УДК 57.01

Андреев В. Л. **Классификационные построения в экологии и систематике.** М.: Наука, 1980.

Книга посвящена выявлению анализу структуры сложных систем в экологии, биогеографии и зоологической систематике. Рассматриваются вопросы формирования цели, выбора признаков, отношений на множествах описаний биологических и географических объектов, мер, критериев и т. п. Обсуждаются способы обработки качественной (нечисловой) информации, использования в классификационных построениях методов многомерного статистического анализа. На конкретных примерах излагаются вычислительные процедуры для реализации методов вручную и на ЭВМ.

Книга рассчитана на экологов, географов, медиков, кибернетиков и лиц, интересующихся приложениями математики в биологии.

Ответственный редактор
кандидат биологических наук
Б. М. СЕМКИН

А $\frac{21005-103}{055(02)-80}$ 772-80, кн. 2; 2001060000

ПРЕДИСЛОВИЕ

В биологии, как и любой другой экспериментальной науке, информацию о внешнем мире приносят наблюдения и измерения. Но после того как фактические данные собраны, первое, что с ними необходимо сделать,— это попытаться разложить их «по полочкам», классифицировать. Таким образом, анализ структуры данных — необходимый этап любого экспериментального исследования.

Долгое время такой анализ осуществлялся на базе словесных рассуждений, умозрительно, или просто на основе неосознанного опыта — интуиции. Поэтому любые успехи, достигнутые в развитии биологической классификации, добывались дорогой ценой. Количественные методы проникали в биологию медленно и не всегда удачно. Тем не менее к настоящему времени они завоевали столь прочные позиции, что привели к образованию специальной дисциплины — «нумерической таксономии», создатели которой противопоставляют ее классической, или ортодоксальной, таксономии, использующей качественные методы.

Такое противопоставление хорошо согласуется с довольно распространенным делением наук на «зрелые», использующие математику, и «незрелые», оперирующие неформальными словесными рассуждениями. Но это — заблуждение, и основано оно, по-видимому, на недостаточном понимании того факта, что развитие науки определяют содержательные идеи, а не форма, в которую они облечены.

Совокупность представлений, на которых базируются идеи, часто называют моделью. Одна из задач настоящей книги заключается в попытке описать модель, объясняющую смысл классификационных построений в экологии и систематике.

Кроме этого, в работе приводится подробное описание методов, которые могут оказаться полезными в практических и теоретических исследованиях биологам и географам. Для каждого метода подобраны конкретные примеры, помогающие выяснить его существо и назначение, а также подробно излагаются вычислительные процедуры, способствующие его практической реализации.

Короче говоря, назначение данной работы — формирование средств эффективного проведения классификационных построений в определенных биологических целях.

Книга написана для биологов и географов. Это и определило компоновку и способ изложения материалов. Важные для математика результаты не рассматриваются, если они не представляют ценности для прикладных задач, и, наоборот, математически малозначащие построения исследуются с особой тщательностью, если это важно для целей экологии и систематики.

Математические статьи и монографии тяжелы или вовсе недоступны для понимания читателям, имеющим «гуманитарное» образование. Поэтому мы сочли необходимым в начале книги подобрать минимум сведений, обеспечивающий понимание основного материала, и изложить их по возможности просто. Все, что требуется для понимания ее первой части,— это общелогическая культура мышления и некоторое терпение. Вторая часть написана в предположении, что читатель знаком с основами математической статистики в пределах обычных курсов биометрии.

Содержание отдельных глав достаточно очевидно из оглавления, поэтому излишне дополнительно характеризовать распределение материалов в книге, тем более что некоторые подробности освещены в параграфе 1.2.

Представления автора о классификационных построениях навеяны чтением работ советских инженеров и математиков: Н. Загоруйко, Г. Лбова, А. Горелика и В. Скрипкина, В. Дружинина и Д. Конторова, Ю. Воронина, Б. Семкина и др. Однако это не означает их полного совпадения с моделями указанных авторов.

ОСНОВНЫЕ ОБОЗНАЧЕНИЯ

$M = \{a, b, \ldots\}$ — множество M, состоящее из элементов a, b, ...;

$\{x \mid \ldots y \ldots\}$ — множество всех таких элементов x, для которых выполнено условие ... y ...;

$x \in B$ — x принадлежит множеству B;

$x \notin B$ — x не принадлежит множеству B;

$\emptyset$ — пустое множество;

$A \cap B = \{x \mid x \in A$ и $x \in B\}$;

$A \cup B = \{x \mid x \in A$ или $x \in B\}$;

$A \setminus B = \{x \mid x \in A$ и $x \notin B\}$;

$a \to b$ — из a следует b;

$B \subseteq C$ — из того, что x принадлежит множеству B следует, что он принадлежит и множеству $C (x \in B \to x \in C)$;

$m(A)$ — количество элементов множества A;

$\langle M, A \rangle$ — отношение A на множестве M, $A \subseteq M \cdot M$;

xAy — x находится в отношении A с y;

$A \cdot B = A \times B = \{(a, b) \mid a \in A, \quad b \in B\}$ — декартово произведение множеств A и B (множество всех возможных пар (a, b), таких, что $a \in A$, $b \in B$);

$B \oplus C = (B \setminus C) \cup (C \setminus B)$;

$\overline{B} = A/B$ для всех подмножеств B рассматриваемого множества A;

$f : M \to L$ — отображение множества M в множество L.

Часть I

ДЕТЕРМИНИСТСКИЕ МЕТОДЫ ПОСТРОЕНИЯ И ИССЛЕДОВАНИЯ СИСТЕМ-КЛАССИФИКАЦИЙ

Глава 1

ВВЕДЕНИЕ В ПРОБЛЕМУ

1.1. Существо и значение проблемы

Мир, окружающий человека, огромен по своему многообразию. «Нельзя войти дважды в одну и ту же реку»,— так характеризовал древнегреческий философ известный факт, что в природе постоянно происходят изменения и что нельзя найти одинаковые предметы или явления. Между тем человек может успешно справляться с этим многообразием и принимать решения даже в непрерывно изменяющейся обстановке. При этом число решений ограниченно, хотя число воспринимаемых состояний среды бесконечно.

Это можно видеть на примере ихтиолога, разбирающего траловые уловы. Несмотря на бесчисленное множество вариантов строения конкретных особей, он, как правило, безошибочно узнает интересующий его вид. Точно так же в обыденной жизни мы, как правило, безошибочно отличаем стул от стола, несмотря на то что мебель может быть и нестандартной.

Такое «узнавание» подготовлено всем предшествующим опытом человека: в процессе деятельности много раз встречаясь с конкретными объектами, он формирует в сознании некоторые абстрактные «образы» объектов и в дальнейшем любой вновь встречающийся предмет или явление сопоставляет с одним из таких «образов», хранящихся в памяти.

На этих примерах мы знакомимся с одним из частных видов классификации: «узнаванием», идентификацией, или, как его называют в математической и инженерной литературе, «распознаванием образов». Как можно видеть, существо распознавания заключается в том, чтобы любой встречающийся объект с наименьшей вероятностью ошибки отнести к одному из заранее сформированных классов. Трудность решения такой задачи обусловлена «размытостью» границ некоторых классов, когда отдельные объекты бывают «одинаково похожи» на представителей разных классов.

Более общий вид классификации включает не только соотнесение объектов к одному из классов, но и одновременное формирование самих «образов», число которых может быть заранее неизвестно. Такого рода сортировка производится на основе осознанного или интуитивного стремления собрать «в кучу» в некотором смысле «схожие» объекты, да еще так, чтобы объекты из разных «куч» (классов) были по возможности «несхожими». В процессе «развала на кучи» процедура идентификации выполняется многократно.

Само стремление собрать похожие объекты в «кучу» вполне понятно: именно классификация позволяет человеку ориентироваться в бесконечном разнообразии окружающего мира. Другими словами, классификация есть средство экономии памяти [57].

В этом смысле формирование классов одного уровня в определенный момент становится невыгодным: их оказывается чрезвычайно много. Тогда следующим этапом может быть формирование «куч» из похожих классов. Именно с этого момента начинается, как мы полагаем, иерархическая классификация. Результатом ее является система классов, находящаяся в отношении иерархии и образующая определенную структуру.

Идя дальше по этому пути, можно рассматривать такие системы как отдельные объекты и, если их оказывается много, строить и изучать структуру системы, элементами которой являются также системы. При этом под изучением структуры подразумеваются определения наиболее важных элементов и связей, показателей ее сложности (информационной емкости), классов эквивалентных (но более просто реализуемых) структур и т. п.

Перечисленные виды классификаций по аналогии с «узнаванием» можно назвать одним из блоков, составляющих мышление. Так что проблемы классификации — это часть проблемы мышления [11].

Наряду с таким общетеоретическим значением проблемы классификации играют фундаментальную роль во многих науках. Некоторые ученые полагают даже, что состояние классификаций изучаемых объектов отражает состояние всей соответствующей науки в целом [16].

Прикладное значение проблемы следует из возможности не только проводить эффективный анализ экспериментальных данных, но и решать задачи, например, по определению границ экологических систем (а следовательно, и разрабатывать стратегию оптимального управления ими), строить прогнозирующий аппарат и осуществлять прогноз относительно совокупности любых элементов экосистемы. Примеры таких приложений приводятся в соответствующих разделах книги.

Несмотря на очевидную важность обсуждаемой проблемы, в отечественной биологии все еще слабо используются успехи, достигнутые на пути ее решения. Необходимость популяризации этих успехов, по мнению автора, заключается в том, чтобы содействовать устранению наметившейся дивергенции представлений

биологов и математиков. Среди первых, например, до сих пор бытует стремление открыть «естественную» систему организмов [35, 48], а создатели «нумерической таксономии» утверждают, что любая классификация будет тем более эффективной, чем больше используется признаков для описания объектов [63]. Нередко теория классификации биологических объектов противопоставляется построениям систем-классификаций объектов небиологического происхождения.

Такие модели противоречат подходам, развиваемым математиками и инженерами в связи с проблемами создания искусственного интеллекта. Многое из того, что ими сделано для теории классификации, попросту неизвестно отечественным биологам. Среди последних, в особенности зоологов, приложения математики в зоогеографии и зоологической систематике представляют собой разрозненные попытки применения отдельных методов. Однако и в этих случаях нередко возникают недоразумения, как плата за веру в непогрешимость результатов, полученных по «сложным» формулам.

1.2. Основные этапы построения и исследования систем-классификаций

Любому изучению природных объектов предшествует формирование цели: бесцельное созерцание нельзя назвать изучением. В смысле этой цели дальнейшая работа состоит в том, чтобы выяснить особенности объектов, что отличает и роднит их.

Итак, первым этапом классификационных построений является глубокое проникновение в суть рассматриваемых явлений и выбор соответствующего принципа классификации. Примеры формирования цели и принципов можно найти в параграфах 3.2, 3.6, а также в описаниях конкретных исследований на протяжении всей книги.

Второй этап — установление списка признаков, подлежащих учету на отдельных объектах. При этом необходимо помнить, что в список должны быть включены такие признаки, которые в полной мере характеризуют изучаемые объекты в смысле заданной цели, а измерение или учет их не составляет принципиальных трудностей. Следует исключать такие из них, которые по сравнению с внутриклассовой изменчивостью мало изменяются при переходе от класса к классу и, следовательно, имеют слабые разделительные свойства. При неизвестном числе классов отбрасываются признаки, имеющие малую дисперсию.

К этому вплотную примыкает вопрос о выборе способа измерения или учета признаков. В конечном итоге качество классификации в значительной мере определяется объемом и качеством исходной информации, а эти последние зависят от длины списка, разделительной ценности признаков и точности измерений на объектах.

Эти вопросы обсуждаются в параграфах 2.5, 3.1, 4.1.

Третий этап — отбор объектов и производство измерений. Самое главное требование, предъявляемое к отбору, заключается в том, чтобы выборка обладала репрезентативностью, т. е. по возможности полно отражала все исследуемые свойства генеральной совокупности. Никакой математический аппарат не позволяет избежать ошибочных выводов в том случае, когда отбор произведен предвзято и выборки оказываются «нетипичными».

Измерения признаков, как указывалось, должны производиться с такой точностью, какую только допускают технические возможности. Грубые измерения влекут за собой потерю информации.

Способы составления выборок подробно описываются в любых пособиях по биометрии и в данной работе почти не затрагиваются.

Четвертый этап — выбор отношений на множестве описаний объектов, а также мер, порождающих отношения, решающих правил и критериев эффективности, производство вычислений.

Понятия «множество», «отношение», «мера» рассматриваются в параграфах 2.1, 2.2. Для статистических моделей распознавания мерам расстояния между выборками, объектами и выборками посвящена глава 7, решающие правила и критерии эффективности рассматриваются в главе 8. Использование бинарных отношений наиболее подробно описано в параграфах 3.3, 3.4, 4.4, а n-арных отношений — в параграфах, связанных с математической логикой: 2.4, 4.2, 4.3.

В указанных разделах приводятся и подробные вычислительные схемы. Примерам использования различных мер в построениях при неизвестном числе классов (кластер-анализ) посвящены параграф 3.5, а также глава 9.

Пятый этап — построение и анализ структурной схемы системы, в которой связи между элементами соответствуют выполнению отношений между ними.

Обычный способ представления структурных схем — с помощью графов (§ 2.3) и дендрограмм (§ 3.5). Понятие системы и структурной схемы вводится в параграфе 2.5. Анализ структурных схем проводится в главе 5. Практические примеры построения структурных схем можно найти в параграфах 3.3, 3.4, 3.5, 9.2, 9.3.

Шестой этап — интерпретации, т. е. перенос полученных утверждений с системы-модели на систему-прототип. Выполнение этого этапа связано иногда с неустранимыми трудностями из-за того, что сложность модели всегда меньше, чем сложность оригинала. Успехом исследования в этом плане является построение простой, но легко интерпретируемой модели.

Вопросы интерпретации рассматриваются на практических примерах книги.

Как можно видеть из приведенного перечня, основные материалы книги посвящены описаниям четвертого и пятого этапов Такая направленность ее обусловлена несколькими причинами

Из них главным является то, что выполнение первых трех этапов почти целиком зависит от исследователя, так как они составляют творческую часть процедуры классификации, которая в значительной степени определяется субъективными причинами и не может быть формализована.

С другой стороны, классификация — это процесс переработки информации, осуществляемый по определенным правилам логики. Это — трудная, хотя и нетворческая часть процедуры. Ее необходимо формализовать для того, чтобы добиться унификации и определенности результатов, их сравнимости, увеличения теоретической и практической значимости. Перспективность этого направления становится особенно очевидной, если учесть, что формализация — обязательный шаг для переработки информации с помощью машин. А это значит, что при наличии типовых программ лучшие достижения в этой области могут быть доступны пользователям независимо от уровня их математической подготовки. Быстрота и точность, которые обеспечивает ЭВМ, освобождают исследователя от рутинных работ в тот период, когда поток биологической информации растет лавинообразно. Кроме того, ЭВМ дают возможность получать результаты, которые не могут быть получены никаким другим путем.

Именно этим аспектам и посвящена основная часть книги, а компоновка ее материалов обусловлена стремлением обеспечить понимание особенностей обработки информации в зависимости от свойств, которыми она обладает.

Вместе с тем, как это будет видно далее, роль субъекта как творческого начала в процедурах классификации выделяется красной нитью на всех этапах. Так что ответственность за конечные результаты построений несет тот, кто создает модели-гипотезы, выбирает способы формирования исходных данных, методику сбора и обработки информации, интерпретирует результаты и сопоставляет их с поставленной целью. В этих проблемах выбора нет однозначных решений, а это означает, что и в данных исследованиях главную роль играют все те же опыт, удача и чувство здравого смысла.

1.3. Некоторые библиографические сведения

Число публикаций, относящихся к распознаванию образов, исследованию сложных систем и структуры данных, выросло настолько, что даже по специальным вопросам издаются сотни работ. Значительно меньшее число статей касается приложений этих вопросов в систематике и совсем мало в экологии. Правда, в последнем издании [63] авторы упоминают, что к 1973 г. в мировой биологической литературе публикуется ежегодно около 200 статей, в той или иной степени затрагивающих вопросы «нумерической таксономии». В этой же книге освещаются многие из затронутых в данной работе проблем и приводится сравнительно подробная библиография.

Из изданий на русском языке, посвященных приложениям математических методов в таксономии, можно указать монографии [7,48], а по применению системного анализа в биологии — книгу Т. Г. Гильманова [18].

Литературу по частным вопросам читатель найдет в конце этой книги. В настоящем же параграфе приводятся ссылки на монографии, в которых имеется сравнительно подробное описание литературных источников.

Теория множеств рассматривается П. Александровым [1] и Э. Мендельсоном [38], а популярное ее изложение дано в брошюрах [26, 44, 49, 56].

Теории графов посвящены книги [8, 52], популярно она изложена в брошюрах [51, 56]. В последней из них приводится еще и популярное и достаточно строгое изложение теории отношений.

Для изучения математической логики можно рекомендовать книги [20, 38, 57], популярное ее изложение приводится в брошюрах [17, 30, 37, 49].

Теория систем рассматривается в книгах [22, 40, 55], анализ структуры данных — в работе [9]. Многомерный статистический анализ является предметом обсуждения книг [13, 33, 50, 45].

Проблемы распознавания образов подробно рассмотрены в монографиях [6, 27, 53], для первичного знакомства полезен справочник [14], популярное изложение ее можно найти у М. Бонгарда [11].

Методам кластер-анализа посвящена монография [24], эти вопросы рассматриваются также в книгах [23, 25]. Идентификации и составлению определителей посвящена главная часть книги [41].

Математический аппарат, используемый нами, заимствован в основном из указанных выше источников.

Глава 2

ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Нами уже неоднократно употреблялись такие термины, как «система», «множество», «отношения», «сходство» и др. Причем до сих пор использовались только обиходные значения этих слов. Переход от качественных расплывчатых понятий к точно формулируемым на языке математики называется экспликацией. Одной из задач ближайших разделов данной работы является экспликация первичных понятий систематики и биогеографии в терминах теории множеств и универсальной алгебры. Поэтому предварительно рассмотрим некоторые понятия, используемые в этих науках.

2.1. Множества

При классификационных построениях все операции производятся над множествами классифицируемых объектов, поэтому прежде всего необходимо уточнить, что будет пониматься под множеством объектов, и обсудить способы описания этих множеств.

Под множеством M, согласно Кантору (см. [40]), понимается любое объединение в одно целое определенных, вполне различимых объектов из нашего восприятия или мысли. Объекты x, которые называются элементами M,— это любые предметы или явления реального мира (или мысленные конструкции). Понятия «множество» и «элемент множества» нельзя определить более строго, так как нет более элементарных понятий, посредством которых они могли бы быть определены.

Примеры: множество видов, обитающих на какой-либо территории, множество индивидуумов, составляющих выборку, множество географических пунктов, в которых учитывались виды, множество измеряемых признаков и т. п.

Для того чтобы указать, что x есть элемент множества M, используется запись: $x \in M$, при этом обычно названия множеств обозначают прописными буквами, а их элементы — строчными. Запись $x \notin M$ читается: «элемент x не принадлежит множеству M».

Множества, состоящие из конечного числа элементов, называются конечными. Если число элементов бесконечно, то множества называются бесконечными, а если количество элементов равно нулю, то множество называется пустым. Пустое множество обозначается ϕ (перечеркнутый кружок).

Для того чтобы показать, что ряд объектов образует множество, обозначения этих объектов заключаются в фигурные скобки. Например,

$$M = \{x, y, z\} \tag{2.1}$$

означает, что множество M состоит из объектов x, y, z. Полезно отметить, что $\{a, b, c\} = \{a, б, a, a, c, б, c\}$, т. е. одинаковые элементы множества в общем случае считаются за один.

Задать некоторое множество можно двумя способами: во-первых, указать полный перечень объектов, принадлежащих этому множеству, например, как в (2.1); во-вторых, указать формальное правило для определения того, принадлежит или не принадлежит какой-либо объект x множеству M. Запись

$$J = \{j \mid j \text{ — целое число, } 1 \leqslant j \leqslant n\} \tag{2.2}$$

читается: «множество J есть множество всех тех и только тех элементов j, которые являются целыми числами, изменяющимися в пределах от 1 до n». Здесь описание множества в фигурных скобках разбивается на две части вертикальной чертой, слева от которой обозначаются все элементы множества (все те и только те j), а справа дается характеристика, или условие, которое всегда истинно для элементов из левой части (которые являются...).

Количественной характеристикой (мерой) конечного множества A является число его элементов, которое обозначается $m(A)$. Теория конечных множеств изучает правила, по которым, зная количество элементов некоторых множеств, можно вычислить количество элементов некоторых других множеств, составленных из первых с помощью определенных операций.

Одной из таких операций является операция «сложения»:

$$C = A \cup B \tag{2.3}$$

(читается: C есть объединение множеств A и B). Результат ее — сумма, или объединение, множеств A и B — это множество C, состоящее из тех и только тех элементов, которые входят либо в A, либо в B (т. е. входят хотя бы в одно из множеств A или B). Например, пусть даны множества $A = \{$щука, окунь, плотва$\}$, $B = \{$лещ, окунь, язь, плотва$\}$, тогда $C = A \cup B = \{$щука, окунь, плотва, лещ, язь$\}$. Иными словами, при объединении множеств общие элементы считаются по одному разу. Нетрудно заметить, что «сложение» множеств отличается от сложения в обычной алгебре. В частности, $A \cup A = A$ или в более общем случае

$$A \cup A \cup \ldots \cup A = A. \tag{2.4}$$

Однако коммутативный и ассоциативный законы сохраняются, как и в алгебре:

$$A \cup B = B \cup A, \quad A \cup (B \cup C) = (A \cup B) \cup C. \tag{2.5}$$

Пересечением множеств A и B называется множество C, в которое входят те и только те элементы, которые входят одновременно в A и B:

$$C = A \cap B$$

(читается: C есть пересечение множеств A и B). Например,

$$\{1, 2, 3\} \cap \{2, 3, 4\} = \{2, 3\}.$$

Пересечение множеств A и B из предыдущего примера:

$$A \cap B = \{\text{окунь, плотва}\}.$$

Пересечение множеств отвечает коммутативному и ассоциативному законам

$$A \cap B = B \cap A, \quad A \cap (B \cap C) = (A \cap B) \cap C, \tag{2.6}$$

а также дистрибутивному закону

$$A \cap (B \cup C) = (A \cap B) \cup (B \cap C).$$

Отметим также, что

$$A \cap A = A, \quad A \cap A \cap \ldots \cap A = A. \tag{2.7}$$

Разностью множеств A и B называется множество

$$C = A \setminus B,$$

состоящее из тех и только тех элементов A, которые не входят в B. Например,

{щука, окунь, плотва} $\setminus$ {лещ, окунь, язь, плотва} = {щука}. Разность множеств отвечает коммутативному и ассоциативному законам. Множество элементов A, которое не входит в множество B, будем обозначать $\bar{B}$.

Если A и B — множества, то говорят, что A содержится в B, и пишут

$$A \subset B \quad (\text{или } B \supset A) \tag{2.8}$$

в том и только в том случае, если каждый элемент A является элементом B. В этом случае A называется подмножеством, а B — надмножеством. Множества A и B равны, если одновременно $A \subset B$, и $B \subset A$.

Пустое множество играет роль нуля в операциях над множествами

$$A \cup \phi = A, \quad A \cap \phi = \phi, \quad A \setminus \phi = A. \tag{2.9}$$

Семейством множеств называется множество, элементы которого сами являются множествами. Для обозначения семейства множеств используются обычно рукописные буквы, например,

$$\mathcal{P} = \{A, B, C, D\}.$$

Если все множества, рассматриваемые в определенной ситуации, являются подмножествами некоторого множества У, то последнее называется универсумом для данной ситуации.

Приведем примеры. Пусть в некоторых географических пунктах a и $б$ учитывается наличие или отсутствие 20 видов животных. Тогда множество, состоящее из всех 20 видов, есть универсум; подмножества, входящие в универсум: A — виды, обнаруженные в a, B — виды, обнаруженные в $б$. Пересечение $A \cap B$ — виды, общие для обоих пунктов; объединение $A \cup B$ — все виды, найденные в a и $б$.

Рассмотрим теперь количество элементов некоторых множеств, получаемых в результате операций над исходными множествами. Нетрудно убедиться, что

$$m(A \cap A) = m(A), \tag{2.10}$$

т. е. количество элементов пересечения множества с самим собой равно количеству элементов этого множества;

$$m(A \cup B) = m(A) + m(B) - m(A \cap B), \tag{2.11}$$

т. е. число элементов объединенного множества равно сумме элементов каждого из объединенных множеств за вычетом числа одинаковых элементов;

$$m(\bar{A} \cap B) = m(B) - m(A \cap B), \tag{2.12}$$

т. е. число элементов, которые принадлежат только множеству B и не принадлежат множеству A, равно числу элементов множества B за вычетом числа общих для A и B элементов.

Аналогично

$$m(A \cap \bar{B}) = m(A) - m(A \cap B). \tag{2.13}$$

Приведем еще ряд полезных для дальнейшего изложения равенств:

$$m(\bar{A} \cup \bar{B}) = m(A) + m(B) - 2m(A \cap B). \tag{2.14}$$

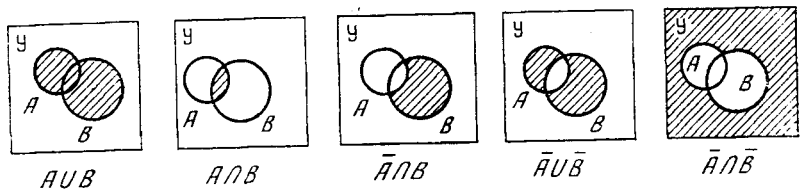


Рис.2.1. Диаграмма Венна

Здесь $m(\bar{A} \cup \bar{B})$ — число несовпадающих элементов, т. е.

$$m(\bar{A} \cup \bar{B}) = m(\bar{A} \cap B) + m(A \cap \bar{B}). \tag{2.15}$$

Наконец, число элементов, не принадлежащих ни A, ни B, по отношению к универсуму составит

$$m(\bar{A} \cap \bar{B}) = m(У) - m(A \cup B). \tag{2.16}$$

Введенные понятия легко иллюстрировать графически с помощью диаграммы Венна (рис. 2.1). Если прямоугольником изобразить множество элементов универсума, а кружками — множества A и B, то содержимое множеств, полученных в результате перечисленных операций (заштриховано), становится ясным из рисунка.

Некоторые другие понятия теории множеств будут введены ниже.

