists in **17** *records.*
Significance Level: Error probability < 0.1

20) *If* **x125** *is* **2**
and **x134** *is* **1**
and **x135** *is* **1**
and **x140** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **1.000**
The rule exists in **16** *records.*
Significance Level: Error probability < 0.1

21) *If* **x125** *is* **2**
and **x133** *is* **1**
and **x135** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **1.000**
The rule exists in **16** *records.*
Significance Level: Error probability < 0.1

22) *If* **x125** *is* **2**
and **x133** *is* **1**
and **x142** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **1.000**
The rule exists in **17** *records.*
Significance Level: Error probability < 0.1

42) *If* **x132** *is* **2**
and **x134** *is* **1**
and **x137** *is* **1**
and **x140** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **1.000**
The rule exists in **15** *records.*
Significance Level: Error probability < 0.1

43) *If* **x142** *is* **2 ... 3** (average = **2**)
and **x144** *is* **2**
Then
Group *is not* **2**
Rule's probability: **0.800**
The rule exists in **24** *records.*
Significance Level: Error probability < 0.1

44) *If* **x125** *is* **2**
and **x133** *is* **1**
and **x141** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **1.000**
The rule exists in **16** *records.*
Significance Level: Error probability < 0.1

45) *If* **x134** *is* **1**
and **x135** *is* **1**
and **x137** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.887**
The rule exists in **55** *records.*
Significance Level: Error probability < 0.1

46) *If* **x125** *is* **2**
and **x130** *is* **1**
and **x134** *is* **1**
and **x140** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **1.000**
The rule exists in **17** *records.*
Significance Level: Error probability < 0.1

47) *If* **x135** *is* **1**
and **x143** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.931**
The rule exists in **27** *records.*
Significance Level: Error probability < 0.1

48) *If* **x125** *is* **2**
and **x133** *is* **1**
and **x137** *is* **1**
and **x144** *is* **1**
Then

Rule's probability: **1.000**
The rule exists in **16** *records.*
Significance Level: Error probability < 0.1

23) *If* **x133** *is* **1**
and **x136** *is* **1**
and **x137** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.891**
The rule exists in **41** *records.*
Significance Level: Error probability < 0.1

24) *If* **x125** *is* **2**
and **x130** *is* **1**
and **x133** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **1.000**
The rule exists in **16** *records.*
Significance Level: Error probability < 0.1

25) *If* **x137** *is* **1**
and **x138** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.873**
The rule exists in **62** *records.*
Significance Level: Error probability < 0.1

Group *is* **2**
Rule's probability: **1.000**
The rule exists in **15** *records.*
Significance Level: Error probability < 0.1

49) *If* **x130** *is* **1**
and **x134** *is* **1**
and **x137** *is* **1**
and **x144** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.887**
The rule exists in **47** *records.*
Significance Level: Error probability < 0.1

50) *If* **x135** *is* **1**
and **x137** *is* **1**
and **x143** *is* **1**
Then
Group *is* **2**
Rule's probability: **0.929**
The rule exists in **26** *records.*
Significance Level: Error probability < 0.1

Результаты обработки данных системой See5 (decision trees)

```
Trial 0:-
Rule 0/1: (cover 86)
        x144 > 1
        -> class 1  [0.705]
Rule 0/2: (cover 105)
        x144 <= 1
        -> class 4  [0.776]
Trial 1:-
Rule 1/1: (cover 23.5)
        x137 <= 1
        x142 > 1
        x144 > 1
        -> class 1  [0.726]
Rule 1/2: (cover 21.0)
        x135 > 1
        x137 <= 1
        x138 <= 1
        x144 > 1
        -> class 1  [0.696]
Rule 1/3: (cover 47.1)
        x137 > 1
        -> class 1  [0.593]
Rule 1/4: (cover 76.3)
        x137 <= 1
        x144 <= 1
        -> class 4  [0.758]
Rule 1/5: (cover 70.1)
        x135 <= 1
        x137 <= 1
        x138 <= 1
        x142 <= 1
        -> class 4  [0.716]
Rule 1/6: (cover 14.0)
```

```
Trial 4:-
Rule 4/1: (cover 12.3)
        x125 <= 1
        x133 <= 1
        x137 > 1
        x141 <= 1
        x144 <= 1
        -> class 1  [0.688]
Rule 4/2: (cover 11.4)
        x133 > 1
        x138 <= 1
        x140 <= 1
        x144 > 1
        -> class 1  [0.672]
Rule 4/3: (cover 21.6)
        x125 <= 1
        x133 <= 1
        x141 > 1
        -> class 1  [0.667]
Rule 4/4: (cover 11.1)
        x125 <= 1
        x130 <= 1
        x140 <= 1
        x144 > 1
        -> class 1  [0.647]
Rule 4/5: (cover 26.3)
        x140 > 1
        -> class 1  [0.568]
Rule 4/6: (cover 36.2)
        x133 <= 1
        x137 <= 1
        x140 <= 1
        x141 <= 1
```

```
Trial 6:-
Rule 6/1: (cover 12.4)
        x134 <= 1
        x138 > 1
        x141 > 1
        -> class 1  [0.642]
Rule 6/2: (cover 12.4)
        x134 <= 1
        x138 <= 1
        x141 > 1
        x142 > 1
        -> class 1  [0.640]
Rule 6/3: (cover 31.9)
        x130 > 1
        x134 <= 1
        x141 <= 1
        -> class 1  [0.629]
Rule 6/4: (cover 22.1)
        x134 <= 1
        x138 <= 1
        x141 > 1
        x142 <= 1
        -> class 4  [0.645]
Rule 6/5: (cover 70.7)
        x130 <= 1
        x134 <= 1
        x141 <= 1
        -> class 4  [0.615]
Rule 6/6: (cover 41.5)
        x134 > 1
        -> class 4  [0.547]
Trial 7:-
Rule 7/1: (cover 18.8)
```

