Б. Г. Миркин

ВВЕДЕНИЕ В АНАЛИЗ ДАННЫХ

УЧЕБНИК И ПРАКТИКУМ
ДЛЯ ВУЗОВ

Рекомендовано Учебно–методическим отделом высшего образования в качестве учебника для студентов высших учебных заведений, обучающихся по инженерно–техническим, естественнонаучным и экономическим направлениям и специальностям

Книга доступна
на образовательной платформе «Юрайт» urait.ru,
а также в мобильном приложении «Юрайт.Библиотека»

Москва ▪ Юрайт ▪ 2020

УДК 51(075.8)
ББК 22.161я73
М63

***Автор*:**

Миркин Борис Григорьевич — доцент, доктор технических наук, профессор кафедры анализа данных и искусственного интеллекта отделения прикладной математики и информатики факультета бизнес-информатики Национального исследовательского университета «Высшая школа экономики», почетный профессор компьютерных наук Лондонского университета.

***Рецензенты*:**

Моттль В. В. — доктор технических наук, профессор Московского физико-технического института, ведущий научный сотрудник вычислительного центра РАН;

Алескеров Ф. Т. — доктор технических наук, руководитель Департамента математики факультета экономики Национального исследовательского университета «Высшая школа экономики».

Миркин, Б. Г.

М63 Введение в анализ данных : учебник и практикум / Б. Г. Миркин. — Москва : Издательство Юрайт, 2020. — 174 с. — (Высшее образование). — Текст : непосредственный.

ISBN 978-5-9916-5009-0

Анализ данных — предмет, порожденный компьютерной революцией, приведшей к накоплению огромного количества конкретных данных о совокупностях объектов, таких как страны или регионы, веб-сайты, работодатели и работники, товары и продавцы. В отличие от классической математической статистики анализ данных не пытается вывести свойства окружающего мира исходя из специально собранных данных, а ориентирован на отыскание каких-либо паттернов, закономерностей, структуры в имеющихся данных.

В данном учебнике, подготовленном на основе большого международного опыта исследований и преподавания, излагаются основные методы анализа данных, относящихся прежде всего к одному или двум изучаемым признакам. Подробно рассмотрены вопросы анализа и интерпретации связей между двумя количественными, двумя качественными, а также качественным и количественным признаками. Из многомерных методов рассмотрены наивный Бэйесовский классификатор и метод *K*-средних для кластерного анализа, включая «интеллектуальную» версию с автоматическим определением числа кластеров и их начального местоположения. Изложение ориентировано на людей, предпочитающих не формулы, а вычисления, и содержит большое количество иллюстративных примеров применения рассматриваемых понятий к анализу реальных данных.

Для студентов бакалавриата и магистратуры инженерно-технических специальностей, также может использоваться для самостоятельного изучения.

УДК 51(075.8)
ББК 22.161я73

ISBN 978-5-9916-5009-0

Оглавление

Для тех, кто боится формул,
но не прочь посчитать.

Предисловие

Настоящее издание — не совсем обычное. Оно написано, чтобы помочь людям, желающим анализировать данные, освоить методы их первичного анализа так, чтобы по возможности обойтись без специальных математических знаний. Основной предмет учебника — методы анализа одномерных и двумерных распределений, тема, которую другие учебники «перепрыгивают», уделяя ей лишь очень небольшое внимание. Во всяком случае, автор не знает других учебников, в которых бы предмет был раскрыт с такой полнотой. «Ничего себе, полнота! — усмехнется или возмутится случайно заглянувший в книгу специалист по математико-статистическим методам. — Не только не полнота, а наоборот, сплошная дырка. Здесь нет практически ни слова о статистике одномерных вероятностных распределений и проверке статистических гипотез о них, а ведь это десятки и сотни страниц пропущенного текста». И это будет тот самый случай, который имел в виду Козьма Прутков, предупреждая, что «специалист подобен флюсу: полнота его односторонняя». Утверждение специалиста, как говорится, верно, но не правильно. Упомянутые разделы действительно пропущены — но не упущены, потому что, по мнению автора, они не входят в базовое содержание анализа данных. Методы проверки статистических гипотез играют роль, и очень важную, в специальном классе ситуаций, когда, например, агроном хочет понять, какой сорт семян, при прочих равных условиях принесет наилучший урожай, или врач пытается определить, дает ли новая методика лечения заметно лучший результат, чем существующая методика. Для ответа на подобные вопросы надо аккуратно поставить эксперимент, получить сопоставимые данные и аккуратно сравнить результаты, принимая во внимание их случайный разброс. Математическая статистика дает методы, позволяющие провести такое сравнение во многих случаях. Но это скорее боковое ответвление, а не магистральная дорога в анализе данных, и поэтому рассказ о таких методах здесь отсутствует.

Анализ данных имеет дело с такими данными, которые оказались в распоряжении исследователя более или менее случайно, не как результат целенаправленного эксперимента, а как результат чьих-то наблюдений или просто статистической сводки. Это могут быть, например, данные о социально-экономическом состоянии регионов России или стран Европы в таком-то году. Или это может быть совокупность сообщений, отправленных членами какой-либо социальной сети в течение определенного промежутка времени. В подобных ситуациях типичные вопросы таковы. Какой смысл можно извлечь из этих данных? Есть ли какая-нибудь структура в данных о рассматриваемом множестве объектов? Могут ли эти признаки помочь в прогнозировании тех? Подобная ситуация скорее харак-

терна для путешественника, чем ученого. Ученый сидит за столом, получает воспроизводимые данные об окружающем мире и старается включить их в грандиозную научную модель этого мира. Путешественник же должен понять, как ему лучше себя вести здесь и сейчас.