2.2. Отношения

Термин «отношение» легко пояснить перечислением некоторых примеров: «быть родственником» (отношение родства), «быть похожим» (отношение сходства), «быть неразличимым» (отношение эквивалентности) и т. п., т. е. отношение — это качественное понятие, отражающее тот факт, что между некоторыми элементами непустого множества существуют (заданы) какие-то связи. Задать отношение на множестве — значит указать между какими элементами оно выполняется.

Если отношение связывает сразу n элементов с указанием, какой из них является первым, какой — вторым и т. д., то оно называется n-арным. При $n = 2$ имеем бинарные отношения. Г. Биргоф (см. [40]) дает более общее определение бинарному отн

шению, понимая под этим любое правило, которое указывает для каждой упорядоченной пары элементов x, $y \in M$, что либо отношение имеет место между x и y, либо оно не имеет места.

В дальнейшем фразу «отношение A задано на множестве M» будем записывать как $\langle A, M\rangle$, а фразу «элемент x находится в отношении A с элементом y» — как xAy. Запись $x\bar{A}y$ означает, что элемент x не находится в отношении A с элементом y. Естественно, что в общем случае $xAy \neq yAx$.

Более точно понятие «отношение» можно определить в терминах теории множеств. Для этого введем некоторые дополнительные понятия.

Декартовым произведением (или просто произведением) множеств X и Y называется множество всех пар x, y, из которых $x \in X$ $y \in Y$; это множество обозначается $X \cdot Y$. По индукции можно определить произведение множеств X_1, X_2, X_3, положив

$$X_1 \cdot X_2 \cdot X_3 = (X_1 \cdot X_2) \cdot X_3 \quad \text{и т. д.}$$

В частности, если $X_1 = X_2 = \ldots X_n = X$, то декартово произведение обозначается X^n и называется n-й степенью множества X.

Произвольное подмножество

$$A \subseteq X_1 \cdot X_2 \ldots X_n \tag{2.17}$$

называется n-арным отношением между множествами X_1, X_2, ... ; X_n. Если $X_1 = X_2 = \ldots X$, то A — n-арное отношение на множестве X. В частности, если

$$A \subseteq X \cdot X = X^2, \tag{2.18}$$

то A называется бинарным отношением на множестве X.

Рассмотрим пример. Положим, что три вида рыб: щука, окунь, плотва — обитают в одном водоеме и находятся в отношении хищничества друг к другу. Сократим их названия до начальных букв и образуем из последних множество $M = \{$щ, о, п$\}$. Множество всех возможных пар: {(щ, щ), (щ, о), (щ, п), (о, щ), (о, о), (о, п), (п, щ), (п, о), (п, п)} — декартово произведение $M \cdot M = M^2$. Рассмотрим подмножество A множества M, состоящее из шести пар:

$$A = \{(\text{щ}, \text{щ}), (\text{щ}, \text{о}), (\text{щ}, \text{п}), (\text{о}, \text{щ}), (\text{о}, \text{о}), (\text{о}, \text{п})\}.$$

Множество A выражает отношение «быть хищником», и никакие другие элементы $M \cdot M$ не принадлежат ему. Здесь связи щAщ и оAо отражают факт каннибализма в реальной системе, а связи щAо и оAщ — факт поедания щукой окуня, а щуки (молоди) — окунем (взрослым).

Итак, отношение — это пара $\langle A, M\rangle$, где M — множество, на котором отношение определено, а A — подмножество пар $M \cdot M$, для которых это отношение выполнено. Множество M называется областью задания отношения A.

Более наглядно отношение хищничества в последнем примере можно представить в виде таблицы (матрицы) размерностью 3 × 3

число строк и столбцов), в которой принадлежность подмножеству отмечается «1», а непринадлежность — «0»:

	щ	о	п
щ	1	1	1
о	1	1	1
п	0	0	0

Заметим, что порядок в парах существен, например, (щ, п) $\in A$, (п, щ) $\notin A$. При матричном изображении отношений эта особенность приводит к несимметричности ее относительно главной диагонали. Отношения подобного типа будем называть несимметричными.

В отличие от этого отношение A называется симметричным, если для любых x, $y \in M$ из xAy следует yAx. Примером симметричного отношения может быть отношение сходства: здесь из факта xAy (x похож на y) следует yAx (y похож на x).

Другая особенность матрицы заключается в том, что на главной диагонали не везде стоят 1: отношение xAx не выполняется для каждого $x \in M$. Отношения подобного типа называются нерефлексивными. В отличие от этого отношение A называется рефлексивным, если xAx для каждого $x \in M$. Если же $x\bar{A}x$ для любого $x \in M$, то отношение называется антирефлексивным. Например, отношение сходства — рефлексивно, отношение «быть братом» — антирефлексивно.

Отметим еще одно свойство отношений. Рассмотрим тройки элементов x, y, $z \in M$. Если для любой тройки из xAy и yAz следует xAz, то отношение называется транзитивным, т. е. если для соседних элементов тройки x, y, z отношение A выполняется, то оно выполняется и для ее крайних элементов. Отношение «быть хищником» в общем случае нетранзитивно: существуют такие тройки x, y, $z \in M$, что xAy (x поедает y) и yAz (y поедает z), но $x\bar{A}z$ (x не поедает z). Соблюдение этого свойства в примере с элементами щ, о, п, чисто случайное. Примером транзитивного отношения является отношение «быть одинаковым», «быть больше» и т. п.

Таким образом, отношение «быть хищником» нерефлексивно, несимметрично и нетранзитивно.

Примером тернарных (тройственных) отношений может являться отношение «быть родителями», представленное упорядоченной тройкой элементов

{(мать, отец), их ребенок}.

В практических задачах часто приходится сопоставлять различные отношения, заданные на одном и том же или на различных множествах. В связи с этим полезно рассмотреть отношения на таких парах x, y, где $x \in M$, а $y \in L$. Отношение f этого вида называется функцией, или отображением, если для каждого $x \in M$ существует единственный элемент $y \in L$, для которого выполнено

xfy. Функция f символически записывается как $f\colon M \to L$ и чита
ется: отображение f множества M в множество L. Элемент y назы
вается в этом случае образом элемента $x \in M$, а элемент x являетс
прообразом для элемента $y = f(x)$.

Из определения $f\colon M \to L$ следует, что каждый элемент $x \in M$
имеет ровно один образ, но не всякий элемент $y \in L$ обязан имет
прообраз. Если же такой прообраз существует, то он не обяза
быть единственным. Рассмотрим пример. Пусть L — множеств
всех реально существующих видов, а M — множество организмов
собранных в некотором географическом пункте, и пусть $f\colon M \to L$ –
отображение, которое каждому организму выборки ставит в соот
ветствие некоторый вид. Ясно, что каждый $x \in M$ принадлежи
к какому-либо виду, но имеются виды, которые не были встречен
в данном пункте. Кроме того, несколько организмов могут при
надлежать к одному и тому же виду.

Отображение $f\colon M \to L$ называется сюръективным, если любо
элемент $y \in L$ имеет хотя бы один прообраз. Например, пусть L –
множество всех известных родов, а M — множество всех извест
ных видов. Отображение $f\colon M \to L$ ставит в соответствие каждому
виду определенный род. Ясно, что каждый род содержит по край
ней мере один вид, следовательно, отображение множества видов
на множество родов сюръективно.

Отображение $f\colon M \to L$ называется инъективным, если для каж
дого $y \in L$ существует не более одного прообраза. Отображение
множества организмов выборки в множество реальных видов
инъективно, так как существуют виды, к которым не принадле
жит ни один из организмов выборки.

Если отображение $f\colon M \to L$ одновременно сюръективно и
инъективно, то оно называется биективным. При этом множества
M и L являются равномощными (содержащими одно и то же коли
чество элементов). Пусть L — множество видов, обнаруженных
в некотором географическом пункте, M — список этих видов, со
стоящий из названий. Отображение $f\colon M \to L$ — биективно, если
каждый реальный вид имеет название.

Введенные понятия поясняются схемой на рис. 2.2.

2.3. Графы

Существует еще один способ задания бинарных отношений на
конечных множествах. Изобразим элементы множества M точками
на плоскости. Если выполнено отношение xAy, где $x, y \in M$, то
проведем стрелку из x в y. Совокупность точек и соединяющих их
линий назовем графом, при этом точки называются вершинами
графа, а соединяющие их линии — ребрами. Если ребра изобра
жаются в виде стрелок, то они называются дугами, а графы в этом
случае называются ориентированными графами, или орграфами.
Симметричные отношения изображаются обычно в виде графов,
а несимметричные — орграфами.

Для примера построим орграф отношения «быть хищником»
множестве $M = \{$щ, о, п$\}$ (рис. 2.3). Дуга, изображаю
ая отношение xAx, называется петлей; на рис. 2.3 имеются две
етли: щАщ и оАо.

В примере все вершины орграфа связаны дугами. Такой граф
азывается связным. В отличие от этого граф называется несвяз
ым, если он состоит из отдельных связ
ых подграфов (компонент) или изоли
ованных вершин.

Два ребра называются смежными,
если они имеют общую вершину. Эти
ребра называются также инцидентными
данной вершине. Степенью вершины на
зывается число инцидентных ей ребер.
Для орграфов различают число входя
щих и число выходящих из вершины
дуг. В этом случае полустепенью исхода
вершины называют число вершин, смеж
ных из данной вершины; полустепенью
захода — число вершин, смежных к дан
ной вершине.

Орграф на рис. 2.3 имеет две верши
ны, щ и о, у которых полустепень исхо
да равна трем: из них выходят стрелки
к трем вершинам. Для вершины п полу
степень исхода равна нулю: нет вершин,
к которым направлены стрелки из п. По
лустепень захода для вершин щ, п, о
равна двум.

Последовательность ребер, при кото
рой конец одного ребра совпадает с на
чалом другого, называется маршрутом.
Путь — это маршрут, у которого все
вершины различны. Если существует
путь от одной вершины к другой, то го
ворят, что вторая вершина достижима
из первой.

Орграф называется сильно связным,
если любые две вершины взаимно дости
жимы, слабо связным — если любые две
вершины соединены полупутем (нет двойных разнонаправленных
стрелок между любыми двумя вершинами), односторонне связ
ным — если для любых двух вершин по крайней мере одна дости
жима из другой. На рис. 2.3 имеем односторонне связный граф.

Двудольный граф (или биграф) — это граф, множество вершин
которого можно разбить на два подмножества таким образом, что
каждое ребро графа соединяет вершины из разных множеств
(на рис. 2.2 изображены двудольные графы).

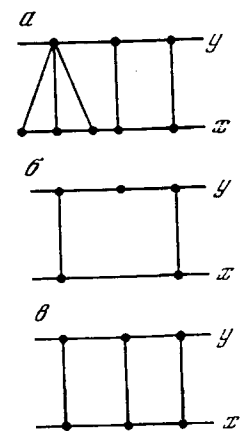


Рис. 2.2 Схема, поясняющая сюръективное (*а*), инъективное (*б*) и биективное (*в*) отображения

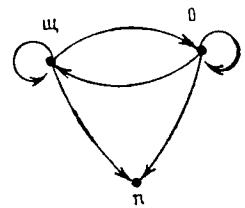


Рис. 2.3. Отношение «быть хищником» на множестве $M = \{$щ, о, п$\}$

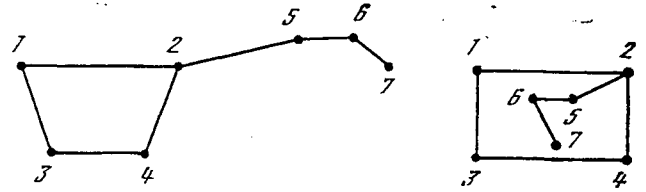


Рис. 2.4. Эквивалентные графы (цифры — вершины графа)

Говорят, что два графа обладают одной и той же структурой, если в одном графе столько же вершин и ребер, сколько в другом, и если ребра одного графа соединяют такие же вершины, что и в первом. Такие графы называются эквивалентными (рис. 2.4).

2.4. Алгебра логики

Элементарную основу алгебры логики составляет алгебра высказываний, в которой изучаются высказывания и операции над ними. При этом под высказыванием понимают всякое утверждение, о котором можно вполне определенно и объективно сказать, истинно оно или ложно.

«Лошади едят овес» — высказывание, которое истинно; «кит — рыба» — высказывание, которое ложно; «будьте внимательны» — не является логическим высказыванием, поскольку относительно этой фразы нельзя сказать, истинна она или ложна.

Обычно истину обозначают единицей, а ложь — нулем. Тогда значение истинности равно либо нулю, либо единице.

По существу можно говорить об отображении f: $M \to L$, где M — множество всех высказываний, а $L = \{0, 1\}$, т. е. об отображении совокупности всех высказываний на множество L, состоящее из двух элементов, один из которых носит название истины, а другой — лжи.

В каждом высказывании имеется объект и предикат (признак, свойство). Чтобы пояснить эти понятия, разберем высказывание «млекопитающие не несут яйца». Здесь объект — млекопитающие, предикат — признак, или свойство, млекопитающих, заключающееся в том, что они не несут яйца. Обозначим значение истинности высказывания y, объект — x, а предикат — f. Тогда приведенное высказывание можно записать как

$$y = f(x) \tag{2.19}$$

(читается: y тогда и только тогда, когда x обладает свойством f). При подстановке вместо x какого-либо конкретного объекта значение $f(x)$ будет либо истинным, либо ложным. Например, f (корова) $= 1$, а f (утконос) $= 0$. Итак, f — это предикат, который задается функцией, или отображением множества всех млекопитающих в двухэлементное множество $L = \{0, 1\}$.

Разберем еще одно высказывание: «два вида не могут сосуществовать в одной экологической нише». Пусть $M = \{x_1, x_2, \ldots, x_n\}$ — множество всех видов, а f — свойство пары каких-либо видов x_i, $x_j \in M$, заключающееся в том, что эта пара не может достаточно долго сосуществовать в одной и той же экологической нише. Тогда $f(x_i, x_j) = 1$, где f есть предикат на множестве M. Здесь мы привели пример двухместной функции, определенной на множестве M со значениями на множестве $L = \{0, 1\}$.

В более общем случае n-арным (n-местным) предикатом, определенным на множестве M, называется отображение n-арной декартовой степени множества M в двухэлементное множество $L = \{0, 1\}$. Короче,

$$f : M^n \to L. \tag{2.20}$$

Вспоминая, что любое подмножество $A \subseteq M$ называется n-арным отношением, определенным на множестве M (см. параграф 2.3), легко установить связь между понятиями n-арного предиката и n-арного отношения, определенных на одном и том же множестве M, а именно подмножество A всех систем $x_1, x_2, \ldots, x_n \in M$, для которых $f(x_1, x_2, \ldots, x_n) = 1$ есть n-арное отношение на множестве M. Так что отношение — это признак, предикат, по которому из заданного множества M^n выделяется подмножество A систем $x_1, x_2, \ldots, x_n$.

Если имеются некоторые высказывания, то из них при помощи логических связок «и», «или», «не», «если..., то...», «тогда и только тогда, когда...» можно образовывать новые высказывания. Например, формулировка транзитивности отношения «быть больше»: «если $a > b$ и $b > c$, то $a > c$» — образовано из трех высказываний: $f(a, b)$, $f(b, c)$, $f(a, c)$, где f — предикат, определенный на множестве натуральных чисел. Значит, транзитивное свойство отношения «быть больше» можно записать в виде предложения

«если $f(a, b)$ и $f(b, c)$, то $f(a, c)$». (2.21)

При конкретных значениях a, b, c предложение (2.21) является высказыванием, а его истинностные значения зависят как от значений a, b, c, так и от используемых связок.

В алгебре логики роль связок играют так называемые логические операции.

Одна из них — операция отрицания — соответствует логической связке «не». Более точно отрицанием высказывания A называется высказывание, обозначаемое $\bar{A}$ (читается: не A), которое истинно, когда A ложно, и ложно, когда A истинно. Например, если A обозначает высказывание «курица — птица», то $\bar{A}$ обозначает предложение «курица — не птица».

Другой операцией является операция конъюнкции, которая соответствует логической связке «и». Конъюнкцией высказываний A и B называется высказывание, обозначаемое $A \cdot B$ (читается A и B), значение которого истинно только тогда, когда оба выска

зывания A и B истинны. Во всех остальных случаях оно ложно. Используя эти понятия, выражение (2.21) можно записать:

«если $f(a, b) \cdot f(b, c)$, то $f(a, c)$». (2.22)

Следующая операция — операция дизъюнкции, которая выполняет роль связки «или». Дизъюнкцией высказываний A и B называется высказывание, обозначаемое $A + B$ (читается: A или B), значение которого только тогда ложно, когда оба высказывания A и B ложны. Во всех остальных случаях оно истинно. Например, высказывание A: «в данной реке обитает щука», высказывание B: «в данной реке обитает окунь». Тогда $C = A + B$ читается: «в данной реке обитает либо щука, либо окунь», а высказывание $D = A \cdot B$: «в данной реке обитают окунь и щука».

Следующая операция — операция импликации — выполняет роль логической связки «если..., то...». Импликацией двух высказываний A и B называется высказывание, обозначаемое $A \rightarrow B$ (читается: A имплицирует B), значение которого ложно только тогда, когда A истинно и B ложно. Во всех остальных случаях оно истинно. Используя это понятие, выражение (2.22) можно записать:

$$f(a, b) \cdot f(b, c) \rightarrow f(a, c). \quad (2.23)$$

Здесь роль высказывания A играет сложное высказывание, состоящее из двух других, соединенных связкой «и». В импликации $A \rightarrow B$ A называется посылкой, а B следствием. При умозаключениях из того, что импликация $A \rightarrow B$ истинна и посылка A истинна, делают вывод, что B истинно.

Еще одна операция — операция эквивалентности — соответствует логической связке «тогда и только тогда, когда...». Эквивалентностью высказываний A и B называется высказывание, обозначаемое $A = B$ (читается: A эквивалентно B), значение которого истинно только тогда, когда одновременно оба высказывания A и B либо истинны, либо ложны. Условия, в которых $A \rightarrow B$ и $B \rightarrow A$, означают, что $A = B$. Например, фразу «нет дыма без огня, а огня без дыма» можно заменить эквивалентной по смыслу фразой: «дым появляется тогда и только тогда, когда есть огонь» или «огонь есть тогда и только тогда, когда есть дым».

Среди всех высказываний выделим такие, которые остаются истинными или ложными безотносительно к тому, какие значения истинности принимают входящие в них элементы, т. е. такие, что для всех элементов из M^n они принимают только одно значение, равное единице, или только одно значение, равное нулю. В первом случае они называются тождественно истинными (ТИ) высказываниями, во втором — тождественно ложными (ТЛ). Легко установить, например, что формулы

$$A \rightarrow A,\ A + \bar{A},\ \overline{A + \bar{A}} \quad (2.24)$$

являются ТИ-высказываниями.

Теперь приведем основные правила алгебры высказываний, которые позволяют решать практические задачи.

1. $A + B = B + A$.
2. $A + (B + C) = (A + B) + C$.
3. $(A + B) \cdot C = A \cdot C + B \cdot C$.
4. $A \cdot B = B \cdot A$.
5. $A \cdot (B \cdot C) = (A \cdot B) \cdot C$.
6. $\bar{\bar{A}} = A$.

Здесь эквивалентности 1—4 выражают коммутативный и ассоциативный законы, эквивалентность 5 — дистрибутивный закон, а 6 называется законом двойного отрицания. Все эти правила аналогичны правилам обычной алгебры. Аналогия в последнем случае: $-(-a) = a$.

7. $A + A = A$.
8. $A \cdot A = A$.

Правило 7 читается: «либо A, либо A есть то же самое, что A» (вспомните: «что в лоб, что по лбу»). Правило 8 читается: «одновременно A и A есть то же самое, что A» (т. е. один и тот же объект не может считаться двойным). Эти правила обладают свойствами идемпотентности (от латинского idem — «тот же самый» и potens — «мощный»).

9. $A + 1 = 1$.
10. $A + 0 = A$.
11. $A \cdot 1 = A$.
12. $A \cdot 0 = 0$.

Поскольку значение истинности дизъюнкции ложно только тогда, когда сразу оба составляющие высказывания ложны, то правило 9 есть ТИ-высказывание: «одно из составляющих всегда истинно». Правило 10 указывает, что значение истинности дизъюнкции в случае, когда одно из составляющих всегда ложно, зависит от второго составляющего: если оно ложно, то и $A + 0$ ложно, если A истинно, то и $A + 0$ истинно.

Для разъяснения правил 11 и 12 напомним, что значение истинности конъюнкции истинно только тогда, когда оба составляющих истинны.

13. $A + \bar{A} = 1$.
14. $A \cdot \bar{A} = 0$.

Правило 13 можно выразить словами: «все есть либо A, либо не A», правило 14 — «ничто не может быть одновременно A и не A».

15. $\overline{A + B} = \bar{A} \cdot \bar{B}$.
16. $\overline{A \cdot B} = \bar{A} + \bar{B}$.

Правило 15 читается: «то, что не есть A или B, равносильно тому, что это не A и не B». Например, заключенную в скобки часть фразы: «бывают организмы, про которые можно сказать, что они (не являются животными (A) или растениями (B))», можно заменить

эквивалентной по смыслу частью: «это и не животные и не растения $(\bar{A}\cdot\bar{B})$».

Правило 16 читается: «то, что не есть одновременно A и B, равносильно тому, что это либо не A, либо не B». Например, «не бывает крокодилов (A), обитающих в голой пустыне (B). Это так же верно, как и то, что либо это не крокодилы, либо это не голая пустыня $(\bar{A}+\bar{B})$». Первое предложение выражает ту мысль, что крокодилы и голая пустыня несовместны: $\bar{A}\cdot\bar{B}$.

Эквивалентности 15 и 16 называются законами Де Моргана и примечательны тем, что позволяют переходить от дизъюнкции к конъюнкции и, наоборот, от конъюнкции к дизъюнкции.

17. $A \rightarrow B = \bar{B} \rightarrow \bar{A}$,

т. е. «если A, то B» равносильно тому, что «если не B, то не A». Например, фраза «если животные живут долго (A), то, значит, они имеют пищу (B)» равносильна высказыванию: «если нет пищи $(\bar{B})$, то животные не живут долго $(\bar{A})$». Эквивалентность 17 называется законом контрапозиции.

18. $A \rightarrow B = (\bar{A} + B = 1)$,

т. е. «если A, то B», равносильно высказыванию: «всегда верно, что или не A, или B». Так, относительно последнего примера $(\bar{B} \rightarrow \bar{A})$ можно сказать: «Всегда верно, что или животные имеют пищу, или они не живут долго $(\bar{B} + \bar{A} = B + \bar{A} = 1)$».

Правило 18, выражающее переход от импликаций к ТИ-высказываниям, играет исключительно важную роль в практических приложениях алгебры логики.

19. $A + \bar{A}\cdot B = A + B$.

Чтобы убедиться в справедливости эквивалентности 19, произведем операцию отрицания над обеими частями. Согласно законам Де Моргана, а также правилам 14 и 10

$$\overline{A + \bar{A}\cdot B} = \bar{A}\cdot(A + \bar{B}) = \bar{A}\cdot A + \bar{A}\cdot\bar{B} = \bar{A}\cdot\bar{B},$$

$$\overline{A + B} = \bar{A}\cdot\bar{B},$$

т. е. получим очевидное равенство, а из понятия эквивалентности следует, что $(A = B) = (\bar{A} = \bar{B})$.

20. $A + A\cdot B = A$.

Это правило называется законом поглощения, так как добавочный член в выражении «поглощается». Эквивалентность 20 следует из того, что $(A\cdot B \rightarrow A) = 1$. Например, утверждение «если в каком-либо водоеме обитают и щука и окунь, то щука там обитает» всегда истинно.

Несмотря на обилие приведенных формул, при практическом использовании они легко запоминаются и уже после решения двух-трех задач становятся такими же привычными, как действия

в арифметике или элементарной алгебре. Примеры практических задач, решение которых основано на приведенных правилах, можно найти в параграфах 4.2, 4.3.

2.5. Системы. Системы-классификации

Опираясь на введенные понятия, определим систему как непустое множество объектов (или несколько таких множеств), между которыми установлены некоторые отношения [40]. Благодаря последним набор элементов рассматривается как целостное единство, обладающее интегративными свойствами и противостоящее окружению, или среде. В качестве среды может рассматриваться система более высокого порядка (надсистема), в которой исследуемая система является лишь элементом. С другой стороны, элементы исследуемой системы могут рассматриваться как системы более низкого порядка (подсистемы).

На основе сказанного систему C можно представить пятеркой

$$C = C(\mathcal{S}, \mathcal{R}, A^{(S)}, A^{(RS)}, A^{(SR)}), \qquad (2.25)$$

члены которой имеют следующее толкование:

множество $\mathcal{S} = \{S_1, \ldots, S_p\}$ — состав системы, где $S_1, \ldots, S_N$ — внутренние элементы C;

множество $\mathcal{R} = \{R_1, \ldots, R_q\}$ — окружающая среда, а $R_1 \ldots R_q$ — внешние элементы C;

множество $A^{(S)}$ — все n-арные отношения на элементах (внутренняя структура системы C);

множества $A^{(RS)}$ и $A^{(SR)}$ — все n-арные отношения между элементами множеств $\mathcal{S}$ и $\mathcal{R}$ (структура связей взаимодействия системы со средой).

Если хотя бы один член пятерки изменяется во времени или пространстве, то система называется динамичной, в противном случае — статичной.

Системы-классификации — это результат классификационных построений на множествах объектов. Примерами систем являются множество описаний объектов с заданным отношением эквивалентности (принадлежности к одному и тому же классу); множество классов с заданным отношением иерархии; множество классификаций с заданным отношением доминирования и т. п.

В данных примерах указаны системы-модели, т. е. некоторые абстрактные аналоги реальных систем, которые значительно проще последних во всех аспектах, кроме самых важных для конкретного рассмотрения.

Каково же соотношение между моделью системы и ее прототипом? Один и тот же объект окружающего мира, скажем организм, с точки зрения анатома — сложная система, состоящая из пищеварительной, выделительной и других подсистем, эколог же может рассматривать организм как элемент более сложной системы — популяции, а технолог оценит его с гастрономической

точки зрения как пригодный или непригодный компонент какого-либо блюда. Все это иллюстрирует ту мысль, что система — это не только сущность, но и отношение к сущности [22].

Такая двойственность (т. е. сочетание субъективного и объективного начал) характерна и для систем-классификаций. Возникает она из-за того, что человек при классификационных построениях учитывает лишь ограниченное число признаков из бесконечного числа возможных. Для бесконечного набора, которым обладает реальный объект, существует и бесконечное множество вариантов выбора ограниченных наборов.