x137 <= 1
x138 > 1
x142 <= 1
x144 > 1
-> class 4 [0.624]
Trial 2:-
Rule 2/1: (cover 27.0)
x140 > 1
-> class 1 [0.626]
Rule 2/2: (cover 164.0)
x140 <= 1
-> class 4 [0.576]
Trial 3:-
Rule 3/1: (cover 1.5)
x125 > 1
x132 > 1
x143 > 1
-> class 1 [0.640]
Rule 3/2: (cover 131.0)
x125 <= 1
x143 > 1
-> class 1 [0.566]
Rule 3/3: (cover 25.5)
x125 > 1
x132 <= 1
-> class 4 [0.812]
Rule 3/4: (cover 30.7)
x143 <= 1
-> class 4 [0.726]

x144 <= 1
-> class 4 [0.793]
Rule 4/7: (cover 30.3)
x125 > 1
x140 <= 1
-> class 4 [0.733]
Rule 4/8: (cover 12.4)
x130 > 1
x133 <= 1
x141 <= 1
x144 > 1
-> class 4 [0.634]
Rule 4/9: (cover 54.8)
x125 <= 1
x133 > 1
x140 <= 1
-> class 4 [0.568]
Trial 5:-
Rule 5/1: (cover 15.3)
x133 <= 1
x136 <= 1
x142 > 1
-> class 1 [0.701]
Rule 5/2: (cover 51.1)
x141 > 1
x144 > 1
-> class 1 [0.663]
Rule 5/3: (cover 28.4)
x136 > 1
-> class 1 [0.569]
Rule 5/4: (cover 91.1)
x136 <= 1
x144 <= 1
-> class 4 [0.675]
Rule 5/5: (cover 121.7)
x141 <= 1
-> class 4 [0.603]

x130 <= 1
x132 <= 1
x134 <= 1
x144 > 1
-> class 1 [0.719]
Rule 7/2: (cover 11.7)
x125 <= 1
x132 > 1
x136 > 1
-> class 1 [0.690]
Rule 7/3: (cover 23.7)
x130 > 1
x132 <= 1
x134 <= 1
-> class 1 [0.645]
Rule 7/4: (cover 15.1)
x125 > 1
x132 > 1
-> class 1 [0.577]
Rule 7/5: (cover 41.4)
x130 <= 1
x132 <= 1
x134 <= 1
x144 <= 1
-> class 4 [0.669]
Rule 7/6: (cover 60.4)
x125 <= 1
x132 > 1
x136 <= 1
-> class 4 [0.582]
Rule 7/7: (cover 11.9)
x132 <= 1
x134 > 1
-> class 4 [0.537]

Отчет системы See5

Evaluation on training data (191 cases):

Trial	Decision Tree		Rules		
	Size	Errors	No	Errors	
0	2	48(25.1%)	2	48(25.1%)	
1	6	52(27.2%)	6	52(27.2%)	
2	9	57(21.8%)	2	81(42.4%)	
3	4	70(36.6%)	4	70(36.6%)	
4	10	51(26.7%)	9	51(26.7%)	
5	9	53(27.7%)	5	52(27.2%)	
6	8	60(31.4%)	6	69(36.1%)	
7	7	63(33.0%)	7	63(33.0%)	
boost		40(20.9%)		40(20.9%)	<<

(a)	(b)	<-classified as
57	27	(a): class 1
13	94	(b): class 2

Таблица 3.10. Сводная таблица результатов обработки клинико-психологических данных различными методами

Метод	Ошибка прогноза для 1-й группы	Ошибка прогноза для 2-й группы	Отказ от прогноза	Кол-во правил
See5	32,1%	12,1%	–	39
WizWhy	23,2%	19,2%	5,2%	327

Заключение

Выбор систем, рассмотренных в данной книге и относящихся к классу «логический анализ данных», не случаен. Эти системы являются достаточно популярными, отражают основные тенденции современного анализа данных и обладают одним из важных свойств – интерпретируемостью результата.

Не рассмотренными в книге, но также весьма существенными для интеллектуального анализа данных являются 2 общих приема, которые используются при работе с комитетами алгоритмов. Это «бустинг» (boosting) и «бэггинг» (bagging – сокращение от bootstrap aggregation). Данные приемы предназначены для повышения «обобщающей способности» получаемых моделей – способности выдавать правильные результаты не только для примеров, участвовавших в процессе обучения, но и для любых новых, не участвовавших в процессе обучения данных. Кратко охарактеризуем эти два приема.