Анализ данных в настоящее время не включает проблематику изучения механизмов получения данных. Все, что нас интересует — это наличие в данных каких-либо общих паттернов. Если удастся такой паттерн обнаружить; если удастся потом убедиться, что он — не артефакт применения метода, а действительно существует в данных; если удастся понять, исходя из наших знаний о явлении, к которому относятся данные, возможную причину возникновения паттерна; и если, наконец, на этой основе удастся предложить новый метод использования явления — вот тогда можно говорить о том, что метод анализа данных работает! Впрочем, вопросы использования результатов анализа данных обычно остаются вне поля зрения специалистов по анализу данных: считается достаточным, чтобы нашелся паттерн, проливающий некий новый свет на явление, к которому относятся данные, чтобы объявить об успешности анализа.

Согласно точке зрения, подробно описанной в более полном учебнике автора [17], имеется два основных способа анализа данных: суммаризация и коррелирование. Суммаризация, как и английский оригинал, означает подытоживание, агрегирование, представление в сжатом виде. Коррелирование — это отыскание связей между различными признаками, описывающими объекты, без каких-либо попыток приписать этим связям причинный характер. Попробуем осветить это понятие чуть подробнее на следующем примере. Наблюдения показали следующую корреляцию: новорожденные дети более активны и восприимчивы у тех матерей, которые ели много рыбы во время беременности. Значит ли это, что именно рыбоедение дает эффект? Да, говорят одни: в рыбе много фосфора, а фосфор — строительный материал мозга. Нет, говорят другие: рыба тут ни при чем. Просто эти женщины — богатые, ведь рыба дорого стоит, особенно в пересчете на калории. А у богатых уход за ребенком лучше, вот он и более активен. Кто же прав? Имеющаяся информация не позволяет прийти к однозначному выводу. Все, что анализ данных может дать — это паттерн, а для выяснения причины паттерна нужны дополнительные данные, в данном случае надо изучать приток фосфора в мозг новорожденного в процессе беременности (очень сложно!) и (или) уровни благосостояния рожениц (значительно проще).

Говоря о полноте изложения материала в данном учебнике, автор имеет в виду полноту раскрытия проблематики суммаризации и коррелирования на уровне одно- и двумерных распределений. Случай одного признака рассмотрен в главе 2. Глава 3 трактует случай, когда в анализ включаются два признака. При этом отдельно проанализированы задачи коррелирования для ситуаций, в которых (а) оба признака количественные, или (б) оба признака категоризованные, или (в) один — категоризованный, а другой — количественный. Во всех трех случаях идея коррелирования проводится, исходя из основной цели — улучшения предсказания значений одного признака по значениям другого. Почему-то эта довольно популярная идея не нашла своего отражения в существующих учебниках. Поэтому изложение автором даже таких довольно традиционных тем, как линейная регрессия (ситуация (а)) и табличная регрессия (ситуация (в)) получается довольно свежим и прагматически ориентированным[1]. Что касается ситуации (б) категори-

[1] Здесь хочется сослаться на мнение рецензента учебника [17]: «Выделю только одно из многих успешных мест учебника: я сомневаюсь, что читатель когда-либо снова встретит такое детальное и превосходное описание корреляционных понятий» (Computing Reviews of ACM, June 2011).

зованных признаков, то здесь использована и вовсе нетрадиционная идея. За счет применения так называемых индексов Кетле удается представить коэффициент хи-квадрат, введенный К. Пирсоном для проверки гипотезы о статистической независимости категоризованных признаков, как меру их корреляции, и на этой основе визуализировать структуру связи между значениями признаков. В других учебниках читателя специально предупреждают: величина хи-квадрат не характеризует уровень связи и не может использоваться для ее оценки; ан нет, согласно представленному подходу — может! Глава 4 дает возможность «одним глазком» взглянуть на методы анализа многомерных данных. Приводятся два очень популярных метода, один для суммаризации (метод *K*-средних кластерного анализа), другой для коррелирования (наивный Бэйесовский[1] классификатор)[2]. Выбор методов определяется не только популярностью, но и возможностью избежать сложных формул и выводов при их объяснении. Все изложение иллюстрируется на примерах конкретных данных, в основном сквозных, которые приводятся в вводной главе 1 вместе с ассоциированными проблемами анализа данных.

Имеющиеся русскоязычные учебники анализа данных либо делают сильный перекос в сторону задач оценки вероятностных распределений и проверки статистических гипотез (как, например, А. С. Айвазян, И. С. Енюков, Л. Д. Мешалкин, 1983; Ю. Н. Тюрин, А. А. Макаров, 2003; М. Б. Лагутин, 2009), либо с места в карьер переходят к реализации методов многомерного анализа данных на каком-либо прикладном пакете программ (В. Н. Калинина, В. И. Соловьев, 2010; А. П. Кулаичев, 2006), либо же слишком специальны (Н. Г. Загоруйко, 1999; Б. Г. Миркин, 1985). Данный учебник не относится ни к одной из этих категорий.

Другие особенности учебника состоят в следующем.