Множество признаков, которое учитывается на объектах, назовем системой описания, а множество значений каждого из учитываемых признаков на конкретных объектах — описаниями этих объектов. Так что аналоги-модели объектов — это система множеств, каждое из которых есть описание.

Пусть $\mathscr{S}' = \{S_1', \ldots, S_n'\}$ — множество реальных объектов, а $\mathscr{S} = \{S_1, \ldots, S_p\}$ — множество их описаний. Тогда отображение α: $\mathscr{S}' \to \mathscr{S}$ называется гомоморфным отображением множества $\mathscr{S}'$ на множество $\mathscr{S}$, если $\mathscr{S}$ имеет тот же состав, что и $\mathscr{S}'$ (но обратное неверно). Аналогично отображение системы $C' = C(\mathscr{S}', \mathscr{R}', A^{(R)}, A^{(R'S)}, A^{(S'R')})$ на систему $C = C(\mathscr{S}, \mathscr{R}, A^{(R)}, A^{(RS)}, A^{(SR)})$ считается заданным, если задана пятерка отображений:

$$\alpha_1 : \mathscr{S}' \to \mathscr{S}; \qquad \alpha_4 : A^{(R'S')} \to A^{(RS)},$$

$$\alpha_2 : \mathscr{R}' \to \mathscr{R},$$

$$\alpha_3 : A^{(R')} \to A^{(R)}, \qquad \alpha_5 : A^{(S'R')} \to A^{(SR)}.$$

Система-модель — это образ системы-оригинала при гомоморфном отображении [19].

Поясняя эти определения содержательным языком, можно сказать, что система-модель содержит меньшее число элементов и связей, чем система-оригинал, но все элементы и связи, которые имеются в модели, правильно копируют прототип.

Возникает вопрос: как выбрать «правильную» модель, «хорошую» систему описания? Этот выбор определяется содержательными целями (суперзадачей) классификационных построений. Формальной процедуры для выбора целей исследований нет (и, по-видимому, не может быть), поэтому нет (и, по-видимому, не может быть) и формальных правил для выбора системы описания.

Чтобы пояснить эту мысль, рассмотрим следующую ситуацию. Допустим, что имеется некоторая совокупность деревянных и стальных шариков, независимо от материала изготовления окрашенных в черный и красный цвета. Если при классификационных построениях учитывать цвет и вес отдельно, то получим две в общем случае не совпадающие классификации, если учитывать совместные значения этих признаков, то получим третий вариант.

Определить, какой из них является «естественным», а какой — «искусственным», не представляется возможным. Каждый вариант объективен в том смысле, что существует в реальной действительности, хотя из-за произвола в выборе принципа классификации (цели) он в то же время и субъективен. Вот почему и система-классификация отражает не только сущность, но и отношение к сущности.

После того как цель сформулирована, система описания выбрана и проведены необходимые измерения, в руках исследователя остается некоторый перечень модельных объектов (описаний), между которыми по предположению должны существовать какие-то связи, или отношения. Если элементы системы изобразить графически как точки на плоскости, а связи (отношения) между элементами как линии или стрелки, то такую диаграмму (граф) будем называть структурной схемой системы. Таким образом, система-классификация — это структурная схема, отображающая некоторую совокупность отношений на множествах объектов.

Глава 3

КЛАССИФИКАЦИИ, ОСНОВАННЫЕ НА КАЧЕСТВЕННЫХ ПРИЗНАКАХ

3.1. Виды измерений

В биологических и географических исследованиях объекты изучения настолько сложны, что практически никогда не изучаются целиком. Обычным приемом их характеризации является организованный по определенным правилам отбор некоторой представительной части, именуемой выборкой из генеральной совокупности. Примером последней может служить множество видов, обитающих в каком-либо регионе в заданный момент времени. Выборки малого объема из такой совокупности позволят скорее всего узнать только наиболее многочисленных представителей. Увеличивая число выборок или их объем, мы почти наверняка будем встречать и «новые» виды, но абсолютно точно видовой состав может быть определен при одновременном отлове абсолютно всех животных из данного региона.

Детерминистский подход основан на абсолютном доверии к выборочным характеристикам, независимо от соотношения объемов выборки и генеральной совокупности.

Несмотря на очевидную несправедливость этого допущения, такой подход во многих случаях является не только вполне приемлемым, но и единственно возможным. По существу речь идет о некотором недостатке информации, могущем повлечь за собой

неточные количественные выводы. Однако качественные обобщения, следующие из полученных характеристик, могут быть вполне справедливыми и, что самое главное, устойчивыми.

Выше указывалось, что объекты, подлежащие классификации, изучаются прежде всего с точки зрения наличия у них характерных свойств или состояний, именуемых признаками. Значения последних могут измеряться с различной точностью.

Самый слабый вид измерения — это измерения в так называемой шкале наименований: указывается только, одинаковы или нет объекты с точки зрения измеряемого признака. Номинальные признаки встречаются довольно часто. Таковы, например, ответы при заполнении полевой анкеты: «пол», «цвет», «запах» и т. п. При грубых измерениях иногда выделяют альтернативные (бинарные) признаки, т. е. такие, которые принимают всего два значения: «есть — нет». Например: «здоров — нездоров», «жив — мертв», «присутствует — отсутствует» и т. п.

Кроме указанных, используются также порядковые или ранговые признаки, которые могут сравниваться только по отношению «больше — меньше» и у которых бессмысленно даже сравнивать длины интервалов между оценками.

Более точные измерения предполагают и большее число значений. Например, вместо двух значений («большой — маленький») можно ввести три: «большой — маленький — средний». Часто таким значениям приписывают баллы, и подобные оценки называются балльными. Значения (градации) балльной шкалы представляют собой ограниченный дискретный ряд чисел, отстоящих друг от друга на одинаковом расстоянии. Обычно это начальный отрезок натурального ряда или часть ряда целых чисел, симметричных относительно нуля ($0, \pm 1, \pm 2, \ldots, \pm n$).

При дальнейшем увеличении точности измерений число значений можно увеличивать, доводя их до максимально осуществимого. Например, от сотен единиц измерения доводить их до десятков, отдельных единиц, долей единиц и т. д.

Условно все виды оценок делят на качественные и количественные. Вслед за Н. Г. Загоруйко [27] качественными будем считать только те из них, которые измеряются в шкале наименований. Количественные признаки могут измеряться приборами со шкалами «порядка», «отношений», «интервалов» и отражают числовую характеристику степени проявления свойств объектов.

В биологических и географических исследованиях часто бывает достаточно качественной информации об объектах, тем более что ее легче получить и не нужно заботиться об использовании приборов, дающих числовые показатели (для многих биогеографических показателей таких приборов просто не существует). Но нередко причина ее использования заключается в том, что существо суперзадачи предполагает использование мер, решающих правил и т. п., с помощью которых может обрабатываться только качественная информация.

Ниже описываются классификационные построения, которы иллюстрируется подход к выбору системы описания, измере ям и способам их использования.

3.2. Формализация задачи обработки видовых списков

В биогеографических исследованиях зачастую анализируемый атериал состоит из списков видов и перечня географических пунк в с указаниями, обнаружен или нет данный вид в каждом из их. Подобного рода описания могут быть весьма обширными и держать сведения о встречаемости многих сотен видов во мно их пунктах. Это делает их труднообозримыми и малодоступными ля анализа на умозрительном уровне.

Более четкие результаты получаются при использовании мате атических методов, специально предназначенных для сжатия нформации и количественной характеристики интегративных войств анализируемого материала. Однако и на этом пути встре аются трудности, связанные с чрезмерным обилием коэффициен ов, используемых, как принято считать, для одних и тех же елей. До последнего времени в биологической литературе не прек ащаются споры о том, какой из двух самых простых коэффици нтов, Жаккара или Чекановского — Сёренсена, следует приме нять для характеристики сходства [13, 32, 36, 54].

К этому следует добавить, что ни перечня целей, ни единой етодики количественной обработки видовых списков до сих пор е создано, а это зачастую приводит к поверхностному анализу анных, собранных с большими тратами сил и времени.

Такое положение во многом обусловлено тем, что исходные биогеографические понятия не были формализованы, а примене ние математики, как указывалось, становится эффективным толь ко при этом условии. Попытки устранить существующие пробелы предпринимались П. Юхач-Надем [59], А. С. Константиновым [32], и их результаты были обобщены и развиты Б. И. Семкиным [46], а также В. Л. Андреевым и Ю. С. Решетниковым [4].

В данном разделе мы предпринимаем попытки сформулировать класс задач, которые возникают при обработке видовых списков, а также подобрать простые, но математически корректные средства их решения в рамках единой содержательной модели.

Видовые списки обычно представляются как матрица, имеющая q столбцов и p строк (порядка $p \cdot q$), причем номеру столбца соответствует название географического пункта R_j, а номеру строки — название вида S_i, $i = 1, 2, \ldots, p$. Информационным содержанием видовых списков являются указания о присутствии или отсутствии каждого из учитываемых видов в каждом исследуемом пункте. Условимся на пересечении j-го столбца и i-й строки помещать «1», если i-й вид присутствует в j-м пункте, и «0», если i-й вид не присутствует в j-м пункте.

Любой j-й столбец матрицы назовем описанием j-го пункта, а любую i-ю строку — описанием i-го вида. В терминах теории множеств

$$\mathcal{R} = \{R_j | _{j \in J}\}, \tag{3.1}$$

где $\mathcal{R}$ — индексированное множество с элементами R_j; $R_j = R_k$, если $k = j$; J — индексное множество:

$$J = \{j \mid j \text{ — целое число, } 1 \leqslant j \leqslant q\}. \tag{3.2}$$

Напомним, что запись типа $A = \{x \mid P(x)\}$ читается так: «множество A, состоящее из всех таких x, которые обладают свойствами P». Следовательно, формула (3.1) читается: «семейство множеств $\mathcal{R}$, состоящее из всех R_j, таких, у которых элементы j принадлежат множеству J».

Аналогично семейство множеств

$$\mathcal{S} = \{S_i \mid i \in I\} \tag{3.3}$$

есть индексированное множество, а I — индексное множество:

$$I = \{i \mid i \text{ — целое число, } 1 \leqslant i \leqslant p\}. \tag{3.4}$$

Индексация позволяет различать множества, состоящие из одинаковых элементов.

Любое $R_j \in \mathcal{R}$ или $S_i \in \mathcal{S}$ можно рассматривать так же, как кортеж (упорядоченный набор) значений «0» и «1», причем наличие кортежей, не имеющих значений «1» ($m(R_j) = 0$), будем оговаривать особо. Кортеж отличается от упорядоченного множества тем, что в первом могут встречаться одинаковые элементы.

Семейство множеств $\mathcal{R}$ или $\mathcal{S}$ с заданными на них отношениями можно рассматривать как системы, в которых связи между элементами образуют определенную структуру. Следовательно, содержание задач по обработке видовых списков включает подбор типов отношений и анализ структуры порождаемых ими систем.

Иногда исследования на элементах $\mathcal{R}$ называют «Q-анализом», а на элементах $\mathcal{S}$ — R-анализом. Такое деление чисто условно и зависит от того, что считать объектами, а что — признаками. Наличие какого-либо вида в географическом пункте можно рассматривать как характеристику этого пункта, а с другой стороны, множество географических пунктов, в которых встречается или нет данный вид, можно рассматривать как характеристику (местообитание) этого вида. Формальные процедуры R- и Q-анализа могут быть одинаковыми.

3.3. «Банальность» и «экзотичность»

Рассмотрим отношения, порождаемые сравнительно малоизвестными мерами включения.

Еще в 1943 г. Г. Г. Симпсон [61] при сравнении фаун континентов измерял включение списков одного региона в списке видов

другого отношением числа общих видов к числу видов в одном из них. Интерпретация этого показателя чрезвычайно проста и становится понятной уже из такого сопоставления: если видовой список одного региона полностью входит в список другого, то его «включение» будет стопроцентным. В более общем случае процент включения может быть различным для каждого из двух сравниваемых списков, и тогда можно говорить, что один из них по составу видов «более оригинален», чем другой. Сопоставляя меры включения всевозможных попарных сочетаний анализируемых списков, можно выяснить интересные закономерности, отражающие некоторые определенные отношения на множестве регионов.

Таблица 3.1

Видовые списки, представленные как семейство множеств

	R_1	R_2	R_3
S_1	0	1	0
S_2	1	1	1
S_3	1	0	0
S_4	0	1	1
S_5	0	0	1
S_6	1	1	1
S_7	1	1	0
S_8	0	1	0
S_9	1	1	0
S_{10}	1	0	1

Эти простые и ясные идеи могут оказаться весьма плодотворными в самых различных ситуациях, поэтому мы попытаемся развить идеи Г. Г. Симпсона в свете современных представлений математики [46] и проиллюстрировать возможность их использования для решения как теоретических, так и чисто утилитарных биологических задач.

Опираясь на ранее введенные понятия, определим меру включения множества N в множество M как частное от деления мер пересечения:

$$W(M; N) = \frac{m(M \cap N)}{m(N)}. \tag{3.5}$$

Обратим внимание на то, что при обозначении меры в скобках названия множеств разделены точкой с запятой: так обозначаются несимметричные меры.

Мера включения множества N в множество M:

$$W(M; N) = \frac{m(M \cap N)}{m(N)}. \tag{3.6}$$

Рассмотрим некоторый условный пример. В табл. 3.1 приведены три видовых списка, в которых учитывались десять видов ($q = 3$, $p = 10$).

Если анализировать описания какой-либо пары объектов, скажем R_1 и R_2, то можно встретить следующие четыре сочетания значений признаков:

1) 0 — 0 (у обоих объектов нет k-го признака);
2) 0 — 1 (у R_1 — нет, у R_2 — есть);
3) 1 — 0 (у R_1 — есть, у R_2 — нет);
4) 1—1 (у R_1 и R_2 — есть).

Число сочетаний каждого типа обозначим соответственно

1) $-m(\bar{R}_1 \cap \bar{R}_2)$; 2) $-m(R_1 \cap \bar{R}_2)$;
3) $-m(\bar{R}_1 \cap R_2)$; 4) $-m(R_1 \cap R_2)$.

По данным табл. 3.1 $m(\bar{R}_1 \cap \bar{R}_2) = 1$, так как такое сочетание значений имеет только один признак S_5; $m(\bar{R}_1 \cap R_2) = 3$ (S_1, S_4, S_8); $m(R_1 \cap \bar{R}_2) = 2$ (S_3, S_{10}); $m(R_1 \cap R_2) = 4$ ($S_2 S_6$, S_7, S_{10}).

Число значений «1» у первого описания $m(R_1) = 6$, у второго — $m(R_2) = 7$; общее число встреченных видов у R_1 и R_2 $m(R_1 \cup R_2) = 9$. На основе этих данных подсчитаем по формулам (3.5), (3.6)

$$W(R_1; R_2) = \frac{m(R_1 \cap R_2)}{m(R_2)} = \frac{4}{7} \approx 57\%,$$

$$W(R_2; R_1) = \frac{m(R_1 \cap R_2)}{m(R_1)} = \frac{4}{6} \approx 67\%.$$

Результаты можно интерпретировать следующим образом. Меры включения первого описания во втором (67%) показывают, что второй объект «оригинальнее», «экзотичнее» первого. Другими словами, описание второго объекта содержит «специфических» признаков больше, чем описание первого, поскольку первое описание включено во второе на 67%, а второе включено в первое на 57%. Легко заметить, что приведенные меры несимметричны, а включение j-го описания в самом себе стопроцентно, так как

$$m(R_j \cap R_j) = m(R_j).$$

При более полном анализе данных табл. 3.1. следует подсчитать меры включения для всех пар объектов. Для этого составим матрицу порядка $q \times q$ и пронумеруем строки и столбцы соответственно номерам изучаемых объектов (табл. 3.2).

Таблица 3.2

Матрица мер включения для данных табл. 3.1

	R_1	R_2	R_3
R_1		67	33
R_2	57		43
R_3	60	60	

Таблица 3.3

Отношение «банальности» B_{60} на множестве $\mathcal{R}$

	R_1	R_2	R_3
R_1		1	0
R_2	0		0
R_3	1	1	

В этой таблице число 57, вторая строка и первый столбец соответствуют $W(R_2; R_1)$, а число 60, третья строка и первый столбец — $W(R_3; R_1)$ и т. д., т. е. индекс при названии первого множества в скобках указывает номер строки, а второго — номер столбца. Таким путем осуществляется полный перебор всех возможных парных сочетаний $W(R_j; R_k)$, $k \in J$. Ясно, что элементы

главной диагонали матрицы всегда равны 1 (100%) и их можно не заполнять.

Данные табл. 3.2 мы используем для анализа отношений «банальности», порождаемых мерами включения. Зададимся некоторым произвольным числом Δ ($0 \leqslant \Delta \leqslant 100$) и каждое значение $W(R_j; R_k) \geqslant \Delta$ заменим единицей, а прочие — нулем. Тогда от матрицы мер включения переходим к матрице отношения «банальности» B_Δ, точнее,

$$\langle B_\Delta; \mathcal{R} \rangle = \{R_j, R_k \in \mathcal{R} \mid W(R_k; R_j) \geqslant \Delta\}, \qquad (3.7)$$

где $j, k \in J$. Запись $R_j B_\Delta R_k$ означает, что описание R_j «банальнее» R_k при пороге Δ, или что R_j и R_k находятся в отношении» «Δ-банальности».

Для данных табл. 3.2 зададимся $\Delta = 60\%$. Тогда получим следующую матрицу отношений «60%-банальности» (табл. 3.3).

Анализируя данные табл. 3.3 по строкам, можно сразу заметить, что описание R_1 «банальнее», R_2, R_2 является самым «экзотичным», а R_3 — самым «банальным» при пороге $\Delta = 60\%$.

Если снизить величину порога до $\Delta = 50\%$, то число единичек увеличится (табл. 3.4).

Таблица 3.4

Отношение «банальности» B_{50} на множестве $\mathcal{R}$

	R_1	R_2	R_3
R_1		1	0
R_2	1		0
R_3	1	1	

Здесь мы отмечаем, что $R_1 B_{50} R_2 = R_2 B_{50} R_1$, т. е. включение описаний R_1 и R_2 друг в друга выше 50%, и, стало быть, они наиболее «похожи» среди прочих пар.

Конечно, в этом примере непосредственный обзор табл. 3.2 и 3.3 не составляет труда, однако при их большей размерности анализ подобного рода был бы затруднительным. В таких случаях целесообразно перейти к графическому изображению матрицы

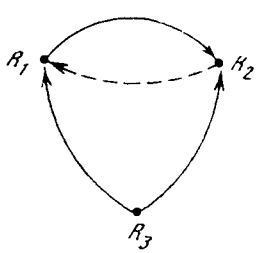


Рис. 3.1. Орграфы отношений «банальности» B_{60} и B_{50} (пунктир)

отношений с помощью графов. В обсуждаемом примере имеется возможность отобразить обе матрицы одновременно на одном рисунке (рис. 3.1).

Здесь получаются связные орграфы, один из которых имеет три дуги, соединяющие три вершины (слабо связный), а другой — четыре дуги, соединяющих эти же вершины (односторонне связ-

ый). Можно заметить, что в первом случае наибольшее число стрелок выходит из вершины R_3, следовательно, соответствующее ей описание самое «банальное», наоборот, в вершину R_2 входит наибольшее число стрелок, значит, соответствующее ей описание является наименее «банальным». Обоюдная направленность дуг между вершинами R_1 и R_2 свидетельствует о большом сходстве соответствующих описаний.

Возникает вопрос: как установить правильную величину порога? При практическом использовании отношения «банальности» B_Δ эту величину подбирают исходя из целевой установки задачи. Обычно же используется самый простой способ перебора серии значений Δ, при которых все существенные связи отражены на графе и сам он не является громоздким. Подбор удобно осуществлять, задавая сначала большие значения Δ, а затем, постепенно снижая их, доводить до такого, при котором все связи все еще остаются легко обозримыми.

Рассмотрим более сложный пример, заимствованный из практики гидробиологических исследований.

В статье Г. Н. Гладких [19] опубликованы данные о наличии 88 видов фитопланктона у юго-восточного побережья о-ва Хонсю в различные сезоны 1967 г. Акватория, на которой собраны пробы, разделяется на два района, условно названных «южным» и «восточным», расположенным к северо-востоку от южного. В южном районе разрезы выполнялись в феврале, мае, августе и ноябре. Спискам видов этих сборов нами присвоены номера соответственно *1*, *2*, *3* и *4*. В восточном районе разрезы проводились в эти же месяцы, за исключением мая, и видовым спискам этих сборов присвоены номера *5*, *6* и *7* (рис. 3.2). Матрица мер включения, подсчитанных по исходным данным, приведена в табл. 3.5, а соответствующий ей орграф отношений «банальности» B_Δ — на рис. 3.2.

Таблица 3.5

Матрица мер включения описаний фитоцена у о-ва Хонсю

	1	2	3	4	5	6	7
1		20	33	0	20	13	20
2	30		20	10	20	10	20
3	15	6		9	12	18	29
4	0	17	50		0	17	50
5	23	15	31	0		15	31
6	17	8	50	8	17		42
7	12	8	38	12	15	19	

Первое, на что следует обратить внимание, это малое число ребер графа даже при таком низком пороге, как $\Delta = 30$. При $\Delta = 50$ граф становится несвязным: вершины R_1, R_2, R_5 изолированы. Эти факты свидетельствуют о малом сходстве описаний и, следовательно, о резкой смене видового состава как по сезонам, так и по районам.

Второе: из южного в восточный входит только одна стрелка, в то время как из восточного в южный — три. Следовательно, видовой состав в южном районе в целом более разнообразен, чем в восточном.

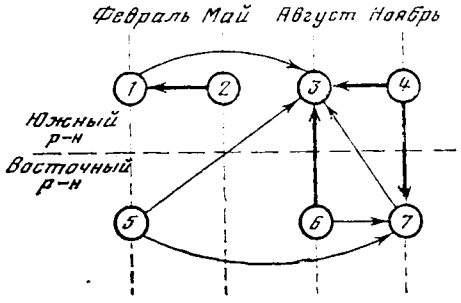


Рис. 3.2. Орграфы отношений «банальности» B_{50} (жирные линии) и B_{30} (тонкие линии) на множестве описаний фитоцена о-ва Хонсю

Третье: наибольшее число стрелок в южном районе входит в точку 3 (август), а в восточном — в точку 7 (ноябрь). Значит, наибольшее разнообразие видов в южном районе приходится на осень, а в восточном — скорее на зиму. Иначе говоря, наблюдается существенная разница сукцессионных изменений по районам.

Четвертое: в южном наибольшее число стрелок выходит из вершины *4*, а в восточном — из вершины *6*. Следовательно, соответствующие им описания являются наиболее «банальными» по видовому составу.

Пятое: видовой состав R_2 (май) является более «экзотичным», чем R_1 (февраль). Значит, началом сериального цикла изменений фитоцена в южном районе нужно считать весну, а не зиму. Наконец, небезынтересно отметить связи вершин 3, 4, 6, 7. Они показывают, в частности, резкое падение видового разнообразия после достижения максимума, а также возможность обмена флоры южного и восточного районов.

На этом примере легко убедиться в высокой эффективности применяемых методов в уплотнении цифровой информации. В самом деле, при непосредственном анализе видовых списков приходится оперировать таблицей $7 \times 88 = 616$ значений типа «0» и «1», а с применением графов — бесхитростным рисунком. Именно это обстоятельство позволяет надеяться, что некоторые усложнения применяемых методов вполне окупятся результатами, которые трудно получить каким-либо иным способом.

Обратимся снова к данным табл. 3.1 и подсчитаем меру включения R_1 в объединении $R_2 \cup R_3$:

$$W(R_1; R_2 \cup R_3) = \frac{m[R_1 \cap (R_2 \cap R_3)]}{m(R_2 \cup R_3)} = \frac{m(R_1 \cap R_2) + m(R_1 \cap R_3) - m(R_1 \cap R_2 \cap R_3)}{m(R_2) + m(R_3) - m(R_2 \cap R_3)}. \quad (3.8)$$

Несмотря на некоторую громоздкость формулы (3.8), практический подсчет по ней не составляет трудностей.

Сначала находим меру объединения $m(R_2 \cup R_3)$. Для этого согласно определению нужно подсчитать общее число имеющихся признаков у R_2 и R_3:

$$m(R_2 \cup R_3) = m(R_2) + m(R_3) - m(R_2 \cap R_3) = 7 + 5 - 3 = 9.$$

Это легко подсчитать и непосредственно, если в табл. 3.1 провести объединение двух последних столбцов, причем в каждой строке ставить единицу только тогда, когда хотя бы в одном из столбцов имеется единица. В результате вместо трех столбцов можно рассматривать два: один из них — R_1, а второй — объединение $R_2 \cup R_3$, но этот случай уже рассматривался с применением формул (3.1) и (3.2). Такой прием (сведение нескольких описаний в два) легко реализуется, и его можно рекомендовать для практического использования. Если же применять второй вариант формулы (3.8), то нужно найти сначала число сочетаний типа «1—1 —1». По данным табл. 3.1 $m(R_1 \cap R_2 \cap R_3) = 2$. Так что

$$W(R_1; R_2 \cup R_3) = \frac{3+3-2}{7+5-3} = \frac{4}{9} \approx 44\%.$$

Пользуясь этими приемами, мы нашли для фитоценотических описаний прибрежных вод о-ва Хонсю:

$$W(R_5 \cup R_6 \cup R_7;\ R_1 \cup \ldots \cup R_4) = 24/54 \approx 45\%, \quad (3.9)$$
$$W(R_1 \cup \ldots \cup R_4;\ R_5 \cup R_6 \cup R_7) = 24/42 \approx 57\%,$$

т. е. южный район оказывается более оригинальным по сравнению с восточным.

Аналогично можно провести объединение не только по сезонам в каждом районе, но и по районам в каждом сезоне.

Может показаться, что для характеристики «оригинальности» объекта лучше вычислять отношение числа его специфических признаков к общему числу имеющихся у него признаков, как это предложено Б. А. Юрцевым [58]. В используемых терминах эта мера выглядит следующим образом:

$$\Phi(R_k; R_j) = \frac{m(\overline{R}_j \cap R_k)}{m(R_j)}. \quad (3.10)$$

Сделаем преобразования:

$$1 - \Phi(R_j; R_k) = \frac{m(R_k) - m(\overline{R}_j \cap R_k)}{m(R_k)} = \frac{m(R_k \cap R_j)}{m(R_k)} = W(R_k; R_j), \quad (3.11)$$

Из (3.11) видно, что две меры дополняют друг друга до единицы и, следовательно, мера (3.10) не приносит дополнительной информации по сравнению с мерой (3.11).