Идея бустинга была предложена в конце 1980-х гг. в контексте вопроса об эквивалентности слабого и сильного обучения. Бустинг реализует процедуру последовательного построения композиции алгоритмов машинного обучения, когда каждый следующий алгоритм стремится компенсировать недостатки композиции всех предыдущих алгоритмов. В течение последних 10 лет бустинг остается одним из наиболее популярных методов ИАД. Основные причины – простота, универсальность, гибкость (возможность построения различных модификаций) и, главное, высокая обобщающая способность.

Бустинг деревьев решений считается одним из наиболее эффективных методов решения задач классификации. В ряде экспериментов наблюдалось практически неограниченное уменьшение частоты ошибок на независимой тестовой выборке по мере наращивания композиции. Более того, качество на тестовой выборке часто продолжало улучшаться даже после достижения безошибочного распознавания всей обучающей выборки. Это изменило существовавшие долгое время представления о том, что для повышения обобщающей способности необходимо ограничивать сложность алгоритмов. На примере бустинга стало понятно, что хорошим качеством могут обладать сколь угодно сложные композиции, если их правильно настраивать.

Теоретическое обоснование эффективности бустинга связано с тем, что взвешенное голосование сглаживает ответы алгоритмов, входящих в комитет. Эффективность бустинга объясняется тем, что по мере добавления базовых алгоритмов увеличиваются отступы обучающих объектов. Причем бустинг продолжает раздвигать классы даже после достижения безошибочной классификации обучающей выборки.

Бэггинг – это метод формирования ансамблей классификаторов с использованием случайной выборки с возвратом или бутстрепа. Он был предложен в 1994 г. При формировании бутстреп-выборок из множества данных случайным образом отбирается несколько подмножеств. Так как отбор производится случайно, набор примеров в этих подмножествах будет различным: некоторые из них могут быть отобраны по несколько раз, а другие – ни разу. Затем на основе каждого подмножества (выборки) строится классификатор. Выходы полученных классификаторов комбинируются (агрегируются) путем голосования или простого усреднения. Считается, что результат будет намного точнее любой одиночной модели, построенной на исходном наборе данных.

В целом в области интеллектуального анализа данных (ИАД), к которой относятся в том числе логические методы, за последнее десятилетие произошли существенные изменения. Слово «интеллектуальный» теперь, скорее, нужно воспринимать в контексте автоматического построения классифицирующих и прогнозирующих моделей. Поиск индивидуально сильных методов и алгоритмов для основной массы специалистов ИАД стал не столь привлекательным – их интересы сместились в сторону умений работать с большими коллективами «слабых» методов и алгоритмов. Вместе с тем представленные в книге инструменты логического анализа данных нередко составляют основу для формирования таких больших коллективов.

Литература

1. *Хитрова*, *А. Н.* Клиническое руководство по ультразвуковой диагностике / под ред. В. В. Митькова. – М., 1996. – Т. 1. – С. 200–256.
2. Информатика в статистике : словарь-справочник. – М. : Финансы и статистика, 1994.
3. *Кильдишев*, *Г. С.* Многомерные группировки / Г. С. Кильдишев, Ю. И. Аболенцев. – М. : Статистика, 1978.
4. Прогнозирование длительности ремиссии при восстановительном лечении больных алкогольной зависимостью на этапе становления ремиссии / О. Ф. Ерышев, Т. Г. Рыбакова, Т. Н. Балашова, Л. А. Дубинина / Санкт-Петербургский психоневрологический ин-т им. В. М. Бехтерева. – СПб., 2006. – 20 с.

Вячеслав Анатольевич ДЮК

ЛОГИЧЕСКИЙ АНАЛИЗ ДАННЫХ

Учебное пособие

Зав. редакцией
литературы по информационным технологиям
и системам связи *О. Е. Гайнутдинова*
Ответственный редактор *Т. С. Спирина*
Корректор *Т. А. Кошелева*
Выпускающий *Т. С. Симонова*

ЛР № 065466 от 21.10.97
Гигиенический сертификат 78.10.07.953.П.1028
от 14.04.2016 г., выдан ЦГСЭН в СПб

Издательство «ЛАНЬ»
lan@lanbook.ru; www.lanbook.com
196105, Санкт-Петербург, пр. Ю. Гагарина, д. 1, лит. А.
Тел./факс: (812) 336-25-09, 412-92-72.
Бесплатный звонок по России: 8-800-700-40-71

Подписано в печать 01.10.19.
Бумага офсетная. Гарнитура Школьная. Формат 60×90 $^{1}/_{8}$.
Печать офсетная. Усл. п. л. 10,00. Тираж 100 экз.

Заказ № 654-19.

Отпечатано в полном соответствии
с качеством предоставленного оригинал-макета
в АО «Т8 Издательские Технологии».
109316, г. Москва, Волгоградский пр., д. 42, к. 5.