Во-первых, основное изложение распределено по трем относительно независимым линиям: «представление», «формулировка» и «вычисление». В данном учебнике буквы «П», «Ф» и «В» в рубрикации подпараграфов означают, что данные подпараграфы относятся к линиям «представление», «формулировка» и «вычисление» соответственно. «Представление» не содержит математических формул и на конкретных данных показывает задачу, метод ее решения, а также комментарии к результатам, когда это необходимо. Напротив, в «формулировке» сосредоточены все математические детали постановки задачи и метода. В «вычислении» объясняется, как провести вычисление с использованием вычислительной среды МатЛаб. Желательно, чтобы читатель имел доступ к этой среде. Учебная версия МатЛаба, особенно в оригинальной неруссифицированной версии, стоит совсем недорого. Использование МатЛаба в контексте рассматриваемых понятий и методов не требует программистских навыков. Азы работы на МатЛабе объясняются в приложении к данной книге. Таким образом, каждый читатель может выбрать такой способ изложения, который ему наиболее подходит.

Во-вторых, применяется четырехуровневая структура самостоятельных заданий, предназначенных для активизации работы читателя:

---

[1] Бэйес Томас (Bayes Thomas, 1702—1761) — английский «непрофессиональный» математик, чья работа стала известна после его смерти. Написание «Бэйес» ближе к английскому произношению фамилии, «Бэйиз», чем укоренившаяся в России форма «Байес». Автор настаивает на переходе к этому более корректному произношению, имея в виду, что читатели данного текста — люди международных контактов, в которых произношение «Байес» неуместно, так как воспроизводит произношение английского слова «bias», означающего «предвзятость».

[2] О популярности этих методов говорит, например, тот факт, что их включили в первую очередь библиотеки программ Махаут для проведения облачных вычислений на так называемых больших данных (URL: http://mahout.apache.org/).

1) «рабочие примеры», которые просто иллюстрируют работу того или иного метода; в начале каждого из них на конкретном примере показывается, как провести расчет и интерпретацию решения, когда это уместно, а затем дается задание для «самостоятельной работы» — повторить то же на других данных. Иногда задание дается в виде «вопроса» с готовым ответом — эти задания тоже следует выполнять самостоятельно; ответ приводится только для сверки;

2) «задания» — более сложные задачи, в которых имеется определенный неформальный элемент, например необходимость создания нового множества данных (по определенному правилу) или же неформальный способ интерпретации;

3) «проекты» — еще более сложные проблемы, в какой-то мере имитирующие научные проекты и требующие проведения небольшого научного исследования;

4) «вопросы» — математические или вычислительные проблемы для тех читателей, которые все же не боятся математики; они, как правило, снабжены ответами — либо в явном виде, либо содержатся в самой формулировке вопроса. Но это не значит, что их не надо решать самостоятельно. Надо. Ответы приводятся лишь для проверки. Всего в учебнике содержатся 27 рабочих примеров, 8 заданий, 6 проектов и 54 вопроса. Большинство решений сопровождается комментариями более общего характера, подчас далеко выходящими за рамки данного случая. Комментарии носят уникальный характер и более не повторяются. Поэтому советую внимательно знакомиться со всеми примерами конкретного анализа.

В-третьих, вводятся самые современные методы вычислительной науки, такие как бутстрэп для оценки доверия к результатам и эволюционные алгоритмы для оптимизации нелинейных критериев.

Кроме того, с учетом современной тенденции уделять и делу время, и потехе час, в учебнике представлено несколько картинок и с полсотни шуток из современного фольклора, с юмористической стороны иллюстрирующих обсуждаемые понятия. В конце каждой главы имеется небольшой раздел «Кстати говоря», в котором размещено некоторое количество анекдотов, связанных с тематикой главы.

Учебник основан на курсах автора для студентов бакалавриата и магистратуры в Биркбек-колледже Лондонского университета (2004—2010), для слушателей Школы анализа данных при Яндексе (2008—2010) и студентов бакалавриата и магистратуры отделений прикладной математики и программной инженерии Национального исследовательского университета Высшей школы экономики (2008—2013). В некоторой мере его содержание следует моему более полному англоязычному учебнику [17].

Хотя основной текст написан так, чтобы его мог освоить человек, не изучавший высшую математику, некоторое знакомство с ней, конечно, полезно. Речь идет прежде всего об азах математического анализа (понятия функции, ее производной, точек минимума), теории вероятностей (частота и условная вероятность, функция плотности) и теории множеств (понятия включения множеств и принадлежности элемента данному множеству).

Теперь несколько слов по существу содержания учебника. В нем четыре главы. В *первой главе* рассматриваются основные типы и примеры задач анализа данных на относительно небольших примерах данных. Во *второй главе* рассматриваются основные понятия одномерного анализа, т.е. анализа индивидуальных признаков. В *третьей главе* рассматриваются основные понятия двумерного анализа, т.е. ана-

лиза пар признаков. В *четвертой главе* рассматриваются два популярных метода многомерного анализа данных: наивный Бэйесовский классификатор и метод *K*-средних кластерного анализа. В заключении на примерах анализа реальных данных показывается, что анализ данных — это далеко не все.