Этот результат следует как частный пример из теоремы Б. И. Семкина и В. И. Двойченкова [47]: две меры, $\mathscr{E}_1$ и $\mathscr{E}_2$, эквивалентны, если они связаны монотонно возрастающей зависимостью, т. е. $\mathscr{E}_1 = \varphi(\mathscr{E}_2)$, где φ — монотонно возрастающая функция; если φ — монотонно убывающая функция, то меры $\mathscr{E}_1$ и $\mathscr{E}_2$ называются коэквивалентными. Нетрудно видеть, что $W(R_k; R_j)$ и $\Phi(R_k; R_j)$ связаны монотонно убывающей зависимостью типа $1 - x$ $(0 \leqslant x \leqslant 1)$, следовательно, они коэквивалентны.

В содержательных терминах эти понятия можно пояснить следующим образом: две меры называются эквивалентными, если при замене одной из них другой качественный результат сравнения останется одним и тем же, т. е. более «банальные» объекты останутся более «банальными», менее «банальные» — менее «банальными».

Главное достоинство упомянутой теоремы заключается в том, что она дает возможность среди бесконечного множества (континуума) мер установить эквивалентные, дающие один и тот же качественный результат.

В частном случае, используя монотонно возрастающую зависимость

$$\varphi(u) = u/(2 - u),$$

можно «изобрести» такую меру включения:

$$W(R_k; R_j)_1 = \frac{m(R_k \cap R_j)}{2m(R_k) - m(R_k \cap R_j)}, \quad (3.12)$$

или же, используя зависимость

$$\varphi(u) = \frac{2u}{1+u},$$

можно получить

$$W(R_k; R_j)_0 = \frac{2m(R_j \cap R_k)}{m(R_k) + m(R_j \cap R_k)}. \quad (3.13)$$

Едва ли стоило «изобретать» эти меры, так как они эквивалентны более просто интерпретируемой мере (3.5). Точно так же нет необходимости использовать и формулу (3.10), если не считать ее интерпретацию более легкой, чем (3.5).

Правда, переход от количественных мер к отношениям, как это можно было видеть на приведенных примерах, связан с потерями информации. Поэтому использование отношений, порожденных коэквивалентными мерами, может оказаться полезным в том смысле, что эти отношения выделяют различные стороны указанной информации и могут иногда удачно дополнять друг друга.

Подобно тому, как это было сделано на примере мер включения, определим отношение «эндемизма» E_Δ как

$$\langle E_\Delta, \mathscr{R}\rangle = \{R_j, R_k \in \mathscr{R} \mid \Phi(R_k; R_j) \geqslant \Delta\}, \qquad (3.14)$$

где $j, k \in J$.

Запись $R_j E_\Delta R_k$ означает, что описание R_j «эндемичнее» R_k при пороге Δ.

Для иллюстрации практического использования введенных отношений продолжим анализ матрицы мер включения описаний фитоцена. Чтобы подсчитать меры специфичности, не нужно прибегать к исходным данным: достаточно все значения табл. 3.5 вычесть из 100 и использовать их для построения графа, отображающего отношения «эндемизма» E_Δ. Окончательные результаты приведены на рис. 3.3.

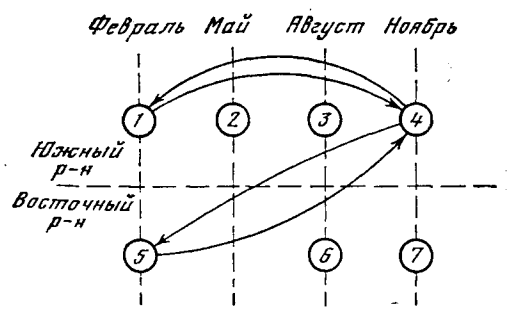


Рис. 3.3. Орграф отношений «эндемичности» E_{100}

Единственное достоинство сопоставления двух последних (3.2 и 3.3) рисунков заключается, пожалуй, в возможности выделения некоторых характерных особенностей или деталей, ускользающих от внимания при анализе только одной из коэквивалентных мер. В данном примере такими деталями является четкое выделение двух пар абсолютно несхожих описаний.

3.4. Сходство и различие

Понятия сходства и различия неоднозначны, и в общем случае разные субъекты вкладывают в них неодинаковое содержание. Об этом можно судить хотя бы по тому факту, что разные исследователи время от времени вводят все новые и новые коэффициенты для количественной характеристики сходства и различия. Если Н. Бейли [7] насчитывал их около 20, то в настоящее время их

стало бесконе[illegible]ного, так как сформулированы несколько правил, по которым «изобретаются» коэффициенты [47].

Одним из первых шагов на пути их упорядочивания является понятие эквивалентности мер, согласно которому, как уже указывалось, эквивалентными называют меры, одинаковые с точностью до монотонных преобразований. Это понятие оказывается полезным еще потому, что приводит к пониманию смысла использования неэквивалентных мер, заключающегося в характеризации различных свойств анализируемого материала.

В самом деле, уплотнение эмпирической информации, достигаемое с помощью математических методов, не дается даром: платой за компактность является тенденциозность извлекаемой информации и сравнительно малая ее часть по сравнению с тем, что содержится в исходных измерениях. Поэтому использование неэквивалентных мер не только желательно, но и необходимо. Выбор же конкретных коэффициентов зависит в первую очередь от суперзадачи — цели конкретного исследования (а также от шкалы измерений). Повторим: формальных правил для выбора целей нет, следовательно, не может быть и формальных правил для выбора подходящей меры сходства.

В математической литературе за меру сходства принимают неотрицательную вещественную функцию $C(R_j, R_k)$, обладающую следующими свойствами:

1) $0 \leqslant C(R_j, R_k) \leqslant 1$ для $k \neq j$;
2) $C(R_j, R_j) = 1$;
3) $C(R_j, R_k) = C(R_k, R_j)$.

Такими свойствами обладает, в частности, континуум эквивалентных мер, представляемых формулой [47]

$$C(R_j, R_k)_u = \frac{2m(R_j \cap R_k)}{(1+u)[m(R_j) + m(R_k)] - 2 \cdot u \cdot m \cdot (R_j \cap R_k)}, \qquad (3.15)$$

где $-1 < u \leqslant \infty$.

Например, при $u = 0$ получаем меру

$$C(R_j, R_k)_0 = \frac{2m(R_j \cap R_k)}{m(R_j) + m(R_k)}, \qquad (3.16)$$

которая численно совпадает с хорошо известным коэффициентом Чекановского—Сёренсена.

При $u = 1$ получаем

$$C(R_j, R_k)_1 = \frac{m(R_j \cap R_k)}{m(R_j \cup R_k)}, \qquad (3.17)$$

которая численно совпадает с коэффициентом Жаккара.

При $u = 3$ получаем меру

$$C(R_j, R_k)_3 = \frac{m(R_j \cap R_k)}{2m(R_j) + 2m(R_k) - 3m(R_j \cap R_k)}, \qquad (3.18)$$

которая численно совпадает с коэффициентом Сокала и Снита.

При $u = -1/2$ получаем меру

$$C(R_j, R_k)_{-1/2} = \frac{4m(R_j \cap R_k)}{m(R_j) + m(R_k) + 2m(R_j \cap R_k)}, \qquad (3.19)$$

которая, по-видимому, еще не использовалась, и т. д.

Заметим, что из формул (3.15) и (3.16) следует эквивалентность коэффициентов Чекановского—Сёренсена и Жаккара, поэтому споры о том, какой коэффициент лучше, следует считать беспредметными. Для перехода от одного из них к другому можно использовать соотношение

$$C(R_j, R_k)_1 = \frac{C(R_j, R_k)_0}{2 - C(R_j, R_k)_0}, \qquad (3.20)$$

не прибегая к анализу исходных данных.

Примером континуума неэквивалентных мер сходства могут служить усредненные меры включения, где усреднение производится с разными «весами». В частности, если сумму $W(R_k; R_j)$ и $W(R_j; R_k)$ разделить пополам, то получим коэффициент

$$K(R_j, R_k) = \frac{1}{2} m(R_j \cap R_k)\left[\frac{1}{m(R_j)} + \frac{1}{m(R_k)}\right], \qquad (3.21)$$

который численно совпадает с хорошо известным коэффициентом Кульчинского, и т. д.

Любые меры, коэквивалентные мерам сходства, будем называть мерами различия. К последним будем относить также и расстояния, обладающие свойствами метрики.

Неотрицательная вещественная функция $D(R_j, R_k)$ называется метрикой, если

1) $D(R_j, R_k) \geqslant 0$ для всех $j, k \in J$;
2) $D(R_j, R_k) = 0$ тогда и только тогда, когда $R_j = R_k$;
3) $D(R_j, R_k) = D(R_k, R_j)$;
4) $D(R_j, R_k) \leqslant D(R_j, R_s) + D(R_k, R_s)$, где $s \in J$.

Существуют и другие меры отдаленности объектов, которые образованы на основе отличных от перечисленных соображений и которые мы также будем называть мерами различия.

Как и в случае мер включения, попарное сходство объектов будем характеризовать матрицей мер. Однако в данном случае в силу ее симметричности относительно главной диагонали достаточно заполнять лишь ее верхний или нижний треугольник. Кроме того, поскольку сходство объекта с самим собой всегда стопроцентно, элементы главной диагонали могут быть опущены. Таким образом, при анализе q описаний матрица парных мер сходства будет иметь в общем случае $q(q-1)/2$ различающихся значений.

Одной из самых простых и легко интерпретируемых мер для характеризации парного сходства, которая имеет обобщение на n описаний [47], является коэффициент Чекановского-Сёренсена

$$C(R_1, \ldots, R_n)_0 = \frac{n\left[\sum^{n} m(R_k) - m(R_1 \cup \ldots \cup R_n)\right]}{(n-1)\sum_k m(R_k)}. \qquad (3.22)$$

Обращаясь к формулам (3.16) и (3.22), можно заметить, что их основу, как и у мер включения, составляют меры пересечения. Поэтому анализ видовых списков удобно начинать с подсчета матрицы значений $m(R_j \cap R_k)$. Поскольку $m(R_j \cap R_k) = m(R_k \cap R_j)$, то матрица мер пересечения симметрична относительно главной диагонали, а элементы последней составляют значения $m(R_j \cap R_j) = m(R_j)$ — число признаков, имеющихся у k-го описания. Таким образом, при анализе q описаний матрица мер пересечения будет иметь в общем случае $q(q+1)/2$ различающихся значений.

Рассмотрим конкретный пример подсчета и использования мер $m(R_j \cap R_k)$.

В работе О. Г. Кусакина [60] приводятся материалы по видовому составу макрофауны литорали островов Курильской гряды: Кунашира, охотское — R_1 и тихоокеанское — R_2 побережья; Шикотана — R_3; Итурупа, охотское — R_4 и тихоокеанское — R_5 побережья; Урупа — R_6; Симушира — R_7; Парамушира — R_8. Списки этих видов состоят из 603 наименований. Материалы были собраны и обработаны специалистами многих научных учреждений страны за период более 20 лет.

Исходная информация, насчитывающая около 5000 измерений, при попарном сравнении описаний может быть представлена в виде матрицы мер пересечений, содержащей всего 36 чисел (табл. 3.6).

Напоминаем, что элемент матрицы, находящийся в j-й строке и k-м столбце, есть $m(R_j \cap R_k)$ — число общих видов в описаниях

Таблица 3.6

Матрица мер пересечения фаунистических описаний восьми районов Курильской гряды

R_1	171							
R_2	95	201						
R_3	121	145	311					
R_4	81	100	143	194				
R_5	56	84	117	106	158			
R_6	36	46	69	69	68	118		
R_7	30	37	67	62	57	65	125	
R_8	48	67	109	97	99	85	93	235

R_j и R_k, а диагональные элементы $m(R_j \cap R_j) = m(R_j)$ — число видов в j-м описании.

Подсчет мер сходства по такой матрице не представляет никаких трудностей. Для примера вычислим коэффициент сходства между третьим и вторым описаниями. Элемент третьей строки и второго столбца имеет значение 145, диагональные элементы соответственно равны 311 и 201, тогда

$$C(R_2, R_3)_0 = \frac{2 \cdot 145}{311 + 201} \approx 57\%.$$

Заодно с этим поясним, как использовать матрицу мер пересечения для подсчета мер включения:

$$W(R_2; R_3) = \frac{145}{311} \approx 47\%; \qquad W(R_2; R_3) = \frac{145}{201} \approx 72\%.$$

Результаты расчета мер сходства приведены в табл. 3.7. Способ ее компоновки иллюстрирует, в каком виде представляются подобные данные для публикаций.

Таблица 3.7

Матрица мер сходства фаунистических описаний восьми районов Курильской гряды

(задана двумя частями верхнего треугольника)

	R_2	R_3	R_4	R_5	R_6	R_7	R_8	
	51	50	44	34	25	20	24	R_1
R_7	52		57	51	47	29	23	31	R_2
R_6	48	53		57	49	32	31	40	R_3
R_5	50	40	49		60	44	43	45	R_4

Отношения сходства и различия, порождаемые мерами, введем аналогично тому, как вводились предыдущие отношения:

$$\langle C_\Delta, \mathcal{R} \rangle = \{R_j, R_k \in \mathcal{R} \mid C(R_j, R_k)_u \geqslant \Delta\}, \tag{3.23}$$

$$\langle D_\Delta, \mathcal{R} \rangle = \{R_j, R_k \in \mathcal{R} \mid D(R_j, R_k)_u \geqslant \Delta\}. \tag{3.24}$$

Графы отношений сходства C_Δ для восьми описаний фауны Курильской гряды приведены на рис. 3.4.

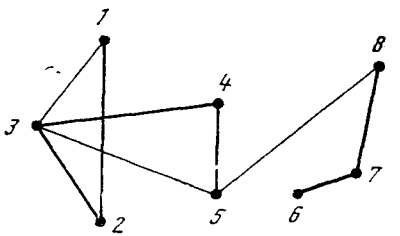
Поскольку матрица отношений сходства симметрична, то соответствующий ей орграф должен иметь все ребра взаимно ориентированными, и поэтому орграф заменяют графом.

При $\Delta = 51$ граф не связный и распадается на две компоненты: H_1 с вершинами *1—5* (описания южных районов) и H_2 с вершинами *6—8* (описания северных районов). При $\Delta = 50$ добавляются три новых ребра и граф становится связным. Вершина *5* является точкой сочленения двух компонент, и это характеризует соответствующий район (океанское побережье о-ва Итуруп) как промежуточный или граничный между двумя подобластями, из

которых южную О. Кусакин [60] называет низкобореальной Айнской, а северную — высокобореальной Берингийской.

Вывод о том, что граница между двумя подобластями проходит через океанское побережье о-ва Итуруп, подтверждается также анализом орграфа отношений «банальности B_{60}» (рис. 3.5).

Из вершины *5* выходит наибольшее число стрелок, и соответствующее ей описание является самым «банальным», причем отношение «банальности» для него выполняется с описаниями обеих подобластей.



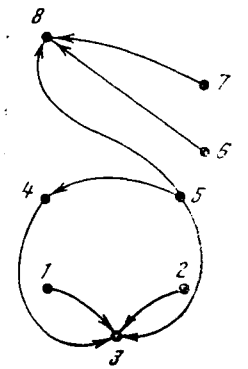


Рис. 3.4. Графы отношений сходства C_{51} (жирные линии) и C_{50} (тонкие линии) на множестве описаний фауны литорали Курильской гряды

Рис. 3.5. Орграф отношений «банальности» на множестве описаний макрофауны литорали Курильской гряды

Примечательно, что в R_5 встречено 158 видов, а в соседних северных районах 118 и 125 видов, т. е. видовой состав океанского побережья о-ва Итуруп не является самым бедным, а именно «банальным», неспецифичным. В противоположность этому самыми специфичными оказались описания R_3 (о-в Шикотан) и R_8 (о-в Парамушир). Интересно, что R_8 не является ни самым «банальным», ни самым бедным: литоральная фауна самого северного острова специфична по сравнению с некоторыми южными районами.

3.5. Отношение иерархии

На примере графа отношений C_{51} (рис. 3.4) можно проиллюстрировать результат разбиения множества $\mathcal{R}$ на два подмножества: $H_1 = \{R_1, \ldots, R_5\}$ и $H_2 = \{R_6, R_7, R_8\}$.

В общем случае под разбиением $\mathcal{H}$ множества $\mathcal{R}$ по отношению A будем понимать представление $\mathcal{R}$ в виде совокупности непустых подмножеств H_k, $k = 1, 2, \ldots, n$, таких, что

$$H_e \cap H_k = \phi, \quad e \neq k, \quad \bigcup_{k=1}^{n} H_k = \mathcal{R}. \tag{3.25}$$

Подмножества H_k будем называть классами.

Если задано разбиение, то элементы, входящие в один и тот же класс, называются эквивалентными (неразличимыми). Поясним, почему в обсуждаемом примере элементы из H_1 или H_2 эквивалентны: в каждом из них рассматривается отношение «быть связным». Граф на рис. 3.4 наглядно иллюстрирует, что в каждой компоненте любая вершина достижима из любой другой, следовательно, между этими вершинами существует путь, а это и означает, что они связны. Нетрудно убедиться, что отношение «быть связным» для вершин связного графа рефлексивно, симметрично и транзитивно. Отношения, обладающие всеми этими тремя свойствами, называются эквивалентностью.

Итак, эквивалентность порождает разбиение, и верно обратное: всякое разбиение устанавливает отношение эквивалентности.

Рассмотрим далее отображение некоторого множества $\mathcal{H}^{(1)}$ в множество $\mathcal{H}^{(2)}$, из которых последнее образовано соединением некоторых классов из $\mathcal{H}^{(1)}$. Отображение f: $\mathcal{H}^{(1)} \to \mathcal{H}^{(2)}$ сюръективно: каждому элементу из $\mathcal{H}^{(2)}$ соответствует хотя бы один элемент из $\mathcal{H}^{(1)}$. Соединение классов есть также класс, более широкий по сравнению с исходным. То обстоятельство, что $H^{(2)}$ является классом более широким, чем $H^{(1)}$, отобразим как $H^{(2)}$ И $H^{(1)}$, а отношение И назовем отношением иерархии (подчинения). Так что эта же запись может читаться как «$H^{(2)}$ подчиняет $H^{(1)}$» или «$H^{(1)}$ подчинен $H^{(2)}$».

Множество $\mathcal{H}^{(2)}$ назовем сгущением $\mathcal{H}^{(1)}$, если хотя бы один из классов $\mathcal{H}^{(2)}$ есть соединение классов из $\mathcal{H}^{(1)}$.

Пусть далее

$$\mathcal{U} = \{\mathcal{H}^{(1)}, \ldots, \mathcal{H}^{(s)}\} \tag{3.26}$$

есть множество разбиений, таких, что $\mathcal{H}^{(k)}$ — сгущение $\mathcal{H}^{(k-1)}$, где $k \in K$,

$$K = \{k \mid k \text{ — целое число, } 1 \leqslant k \leqslant s\}.$$

Тогда в предельном случае $\mathcal{H}^{(1)}$ состоит из всех классов, содержащих ровно по одному элементу, а $\mathcal{H}^{(s)}$ — из одного класса, совпадающего с $\mathcal{R}$.

Множество $\mathcal{U}$ есть иерархическая система, состоящая из s уровней. Номеру каждого уровня можно сопоставить его ранг (так как K — упорядоченное множество), а название всех классов одного ранга считать категорией.

При практических построениях иерархических классификаций удобно использовать различные приемы конструирования дендрограмм [24, 63] — графического способа изображения системы (3.26). Последовательный процесс образования сгущений начинается с рассмотрения q объектов, принадлежащих разбиению $\mathcal{H}^{(1)}$. Иначе говоря, на первом шаге каждый объект из заданного множества считается классом. Затем два наиболее схожих объекта объединяются в один класс и общее число последних становится равным $q - 1$. Эти классы принадлежат разбиению $\mathcal{H}^{(2)}$, являющемуся сгущением $\mathcal{H}^{(1)}$. Если имеется n одинаково схожих объектов, то объединяются любые два из них. Среди оставшихся снова отыскиваются наиболее схожие, которые объединяются, и так действуем до тех пор, пока все объекты не попадут в один класс $\mathcal{H}^{(3)}$.

Существует много способов построения дендрограммы, из которых рассмотрим простейшие, основанные на использовании матрицы мер сходства.

Воспользуемся данными табл. 3.7 и отыщем в ней наибольшее значение. Таковым оказывается $C\,(R_5, R_4)_0 = 60$, значит, наиболее схожими являются описания R_5 и R_4, которые и объединяем в один класс $H_4^{(1)} = \{R_4, R_5\}$. Условимся подстрочный индекс для обозначения более широкого класса выбирать минимальным на множестве индексов входящих в него классов.

Теперь определим сходство этого нового класса со всеми остальными как попарные расстояния между двумя множествами. Для этого выписываем отдельно элементы четвертой строки и четвертого столбца, а также элементы пятой строки и пятого столбца. Получаем

$C(R_j, R_4)$:	44	51	57	·	60	44	43	46
$C(R_j, R_5)$:	34	47	50	60	·	49	40	50.

Здесь точками отмечены диагональные элементы. Сравнивая поэлементно оба массива, отбираем каждый раз элемент с наибольшим значением и формируем из них новый массив (элементы в квадратике опускаем):

$$C(R_j, R_{4\cdot5}): 44 \quad 51 \quad 57 \quad \cdot \quad 49 \quad 43 \quad 50.$$

Этот массив и вписываем на место четвертой строки и четвертого столбца, а элементы пятой строки и пятого столбца вычеркиваем. В результате получаем новую матрицу

·					
51	·				
50	57	·			
44	51	57	·		
25	29	32	49	·	
20	23	31	43	53	·
24	31	40	50	48	52

В ней наибольшие значения имеют элементы $C\,(R_2, R_3)$ $C\,(R_{4\cdot5}, R_5)$. Для определенности выбираем первый из них

а вторые строку и столбец объединяем с третьими строкой и столбцом:

$C(R_j, R_2)$: 51	•	57	51	29	23	31
$C(R_j, R_3)$: 50	57	•	57	32	31	40
$C(R_j, R_{2\cdot3})$: 51		•	57	32	31	40

Новая матрица имеет вид

•					
51	•				
44	57	•			
25	32	49	•		
20	31	43	53	•	
24	40	50	48	52	•

В ней наибольший элемент $C(R_{4\cdot5}, R_{2\cdot3}) = 57$. Получаем новое название второй строки: $R_{4\cdot5\cdot2\cdot3}$ и т. д.

Выпишем результаты каждого шага в виде перечисления индексов в том порядке, в каком объединялись классы, а также уровни сходства, на которых это объединение происходило:

$$4\cdot5 - 60$$
$$2\cdot3 - 57$$
$$4\cdot5\cdot2\cdot3 - 57$$
$$6\cdot7 - 53$$
$$6\cdot7\cdot8 - 52$$
$$4\cdot5\cdot2\cdot3\cdot1 - 51$$
$$4\cdot5\cdot2\cdot3\cdot1\cdot6\cdot7\cdot8 - 50$$

Эти результаты используются для построения дендрограммы следующим образом.

Проведем вертикальную линию на плоскости и проградуируем ее в пределах изменения уровней сходства, в данном случае от 100 до 50 (рис. 3.6). На уровне 100 расположим по горизонтали восемь точек и пронумеруем их в том порядке, который указан в последней строке вспомогательной таблицы. Далее, следуя этой таблице, соединим точки по вертикали: первыми соединяются вершины *4* и *5*, причем ордината точки сочленения равна 60, а абсцисса — середине расстояния между этими точками. Прочие построения довольно очевидны.

Дендрограмма делает наглядной структуру иерархической системы сходства. В данном примере наибольшим сходством обладают R_4 и R_5, наименьшим — классы $H_1^{(8)} = \{R_4, R_5, R_3, R_2, R_1\}$ и $H_5^{(8)} = \{R_6, R_7, R_8\}$. Вершина R_5, как можно видеть, попадает в один класс с описаниями низкобореальной подобласти.

Отметим одно немаловажное обстоятельство, касающееся способа определения сходства каждого вновь образованного класса со всеми остальными. В описываемом примере это был просто выбор максимального значения из двух, соответствующих новому объединению, но иногда более полезными оказываются иные приемы.

В обзоре [24] приводятся шесть наиболее употребительных методов, которые описываются единой формулой:

$$G(H_j, H_k) = \alpha_u G(H_j, H_u) + \alpha_e G(H_j, H_e) + \beta G(H_u, H_e) + \gamma [G(H_j, H_u) - G(H_j, H_e)], \quad (3.27)$$

где $G(H_j, H_k)$ означает меру сходства или различия классов H_j и $H_k = \{H_u, H_e\}$.

Параметры α_u, α_e, β, γ задают вид процесса:

1) минимальные значения: $\alpha_u = \alpha_e = 1/2$, $\beta = 0$, $\gamma = -1/2$;
2) максимальные значения: $\alpha_u = \alpha_e = 1/2$, $\beta = 0$, $\gamma = 1/2$;
3) медиана: $\alpha_u = \alpha_e = 1/2$, $\beta = 0$, $\gamma = 0$;
4) среднее группы: $\alpha_u = n_u/n_k$, $\alpha_e = n_e/n_k$, $\beta = \gamma = 0$;
5) центроидный метод: $\alpha_u = n_u/n_k$, $\alpha_e = n_e/n_k$, $\beta = -\alpha_u\alpha_e, \gamma = 0$;
6) метод Уорда:

$$\alpha_u = \frac{n_j + n_u}{n_j + n_k}, \quad \alpha_e = \frac{n_j + n_e}{n_j + n_k}, \quad \beta = -\frac{n_j}{n_j + n_k}, \quad \gamma = 0.$$

Самый предпочтительный метод, по нашему мнению,— метод средней

$$G(H_j, H_k) = \frac{n_u}{n_k} G(H_j, H_u) + \frac{n_e}{n_k} G(H_j, H_e),$$

где n_u, n_k — число объектов соответственно u-го и k-го классов; $n_k = n_u + n_e$. Из формулы видно, что вместо выбора максимального значения, как было показано в примере, из элементов двух сравниваемых строк $C(R_j, R_u)$, $C(R_j, R_e)$ находим средневзвешенное значение.

Вид дендрограммы существенно зависит от выбора мер сходства и метода кластеризации, из которых наиболее сильное влияние оказывает мера сходства.

Чтобы проиллюстрировать это влияние, используем данные матрицы мер пересечения (см. табл. 3.6) для построения дендрограммы на основе неэквивалентной (3.16) меры:

$$D(R_j, R_k) = m(R_j) + m(R_k) - 2m(R_j \cap R_k) \quad (3.28)$$

и метода медианы (рис. 3.7).