В результате изучения материала учебника студент будет:

знать

- основные понятия анализа данных и смежных дисциплин;
- основные понятия и методы визуализации и анализа индивидуальных признаков;
- основные понятия и методы анализа и визуализации пар признаков, включая ситуации, когда оба признака количественные, оба признака номинальные или один признак количественный, а второй — номинальный;
- наивный метод Бэйеса для классификации многомерных объектов, его обоснование и способы оценки точности прогноза;
- метод *K*-средних для кластерного анализа данных, его критерий и интеллектуальную версию, основанную на автоматизации выбора числа кластеров и их начальных центров;

уметь

- производить предварительное преобразование данных путем бинарного перекодирования категорий номинальных признаков, центрирования и нормализации признаков;
- производить анализ и визуализацию распределений индивидуальных признаков, включая использование вычислительного метода бутстрэп для построения доверительного интервала среднего значения;
- использовать методы анализа и визуализации распределений пар признаков, включая линейную регрессию для пар количественных признаков, табличную регрессию количественного признака по номинальному признаку и анализ структуры таблицы сопряженности для пар номинальных признаков;
- использовать наивный метод Бэйеса для классификации документов и оценивать точность получаемого прогноза;
- использовать метод *K*-средних, а также его интеллектуальную версию, для кластерного анализа данных;

владеть навыками

- компьютерного представления и предварительной обработки реальных данных и метаданных размерами до нескольких десятков признаков и нескольких сот объектов;
- использования МатЛаба или другой вычислительной среды для анализа и визуализации распределений индивидуальных признаков на реальных данных;
- использования МатЛаба или другой вычислительной среды для анализа и визуализации связей между парами признаков на реальных данных;
- использования МатЛаба или другой вычислительной среды для кластерного анализа многомерных реальных данных методом *K*-средних и его интеллектуальной версией.

Таким образом, в результате изучения представленного материала студент должен ***быть компетентным*** в понимании основных понятий и методов, связанных с анализом данных, прежде всего в разрезе отдельных признаков или пар признаков, а также умении их применять для анализа реальных данных с использованием вычислений на современных вычислительных устройствах. Это относится и к представленным многомерным методам: наивному Бэйесовскому классифи-

катору и методу *K*-средних кластерного анализа с его интеллектуализированной версией. Более подробно компетенции описаны в аннотациях к отдельным главам.

Учебник ориентирован на использование в курсах анализа данных, математической статистики и машинного обучения в бакалавриате инженерных специальностей — прикладной математики, информатики, программной инженерии, а также в курсах количественных методов для неинженерных специальностей — экономики, социологии, менеджмента, географии, филологии и пр. Для неинженерных специальностей учебник может быть рекомендован к использованию и в магистерских программах. Кроме того, учебник может быть использован для самостоятельного изучения теми, кто по характеру своей деятельности хотел бы использовать данные и методы их анализа.

В заключение хочу выразить благодарность моим коллегам по работе в НИУ ВШЭ, сделавшими возможной и приятной работу над данным учебником, за внимание и поддержку. Речь идет прежде всего о кафедре анализа данных и искусственного интеллекта, лаборатории интеллектуальных систем и структурного анализа (заведующий кафедрой С. О. Кузнецов) и международной лаборатории анализа и выбора решений (руководитель Ф. Т. Алескеров). Е. Л. Черняк взяла на себя один из циклов вычитки рукописи. Рекомендации и многочисленные поправки редактора издательства «Юрайт» Е. В. Ткаченко были учтены при доработке рукописи. Конечно, все остающиеся ошибки — всецело на моей ответственности.

Глава 1
ЧТО ТАКОЕ АНАЛИЗ ДАННЫХ?

В этой вводной главе рассказывается, что такое анализ данных и чем он отличается от математической статистики. Приводятся примеры задач анализа данных. Вводятся два основных типа задач анализа данных: суммаризация и коррелирование. Дается представление о сходных дисциплинах, возникших в связи с развитием и распространением вычислительной техники.

После изучения данной главы студент будет:

знать

- понятие таблицы данных «объект-признак»;
- основные типы задач анализа данных;
- основные типы шкал признаков;
- отличия в подходах математической статистики и анализа данных;
- понятие о связанных подходах;

уметь

- формировать таблицы данных и формулировать типовые задачи их анализа;
- переводить данные в количественный формат путем сведения номинальных признаков к совокупностям бинарных;

владеть навыками

- отыскания таблиц данных, связанных с той или иной содержательной проблемой, в Интернете;
- предварительного анализа таблиц данных на предмет выявления возможных задач, относящихся к суммаризации и коррелированию признаков;
- перевода данных в количественный формат путем сведения номинальных признаков к совокупностям бинарных признаков.

1.1. Иллюстративные проблемы анализа данных

Прежде чем переходить к обсуждению методов, рассмотрим примеры данных и задач, связанных с их анализом. Чтобы проиллюстрировать мысль, что анализ данных может применяться не только к большим множествам данных, но и к малым, первый пример взят намеренно очень небольшим. Это также полезно с точки зрения того, что данные можно увидеть такими, как они есть, просто глядя на таблицу.

Пример 1.1. Компании

В табл. 1.1 приводятся данные о восьми компаниях в разрезе следующих пяти признаков:

1) доход — годовой доход, млрд руб.;
2) доля Р — доля рынка, %;
3) ОП — число основных потребителей;
4) Инт — есть ли ведение бизнеса по Интернету (+) или нет (–);

5) сектор экономики — превалирующая отрасль народного хозяйства: а) химия; б) металлургия; в) торговля.

Обратим внимание на терминологию: графы табл. 1.1, как и других таблиц данных, соответствуют *признакам*, строки — *объектам*, а в самой таблице находятся *значения* признаков.

Таблица 1.1

Данные о компаниях

Компания	Доход, млрд руб.	Доля Р, %	ОП	Инт (да/нет)	Сектор экономики
Авер	19,0	21,85	2	–	Химия
Ант	29,4	18,00	3	–	Химия
Астон	23,9	19,00	3	–	Металлургия
Бмарт	18,4	13,95	2	+	Химия
Брек	25,7	11,15	3	+	Металлургия
Бумо	12,1	8,45	2	+	Металлургия
Виж	23,9	15,10	4	+	Торговля
Вурд	27,2	29,00	5	+	Торговля

Примечание. Совокупность восьми компаний охарактеризована пятью разнотипными признаками. Имена компаний отражают основные группы производимой продукции (либо А, либо Б, либо В).

Хотя сами данные носят чисто иллюстративный характер, следующие вычислительные проблемы довольно типичны в анализе данных:

- однородность. Есть ли в данных аномально большие или аномально малые компании? Можно ли утверждать, что в данных механически объединены разные типы компаний?
- визуализация. Отображение компании на экране так, чтобы расстояния между их позициями отражали сходство между ними: чем более похожи компании, тем ближе позиции;
- если кластеризовать компании в группы по степени сходства, будут ли кластеры соответствовать основным группам продукции, А, Б и В? Если да, то какие признаки будут наиболее весомы?
- можно ли вывести какие-либо точные правила для атрибуции вида продукции исходя из данных признаков? (Эти правила затем можно применить и для компаний, не вошедших в данную таблицу (прогноз));
- можно ли обнаружить какую-либо взаимосвязь между структурными признаками компаний (три признака справа в таблице) и признаками рыночной активности (доход и доля рынка)?

Имеется серьезная структурная разница между признаками в табл. 1.1: не все они количественные! На самом деле в табл. 1.1 представлены все три обычно рассматриваемые типа шкал:

1) *количественные*, т.е. такие, для которых осмысленна операция усреднения их значений. В табл. 1.1 таковыми являются признаки «Доход», «Доля Р» и «ОП»;

2) *бинарные*, т.е. такие, которые допускают только значения, соответствующие ответам «да» и «нет»; таков признак «Инт» в табл. 1.1;

3) *номинальные*, т.е. такие, которые допускают несколько непересекающихся категорий, такие как «Сектор» в табл. 1.1.

Обычно бинарные и номинальные признаки рассматривают как неколичественные, т.е. «категоризованные». На самом деле их можно перевести в «количе-

ственный» формат, приписывая 1 ответу «да» и 0 — ответу «нет». Для бинарных признаков эта перекодировка тривиальна. Номинальный признак должен быть сначала «расщеплен» в систему бинарных признаков, соответствующих отдельным категориям. Например, признак «Сектор» в табл. 1.1 будет заменен на три бинарных признака, соответствующих его категориям (табл. 1.2):

- Сектор химии;
- Сектор металлургии;
- Сектор торговли.

Такие бинарные признаки часто называют фиктивными, или «дамми» (от англ. *dummy* — чурбан).

Таблица 1.2

Данные о компаниях из табл. 1.1, преобразованные в количественный формат

Номер	Доход	Доля Р	ОП	Инт	Хим	Мета	Торг
1	19,0	21,85	2	0	1	0	0
2	29,4	18,00	3	0	1	0	0
3	23,9	19,00	3	0	0	1	0
4	18,4	13,95	2	1	1	0	0
5	25,7	11,15	3	1	0	1	0
6	12,1	8,45	2	1	0	1	0
7	23,9	15,10	4	1	0	0	1
8	27,2	29,00	5	1	0	0	1
Среднее	22,4	17,06	3,0	0,625	0,375	0,375	0,25

Взятие среднего значения у бинарного 1/0 признака вполне осмысленно: среднее значение есть не что иное как доля данной категории в множестве объектов! Этот факт, а также ему подобные, приводят к тому, что в данном тексте 1/0 признак считается количественным.

Пример 1.2. Ирисы

Ирисы — пожалуй, самая популярная таблица данных (табл. 1.3). Она характеризует коллекцию, собранную ботаником Э. Андерсоном и использованную Р. Фишером (1936) в его основополагающей статье о дискриминантном анализе. В ней представлены 150 цветков ириса, относящихся к трем видам: I. *Iris setosa* (диплоид), II. *Iris versicolor* (тетраплоид) и III. *Iris virginica* (гексаплоид), по 50 экземпляров из каждого. Признаки таблицы относятся к измерениям длины и ширины чашелистика (w1, w2), а также лепестков (w3, w4) (рис. 1.1).

Виды определяются генотипом, а признаки — фенотипом. Возникает вопрос, можно ли описать виды в терминах измеренных признаков. Хорошо известно, что вид I довольно легко отделяется от остальных, тогда как виды II и III перемешаны (возможно, из-за ошибок Андерсона в определении видов). Вместе с тем ботаники понимают, что виды можно неплохо отличить по форме лепестка, которая в определенной степени отражается площадью прямоугольника, т.е. произведения $w3 \cdot w4$.