Мера (3.28) численно совпадает с (2.14). В отличие от ранее использованной (3.16) она характеризует различие, однако все коэквивалентные ей меры сходства неэквивалентны (3.16).

Дело в том, что интуитивный смысл, вкладываемый в понятие сходства, в обоих случаях существенно различается. Если в первом из них наиболее сходными объектами являются имеющие

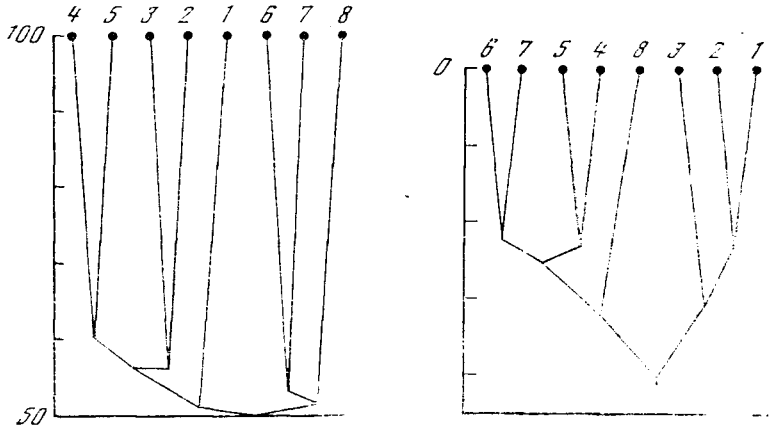


Рис. 3.6. Дендрограмма сходства описаний фауны литорали Курильской гряды

Рис. 3.7. Дендрограмма различия описаний фауны литорали Курильских островов

наибольшее число совпадений по присутствующим признакам, то во втором — имеющие наибольшее число совпадений одновременно по присутствующим и отсутствующим признакам.

Предположим, что анализируются два описания географических пунктов, в которых не встречено ни одного из учитываемых видов. Тогда в первом случае коэффициент сходства — неопределенность вида $^0/_0$, во втором — сходство стопроцентно. Обе меры не лишены практического смысла, однако каждая из них характеризует различные свойства одного и того же материала. При подробном анализе нелишне использовать несколько неэквивалентных мер, выбирая только легко интерпретируемые.

Применение меры (3.28) в данном примере принесло неожиданные результаты: описания R_4 и R_5 относятся к высокобореальной Берингийской подобласти. Однако эта неожиданность легко объясняется: соответствующие географические пункты по составу фауны схожи с низкобореальной, а по ее бедности — с высокобореальной подобластью.

Уместно отметить, что при четких различиях неэквивалентные меры дают качественно одни и те же результаты, а при нечетких — описания «промежуточного» характера вносят противоречия. Это и наблюдается в данном примере.

3.6. Замечания о формализации задачи классификационных построений в зоологической систематике

Прежде чем переходить к обсуждению практических приложений обсуждаемых методов в зоологической систематике, следует заметить, что фундаментальные понятия этой области знания не установились и сильно различаются у сторонников новых и традиционных направлений. Это создает известные трудности в формулировании целей и трактовке результатов конкретных исследований. Кроме того, все известные нам подходы страдают тем недостатком, что в них не оговаривается с достаточной полнотой необходимость соблюдения принципа соразмерности между строгостью выводов и строгостью посылок. Часто использование «сложных» формул создает иллюзии объективности и точности результатов, хотя эти качества теряются, как только строгие выводы переносятся на ситуации с нестрого определенными понятиями.

В одной из самых популярных в СССР работ традиционного направления [35] понятие классификации вводится по Симпсону: «зоологическая классификация — это упорядочение * животных в группы (или серии) на основании их взаимоотношений» и тут же добавляется, что процесс классификации сильно отличается от процесса определения (идентификации). Такого же мнения придерживаются Снит и Сокал [62]. Мы считаем такое понимание недостаточно корректным, так как в практической деятельности зоологи часто оперируют не самими животными, а их описаниями (моделями) и описаниями не всех животных, а только определенным образом выбранных из общей совокупности. Множество реальных объектов затем отображается в множество выделенных классов на основе «узнавания» — идентификации.

Процесс сортировки выборочных единиц по классам заключается иногда в том, что любая единица из этой выборки сопоставляется с каждым из заранее сформированных классов (если она берется не первой) и либо относится к одному из них, либо образует новый класс. Так что идентификация в этом процессе выполняется многократно. Она является средством пополнения наших знаний о классах и составной частью процесса классификации.

Подобная же неоднозначность вытекает из определения «таксона», «таксономии», «категории» и др. Многие недоразумения возникают из-за совмещения двух разных понятий: реальный организм и его описание — в одном: «операциональная таксономическая единица». В связи с этим мы, как и ранее, вновь вводимые термины будем по возможности тщательно пояснять.

Если на основе ранее введенных понятий под объектом понимать организм (вместо географического пункта), а под при-

* В цитируемой работе — «разделение», хотя Снит и Сокал, цитируя Симпсона, употребляют слово «ordering».

знаком — свойство или состояние организма (вместо учитываемых видов), то все ранее использованные процедуры и правила становятся справедливыми и для классификационных построений в зоологической систематике. Поэтому нет необходимости повторять введенные выше определения, а также изменять обозначения. Вместо этого мы используем конкретные материалы главным образом для дальнейшего изучения структуры систем-классификаций.

Глава 4

КЛАССИФИКАЦИИ, ОСНОВАННЫЕ НА СМЕШАННЫХ И КОЛИЧЕСТВЕННЫХ ПРИЗНАКАХ

4.1. Некоторые трудности

Во многих задачах встречается сочетание как качественных, так и количественных сведений. К сожалению, в настоящее время не разработаны подходы к комплексному анализу данных, поэтому в таких случаях необходим переход к одному типу, числовому или качественному. Эта процедура требует известной осторожности, и ее следует выполнять с учетом того, какие преобразования допустимы для той или иной шкалы, а также какие меры предполагается использовать в конкретной задаче.

Если анализ осуществляется с помощью сведения всех показателей к количественным, то в качественные оценки привносится дополнительная, искажающая информация. Р. Дуда и П. Харт пишут по этому поводу: «Как нужно обрабатывать векторы, чьи компоненты содержат смесь из номинальных, порядковых, интервальных и относительных шкал? В конечном счете нет методологического ответа на эти вопросы. Когда пользователь выбирает некоторую функцию подобия или нормирует свои данные каким-либо конкретным методом, он вводит информацию, которая задает процедуру» [23, с. 238].

Примером перехода от комплексных к количественным переменным является использование метрики:

$$d(R_j, R_k) = \left[\sum_m \left(\frac{x_{mj} - x_{mk}}{\sigma_m}\right)^2\right]^{1/2}, \qquad (4.1)$$

где x_{mj} — значение m-го признака у j-го объекта;

$$\sigma_m^2 = \frac{1}{q-1}\left[\sum x_{mk}^2 - \frac{1}{q}\left(\sum x_{mk}\right)^2\right];$$

σ_m — стандартное отклонение m-го признака.

При другом подходе комплексная обработка производится с помощью сведения числовых показателей к качественному виду. При этом часть информации теряется, что может быть еще более нежелательным, чем введение дополнительной информации. Однако если выводы, полученные на основе количественной обработки данных, совпадут с выводами качественной обработки, то с большой долей уверенности можно утверждать, что они действительно основаны на исходных данных, а не на методе их извлечения [39].

Обычно при переходе к качественным признакам прибегают к следующим преобразованиям: весь диапазон значений количественного признака разбивается на некоторое число градаций и каждой градации ставится в соответствие один разряд двоичного числа. Если признак принял значение какой-то градации, то этому разряду приписывается значение «1».

Например, если диапазон значений признака «длина» укладывается в интервале от 5 до 11 см, то данный интервал можно разбить, скажем, на три градации: от 5 до 7, от 7 до 9 и от 9 до 11 см. Эти градации и следует считать новыми качественными признаками, и если исследуемый конкретный объект имеет длину 6 см, то его описание 1 0 0, а если 7 см, то 0 1 0 и т. д.

Такое представление чисел не сохраняет линейную упорядоченность, т. е. теряется информация о мере «близости» различных качественных состояний. Чтобы устранить этот эффект (в тех случаях, когда он нежелателен), прибегают к следующему приему: если признак принимает значение k-й градации, то все разряды двоичного числа, предшествующие k-му, также принимают значение 1. В таком случае объекту с размерами 6 см нужно сопоставить описание 1 0 0, а с длиной 7 см — 1 1 0.

Требование линейной упорядоченности множества значений особенно важно для признаков, измеренных по шкале порядка.

Примеры использования смешанных признаков приводятся в следующем параграфе.

4.2. Алгебра логики как средство прогнозирования (распознавания)

Применение формул алгебры логики для целей распознавания покажем сначала на некотором условном примере.

Из отчета промысловой разведки следовало, что в тех местах, где температура воды была сравнительно низкой или где встречались поля фитопланктона, косяки сельди не встречались. Но в тех местах, где замечались поля зоопланктона, всегда встречались и косяки сельди.

На основе этой информации требуется установить:

1. Можно ли сделать вывод о том, что в тех местах, где нет одновременно ни фитопланктона, ни зоопланктона, не стоит ожидать косяков сельди?

2. Следует ли из отчета, что фитопланктон встречается только в холодной воде?

3. Можно ли полагать, что там, где встречается зоопланктон, фитопланктон не может быть встречен?

4. Как изменятся ответы, если в ответ дополнительно внести фразу: «Где не было зоопланктона, там сельди не было?»

На этом элементарном примере мы попытаемся последовательно раскрыть всю технику а) построения прогнозирующего аппарата и б) осуществления собственно прогноза.

Для решения задачи а) необходимо исходное сообщение (текст) формализовать (эксплицировать), т. е. переложить на язык алгебры логики, а затем всю информацию отчета записать в виде единого ТИ-высказывания.

Процесс экспликации начинается с установления всех объектов, фигурирующих в системе и важных для осуществления прогноза. Нетрудно видеть, что таких объектов четыре: температура воды (Х), поля фитопланктона (Ф) и зоопланктона (З), косяки сельди (С).

Относительно этих объектов, следуя тексту отчета, можно построить следующие высказывания:

если (вода холодная (Х) или наблюдались поля фитопланктона (Ф), то (косяки сельди не встречались ($\overline{С}$));

если (встречались поля зоопланктона (З)), то (сельдь всегда встречалась (С)).

Оба высказывания — импликации («если..., то...»). Первое высказывание сложное и состоит из двух простых, соединенных связкой «или». В принятых обозначениях можно записать

$$\begin{aligned} &Х + Ф \to \overline{С} \quad (\text{"если Х или Ф, то } \overline{С}\text{"}) \\ &З \to С \quad (\text{"если З, то С"}). \end{aligned} \tag{4.2}$$

При записи импликации, скажем $A \to B$ («если A, то B»), всегда необходимо проверять, не следует ли из содержания текста, что и $B \to A$. Если это так, то $A = B$.

Следующим шагом построения прогнозирующей системы является переход от импликаций к ТИ-высказываниям. На основе правила 18 (см. с. 24)

$$\overline{Х + Ф} + \overline{С} = 1,$$
$$\overline{З} + С = 1$$

или на основе законов Де Моргана

$$\begin{aligned} &\overline{Х}\cdot\overline{Ф} + \overline{С} = 1, \\ &\overline{З} + С = 1. \end{aligned} \tag{4.3}$$

Теперь заметим, что поскольку обе формулы — ТИ-высказывания, то при логическом умножении их левых частей получим также ТИ-

высказывание. Умножение ведем по следующей схеме:

$$\frac{\begin{array}{l}\overline{Х}\cdot\overline{Ф} + \overline{С} \\ \overline{З} + С\end{array}}{\overline{Х}\cdot\overline{Ф}\cdot\overline{З} + \overline{З}\cdot\overline{С} + С\cdot\overline{Х}\cdot\overline{Ф} + С\cdot\overline{С}},$$

т. е. правые части опускаем, а левые перемножаем по правилам обычной алгебры. Обращаясь к правилам 1—20 (с. 23—24), проверяем, нельзя ли упростить ответ.

Ранее указывалось, что $A\cdot\overline{A} = 0$, аналогично $С\cdot\overline{С} = 0$, и этот член опускаем из результата. Оставшаяся часть не поддается дальнейшему упрощению. Следовательно, всю информацию, содержащуюся в тексте, можно записать формулой

$$\overline{Х}\cdot\overline{Ф}\cdot\overline{З} + \overline{З}\cdot\overline{С} + \overline{Х}\cdot\overline{Ф}\cdot С = 1. \tag{4.5}$$

Этим закончено построение прогнозирующей системы, которая считается готовой для осуществления собственно прогноза относительно любой переменной, фигурирующей в схеме высказываний.

Переходим к ответам на вопросы. Формализуем первый из них. Он задает ситуацию: $\overline{Ф}\cdot\overline{З} = 1$, т. е. такую, в которой нет ни фито-, ни зоопланктона, а относительно прочих переменных, в частности С, ничего неизвестно. Требуется узнать, что можно сказать о заданной ситуации на основе всей имеющейся информации (4.5). Выражение

$$\overline{Ф}\cdot\overline{З} = 1 \tag{4.6}$$

есть ТИ-высказывание, поскольку предполагается, что такая ситуация в действительности (в момент прогноза) имеет место. Для ответа перемножим левые части (4.5) и (4.6):

$$\frac{\begin{array}{l}\overline{Х}\cdot\overline{Ф}\cdot\overline{З} + \overline{З}\cdot\overline{С} + \overline{Х}\cdot\overline{Ф}\cdot С \\ \overline{Ф}\cdot\overline{З}\end{array}}{\overline{Х}\cdot\overline{Ф}\cdot\overline{Ф}\cdot\overline{З}\cdot\overline{З} + \overline{Ф}\cdot\overline{З}\cdot\overline{З}\cdot\overline{С} + \overline{Х}\cdot\overline{Ф}\cdot\overline{Ф}\cdot\overline{З}\cdot С}.$$

На основе правила 2 уберем двойные обозначения, получим

$$\overline{Х}\cdot\overline{Ф}\cdot\overline{З} + \overline{Ф}\cdot\overline{З}\cdot\overline{С} + \overline{Х}\cdot\overline{Ф}\cdot\overline{З}\cdot С = 1.$$

Третье слагаемое содержит все члены первого, поэтому согласно закону «поглощения»

$$\overline{Х}\cdot\overline{Ф}\cdot\overline{З} + \overline{Ф}\cdot\overline{З}\cdot\overline{С} = 1. \tag{4.7}$$

Дальнейшее упрощение невозможно. Дешифруем ответ. Поскольку ситуация $\overline{Ф}\cdot\overline{З} = 1$ задана, то он будет выглядеть так:

$$\overline{Ф}\cdot\overline{З} \to \overline{Х} + \overline{С},$$

т. е. из (4.7) множитель $\bar{Ф}\cdot\bar{З}$ выносится в левую часть импликации. На основе правила 19

$$\bar{Ф}\cdot\bar{З} \to \bar{Х} + Х\cdot\bar{С}.$$

Итак, в заданной ситуации следует ожидать либо нехолодную воду и относительно сельди ничего нельзя сказать (она может быть или не быть), или косяки сельди не будут встречены, но тогда вода должна быть обязательно холодной. Ответ на первый вопрос отрицательный.

Для ответа на второй вопрос задаемся ситуацией $Ф = 1$ и перемножаем левые части этой формулы и (4.5):

$$\frac{\bar{Х}\cdot\bar{Ф}\cdot\bar{З} + \bar{З}\cdot\bar{С} + \bar{Х}\cdot\bar{Ф}\cdot С}{Ф}$$

$$\frac{}{0 + Ф\cdot\bar{З}\cdot\bar{С} + 0}.$$

Первое и третье слагаемые ложны ($Ф\cdot\bar{Ф} = 0$), поэтому ответ

$$Ф \to \bar{З}\cdot\bar{С},$$

т. е. в заданной ситуации не следует ожидать ни зоопланктона, ни сельди, а относительно температуры воды ничего нельзя сказать. Ответ на второй вопрос отрицательный.

В третьем вопросе задана ситуация $З = 1$; перемножив это выражение с (4.5), получим: $З \to \bar{Х}\cdot\bar{Ф}\cdot С$. Ответ утвердительный.

Четвертый вопрос требует введения в (4.5) дополнительной информации: $\bar{З} \to \bar{С} = (З + \bar{С} = 1)$. Перемножив это выражение с (4.5), получим для содержания всего отчета

$$\bar{Х}\cdot\bar{Ф}\cdot З\cdot С + \bar{З}\cdot\bar{С} = 1. \qquad (4.8)$$

Поступая, как и ранее, устанавливаем

$$\bar{Ф}\cdot\bar{З} \to \bar{С},\quad Ф \to \bar{З}\cdot\bar{С},\quad З \to \bar{Х}\cdot\bar{Ф}\cdot С.$$

На первый и третий вопросы ответы утвердительные.

На данном примере показано, что в развиваемом подходе процесс прогнозирования распадается на два этапа: создание прогнозирующей системы (этап «обучения») и осуществление собственно прогноза. Подход позволяет проводить «дообучение» системы каждый раз, как появляется новая информация или осуществляемый прогноз оказывается неверным (см. ниже). Оба этапа распадаются на несколько шагов.

Продемонстрируем их выполнение на более сложном примере, заимствованном из практики медико-географических исследований. Речь идет о прогнозировании численности массовых видов клещей — переносчиков вируса клещевого энцефалита в Приморье.

Шаг 1. Формирование списка переменных (элементов системы). Для прогнозирования численности массовых видов клещей в Приморье важно учитывать следующие переменные (табл. 4.1).

Таблица 4.1

Список переменных (упрощенный вариант)

Название признака	Обозначение признака	Градации
Виды деревьев:		
1. Светлохвойные	К	Есть — нет
2. Темнохвойные	Т	»
3. Широколиственные	Ш	»
4. Мелколиственные	М	»
5. Дубняки	Д	»
6—9. Густота древостоя	$П_i$	Большая — малая
10. Пятнистость леса	$п$	»
11. Кустарники лесные	Л	Есть — нет
12. Кустарники открытых мест	Ө	»
13—16. Травянистый покров	$Тр_i$	Редкий, средний, густой
17—20. Влажность	W_i	Низкая, средняя, высокая
21—23. Характер рельефа	g_i	Долина, склон, плакор
24—27. Удаленность от моря, км	S_i	50, 150, 150
28—31. Высота над уровнем моря, м	h_i	200, 200—700, 700—1000
32—35. Мелкие млекопитающие, на 100 л особей на 100 ловушек за сутки	$Гр_i$	10, 10—30, 30
36—39. Птицы, число особей на 1 км^2	$Пт_i$	200, 200—600, 600
40—43. Средние млекопитающие, численность	$Ср_i$	Мало, средне, много
44—47. Крупные млекопитающие, численность	$Кр_i$	Мало, средне, много
Численность клещей		
48—51. *Ixodes persulcatus*	p_i	200, 200—600, 600
52—55. *Haemophysalis Japonica*	l_i	100, 100—300, 300
56—59. *H. concinna*	c_i	100, 100—400, 400
60—63. *Dermacentos silvarum*	S_{v_i}	100, 100—400, 400

В зависимости от конкретной задачи физический состав переменных может быть самым различным, число градаций каждого признака также устанавливается в зависимости от конкретной цели и может быть самым различным. В данном примере приводится лишь сокращенный вариант прогнозирующей системы, разрабатываемой нами совместно с сотрудниками Лаборатории медицинской географии ТИГ ДВНЦ АН СССР.

Шаг 2. Формализация исходных сведений о связях между переменными. В отношении *I. persulcatus* можно определить условия, в которых этот вид в Приморье а) не достигает высокой численности, б) не достигает даже средней градации, в) практически не встречается.

На обычном языке условия типа «а» характеризуются следующим образом: «лес не содержит ни светло-, ни темнохвойных, ни широколиственных видов деревьев; дубняки или широколиственные леса с несильной густотой; мелколиственные леса; слабая густота леса; близость моря или низкая влажность».

Условия типа «б»: «или дубняки, или мелколиственные, или широколиственные виды деревьев, образующих редколесье; пятнистые дубняки со средней густотой или пятнистые мелколиственные леса со средней густотой; близко или рядом море; дубняки с несильной густотой».

Условия типа «в»: «нет одновременно ни грызунов, ни птиц, или: ни средних, ни крупных прокормителей; сухо или нет леса».

При переходе к языку алгебры логики можно сразу отметить, что каждый тип условий составляет импликацию, а именно а $\rightarrow \bar{p}_3$, б $\rightarrow p_0 + p_1$, в $\rightarrow p_0$. На основе принятых обозначений запишем содержание импликаций как

$$\overline{\text{Т}}\cdot\overline{\text{Ш}}\cdot\overline{\text{К}} + (\text{Д} + \text{Ш}) + \overline{\Pi}_3 + \text{М} + \Pi_1 + S_1 + W_1 \rightarrow \bar{p}_3, \quad (4.9)$$

$$(\text{Д} + \text{М} + \text{Ш})\cdot\pi\,\Pi_1 + (\text{Д} + \text{М})\cdot\pi\cdot\Pi_2 + S_0 + S_1 + \text{Д}\cdot\overline{\Pi}_3 \rightarrow$$
$$\rightarrow p_0 + p_1,$$

$$\text{Гр}_0\text{Пт}_0 + \text{Ср}_0\text{Кр}_0 + W_0 + \Pi_0 \rightarrow p_0.$$

Как правило, исходная информация записывается в виде импликаций, но не обязательно. Например, в отношении переменной p_i можно утверждать, что $p_0\cdot p_1 = 0$, $p_0\cdot p_2 = 0$, $p_0\cdot p_3 = 0$ и т. д., т. е. элементы p_i попарно несовместны. Кроме того, всегда имеется какое-либо одно состояние: $p_0 + p_1 + p_2 + p_3 = 1$. Так что

$$p_0\cdot\bar{p}_1\cdot\bar{p}_2\cdot\bar{p}_3 + \bar{p}_0\cdot p_1\cdot\bar{p}_2\cdot\bar{p}_3 + \bar{p}_0\cdot\bar{p}_1\cdot p_2\cdot\bar{p}_3 + \bar{p}_0\cdot\bar{p}_1\bar{p}_2\cdot p_3 = 1. \quad (4.10)$$

Напомним, что при составлении импликаций $A \rightarrow B$ необходимо проверять, соблюдается ли и $B \rightarrow A$. Если это так, то $B \rightarrow A$ добавляется к исходной информации. Условия типа (4.10) можно и не добавлять: они легко запоминаются, поэтому ошибок можно избежать. При машинном счете их можно предусмотреть при расшифровке ответов.

Шаг 3. Переход к форме ТИ-высказываний. Используя правила 1—20, исходную информацию можно записать как

1) $\text{Т} + \text{К} + \text{Ш} + \bar{p}_3$,
2) $\overline{\text{М}}\cdot\overline{\Pi}_1\cdot\overline{S}_1\cdot\overline{W}_1 + \bar{p}_3$,
3) $\overline{\text{Д}}\cdot\overline{\text{Ш}} + \Pi_3 + \bar{p}_3$,
4) $\overline{\text{Д}}\cdot\overline{\text{М}}\cdot\overline{\text{Ш}} + \overline{\pi} + \overline{\Pi}_1 + p_0 + p_1$,
5) $\overline{\text{Д}}\cdot\overline{\text{Ш}} + \overline{\Pi}_2 + \pi + p_0 + p_1$,
6) $\overline{\text{Д}} + \Pi_3 + p_0 + p_1$,
7) $\overline{S}_0\cdot\overline{S}_1 + p_0 + p_1$,
8) $\overline{\text{Гр}_0} + \overline{\text{Пт}_0} + p_0$,
9) $\overline{\text{Ср}_0} + \overline{\text{Кр}_0} + p_0$,
10) $\overline{W}_0\cdot\overline{\Pi}_0 + p_0$.

Правая часть каждого выражения стереотипна (=1), поэтому опущена. Кроме того, число высказываний увеличено по сравнению с числом исходных в соответствии с требованием приемлемой простоты каждого из них. А это увеличение становится возможным потому, что высказывание $A + B \rightarrow C$ равносильно $A \rightarrow C$, $B \rightarrow C$.

Этим шагом заканчивается этап «обучения», и система считается подготовленной к осуществлению прогноза.

Шаг 4. Осуществление прогноза. Допустим, что некоторый географический пункт Приморья характеризуется следующими условиями: «лес широколиственный, пятнистый, светлохвойные, темнохвойные, мелколиственные виды деревьев и дубняки отсутствуют; густота леса средняя, влажность большая, близко к морю, грызунов и птиц много, средних и крупных прокормителей мало». Требуется определить, какую численность *I. persulcatus* следует ожидать в заданных условиях.

Поскольку перечисленные условия считаются действительно существующими, то можно записать

$$\text{Ш}\cdot\pi\cdot\overline{\text{К}}\cdot\overline{\text{Т}}\cdot\overline{\text{М}}\cdot\overline{\text{Д}}\cdot\Pi_2\cdot W_3\cdot S_0\cdot\text{Гр}_3\cdot\text{Пт}_3\cdot\text{Ср}_1\cdot\text{Кр}_1 = 1.$$

Для осуществления прогноза левую часть (4.11) умножаем на каждое из 10 ТИ-высказываний, а полученные результаты снова перемножаем по правилам конъюнкции. Чтобы не переписывать каждый раз условия (4.11), обозначим их как X, тогда результаты перемножений будут

1) $0 + 0 + 0 + X + X\bar{p}_3 = X$,
2) $X + X\cdot\bar{p}_3 = X$,
3) $0 + 0 + X\cdot\bar{p}_0 = X\cdot\bar{p}_3$,
4) $0 + 0 + X + X\cdot(p_0 + p_1) = X$,
5) $X + 0 + 0 + X\cdot(p_0 + p_1) = X$,
6) $X + 0 + X\cdot(p_0 + p_1) = X$,
7) $X + X\cdot(p_0 + p_1) = X\cdot(p_0 + p_1)$,
8) $0 + X + X\cdot p_0 = X$,
9) $X + X + X\cdot p_0 = X$,
10) $X + X\cdot p_0 = X$.

Перемножение правых частей дает окончательно: $X\cdot(p_0 + p_1)\cdot\bar{p}_3$, или

$$X \rightarrow p_0 + p_1,$$

т. е. из истинности условий X следует истинность $p_0 + p_1$.

Итак, в заданной ситуации следует ожидать либо низкую, либо нулевую численность интересующего вида.