Из проблем анализа данных, относящихся к этой таблице, укажем следующие:

- существует ли признак или пара признаков, распределение которых информативно с точки зрения описания трех видов ириса?
- визуализация данных на экране так, чтобы похожие объекты отображались близкими друг к другу точками;
- анализ связи между разными измерениями, включая возможность предсказания, скажем, размеров лепестка по размерам чашелистика;
- возможность сведения всех признаков в единый непосредственно неизмеримый признак «размер» цветка.

Рис. 1.1. **Чашелистик (*sepal*) и лепесток (*petal*) в цветке ириса**

Таблица 1.3

Ирисы (в строках — информация о 150 экземпляров ириса, измеренных по четырем признакам в разрезе трех видов)

№	I Iris setosa	II Iris versicolor	III Iris virginica
	w1 w2 w3 w4	w1 w2 w3 w4	w1 w2 w3 w4
1	5.1 3.5 1.4 0.3	6.4 3.2 4.5 1.5	6.3 3.3 6.0 2.5
2	4.4 3.2 1.3 0.2	5.5 2.4 3.8 1.1	6.7 3.3 5.7 2.1
3	4.4 3.0 1.3 0.2	5.7 2.9 4.2 1.3	7.2 3.6 6.1 2.5
4	5.0 3.5 1.6 0.6	5.7 3.0 4.2 1.2	7.7 3.8 6.7 2.2
5	5.1 3.8 1.6 0.2	5.6 2.9 3.6 1.3	7.2 3.0 5.8 1.6
6	4.9 3.1 1.5 0.2	7.0 3.2 4.7 1.4	7.4 2.8 6.1 1.9
7	5.0 3.2 1.2 0.2	6.8 2.8 4.8 1.4	7.6 3.0 6.6 2.1
8	4.6 3.2 1.4 0.2	6.1 2.8 4.7 1.2	7.7 2.8 6.7 2.0
9	5.0 3.3 1.4 0.2	4.9 2.4 3.3 1.0	6.2 3.4 5.4 2.3
10	4.8 3.4 1.9 0.2	5.8 2.7 3.9 1.2	7.7 3.0 6.1 2.3
11	4.8 3.0 1.4 0.1	5.8 2.6 4.0 1.2	6.8 3.0 5.5 2.1
12	5.0 3.5 1.3 0.3	5.5 2.4 3.7 1.0	6.4 2.7 5.3 1.9
13	5.1 3.3 1.7 0.5	6.7 3.0 5.0 1.7	5.7 2.5 5.0 2.0
14	5.0 3.4 1.5 0.2	5.7 2.8 4.1 1.3	6.9 3.1 5.1 2.3
15	5.1 3.8 1.9 0.4	6.7 3.1 4.4 1.4	5.9 3.0 5.1 1.8
16	4.9 3.0 1.4 0.2	5.5 2.3 4.0 1.3	6.3 3.4 5.6 2.4
17	5.3 3.7 1.5 0.2	5.1 2.5 3.0 1.1	5.8 2.7 5.1 1.9
18	4.3 3.0 1.1 0.1	6.6 2.9 4.6 1.3	6.3 2.7 4.9 1.8
19	5.5 3.5 1.3 0.2	5.0 2.3 3.3 1.0	6.0 3.0 4.8 1.8
20	4.8 3.4 1.6 0.2	6.9 3.1 4.9 1.5	7.2 3.2 6.0 1.8
21	5.2 3.4 1.4 0.2	5.0 2.0 3.5 1.0	6.2 2.8 4.8 1.8
22	4.8 3.1 1.6 0.2	5.6 3.0 4.5 1.5	6.9 3.1 5.4 2.1
23	4.9 3.6 1.4 0.1	5.6 3.0 4.1 1.3	6.7 3.1 5.6 2.4
24	4.6 3.1 1.5 0.2	5.8 2.7 4.1 1.0	6.4 3.1 5.5 1.8
25	5.7 4.4 1.5 0.4	6.3 2.3 4.4 1.3	5.8 2.7 5.1 1.9
26	5.7 3.8 1.7 0.3	6.1 3.0 4.6 1.4	6.1 3.0 4.9 1.8
27	4.8 3.0 1.4 0.3	5.9 3.0 4.2 1.5	6.0 2.2 5.0 1.5
28	5.2 4.1 1.5 0.1	6.0 2.7 5.1 1.6	6.4 3.2 5.3 2.3