Шаг 5. «Дообучение» системы. В процессе исследований получена следующая дополнительная информация о численности массовых видов клещей Приморья:

H. japonica

1) $П_1 + Д \to \bar{j}_3$, 2) $W_0 + М + \pi \cdot (П_0 + П_1) \to j_0 + j_1$,

3) $П_0 + \theta + Т + Пт_0 \cdot Ср_0 + h_3 + Ср_0 \cdot Кр_0 \to j_0$;

H. concinna

1) $\bar{g} \to c_3$,

2) $h_2 + W_0 + W_1 + \bar{g}\,(Ш + К) + \pi \cdot П_1 + П_2 + c_0 + c_1$,

3) $Гр_0 \cdot Пт_0 \cdot Ср_0 + Пт_0 \cdot Ср_0 \cdot Кр_0 + К \cdot Ш \cdot \bar{g} + Т + Тр_0 + П_3 + h_3 + c_0$;

D. silvarum

1) $g + W_3 + \bar{\pi} + К + Т + (Ш + М + Д)\,(П_2 + П_3) + h_3 + h_2 + Sv_0 + Sv_1$,

2) $П_3 + Гр_0 + Кр_0 \to Sv_0$.

Приводя эти высказывания к форме ТИ и добавляя эту информацию к ранее имевшейся, снова получаем «обученную» систему, готовую к осуществлению прогноза в отношении любой переменной, введенной в рассмотрение.

Определим численность всех видов клещей в заданных условиях. В исходной информации связи между разными видами клещей не указаны, поэтому нет необходимости выражение (4.11) сопоставлять со всей информацией, достаточно найти конъюнкции для описаний каждого вида в отдельности. Поступая, как и ранее, получим

$$X \to j_0 + j_1 \cdot \bar{\theta} \cdot \bar{h}_3, \qquad X \to g + c_0 + c_1,$$

$$X \to Sv_0 + Sv_1.$$

Другими словами, в заданной ситуации следует ожидать, что численность *H. japonica* будет либо нулевой, либо низкой, но в последнем случае не должно быть одновременно ни кустов открытых мест, ни большой высоты над уровнем моря (эти признаки не задавались для прогноза). Численность *H. concinna* будет либо нулевой, либо малой, но если имеет место долина, то она может быть любой. Наконец, численность *D. silvarum* будет либо нулевой, либо малой.

Специальные исследования, предпринятые после составления прогноза на опытном участке (о-в Рейнеке), показали полное соответствие прогноза и действительности, хотя, по мнению специалистов, о-в Рейнеке не является типичным в смысле материала «обучения».

По описываемой схеме можно проводить прогноз в отношении любой переменной (или их комплекса), участвующей в рассмотрении. В заданной ситуации примера отсутствовали сведения о численности клещей, высоте над уровнем моря, наличии кустарников

и др. Если задавать еще менее определенные условия, то прогноз будет менее определенным и содержать несколько вариантов ответа, соединенных связкой «или». Эта неопределенность тем больше, чем меньше связей между всеми переменными в исходной информации. Пусть, например,

$$П_3 \to Тр_0 \cdot (Л + \theta).$$

Тогда в заданной ситуации информация о наличии кустарников и травянистого покрова избыточна, если в ней участвует переменная $П_3$.

Развиваемый подход требует от биологов и географов некоторых навыков в обращении с формальными логическими правилами, но зато позволяет учесть не только опытную (полевые и экспериментальные наблюдения) информацию, но также и сведения самого общего характера. Использование вычислительных машин при таком подходе не всегда обязательно.

4.3. Алгоритмы автоматического построения прогнозирующей системы и осуществления прогноза

В отличие от изложенного выше другой подход не требует от биологов и географов знаний математической логики, но позволяет учесть только экспериментальную информацию. Создание прогнозирующего аппарата и осуществление прогноза в этом случае проводится в вычислительном центре, а участие биологов и географов сводится к заполнению анкет в соответствии с краткой инструкцией.

Инструкция для составления анкеты

1. Устанавливаются цель исследования и список переменных, которые исследователь считает важными для достижения цели (см. шаг 1).

2. Вводится понятие «географический пункт» (масштаб), под которым для разных целей можно подразумевать, например, станции, острова, материки и т. п.

3. Для материала «обучения» отбираются те пункты, в которых могут быть измерены значения всех переменных. В анкету эти сведения заносятся в виде одного из трех символов: «0» — нет, «1» — есть, «*» — может быть и может не быть (безразличная характеристика).

Пусть, например, введены три признака: S_1, S_2, S_3 — и четыре географических пункта: а, б, с, в. Тогда возможный вариант анкеты

	а	б	с	в	е	ж
S_1	0	1	1	0	?	1
S_2	0	*	1	0	0	?
S_3	1	0	1	1	*	?

ополнительных столбцах «е» и «ж» указаны все известные усло-
в пунктах, для которых осуществляется прогноз. Прогнозируе-
переменные отмечаются в анкете (?).

Алгоритмы обработки анкеты на ЭВМ

1. Данные анкеты представляются двухмерным массивом A_{ij}, $= 1, 2, \ldots, q$, $i = 1, 2, \ldots, p$, p — число признаков, q — ло объектов.

2. Приведение данных к тупиковой форме. Каждый столбец j авнивают со столбцом k по правилам:

— если они полностью совпадают, то один из них отбрасы-ется;

— если k содержит все значения «0» или «1», совпадающие с j, k отбрасывается;

— если j и k не совпадают только в одном разряде и ни одно них не есть «*», то один из столбцов отбрасывается, а во втором совпадающее значение заменяется «*»;

— если j и k не совпадают только в одном разряде, а прочие ачения такие, что $k \to j$, то вместо несовпадающего разряда в k носится «*».

Процедура ведется до тех пор, пока дальнейшие упрощения анут невозможными. Полученная матрица сохраняется. Возмо-ен переход к разреженным матрицам, если «*» превалирует над рочими значениями.

3. Для осуществления прогноза столбец с заданной ситуацией равнивается со столбцами получившейся A_{ij}:

— если $\alpha = k$, то α попадает в таблицу результатов B_{ij}, столбец k вычеркивается;

— если $\alpha \to k$, то α попадает в B_{ij} и вычеркивается из A_{ij};

— если $k \to \alpha$, то k попадает в B_{ij} и вычеркивается из A_{ij}.

Процедура выполняется до тех пор, пока все столбцы A_{ij} не удут сопоставлены с α.

4. Если α не сохранилось, то выполняется 5, иначе произво-дится перемножение α и всех столбцов оставшейся A_{ij} по пра-вилам:

— если α и k не совпадают хотя бы в одном разряде, то их не перемножают и берется столбец $k + 1$; иначе их перемножают по правилам: $0 \cdot 0 = 0$, $0 \cdot * = 0$, $1 \cdot * = * \cdot 1 = 1$, $1 \cdot 1 = 1$, $* \cdot * = *$.

5. Оставшаяся A_{ij} дополняет B_{ij}. Для B_{ij} выполняется *2*, а затем 6.

6. Оставшаяся B_{ij} оформляется как ответ.

Дополнительная информация записывается в виде добавочных столбцов A_{ij} после того, как выполнено *2*, а затем, начиная с 2, все пункты выполняются заново. Неверный прогноз рассматривается как дополнительная информация.

Описание алгоритмов станет более понятным, если учесть, что каждый столбец анкеты — конъюнкция переменных, причем «1» отмечается значение, которое выполняется для данного объекта, «0» — которое не выполняется, «*» — переменная, которая не участвует в конъюнкции. Разные столбцы анкеты представляют дизъюнкции, а их совокупность — ТИ-высказывание, поскольку считается, что в анкету вошли все связи, необходимые для дальнейшего анализа (полная группа событий).

Данные анкеты примера можно записать следующим образом:

$$\bar{S}_1 \cdot \bar{S}_2 \cdot S_3 + S_1 \cdot \bar{S}_3 + S_1 \cdot S_2 \cdot S_3 + \bar{S}_1 \cdot \bar{S}_2 \cdot S_3 = 1. \qquad (4.11)$$

Приведение данных к тупиковой форме заключается: 1) в замене всех одинаковых столбцов (слагаемых) одним, 2) в выполнении закона поглощения для всех столбцов, 3) в выполнении правила: $A + \bar{A} \cdot B = A + B$ и т. д.

Дополнительная информация, так же как и неверный прогноз, позволяет считать, что данные анкеты не представляют полной группы событий, поэтому анкета и дополняется добавочными столбцами.

4.4. Дескриптивные множества и использование количественных данных

Б. И. Семкин [46] вводит понятие дескриптивных множеств и определяет меры пересечения и объединения двух множеств, R_1 и R_2:

$$m(R_1 \cap R_2) = \sum_k \min(x_{k1}, x_{k2}), \qquad (4.12)$$

$$m(R_1 \cup R_2) = \sum_k \max(x_{k1}, x_{k2}). \qquad (4.13)$$

Универсумом можно считать

$$m(\text{У}) = \sum_k \max(x_{k1}, \ldots, x_{kq}). \qquad (4.14)$$

При таком подходе многие введенные ранее соотношения становятся справедливыми для описаний, состоящих из количественных признаков. Пусть даны два описания:

R_1 : 60,0 20,5 80,0 18,0,

R_2 : 55,0 30,0 90,0 10,0

и нужно найти меру включения R_1 в R_2.

Используя (3.5) и (4.12), получим

$$W(R_1; R_2) = \frac{m(R_1 \cap R_2)}{m(R_2)} = \frac{\sum \min(x_{k1}, x_{k2})}{\sum x_{k2}} =$$

$$= \frac{55 + 20{,}5 + 80 + 10}{55 + 30 + 90 + 10} \approx 85\%.$$

Как можно заметить, в числителе подсчитывается сумма минимальных значений, а в знаменателе — сумма компонент одного из векторов.

Приведем конкретный пример из практики биогеографических исследований. В цитированной ранее работе [60] приводятся сведения, на основании которых мы составили следующую таблицу (табл. 4.2).

Таблица 4.2

Распределение видов макрофауны литорали Курильских островов по типам ареалов

	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8
WB	28,0	39,8	41,2	44,9	51,9	41,5	34,4	41,3
HB	10,5	13,4	17,0	23,7	27,2	45,0	59,2	49,0
LB	52,1	36,8	34,4	24,2	10,8	5,9	0,8	1,7
W	9,4	10,0	7,4	7,7	10,1	7,6	5,6	8,1
N	171	201	311	194	158	118	125	235

Примечание. WB — широкобореальные виды, HB — высокобореальные виды, LB — низкобореальные виды, W — виды-убиквисты, N — общее число видов в данном районе.

В этой таблице приведены новые описания восьми районов Курильских островов: названиям столбцов таблицы присвоены соответствующие номера географических пунктов (см. с. 41), а

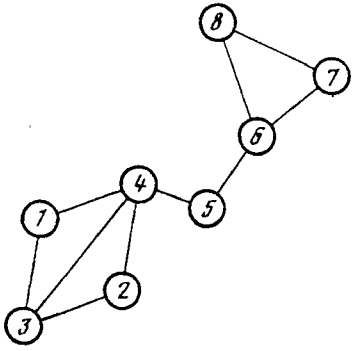


Рис. 4.1. Граф отношений «банальности» для типов ареалов макрофауны литорали Курильских островов

названиям строк (признаков) — характеристики ареалов. Значения признаков — процент видов определенного типа ареала, обитающих в данном географическом пункте. Отметим, что сумма элементов каждого столбца равна 100, поэтому матрица мер парных пересечений на главной диагонали имеет одинаковые значения. Следовательно, соответствующая матрица мер включения симметрична.

Не останавливаясь на промежуточных результатах, приведем граф отношения «банальности» для данных табл. 4.2 (рис. 4.1).

Здесь даже б[illegible]го взгляда достаточно, чтобы найти подтверждение ранее сде[illegible]ным выводам: точка *5* (океанское побережье о-ва Итуруп) принадлежит пограничной зоне между Айнской и Берингийской подобластями. На графе это отражается в том, что ребра, инцидентные вершине *5*, являются мостом между двумя блоками — связными компонентами.

Глава 5

АНАЛИЗ СТРУКТУРНЫХ СХЕМ

5.1. Декомпозиция схемы на сильносвязные и слабосвязные компоненты

Сильносвязными компонентами системы называются такие подсистемы, все составные части которых благодаря обратным связям взаимно достижимы: слабосвязными — такие подсистемы, все элементы которых связаны неориентированным путем. Произвести декомпозицию системы — значит указать, какие связи следует удалить, чтобы система распалась на сильносвязные и слабосвязные компоненты.

Исходными в алгебраической методике исследования графов (структурных схем) являются матрицы смежностей, т. е. матрицы непосредственных связей. Точнее, в матрице смежностей A элемент $a_{ij} = 1$, если в соответствующем орграфе между i-й и j-й вершинами имеется дуга, и $a_{ij} = 0$ в противном случае.

В статье [4] приводится орграф отношений «85%-ной банальности» для восьми описаний ихтиофауны некоторых рек Сибири (рис. 5.1).

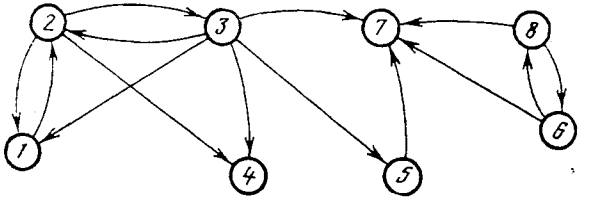


Рис. 5.1. Орграф отношений «85%-ной банальности» для 8 описаний ихтиофауны рек Сибири

Направление стрелок в данном примере указывает на возможные пути расселения рыб в древние времена, поэтому представляет интерес выяснить, какие пути являются «критическими», перекрытие которых привело бы к распаду системы.

Прежде чем решать эту задачу, введем операцию транспонирования матрицы: матрица A^{T} называется транспонированной по отношению к матрице A, если все строки A суть столбцы A^{T}.

Сформулируем следующее утверждение: если R_j — вершина, то сильносвязные компоненты орграфа, содержащие R_j, определяются двойками j-й строки матрицы $A + A^{\mathrm{T}}$. Если все строки $A + A^{\mathrm{T}}$ содержат двойки, то орграф сильносвязный и не распадается на компоненты. В соответствии с этим исследуем данные примера

$$
\begin{array}{c}
\mathrm{A} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \mathrm{A^T} \\
\begin{Vmatrix}
. & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & . & 1 & 1 & 0 & 0 & 0 & 0 \\
1 & 1 & . & 1 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & . & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & . & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & . & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & . & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 1 & .
\end{Vmatrix}
\dotplus
\begin{Vmatrix}
. & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
1 & . & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & . & 0 & 0 & 0 & 0 & 0 \\
0 & 1 & 1 & . & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & . & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & . & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & . & 1 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & .
\end{Vmatrix}
= \\
\mathrm{A \dotplus A^T} \\
= \begin{Vmatrix}
. & 2 & 1 & 0 & 0 & 0 & 0 & 0 \\
2 & . & 2 & 1 & 0 & 0 & 0 & 0 \\
1 & 2 & . & 1 & 1 & 0 & 0 & 0 \\
0 & 1 & 1 & . & 0 & 0 & 0 & 0 \\
0 & 0 & 1 & 0 & . & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & . & 1 & 2 \\
0 & 0 & 0 & 0 & 1 & 1 & . & 1 \\
0 & 0 & 0 & 0 & 0 & 2 & 1 & .
\end{Vmatrix}.
\end{array}
\qquad (5.1)
$$

В (5.1) матрица A — это матрица смежностей орграфа, изображенного на рис. 5.1. В ней сумма элементов i-й строки равна полустепени исхода i-й вершины, а сумма элементов j-го столбца — полустепени захода j-й вершины. Это правило оказывается полезным для проверки соответствия орграфа и матрицы смежностей. Например, сумма элементов второй строки A равна трем, значит, на орграфе имеются три вершины, к которым из вершины *2* направлены стрелки. В матрице A^{T} сумма элементов i-го столбца равна сумме элементов i-й строки матрицы A. Сложение матриц A и A^{T} ведется поэлементно: $a_{ij} + a^{\mathrm{T}}_{ij}$.

Результаты сложения позволяют выделить две сильносвязные компоненты: $u_1 = \{1, 2, 3\}$ и $u_2 = \{6, 8\}$.

Теперь, чтобы выделить слабые компоненты, удаляем подсистемы n_1 и n_2 из матрицы $A + A^{\mathrm{T}}$, получаем сокращенную матрицу

$A' + A'^{\mathrm{T}}$:

$$
\begin{array}{c|ccc}
 & \cdot A & & \\
4 & . & 0 & 0 \\
5 & 0 & . & 1 \\
7 & 0 & 0 & .
\end{array}
\quad
\begin{array}{c|ccc}
 & A'^{\mathrm{T}} & & \\
 & . & 0 & 0 \\
 & 0 & . & 0 \\
 & 0 & 1 & .
\end{array}
\quad
\begin{array}{c|ccc}
 & A' + A'^{\mathrm{T}} & & \\
 & . & 0 & 0 \\
 & 0 & . & 1 \\
 & 0 & 1 & .
\end{array}
\qquad (5.2)
$$

Слабосвязные компоненты орграфа, содержащие i-ю вершину, определяются единичными элементами i-й строки матрицы $A' \dotplus A'^{\mathrm{T}}$. В данном примере замечаем, что единичные элементы имеют 5-я и 7-я строки: соответствующие вершины орграфа принадлежат слабосвязной компоненте $\alpha_1 = \{5, 7\}$.

Для определения связей, удаление которых приводит к разрушению системы, упорядочиваем строки и столбцы матрицы смежностей A в соответствии с выделенными подсистемами:

$$
\begin{array}{c|cc|ccc|cc|c}
 & 1 & 2 & 3 & 6 & 8 & 5 & 7 & 4 & \\
1 & . & 1 & . & . & . & . & . & . & \\
2 & 1 & . & 1 & . & . & . & . & 1 & u_1 \\
3 & 1 & 1 & . & . & . & 1 & . & 1 & \\
\hline
6 & . & . & . & . & 1 & . & 1 & . & \\
8 & . & . & . & 1 & . & . & 1 & . & u_2 \\
\hline
5 & . & . & . & . & . & . & 1 & . & \alpha_1 \\
\hline
7 & . & . & . & . & . & . & . & . & \\
4 & . & . & . & . & . & . & . & . &
\end{array}
\qquad (5.3)
$$

Единичные элементы (нули для наглядности опущены) за пределами очерченных квадратиков соответствуют искомым связям: a_{24}, a_{35}, a_{34}, a_{67}, a_{87}.

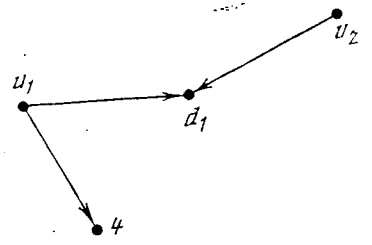


Рис. 5.2. Орграф сокращенной системы

После декомпозиции систему можно представить как сокращенную (рис. 5.2), в которой элементами являются выделенные подсистемы. Орграф сокращенной системы является удобной моделью для решения многих последующих информационных и диагностических задач.

5.2. Наикратчайшие и наидлиннейшие пути

Матричное представление графов хотя и менее наглядно, чем рафическое, обладает определенными преимуществами: дает воз-ожность использовать ЭВМ и, следовательно, позволяет автомати-ировать трудоемкие процессы их анализа. В предыдущем пара-рафе была введена операция сложения матриц, сейчас же введем ще одну операцию: произведением матриц A и B называется атрица C:

$$c_{ik} = \sum_j a_{ij} b_{jk}, \tag{5.4}$$

которой элемент, стоящий в i-й строке и в k-м столбце, равен сум-е произведений элементов i-й строки матрицы A и k-го столбца атрицы B.

Пусть даны матрицы

$$A = \begin{Vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{Vmatrix}, \qquad B = \begin{Vmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{Vmatrix}.$$

Элементы первой строки матрицы $C = A \cdot B$ подсчитываются как

$$c_{11} = a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31},$$
$$c_{12} = a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32},$$
$$c_{13} = a_{11}b_{13} + a_{12}b_{23} + a_{13}b_{33}.$$

Аналогично подсчитываются элементы последующих строк.

Перейдем теперь к анализу матриц смежностей, которые несут полную информацию о графах. В частности, степени матрицы смежностей A орграфа дают полную информацию о числе маршрутов, идущих от одной вершины в другую: элемент a_{ij}^n матрицы A^n равен числу маршрутов длины n, идущих из вершины R_i в вершину R_j.

Кроме матрицы смежностей, с орграфом связаны матрица достижимостей, матрица расстояний D и матрица обходов V. В матрице достижимостей R элемент $r_{ij} = 1$, если i-я вершина достижима из j-й, и равен нулю в противном случае. В матрице D элемент d_{ij} равен расстоянию (длине кратчайшего пути) из i-й вершины в j-ю, если же такого пути не существует, то $d_{ij} = \infty$. В матрице V элемент v_{ij} равен длине наиболее длинного пути из i-й вершины в j-ю, если же такого пути нет, то $v_{ij} = \infty$.

Элементы матриц достижимостей и расстояний связаны с элементами матрицы смежностей следующими соотношениями:

1) $r_{ij} = 1$ тогда и только тогда, когда $a_{ij}^n > 0$ для некоторого n;

2) d_{ij} равно наименьшему из чисел n, для которых $a_{ij}^n > 0$, и бесконечности, если таких чисел нет.

Эффективных методов для нахождения элементов матрицы обходов не существует.

Используем утверждения 1) и 2) для нахождения элементов наикратчайшего пути из i-й вершины в j-ю. Для иллюстрации вычислительных процедур обратимся снова к орграфу на рис. 5.1 и определим наикратчайший путь из вершины *1* в вершину *7*:

A

	1	2	3	4	5	6	7	8
1	.	1	.	.	.	.	.	.
2	1	.	1	1	.	.	.	.
3	1	1	.	1	1	.	1	.
4	.	.	.	.	.	.	.	.
5	.	.	.	.	.	.	1	.
6	.	.	.	.	.	.	1	1
7	.	.	.	.	.	.	.	.
8	.	.	.	.	.	.	1	.

A^2

	.	0	1	1	.	.	.	.
	1	.	.	1	1	.	1	.
	1	1	.	1	.	.	1	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	1	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.

A^3

	.	2	.	1	1	.	1	.
	1	.	1	1	.	.	1	.
	2	2	.	2	1	.	1	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.
	.	.	.	.	.	.	.	.

(5.5)

Здесь, как и ранее, нули заменены точками. В A элемент $a_{17} = 0$, следовательно, вершины *1* и *7* соединены путем, большим 1; в A^2 этот же элемент равен нулю, следовательно, искомый путь больше 2; в A^3 элемент $a_{17} = 1$, откуда заключаем, что наикратчайший путь из вершины *1* в вершину *7* равен 3 (показателю степени A^3). A^3 можно рассматривать также как перечисление маршрутов длиной 3 между всеми вершинами. Однако, чтобы получить матрицу достижимостей, возведение в степень следует вести до тех пор, пока показатель степени не станет равным порядку матрицы ($n = 8$).

Разберем еще один способ определения наикратчайших, а также наидлиннейших путей в графе. Пусть требуется найти оба эти пути от вершины *1* к вершине *7*. Все возможные пути от точки *1* можно представить (рис. 5.3) в виде дерева: графа, не имеющего замкнутых путей (циклов). Способ построения де-

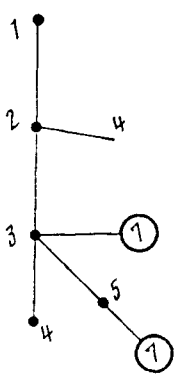


Рис. 5.3. Дерево для определения маршрутов от вершины *1* к вершине *7*

рева становится ясным из рисунка: на первом шаге определяем все вершины, в которые идут пути длиной 1 из вершины *1* (обратные связи игнорируем). Таковых только одна: вершина *2*. Соединяем точки *1* и *2* прямой линией (ребром дерева). Далее отыскиваем все вершины, к которым идут пути длиной 1 от вершины *2*. Таковых две: *3* и *4*. Соединяем точки *2* и *3*, а также *2* и *4* и т. д. Закончив построение, убеждаемся, что наикратчайший путь — $1 \to 2 \to 3 \to 7$, наидлиннейший — $1 \to 2 \to 3 \to 5 \to 7$, причем $\min\ d_{17} = 3$, $\max\ d_{17} = 4$.

5.3. Ранжирование элементов системы в порядке их значимости

Продолжим разбор материалов в статье [4] и рассмотрим отношения «90- и 70%-ной банальности» на множестве описаний пресноводной ихтиофауны

$$\begin{array}{c|cccc} & 5 & 6 & 7 & 8 \\ 5 & . & 1 & 1 & 1 \\ 6 & 1 & . & 1 & 2 \\ 7 & 1 & 1 & . & 2 \\ 8 & 1 & 2 & 2 & . \end{array} . \qquad (5.6)$$

В матрице смежностей (5.6) связи при $\Delta = 70\%$ отмечены 1, связи при $\Delta = 90\%$ — 2. Точками отмечены диагональные элементы. При $\Delta = 90\%$ матрица смежностей примечательна тем, что иллюстрирует нетранзитивность отношения сходства. В самом деле, анализируя связи тройки элементов *6—8—7*, можно заметить, что $R_6 C_{90} R_8$ и $R_8 C_{90} R_7$, но $R_6 \bar{C}_{90} R_7$, т. е. в смысле 90% сходства описания R_6 и R_8 схожи, R_8 и R_7 схожи, но R_6 и R_7 не схожи!

Подобная ситуация часто встречается в спортивных играх, где можно наблюдать парадоксы типа: «чемпион проиграл самому слабому». Подобные парадоксы сильно затрудняют поиск лидера, и математиками были предприняты поиски объективной процедуры выявления лидера [8, 39].

Применительно к матрице (5.6) выявление «лидера» означает поиск элемента, наиболее схожего со всеми остальными. Но вообще эту задачу можно рассматривать несколько шире, считая, что для заданной матрицы отношений необходимо установить для каждого элемента степень его значимости — ранг, который показывал бы, какое место занимает этот элемент по степени проявления изучаемого свойства среди всех остальных элементов.

Если изучаемые отношения являются отношениями сходства, то выделение «лидера» означает в содержательном смысле выделение из общего числа объектов наиболее «типичного». Не вдаваясь в теоретические обоснования процедуры «поиск лидера», которые можно найти в специальных источниках [8, 39], опишем вычислительную схему (алгоритм) ее осуществления.

Упрощенный метод нахождения рангов заключается в следующем [40]. Исходная матрица A возводится в некоторую степень (невысокую — вторую, четвертую), затем суммы элементов строк полученной матрицы A^n делятся на сумму всех элементов этой же матрицы:

$$\omega_i = \frac{\sum_j a_{ij}^n}{\sum_i \sum_j a_{ij}^n}, \qquad (5.7)$$

где $n = 3 \div 4$.