Окончание табл. 1.3

№	I Iris setosa	II Iris versicolor	III Iris virginica
	w1 w2 w3 w4	w1 w2 w3 w4	w1 w2 w3 w4
29	4.7 3.2 1.6 0.2	5.6 2.5 3.9 1.1	5.8 2.8 5.1 2.4
30	4.5 2.3 1.3 0.3	6.7 3.1 4.7 1.5	6.9 3.2 5.7 2.3
31	5.4 3.4 1.7 0.2	6.2 2.2 4.5 1.5	6.7 3.0 5.2 2.3
32	5.0 3.0 1.6 0.2	5.9 3.2 4.8 1.8	7.7 2.6 6.9 2.3
33	4.6 3.4 1.4 0.3	6.3 2.5 4.9 1.5	6.3 2.8 5.1 1.5
34	5.4 3.9 1.3 0.4	6.0 2.9 4.5 1.5	6.5 3.0 5.2 2.0
35	5.0 3.6 1.4 0.2	5.6 2.7 4.2 1.3	7.9 3.8 6.4 2.0
36	5.4 3.9 1.7 0.4	6.2 2.9 4.3 1.3	6.1 2.6 5.6 1.4
37	4.6 3.6 1.0 0.2	6.0 3.4 4.5 1.6	6.4 2.8 5.6 2.1
38	5.1 3.8 1.5 0.3	6.5 2.8 4.6 1.5	6.3 2.5 5.0 1.9
39	5.8 4.0 1.2 0.2	5.7 2.8 4.5 1.3	4.9 2.5 4.5 1.7
40	5.4 3.7 1.5 0.2	6.1 2.9 4.7 1.4	6.8 3.2 5.9 2.3
41	5.0 3.4 1.6 0.4	5.5 2.5 4.0 1.3	7.1 3.0 5.9 2.1
42	5.4 3.4 1.5 0.4	5.5 2.6 4.4 1.2	6.7 3.3 5.7 2.5
43	5.1 3.7 1.5 0.4	5.4 3.0 4.5 1.5	6.3 2.9 5.6 1.8
44	4.4 2.9 1.4 0.2	6.3 3.3 4.7 1.6	6.5 3.0 5.5 1.8
45	5.5 4.2 1.4 0.2	5.2 2.7 3.9 1.4	6.5 3.0 5.8 2.2
46	5.1 3.4 1.5 0.2	6.4 2.9 4.3 1.3	7.3 2.9 6.3 1.8
47	4.7 3.2 1.3 0.2	6.6 3.0 4.4 1.4	6.7 2.5 5.8 1.8
48	4.9 3.1 1.5 0.1	5.7 2.6 3.5 1.0	5.6 2.8 4.9 2.0
49	5.2 3.5 1.5 0.2	6.1 2.8 4.0 1.3	6.4 2.8 5.6 2.2
50	5.1 3.5 1.4 0.2	6.0 2.2 4.0 1.0	6.5 3.2 5.1 2.0

Пример 1.3. Компьютерные атаки

С учетом растущего значения компьютерных сетей возрастает опасность их атак, приводящих к нарушению функционирования сетей. Простейший вид атаки — отказ от обслуживания [denial of service DoS]. Такая атака причиняется командами, которые приводят к тому, что какой-либо ресурс — процессор, память, входное устройство — оказывается перегружен и не может обслуживать нормальные запросы. Две такие атаки в табл. 1.4 помечены как «apache2» и «smurf».

Таблица 1.4

Данные о компьютерных атаках

Pr	BySD	SH	SS	SE	RE	A	Pr	ByS	SH	SS	SE	RE	A
Tcp	62344	16	16	0	0.94	Ap	Tcp	287	14	14	0	0	no
Tcp	60884	17	17	0.06	0.88	Ap	Tcp	308	1	1	0	0	no
Tcp	59424	18	18	0.06	0.89	Ap	Tcp	284	5	5	0	0	no
Tcp	59424	19	19	0.05	0.89	Ap	Udp	105	2	2	0	0	no
Tcp	59424	20	20	0.05	0.9	Ap	Udp	105	2	2	0	0	no
Tcp	75484	21	21	0.05	0.9	Ap	Udp	105	2	2	0	0	no
Tcp	76944	22	22	0.05	0.91	Ap	Udp	105	2	2	0	0	no
Tcp	59424	23	23	0.04	0.91	Ap	Udp	105	2	2	0	0	no
Tcp	57964	24	24	0.04	0.92	Ap	Udp	44	3	8	0	0	no
Tcp	59424	25	25	0.04	0.92	Ap	Udp	44	6	11	0	0	no
Tcp	0	40	40	1	0	Ap	Udp	42	5	8	0	0	no
Tcp	0	41	41	1	0	Ap	Udp	105	2	2	0	0	no

Продолжение табл. 1.4

Pr	BySD	SH	SS	SE	RE	A	Pr	ByS	SH	SS	SE	RE	A
Tcp	0	42	42	1	0	Ap	Udp	105	2	2	0	0	no
Tcp	0	43	43	1	0	Ap	Udp	42	2	3	0	0	no
Tcp	0	44	44	1	0	Ap	Udp	105	1	1	0	0	no
Tcp	0	45	45	1	0	Ap	Udp	105	1	1	0	0	no
Tcp	0	46	46	1	0	Ap	Udp	44	2	4	0	0	no
Tcp	0	47	47	1	0	Ap	Udp	105	1	1	0	0	no
Tcp	0	48	48	1	0	Ap	Udp	105	1	1	0	0	no
Tcp	0	49	49	1	0	Ap	Udp	44	3	14	0	0	no
Tcp	0	40	40	0.62	0.35	Ap	Udp	105	1	1	0	0	no
Tcp	0	41	41	0.63	0.34	Ap	Udp	105	1	1	0	0	no
Tcp	0	42	42	0.64	0.33	Ap	Udp	45	3	6	0	0	no
Tcp	258	5	5	0	0	No	Udp	45	3	6	0	0	no
Tcp	316	13	14	0	0	No	Udp	105	1	1	0	0	no
Tcp	287	7	7	0	0	No	Udp	34	5	9	0	0	no
Tcp	380	3	3	0	0	No	Udp	105	1	1	0	0	no
Tcp	298	2	2	0	0	No	Udp	105	1	1	0	0	no
Tcp	285	10	10	0	0	No	Udp	105	1	1	0	0	no
Tcp	284	20	20	0	0	No	Tcp	0	482	1	0.05	.95	sa
Tcp	314	8	8	0	0	No	Tcp	0	482	1	0.05	.95	sa
Tcp	303	18	18	0	0	No	Tcp	0	482	1	0.05	.95	sa
Tcp	325	28	28	0	0	No	Tcp	0	482	1	0.05	.95	sa
Tcp	232	1	1	0	0	No	Tcp	0	482	1	0.05	.95	sa
Tcp	295	4	4	0	0	No	Tcp	0	482	1	0.05	.95	sa
Tcp	293	13	14	0	0	No	Tcp	0	482	1	0.06	.94	sa
Tcp	305	1	8	0	0	No	Tcp	0	482	1	0.06	.94	sa
Tcp	348	4	4	0	0	No	Tcp	0	482	1	0.06	.94	sa
Tcp	309	6	6	0	0	No	Tcp	0	483	1	0.06	.94	sa
Tcp	293	8	8	0	0	No	Tcp	0	510	1	0.04	.96	sa
Tcp	277	1	8	0	0	no	Icmp	1032	509	509	0	0	sm
Tcp	296	13	14	0	0	no	Icmp	1032	510	510	0	0	sm
Tcp	286	3	6	0	0	no	Icmp	1032	510	510	0	0	sm
Tcp	311	5	5	0	0	no	Icmp	1032	511	511	0	0	sm
Tcp	305	9	15	0	0	no	Icmp	1032	511	511	0	0	sm
Tcp	295	11	25	0	0	no	Icmp	1032	494	494	0	0	sm
Tcp	511	1	4	0	0	no	Icmp	1032	509	509	0	0	sm
Tcp	239	12	14	0	0	no	Icmp	1032	509	509	0	0	sm

Окончание табл. 1.4

Pr	BySD	SH	SS	SE	RE	A	Pr	ByS	SH	SS	SE	RE	A
Tcp	5	1	1	0	0	no	Icmp	1032	510	510	0	0	sm
Tcp	288	4	4	0	0	no	Icmp	1032	511	511	0	0	sm

Атака 'apache2' нацелена на популярный веб-сервер открытого пользоваиия, Apache HTTP Server, заставляя его посылать клиенту огромное количество «пустых» запросов и переполняя буферные каналы. Атака 'smurf' посылает фальшивые ответные послания, «пинги», по различным адресам, заставляя соответствующие компьютеры отвечать «пингами» же по обратным адресам. Если главный адрес «подделан» нарушителем, возникает шквал «пингов», направляемых какому-либо серверу из группы компьютеров сети, переполняя его входные каналы. Такая атака может начинаться с «разведки», отыскивающей проблемы в сети. Популярный софтвер для проведения разведки называется SAINT [Security Administrator's Integrated Network Tool].

Таблица 1.4 — случайная выборка из данных, синтезированных в одной из лабораторий Массачусетского технологического института (Бостон, США, www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/ intex.html). Признаки характеризуют пакет и его источник.

1. Pr — тип протокола, в данном случае один из трех: tcp, icmp, udp (этот признак — номинальный).
2. BySD — число байтов в пакете.
3. SH — количество соединений источника за последние две секунды.
4. SS — число соединений с тем же сервером за последние две секунды.
5. SE — процент ошибочных соединений.
6. RE — процент соединений с отказом обслуживания.
7. A — тип атаки (ap — apache2, sa — saint, sm — smurf, и отсутствие атаки — no).

Из сотни объектов в табл. 1.4 первые 23 — атаки сервера apache2, пакеты 24—69 — нормальные, следующие одиннадцать, 80—90, соответствуют разведке SAINT, и последние десять, 91—100, — атаки smurf.

Примеры проблем анализа данных, которые можно исследовать с использованием табл. 1.4:

- найти признаки пакетов, которые можно использовать для того, чтобы определить, нормально ли функционирует сеть или же она атакована;
- выяснить, есть ли связь между используемым протоколом и типом атаки;
- визуализировать данные так, чтобы близость точек соответствовала похожести значений признаков соответствующих объектов.

Пример 1.4. Малые города английского побережья

В табл. 1.5 приведены иллюстративные данные о 45 малых городах юго-запада Англии. Для целей социального планирования имеет смысл выделить сравнительно небольшую их группу так, чтобы каждый попавший в нее город представлял весь «кластер» похожих на него городов. В таблице города упорядочены по числу жителей. Например, 21 самых малых городов имеют меньше 4000 жителей каждый. Число 4000 взято в качестве разделителя не случайно. Во-первых, оно круглое. Во-вторых, оно помечает значительный разрыв в более чем 1300 между Кингскерсвеллом (3672 жителя) и следующим по численности городом Луе (5022 жителя). Следующий большой разрыв — после Лискярда (7044), отделяющего 9 городов среднего размера от двух групп больших городов, насчитывающих соответственно 6 и 9 городов соответственно. Разделитель между двумя последними группами — между Тавистоком (10 222) и Бодмином (12 553). Так мы получим три или четыре группы городков, которые можно использовать укрупненно при социальном мониторинге. А достаточно ли однородны эти группы с точки зрения других имеющихся признаков? В работе [18] показано, что более однородны в этом множестве не четыре, а семь кластеров: большие города