Значение ω_i и есть «вес» i-го элемента. Полученные «веса» ранжируются: на первое место помещается наибольшее значение, на второе — наибольшее среди оставшихся и т. д. Ранжирование определяет на множестве неодинаковых описаний отношение строгого порядка, которое будем называть отношением доминирования $\langle Д, \mathscr{R} \rangle$.

Считая диагональные элементы матрицы (5.6) как 1, получим

$$A = \begin{Vmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 \\ 1 & 1 & 1 & 2 \\ 1 & 2 & 2 & 1 \end{Vmatrix}, \quad A^2 = \begin{Vmatrix} 4 & 5 & 5 & 6 \\ 5 & 7 & 7 & 7 \\ 5 & 7 & 7 & 7 \\ 6 & 7 & 7 & 10 \end{Vmatrix}, \quad A^4 = \begin{Vmatrix} 102 & 132 & 132 & 84 \\ 132 & 172 & 172 & 198 \\ 132 & 172 & 172 & 198 \\ 154 & 198 & 198 & 234 \end{Vmatrix},$$

$$\sum_j a_{ij} = \begin{Vmatrix} 450 \\ 674 \\ 674 \\ 784 \end{Vmatrix}, \quad \omega_i = \begin{Vmatrix} 0{,}174 \\ 0{,}261 \\ 0{,}261 \\ 0{,}304 \end{Vmatrix}. \qquad (5.8)$$

Вектор-столбец с элементами ω_i показывает, что наибольший «вес» ($\omega_8 = 0{,}304$) имеет элемент 8, второе место занимают 6-й и 7-й элементы и последнее — элемент 5.

В данном примере, предназначенном для иллюстрации характера вычислений, заметно, что исходную матрицу A можно было и не возводить в четвертую степень, чтобы получить правильный ответ, однако в общем случае (см. параграф 5.5) суммированием элементов по строкам исходной матрицы нельзя получить даже приблизительного результата.

Более точный способ нахождения ранга i-го элемента описывается итерационной формулой

$$b^{(t)} = A b^{(t-1)}, \qquad (5.9)$$

где $t = 1, 2, \ldots$; $b^{(0)} = (1,1,\ldots, 1)$; A — матрица попарных мер, или бинарных отношений (смежностей). Итерации ведутся до тех пор, пока значения $b^{(t)}$ не стабилизируются. При этом значения

$b^{(t)}$, поскольку они постоянно увеличиваются, целесообразно нормировать:

$$b_i'^{(t)} = \frac{b_i^{(t)}}{\sqrt{\sum (b_i^{(t)})^2}}. \qquad (5.10)$$

Вычислительную схему опишем на примере матрицы A из (5.8). Вектор можно рассматривать как матрицу, имеющую всего один столбец. Тогда на первом шаге

$$\begin{matrix} A & & & & & b^{(0)} & b^{(1)} & b'^{(1)} \end{matrix}$$
$$\begin{Vmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 \\ 1 & 1 & 1 & 2 \\ 1 & 2 & 2 & 1 \end{Vmatrix} \times \begin{Vmatrix} 1 \\ 1 \\ 1 \\ 1 \end{Vmatrix} = \begin{Vmatrix} 4 \\ 5 \\ 5 \\ 6 \end{Vmatrix}, \quad \begin{Vmatrix} 0{,}39 \\ 0{,}50 \\ 0{,}50 \\ 0{,}59 \end{Vmatrix}, \quad \sqrt{\sum (b_i^{(1)})^2} = \sqrt{102} = 10{,}1; \qquad (5.11)$$

на втором шаге

$$\begin{matrix} A & & & & & b'^{(1)} & b'^{(2)} \end{matrix}$$
$$\begin{Vmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 2 \\ 1 & 1 & 1 & 2 \\ 1 & 2 & 2 & 1 \end{Vmatrix} \times \begin{Vmatrix} 0{,}39 \\ 0{,}50 \\ 0{,}50 \\ 0{,}59 \end{Vmatrix} = \begin{Vmatrix} 0{,}39 \\ 0{,}50 \\ 0{,}50 \\ 0{,}59 \end{Vmatrix}. \qquad (5.12)$$

Замечаем, что уже на втором шаге значения $b_i'^{(2)}$ стабилизировались, т. е. одинаковы со значениями $b_i'^{(1)}$. Наибольший вес, как и в предыдущем случае, имеет элемент 8, наименьший — 5.

Отметим, что имеются системы, ранжирование элементов которых не имеет практического смысла. Это прежде всего такие структуры, которые порождаются транзитивными отношениями порядка. Например, в иерархических системах степень значимости элемента и так очевидна: она определяется уровнем иерархии. Таким образом, определение ранга, или функциональной нагрузки элемента, целесообразно проводить только для нетранзитивных отношений и структур типа сетей.

5.4. Информационный критерий сложности структурной схемы

Структурные свойства графа можно характеризовать разнообразием связей каждой вершины. При «однообразной» структуре все вершины имеют одинаковое число инцидентных ребер, и, наоборот, наибольшее разнообразие наблюдается в случае, когда каждая вершина имеет свои особенности. Структурное (топологическое) разнообразие удобно оценивать с помощью количественных показателей, заимствованных из теории информации.

Пусть Г — конечный граф со множеством вершин H и множеством ребер A. Подмножество H_i всех вершин, имеющих одинаковое число инцидентных ребер, образует класс эквивалентности.

В более общем случае (у орграфа) эквивалентные вершины не различаются по двум показателям: числу исходящих и числу заходящих дуг. Эквивалентность порождает разбиение множества H на классы H_i, причем

$$m(H_1 \cup H_2 \cup \ldots \cup H_s) = m(H) = n, \qquad m(H_i) = n_i, \qquad (5.13)$$

где n — число вершин графа; s — число классов.

Отдельный класс можно рассматривать как событие, а совокупность классов — как полную группу событий. Тогда их частоты определяются отношением n_i/n. Энтропия такой схемы событий дает меру разнообразия данного разбиения множества H [40]:

$$I(\Gamma) = -\sum \frac{n_i}{n} \lg \frac{n_i}{n}. \qquad (5.14)$$

Рассмотрим граф отношений «50%-ного сходства» на рис. 3.4, матрица смежностей которого имеет вид

$$A = \begin{Vmatrix} 1 & 1 & 1 & . & . & . & . & . \\ 1 & 1 & 1 & . & . & . & . & . \\ 1 & 1 & 1 & 1 & 1 & . & . & . \\ . & . & 1 & 1 & 1 & . & . & . \\ . & . & 1 & 1 & 1 & . & . & 1 \\ . & . & . & . & . & 1 & 1 & . \\ . & . & . & . & . & 1 & 1 & 1 \\ . & . & . & . & 1 & . & 1 & 1 \end{Vmatrix}, \quad \Omega = \begin{Vmatrix} 3 \\ 3 \\ 5 \\ 3 \\ 4 \\ 2 \\ 3 \\ 3 \end{Vmatrix}. \qquad (5.15)$$

В матрице Ω указаны суммы элементов A по строкам: i-й элемент Ω соответствует числу ребер, инцидентных i-й вершине графа (в данном случае — степени i-й вершины). Отношение сходства симметрично, поэтому сумма элементов i-й строки равна сумме элементов i-го столбца.

Одинаковые элементы Ω указывают эквивалентные вершины. Разобьем множество всех вершин на классы

$$H_1 = \{6\}, \quad H_2 = \{1, 2, 4, 8\}, \quad H_3 = \{5\}, \quad H_4 = \{3\}. \qquad (5.16)$$

В класс H_1 входят все вершины, степень которых равна 2, число таких вершин $m(H_1) = 1$ (вершина *6*). В класс H_2 входят все вершины, имеющие степень 3, число таких вершин $m(H_2) = 4$ и т. д. Общее число вершин $m(H) = 8$, следовательно,

$$I(\Gamma) = -\left[\frac{m(H_1)}{m(H)} \lg \frac{m(H_1)}{m(H)} + \ldots + \frac{m(H_4)}{m(H)} \lg \frac{m(H_4)}{m(H)}\right] =$$
$$= -\left(3 \cdot \frac{1}{8} \lg \frac{1}{8} + \frac{4}{8} \lg \frac{4}{8}\right) = 0{,}49.$$

Рассмотрим теперь схему вычислений для орграфов на примере рис. 3.5. Матрица смежностей орграфа

$$A = \begin{Vmatrix} 1 & . & 1 & . & . & . & . & . \\ . & 1 & 1 & . & . & . & . & . \\ . & . & 1 & . & . & . & . & . \\ . & . & 1 & 1 & . & . & . & . \\ . & . & 1 & 1 & 1 & . & . & 1 \\ . & . & . & . & . & 1 & . & 1 \\ . & . & . & . & . & . & 1 & 1 \\ . & . & . & . & . & . & . & 1 \end{Vmatrix}, \quad \Omega = \begin{Vmatrix} 2 & 1 \\ 2 & 1 \\ 1 & 5 \\ 2 & 2 \\ 3 & 1 \\ 2 & 1 \\ 2 & 1 \\ 1 & 4 \end{Vmatrix}. \tag{5.17}$$

В матрице Ω числа первого столбца совпадают с полустепенью исхода (суммы элементов A по строкам), второго — с полустепенью захода (суммы элементов A по столбцам) каждой вершины.

Соответственно одинаковым строкам Ω разобьем множество вершин H на классы

$$H_1 = \{1, 2, 6, 7\}, \quad H_2 = \{3\}, \quad H_3 = \{4\}, \quad H_4 = \{5\}, \quad H_5 = \{8\}. \tag{5.18}$$

Поступая, как и ранее, получим

$$I(\Gamma) = -\left[\tfrac{1}{2}\lg\tfrac{1}{2} + 4\left(\tfrac{1}{8}\lg\tfrac{1}{8}\right)\right] = 0{,}60,$$

откуда видно, что количество структурной информации у второго графа существенно больше, чем у первого.

Минимального разнообразия степеней вершин следует ожидать, когда все вершины эквивалентны и образуют единственный класс, максимального — когда число классов равно числу вершин. В первом случае

$$I(\Gamma)_{\min} = -n\left(\tfrac{1}{n}\lg\tfrac{n}{n}\right) = 0,$$

во втором

$$I(\Gamma)_{\max} = -n\left(\tfrac{1}{n}\lg\tfrac{1}{n}\right) = \lg n.$$

Последняя величина дает возможность сравнивать графы с разным числом вершин, используя относительный показатель, изменяющийся в пределах от нуля до единицы:

$$\hat{I}(\Gamma) = \frac{I(\Gamma)}{I(\Gamma)_{\max}} = 1 - \frac{\sum n_i \lg n_i}{n \lg n}. \tag{5.19}$$

5.5. Таксономия структур

Назовем таксономией (от греческого «таксис» — расположение по порядку, «номос» — закон) любое транзитивное и антирефлексивное отношение. Напомним, что антирефлексивное отношение —

это такое отношение, когда из факта выполнения xAy следует, что $y \neq x$ для любой пары $x, y \in M$, $A \subseteq M$. Приведенные выше примеры отношений иерархии и доминирования отвечают требованиям транзитивности и антирефлексивности, следовательно, являются частным случаем таксономии.

Рассмотрим более сложные примеры практического использования отношений подобного рода.

При изучении отношений иерархии можно было заметить, что матрица сходства содержит всю информацию о дендрограмме, однако восстановить по последней матрицу сходства в общем случае не удается. Другими словами, при построении дендрограммы происходит потеря информации. Поэтому в тех случаях, когда возникает необходимость сравнения нескольких матриц мер сходства, сопоставление дендрограмм оказывается неэффективным.

Для этих целей следует использовать поэлементное сравнение матриц на основе какой-либо меры. В частности, для характеризации различий и сходства двух матриц, скажем F и G, можно использовать евклидову метрику

$$d(F, G) = \left[\sum_{i,j}(f_{ij} - g_{ij})^2\right]^{1/2}, \tag{5.20}$$

где f_{ij}, g_{ij} — элементы i-й строки и j-го столбца соответственно матриц F и G.

Аналогично если A и B — матрицы смежностей, то расстояние в пространстве отношений

$$d(A, B) = \sum_{i,j} |a_{ij} - b_{ij}| \tag{5.21}$$

есть количество несовпадений элементов обеих матриц.

Разумеется, для симметричных матриц достаточно учитывать элементы нижнего (верхнего) треугольника.

Приведем примеры. При анализе морфологического строения яйцеклеток 10 видов пресноводных рыб учитывалась следующая информация об ооцитах на стадии их большого роста: диаметры клеток и их ядер, окраска, форма, размеры вакуолей и желточных гранул, количество и расположение их в толще цитоплазмы, толщина и структура оболочек яйцеклеток и др., всего 95 двоичных признаков.

Стадия большого роста ооцитов разделяется на шесть хорошо различимых фаз, которые условно называются D_1, D_2, D_3, E_1, E_2, E_3 [29]. Десять видов, на представителях которых делались измерения*, принадлежат семейству карповых: 1 — горбушка, 2 — уклей, 3 — верхогляд, 4 — монгольский краснопер, 5 — ханкайская востробрюшка, 6 — корейская востробрюшка, 7 — конь пестрый, 8 — угай, 9 — серебряный карась, 10 — сазан. Латинские названия их не приводим, так как в рассматриваемых

* Все измерения выполнены доцентом кафедры ихтиологии и гидробиологии ДВГУ В. Н. Иванковым.

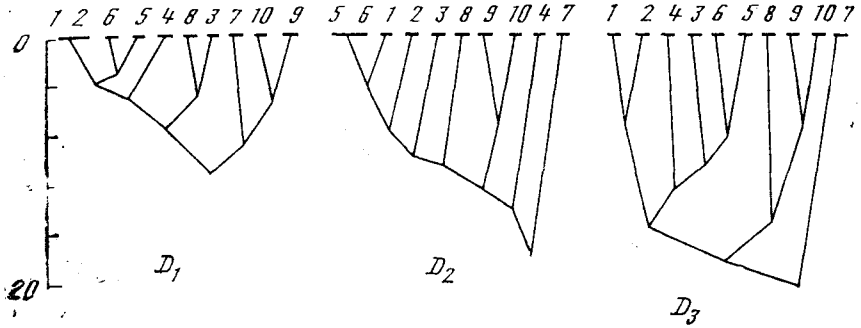


Рис. 5.4. Классификации ооцитов в процессе развития у 10 видов пресноводных рыб

аспектах это несущественно, а вместо названий в дальнейшем будем использовать присвоенные им номера.

Суперзадачу классификационных построений в данном исследовании можно определить как попытку установить отношения филогенетического родства указанных видов рыб на основе анализа морфологического строения их яйцеклеток на стадии большого роста*. Выбор такой цели основан на представлении о соответствии (гомоморфном отображении) структуры родственных отношений видов структуре сходства строения яйцеклеток, которое сравнительно слабо зависит от экологических условий.

В качестве меры различий отдельных описаний использовалась величина

$$D(R_j, R_k) = m(\bar{R}_j \cup \bar{R}_k), \qquad (5.22)$$

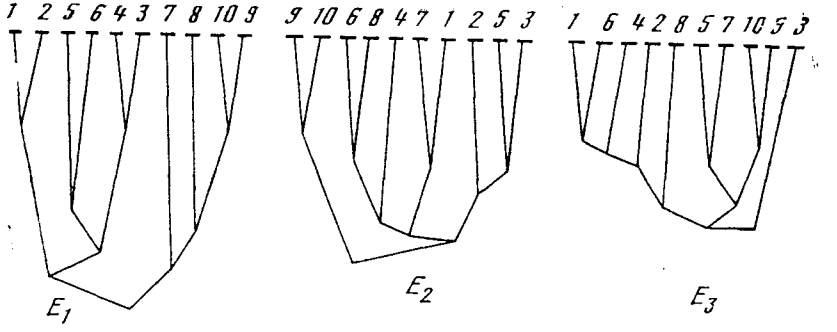
равная числу несовпадающих признаков. Матрицы мер различия для каждой фазы роста послужили основой для графического изображения сходства описаний в виде дендрограмм (рис. 5.4).

Из рис. 5.4 можно заметить, что для большинства рассматриваемых фаз наблюдается сравнительно сильное сходство описаний видов 9 и 10, 1 и 2, 5 и 6. Однако прочие соотношения менее постоянны в течение оогенеза. Наибольшая дивергенция обнаруживается на фазах D_3 и E_1, наименьшая — на крайних фазах: D_1 и E_3. Это обстоятельство порождает следующие задачи: определить иерархию различий полученных классификаций и на их основе выделить наиболее типичный вариант.

Первая задача легко решается, если подсчитать матрицу парных мер (5.20) для всех шести классификаций, представляющих разные фазы развития ооцитов. Результаты расчетов приведены на дендрограмме рис. 5.5.

Оказалось, что наименьшие различия в указанном смысле обнаруживают варианты, соответствующие D_3 и E_2, наибольшие

* Формулировка цели принадлежит В. Н. Иванкову.



расхождения со всеми остальными имеет вариант, представленный фазой D_1. Довольно близки между собой варианты D_2 и E_3, но в целом они имеют большие различия со всеми остальными, чем первая пара. Таксономия структур легко прослеживается на дендрограмме рис. 5.5 как отношение иерархии по признаку подчинения на множестве всех шести вариантов классификации.

Чтобы решить вторую задачу, необходимо ответить на вопрос: какой из представленных вариантов является наиболее схожим со всеми остальными? Непосредственный анализ рис. 5.4 показывает, что наименее схожи со всеми остальными классификации E_1 и D_1, но выбор среди оставшихся наиболее схожего со всеми вариантами затруднен. Ответ на этот вопрос не может быть получен и по данным рис. 5.5. Здесь можно выделить только наиболее похожую пару и можно лишь предполагать, что искомым вариантом является или D_2, или E_3, расположенные в середине рис. 5.5.

Для точного решения поставленной задачи следует обратиться к процедуре поиска «лидера», используя матрицу мер (5.20).

После 10 итераций, проведенные на ЭВМ, были получены следующие значения элементов вектора «весов» нагрузок) для каждой классификации:

$$\|\omega_i\| = \|0{,}259 \quad 0{,}128 \quad 0{,}136$$
$$0{,}229 \quad 0{,}129 \quad 0{,}119\|. \qquad (5.23)$$

Наибольший «вес» получили варианты, соответствующие фазам D_1 и E_1: они имеют наибольшие отличия от всех остальных, наименьший «вес» вариантов, соответствующих фазам E_3 и D_2: они наименее различны со всеми. Очевидно, что в приведенном анализе матрицы мер различий наибольшие «веса» соответствуют «выдающимся» элементам, а наименьшие — «ординарным» (типичным) элементам.

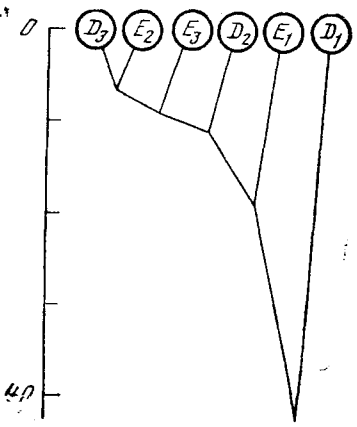


Рис. 5.5. Дендрограмма различий шести классификаций описаний строения яйцеклеток рыб

Значения вектора (5.23) задают отношение доминирования по признаку различия множества классификаций. Это отношение, как указывалось, транзитивно и антисимметрично, следовательно, является таксономией.

После тщательного обдумывания полученных результатов было принято решение отыскать среди всех вариантов не самый «типичный», а тот, в котором дифференциация клеточных структур выражена наиболее четко, и именно этот вариант считать пригодным для использования в целях систематики.

Из физического содержания задачи следует, что наиболее дифференцированная фаза — такая, в которой исследуемые объекты имеют наибольшее число учитываемых признаков. Это обстоятельство отражается на дендрограмме таким образом, что она становится сравнительно сильно вытянутой в длину. Обращаясь к рис. 5.4, можно видеть, что таким требованиям отвечает вариант для E_1, но, как отмечалось, при построении дендрограмм происходит потеря информации. Поэтому в данном случае желательно решить вопрос на основе анализа непосредственно матриц мер различия.

Представляя нижние треугольники шести матриц мер попарных различий в виде векторов-описаний каждой классификации и используя формулы (4.12) и (4.13) для дескриптивных множеств, мы нашли матрицу мер включения (табл. 5.1).

Таблица 5.1

Матрица мер включения классификаций ооцитов в процессе роста

	D_1	D_2	D_3	E_1	E_2	E_3
D_1		99	100	99	100	100
D_2	58		98	98	95	89
D_3	46	77		97	90	79
E_1	40	67	83		81	69
E_2	48	78	94	97		82
E_3	56	84	95	97	95	

Меры включения в данном случае показывают, что если в одной из сравниваемых матриц различий все значения окажутся меньшими, чем во второй, то первая будет включена во вторую на 100%. В табл. 5.1 классификация для D_1 включена в D_3, E_2 и E_3 на 100%. Граф отношений «95%-ной банальности» приведен на рис. 5.6.

Легко видеть, что наибольшее число стрелок выходит из вершины E_3, следовательно, соответствующая фаза является наименее дифференцированной: в ней все различия почти на 95% всего состава меньше, чем во всех других вариантах. Наоборот, наибольшее число стрелок входит в вершину E_1, и это свидетельствует о самых больших различиях между всеми описаниями соответствующей фазы. Одинаковое число входящих и выходящих стрелок имеют вершины D_3, E_2 и D_2, E_3, которые в этом смысле симметричны относительно E_1.

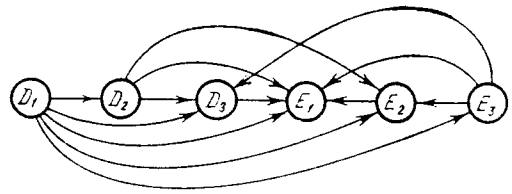


Рис. 5.6. Орграф отношений «95%-ной банальности» для шести вариантов таксономии описаний яйцеклеток рыб

Для более точного, аналитического решения этой задачи мы применили процедуру поиска «лидера», используя данные табл. 5.1 и получили следующий кортеж «весов» для каждого варианта 197, 174, 157, 141, 160, 171, т. е. и в этом случае выводы, полученные на основе рис. 5.6, полностью подтвердились: наименьший вес получил вариант для фазы E_1 и практически одинаковые «веса» — варианты, смежные с E_1. Таким образом, для цитоморфологических исследований в обсуждаемом аспекте следует отбирать яйцеклетки на фазе E_1.

Глава 6

ИДЕНТИФИКАЦИЯ. МИНИМИЗАЦИЯ ОПИСАНИЙ

6.1. Решающие правила

При решении задач идентификации должны быть заданы перечень учитываемых признаков (система описания), алфавит (перечень) классов, конкретные представители этих классов или же их «эталоны» — типичные представители. Любое вновь рассматриваемое описание R объекта R' на основе меры близости должно быть отнесено к одному из классов.

Пусть заданы два класса, H_1 и H_2, и их эталоны: $\hat{R}_1$: 1, 1, 0 $\hat{R}_2$: 0, 0, 1; а также объект R : 1, 0, 0, который следует отнести к одному из двух классов: H_1 или H_2.

Уже непосредственный обзор данных этой задачи убеждает в том, что R имеет с $\hat{R}_1$ число общих признаков большее, чем с $\hat{R}_2$, поэтому R следует отнести к H_1. Действительно, задавая меру сходства

$$C(R, \hat{R}_k)_0 = \frac{2m(R \cap \hat{R}_k)}{m(R) + m(\hat{R}_k)} = \frac{2 \cdot \sum x_i . \hat{x}_{ik}}{\sum x_i . + \sum \hat{x}_{ik}}, \qquad (6.1)$$

определяем

$$C(R, \hat{R}_1)_0 = \frac{2 \cdot 1}{5 + 1} \approx 0{,}67, \qquad C(R, \hat{R}_2) = \frac{2 \cdot 0}{1 + 1} = 0,$$

т. е. $R \in H_1$.

Общий итог этих построений можно сформулировать в виде решающего правила: если $C(R, \hat{R}_1)_0 > C(R, \hat{R}_2)_0$, то R следует отнести к первому классу, а если $C(R, \hat{R}_1)_0 < C(R, \hat{R}_2)$, то ко второму. При $C(R, \hat{R}_1.)_0 = C(R, \hat{R}_2)_0$ однозначного ответа не существует: распознаваемый объект лежит на границе, разделяющей классы.

Если число классов больше двух, то решающее правило остается по существу таким же: распознаваемый объект относится к тому классу, сходство с которым у объекта наибольшее.

Рассмотрим теперь ситуацию, когда классы H_u заданы не эталонами, а полным составом. Пусть $H_1 = \{R_1, R_2, R_3\}$, $H_2 = \{R_4, R_5, R_6\}$, причем

$$\begin{aligned} &R_1\colon 1\,1\,1\,0\,1\,1\,0\,0, \quad && R_4\colon 1\,0\,0\,1\,0\,0\,0\,1,\\ &R_2\colon 1\,1\,1\,1\,1\,1\,0\,0, \quad && R_5\colon 1\,0\,0\,1\,0\,0\,1\,0;\\ &R_3\colon 1\,1\,1\,0\,0\,1\,0\,0, \quad && R_6\colon 0\,0\,0\,1\,0\,0\,0\,1,\\ &R\colon 1\,0\,1\,0\,1\,0\,1\,0. \end{aligned} \qquad (6.2)$$

Для того чтобы определить, к какому из двух классов относится R, подсчитаем какую-либо меру сходства или различия описания R со всеми представителями первого и второго классов. Пусть мера различия есть число несовпадающих значений признаков

$$D(R, R_j) = m(\bar{R} \cup \bar{R}_j) = \sum_{j}^{q_1} \sum_{i}^{p} x_i . \oplus x_{ij}, \qquad (6.3)$$

где $x_i.$, x_{ij} — значения признаков соответственно R и R_j. Тогда для всех представителей обоих классов получим соответственно: $D(R, R_1) = 3$, $D(R, R_2) = 4$, $D(R, R_3) = 4$, $D(R, R_4) = 5$, $D(R, R_5) = 3$, $D(R, R_6) = 6$. Найдем среднее «расстояние» по каждому из классов

$$\hat{D}_1(R, R_j) = 3{,}67, \qquad \hat{D}_2(R, R_j) = 4{,}67, \qquad (6.4)$$

откуда видно, что в среднем различия данного описания с представителями H_1 наименьшие и его следует отнести именно к этому

классу. Р ешающее правило для данного случая: объект относятся му классу, до которого среднее «расстояние» м нимальнс

Принадлежность распознаваемого объекта к одному из кла сов определяется тем надежней, чем сильнее различаются объект разных классов по сравнению с различиями внутри классов. Дл рассматриваемого примера внутриклассовые различия можн представить как среднее «расстояние» описаний, составляющи H_1 или H_2:

$$\begin{aligned} &D(R_1, R_2) = 1, \quad D(R_1\ R_3) = 1, \quad D(R_2, R_3) = 2,\\ &\hat{D}_{H_1}(R_k, R_j) = 1{,}33,\\ &D(R_4, R_5) = 1, \quad D(R_4, R_6) = 2, \quad D(R_5, R_6) = 2,\\ &\hat{D}_{H_2}(R_k, R_j) = 1{,}67. \end{aligned} \qquad (6$$

Усредняя по всем классам, получим $\hat{D}_B(R_k, R_j) = 1{,}5$— оцен внутриклассовых различий H_1 и H_2.

Межклассовые различия есть отличия каждого объекта одного класса от всех объектов другого:

$$\begin{aligned} &D(R_1, R_4) = 6, \quad D(R_1, R_5) = 6, \quad D(R_1, R_6) = 7,\\ &D(R_2, R_4) = 5, \quad D(R_2, R_5) = 5, \quad D(R_2, R_6) = 6,\\ &D(R_3, R_4) = 5, \quad D(R_3, R_5) = 5, \quad D(R_3, R_6) = 6, \end{aligned}$$

$$\hat{D}_M(R_k, R_j) = \frac{51}{9}. \qquad (6$$

Отношение

$$\frac{\hat{D}_B(R_k, R_j)}{\hat{D}_M(R_k, R_j)} \approx 0{,}26$$

показывает, что внутриклассовые различия составляют вс лишь 26% межклассовых, т. е. различимость объектов в дан случае вполне приемлема для многих целей.

Не все признаки играют одинаковую роль в различении кл сов, поэтому желательно оценить информативность каждого них в отдельности. Простейшим критерием полезности может с жить оценка «вклада», вносимого признаком в межклассовые л личия. Так, для условий предыдущего примера можно расс тать, какая доля этих различий приходится на каждый приз в отдельности.

Ранее показано (6.6), что межклассовое «расстоян $D_M(R_k, R_j) = 51$ указывает общее число несовпадающих значе при сравнении каждого описания из H_1 с каждым описанием из Легко установить, что на долю первого признака приходится 3 совпадения, на долю второго — 9 и далее соответственно 9, 6 9, 3, 6. Итак, наиболее значимыми оказались признаки 2, (доля вклада в общее «расстояние» каждого из них наибольш

равна $9/51 \approx 0{,}18$), наименее значимые признаки 1 и 7 (доля вклада у них наименьшая: $3/51 \approx 0{,}06$).

Полученные результаты дают возможность приписать каждому признаку «вес» (нагрузку), численно равный доле вклада в общее «расстояние»: $\| \omega_i \| = \| 0{,}06\ 0{,}18\ 0{,}18\ 0{,}12\ 0{,}12\ 0{,}18\ 0{,}06 \|$.

Распознавание нового объекта можно теперь проводить с использованием полученных нагрузок

$$C_{H_1}(R, R_j) = \sum_j^{q_1} \sum_i^{p} \omega_i x_i . x_{ij} = 0{,}96,$$
$$C_{H_2}(R, R_j) = \sum_j \sum_i \omega_i x_i . x_{ij} = 0{,}18, \tag{6.7}$$

где $C(R, R_j)$ — мера сходства, учитывающая только совпадающие значения, т. е. согласно решающему правилу принадлежность распознаваемого вектора $\mathbf{R}$ из (6.2) к классу H_1 очевидна.

Такой же результат можно получить и при использовании некоторых других мер сходства или различия, например

$$\widehat{C}_{H_1}(R, R_j)_0 = \frac{2}{q_1} \sum_j^{q_1} \frac{\sum \omega_i x_i . x_{ij}}{\sum_i \omega_i x_i . \sum \omega_i x_{ij}} =$$
$$= \frac{2}{3}\left[\frac{0{,}36}{0{,}42 + 0{,}72} + \frac{0{,}36}{0{,}42 + 0{,}84} + \frac{0{,}24}{0{,}42 + 0{,}60} \right] = 0{,}56,$$
$$\widehat{C}_{H_2}(R, R_j)_0 = \frac{2}{q_2} \sum_j^{q_1+q_2} \frac{\sum \omega_i x_i . x_{ij}}{\sum \omega_i x_i . \sum \omega_i x_{ij}} =$$
$$= \frac{2}{3}\left[\frac{0{,}06}{0{,}42 + 0{,}30} + \frac{0{,}12}{0{,}42 + 0{,}24} + 0 \right] = 0{,}17.$$

В случае количественных значений признаков меры можно подбирать в соответствии с указаниями параграфа 5.5.

6.2. Тупиковые тесты. Допустимые и компактные определители. Оптимальные ключи

Существуют и другие принципы минимизации и отбора информативных признаков. Рассмотрим только некоторые, полезные для решения биологических задач, в частности для составления определителей и ключей.

Пусть имеется q описаний, каждое из которых характеризует представителей какой-либо одной категории (например, видов):

$$\begin{aligned} &\mathscr{R} = \{R_j \mid j \in J\}, \quad J = \{j \mid j - \text{целое число}, 1 \leqslant j \leqslant q\}, \\ &R_j = \{x_{ij} \mid i \in I,\ j \in J\}, \\ &\mathscr{S} = \{S_i \mid i \in I\}, \quad I = \{i \mid i - \text{целое число}, 1 \leqslant i \leqslant p\}, \\ &S_i = \{x_{ij} \mid i \in I, j \in J\}, \\ &x_{ij} \in \{0, 1, *\}. \end{aligned} \tag{6.8}$$

Как и ранее, $\mathscr{R}$ — это алфавит описаний, R_j — j-е описание объекта R_j', $\mathscr{S}$ — алфавит признаков (система описания), S_i — i-й признак, x_{ij} — одно из трех значений i-го признака у j-го объекта: $x_{ij} = 0$, когда i-го признака нет, $x_{ij} = 1$, если i-й признак есть, и $x_{ij} = *$, когда i-й признак безразличный (может быть и может не быть).

Множество $\mathscr{R}$ таблично представляется в виде матрицы T, имеющей p строк и q столбцов, а на пересечении i-й строки j-го столбца помещаются значения x_{ij}.

Введем операцию $A \oplus B$ над элементами множества $\{0, 1, *\}$:

$$\begin{aligned} &1 \oplus 1 = 0, \quad 1 \oplus 0 = 0 \oplus 1 = 1, \quad 0 \oplus 0 = 0, \\ &* \oplus 1 = 1 \oplus * = 0 \oplus * = * \oplus 0 = * \oplus * = 0. \end{aligned} \tag{6.9}$$

Признак S_i называется различающим на множестве $\mathscr{R}$, если найдутся такие k и j ($k \neq j$), что $x_{ik} \oplus x_{ij} = 1$. Описания R_j и R_k называются различимыми, если имеют хотя бы один различающий признак. Таблица T называется допустимой, если все столбцы в ней попарно различимы.

Для многих практических целей интересно выяснить: какие из переменных S_i можно удалить, не нарушив различимости описаний R_j?

Набор признаков

$$T_u = \{S_{m_1}, \ldots, S_{m_t}\}, \quad m_1, \ldots, m_t \in I, \quad u = 1, \ldots, r \tag{6.10}$$

называется тестом таблицы T, если после удаления всех признаков, не вошедших в T_u, все $R_j \in \mathscr{R}$ остаются различимыми. Тест называется тупиковым, если при удалении хотя бы одного $S_i \in T_u$ условия различимости R_j нарушаются.

Алгоритмы выявления всех тупиковых тестов таблицы T приводятся в статье [21]. Существо одного из этих алгоритмов можно пояснить с использованием операции $A \ominus B$, позволяющей установить различающие признаки на каждой паре описаний R_j и R_k ($k \neq j$).

Произведем попарное сравнение всех объектов из T и составим вспомогательную таблицу со столбцами

$$R_{jk} = R_j \ominus R_k = \{x_{1j} \oplus x_{1k}, \ldots x_{pj} \oplus x_{pk}\}. \tag{6.11}$$

В каждом R_{jk} выделим номера $\varkappa_1, \ldots, \varkappa_\varepsilon$ различающих признаков и сопоставим R_{jk} дизъюнкцию

$$V_{jk} = (S_{\varkappa_1}, \ldots, S_{\varkappa_\varepsilon})_{jk}, \quad \varkappa_1, \ldots, \varkappa_\varepsilon \in I, \tag{6.12}$$

а разным R_{jk} — конъюнкцию

$$\prod_{j=1}^{q-1} \prod_{k=j+1}^{q} V_{jk}. \tag{6.13}$$

Рассматривая $S_{\varkappa_1}, \ldots, S_{\varkappa_\varepsilon}$ как логические переменные, приведем (6.12) к виду ΣП и упростим выражения типа $A \cdot A = A$, $A + A = A$, $A + A \cdot B = A$. Тогда (6.12) будет иметь вид

$$\sum^{\tau} S_{m_1}, \ldots, S_{m_\eta}. \tag{6.14}$$

Каждому слагаемому в (6.14) соответствует тупиковый тест T. Кроме того, (6.14) указывает все тупиковые тесты таблицы T.

Пример. На множестве описаний $\mathcal{R}$, заданном в виде таблицы

$$T = \begin{array}{c|cccc} & R_1 & R_2 & R_3 & R_4 \\ \hline S_1 & 0 & 0 & 0 & 0 \\ S_2 & 1 & 0 & 1 & 0 \\ S_3 & 0 & 1 & 1 & 1 \\ S_4 & 1 & 1 & 1 & 0 \\ S_5 & 0 & 1 & 0 & 1 \end{array} \tag{6.15}$$

определить все тупиковые тесты.

Прежде чем решать эту задачу, убеждаемся в том, что таблица (6.15) допустима, т. е. не содержит одинаковых столбцов. Если T не является допустимой, то один из неразличимых столбцов должен быть удален.

Проведем попарное сопоставление R_j и R_k $(k \neq j)$, отмечая различающие признаки «1», а прочие «0»:

$$\begin{array}{c|cccccc} & R_{12} & R_{13} & R_{14} & R_{23} & R_{24} & R_{34} \\ \hline S_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ S_2 & 1 & 0 & 1 & 1 & 0 & 1 \\ S_3 & 1 & 1 & 1 & 0 & 0 & 0 \\ S_4 & 0 & 0 & 1 & 0 & 1 & 1 \\ S_5 & 1 & 0 & 1 & 1 & 0 & 1 \\ \hline m\,(R_{jk}) & 3 & 1 & 4 & 2 & 1 & 3 \end{array} \tag{6.16}$$

В нижней строке поместим значения $m\,(R_{jk})$ — число единичек в каждом столбце.

В содержательном смысле значения столбцов $R_{jk} = 1$ означают: описания R_1 и R_2 различимы по признакам S_2, S_3 и S_5, следовательно, один из них — либо S_2, либо S_3, либо S_5 — может быть включен в тест; описания R_1 и R_3 различимы по признаку S_3 и т. д. Для сопоставления теста необходимо учитывать все столбцы (6.16) одновременно, поэтому

$$\underset{R_{12}}{(S_2 + S_3 + S_5)} \cdot \underset{R_{13}}{S_3} \cdot \underset{R_{14}}{(S_2 + S_3 + S_4 + S_5)} \underset{R_{23}}{(S_2 + S_5)} \cdot \underset{R_{24}}{S_4} \cdot \underset{R_{34}}{(S_2 + S_4 + S_5)}. \tag{6.17}$$

Раскрывая скобки, получим все тесты таблицы T, а произведя упрощения: $A \cdot A = A$, $A + A = A$, $A + A \cdot B = A$, найдем все ее тупиковые тесты

$$S_2 \cdot S_3 \cdot S_4 + S_3 \cdot S_4 \cdot S_5. \tag{6.18}$$

число которых $\tau = 2$. Из них первый содержит признаки S_2, S_3, S_4, второй — $S_3 S_4 S_5$. В самом деле, если учитывать каждый набор в отдельности, получим таблицы

$$T_1 = \begin{array}{c|cccc} & R_1 & R_2 & R_3 & R_4 \\ \hline S_2 & 1 & 0 & 1 & 0 \\ S_3 & 0 & 1 & 1 & 1 \\ S_4 & 1 & 1 & 1 & 0 \end{array} \qquad T_2 = \begin{array}{c|cccc} & R_1 & R_2 & R_3 & R_4 \\ \hline S_3 & 0 & 1 & 1 & 1 \\ S_4 & 1 & 1 & 1 & 0 \\ S_5 & 0 & 1 & 0 & 1 \end{array} \tag{6.19}$$

в каждой из которых все столбцы различимы, а удаление хотя бы одной строки нарушает условия различимости. Следовательно, T_1 и T_2 — тупиковые тесты.

Заметим, что признак S_1 не является различающим и не вошел ни в один тупиковый тест. Признак S_2 входит в T_1, S_3 — в T_1 и T_2 и т. д. Обозначим τ_i число тупиковых тестов, в которые входит i-й признак, тогда

$$\omega_i = \tau_i / \tau \tag{6.20}$$

называется информационным весом i-го признака. Для таблицы (6.15) $\omega_1 = 0$, $\omega_2 = 1/2$, $\omega_3 = 1$, $\omega_4 = 1$, $\omega_5 = 1/2$. Наибольший вес имеют признаки S_3 и S_4.

В данном примере громоздкими действиями оказываются раскрытие скобок и упрощение (6.17). Многих трудностей можно избежать, если провести предварительное упрощение таблицы (6.16). В нижней строке $m\,(R_{jk})$ выделим минимальное значение. Таковым оказывается «1», которая встречается два раза: $m\,(R_{13}) = 1$ и $m\,(R_{24}) = 1$. Рассмотрим столбцы, соответствующие минимальным значениям. В R_{13} значение «1» соответствует признаку S_3: вычеркиваем из (6.16) все столбцы, у которых S_3 имеет значение «1», т. е. R_{12} и R_{14}. Во втором из отобранных столбцов — R_{24} — значение «1» соответствует S_4, поэтому вычеркиваем из (6.16) все столбцы, которые имеют «1» в четвертой строке (столбец R_{34}). Остаются: R_{13}, R_{23}, R_{24}, не имеющие «1» в одинаковых строках. Дальнейшее упрощение невозможно, поэтому анализируем содержание оставшихся столбцов:

$$\underset{R_{13}}{S_3} \cdot (\underset{R_{23}}{S_2} + \underset{R_{24}}{S_3}) \cdot S_4 = S_2 \cdot S_3 \cdot S_4 + S_3 \cdot S_4 \cdot S_5,$$

т. е. получим все тупиковые тесты.

Вычеркнуть столбец R_{jk} возможно в том случае, если $R_{jk} \to R_{me}$: все значения $x_{ij} \oplus x_{ik} = 1$ в R_{jk} имеются в R_{me}. Убедимся на примере R_{12} и R_{13} в том, что $R_{12} \to R_{13}$ и первый из них

может быть вычеркнут:

$$(S_1 + S_2 + S_5)\cdot S_3 = \underset{R_{12}}{S_2\cdot S_3} + \underset{R_{13}}{S_3} + S_3\cdot S_5 = S_3$$

(использован закон поглощения: $A + A\cdot B = A$). Кроме того, $(A + B)\cdot(A + B) = A + B$, следовательно, если во вспомогательной таблице имеется несколько одинаковых столбцов, то все они, кроме одного, вычеркиваются. При большой размерности таблицы T можно воспользоваться алгоритмами приведения к тупиковой форме (см. параграф 4.3).

Введем дополнительные понятия.

1. Любой тест, в том числе и допустимая таблица, есть допустимый определитель.

2. Любой набор допустимых определителей называется комплектом допустимых определителей.

3. Комплект допустимых определителей, в который входят все различающие признаки $S_i \in \mathcal{S}$, называется полным.

4. В полном комплекте допустимый определитель называется основным, если он имеет наибольший средний вес $\hat{\omega}_{max}$ признаков, остальные допустимые определители комплекта называются вспомогательными. Если одинаковые $\hat{\omega}_{max}$ имеют несколько определителей комплекта, то любой из них относится к основному, а остальные — к вспомогательным.

Частным случаем допустимых определителей являются компактные определители, формальной основой которых служат тупиковые тесты. Различия допустимых и компактных определителей обусловлены следующим: из допустимости таблицы следует, что она есть тест, и обратно: тест есть допустимая таблица; тупиковый тест есть допустимая таблица, но обратное неверно.

Приведем пример составления компактных определителей. В работе [43] приводится определитель подсемейств муравьев, обитающих в СССР.

1(2) Стебелек двучлениковый (S_1). Жало имеется [(S_2) . . . Mirmicinae (R_1)

2(1) Стебелек одночлениковый ($\bar{S}_1$) 3

3(4) Первый членик брюшка отделен от остальных перетяжкой (S_3). Жало хорошо развито (S_2) . . . . . Ponerinae (R_2)

4(3) Брюшко без перетяжки ($\bar{S}_3$). Жало отсутствует ($\bar{S}_2$) 5

5(6) Анальное отверстие вытянуто в трубочку (S_4). Шпоры средних и задних голеней самцов простые (S_5) . . . . . Formicinae (R_3)

6(5) Анальное отверстие без трубочки ($\bar{S}_4$). Шпоры задних голеней самцов гребенчатые ($\bar{S}_5$) . . . . . Dolichoderinae (R_4)

На основе правил формализации (6.8) составляем основную и вспомогательную таблицы

	R_1	R_2	R_3	R_4
S_1	1	0	0	0
S_2	1	0	0	0
S_3	*	1	0	0
S_4	*	*	1	0
S_5	*	*	1	0

	R_{12}	R_{13}	R_{14}	R_{23}	R_{24}	R_{34}
S_1	1	1	1	0	0	0
S_2	0	1	1	1	1	0
S_3	0	0	0	1	1	0
S_4	0	0	0	0	0	1
S_5	0	0	0	0	0	1
$m(R_{jk})$	1	2	2	2	2	2

(6.21)

В таблице T нет неразличимых столбцов, следовательно, она допустимая. Во вспомогательной таблице наименьшее значение $m(R_{jk}) = 1$ соответствует признаку S_1: вычеркиваем все столбцы, содержащие 1 в первой строке (R_{13}, R_{14}). Кроме того, $R_{23} = R_{24}$, поэтому удаляем любой из них (R_{24}). В остальных столбцах значения «1» соответствуют следующим строкам:

$$\underset{R_{12}}{S_1\cdot(S_2 + S_3)}\cdot\underset{R_{23}}{(S_1 + S_3)} = \underset{R_{34}}{S_1\cdot S_3\cdot S_4} + S_1\cdot S_2\cdot S_5 + S_1\cdot S_3\cdot S_4 + S_1\cdot S_3\cdot S_5.$$

Полный комплект состоит из четырех компактных определителей, каждый из которых использует только три из пяти исходных признаков:

$$T_1 = \begin{array}{c|cccc} & R_1 & R_2 & R_3 & R_4 \\ \hline S_1 & 1 & 0 & 0 & 0 \\ S_2 & 1 & 1 & 0 & 0 \\ S_4 & * & * & 1 & 0 \end{array} \qquad T_2 = \begin{array}{c|cccc} & R_1 & R_2 & R_3 & R_4 \\ \hline S_1 & 1 & 0 & 0 & 0 \\ S_2 & 1 & 1 & 0 & 0 \\ S_5 & * & * & 0 & 0 \end{array}$$

$$T_3 = \begin{array}{c|cccc} & & & & \\ \hline S_1 & 1 & 0 & 0 & 0 \\ S_3 & * & 1 & 0 & 0 \\ S_4 & * & * & 1 & 0 \end{array} \qquad T_4 = \begin{array}{c|cccc} & & & & \\ \hline S_1 & 1 & 0 & 0 & 0 \\ S_3 & * & 1 & 0 & 0 \\ S_5 & * & * & 1 & 0 \end{array}$$

(6.22)

Информационные веса признаков: $\omega_1 = 1$, $\omega_2 = \omega_3 = \omega_4 = \omega_5 = 0{,}5$. Средний вес у всех T_u одинаков ($\hat{\omega}_{max} = 0{,}67$), следовательно, любой из определителей (6.22) может считаться основным.

Тесты являются лишь формальной основой определителей. Поэтому при их составлении главное внимание уделяется структуре, в то время как при сопоставлении определителей важно учитывать не только структуру, но и содержательный смысл составляющих элементов (признаков).

Тесты, различающиеся только названиями строк, изоморфны, а соответствующие им определители называются подобными. Пары T_1 и T_2, T_3 и T_4 в (6.22) изоморфны, так как имеют неразличимые структуры.

При отборе определителей из полного комплекта следует отдавать предпочтение таким, которые имеют: 1) легко и надежно определяемые признаки; 2) наибольшее значение $\hat{\omega}_{max}$; 3) наибольшее число подобных определителей; 4) наименьшее число признаков; 5) наименьшее число значений «*». Некоторые из этих требований противоречивы, поэтому компромиссное решение нужно находить, исходя из содержательных аспектов конкретной задачи.

В (6.22) каждая T_u имеет одинаковые значения $\hat{\omega}_{max}$ и число изоморфных таблиц, однако среднее на столбец число значений «*» для T_1 и T_2 равно 0,5, а для T_3 и T_4 — 0,67. Следовательно, выбираем первую пару: T_1 и T_2. Эти таблицы различаются только названием последней строки (S_4 и S_5). Поскольку значения S_5 определяются на реальных объектах труднее и менее надежно, чем значения S_4, то окончательно выбираем T_1.

Определитель есть своего рода хранилище информации, отобранной в соответствии с заданной целью. Ключ — это способ (алгоритм) отыскания некоторой, нужной в данный момент, ее части.

Компактные определители по сравнению с допустимыми предназначены для облегчения поиска нужной информации за счет устранения избыточности. К сожалению, на практике это устранение неизбежно приводит к потере надежности распознавания. Поэтому оптимальный определитель — это такой, который обеспечивает максимальную надежность поиска при минимальном объеме.

Составление ключа также является задачей с противоречивыми требованиями: обеспечить краткость и простоту использования, с одной стороны, и точность — с другой. В общем случае эта задача сводится к поиску экстремума некоторого достаточно сложного функционала.

Не останавливаясь подробно на методологии решения подобных проблем, приведем алгоритм построения ключа, использующего минимальное число шагов (вопросов) и обеспечивающего максимальную точность определения.

Из общих положений теории информации [15] следует, что если имеется q объектов, заданных бинарными описаниями, то минимальное число признаков, обеспечивающиее их безошибочное распознавание, находится по уравнению

$$p_{min} = \frac{\lg q}{\lg 2}. \qquad (6.23)$$

Например, для различения восьми объектов ($q = 8$) требуется не менее трех бинарных признаков, для 2000 объектов — не менее 15 и т. д. Процедура отбора переменных в этом случае сводится к тому, что на первом шаге выбирается признак, делящий совокупность объектов пополам (половина объектов имеет значение этого признака, равное единице, а другая половина — равное нулю), вторым отбирается такой, который делит одновременно обе полученные подсовокупности пополам и т. д. Однако на практике редко удается точно выполнить эти правила, поэтому решение ищется лишь по возможности близким к оптимальному.

Для примера проанализируем T_1 из (6.22). При $q = 4$ согласно (6.23) $p_{min}=2$. Тем не менее имеющаяся информация не позволяет найти решение при таком числе переменных. Признак S_2 делит все четыре подсемейства на две равных совокупности: R_1, R_2 (жало имеется) и R_3, R_4 (жало отсутствует). Признак S_1 делит подсемейства R_1 и R_2, а признак S_4 — R_3 и R_4. Таким образом, в данном примере минимальное число разделяющих признаков равно трем. Оптимальный ключ содержит всего два вопроса: 1) имеется ли в заданном наборе S_2 или $\bar{S}_2$; 2) если S_2, то выясняется значение S_1, а если $\bar{S}_2$, то значение S_4.

Для повышения надежности определения можно использовать следующий прием. T_1 и T_2 имеют эквивалентные структуры из-за того, что в допустимой таблице (6.21) четвертая и пятая строки одинаковы. Это дает возможность объединить T_1 и T_2 в одном определителе

$$T^T_{1,2} = \begin{array}{c|ccc} & S_2 & S_1 & S_{4,5} \\ \hline R_1 & 1 & 1 & * \\ R_2 & 1 & 0 & * \\ R_3 & 0 & 0 & 1 \\ R_4 & 0 & 0 & 0 \end{array}, \qquad (6.24)$$

где индекс T означает операцию транспонирования.

Признаки в таблице (6.24) перечислены в том порядке (слева направо), которого требует ключ: на первом месте стоит S_2. Запись $S_{4,5}$ означает, что в данном столбце помещаются значения S_4 и S_5, которые одинаково изменяются при переходе от объекта к объекту. Это означает также, что для распознавания объекта можно использовать либо S_4, либо S_5, либо S_4 и S_5 одновременно. Прямоугольником обведены значения, знание которых при распознавании желательно, но необязательно.

Допустим, требуется распознать объект

$$S_1 S_2 S_3 S_4 S_5$$
$$R = (?\ 0\ 0\ ?\ 1).$$

Вопросительным знаком отмечены неизвестные значения соответствующих признаков: в распознаваемом объекте неизвестны значения S_1 и S_4 ($x_1. = ?$, $x_4. = ?$). Согласно (6.24) первым анализируем признак S_2: поскольку $x_2. = 0$, то искомый образ находится в нижней половине определителя (R_3 или R_4). В этой половине важно значение S_4 или S_5. Поскольку значение S_4 неизвестно, используем $x_5. = 1$ и устанавливаем, что объект R относится к R_3 (подсемейство Formicinae).

Данный пример иллюстрирует только идею составления ключей, но не трудности, которые встречаются на практике. Поэтому могут оказаться полезными некоторые упрощенные процедуры минимизации описаний. Существо одной из них сводится к следующему.

На первом шаге определяется признак, дающий максимальную энтропию на множестве описаний:

$$E_i = \frac{1}{n}\left[n_1 \lg\left(\frac{n_1}{n}\right) + n_0 \lg\left(\frac{n_0}{n}\right)\right], \qquad (6.25)$$

где n — объем совокупности; n_1 — число описаний, имеющих значение 1 у i-го признака; $n_0 = n - n_1$.

В соответствии с этим исходная совокупность разбивается на две, имеющие n_1 и n_0 описаний. Затем точно таким же образом каждая подсовокупность разбивается на две и так до тех пор, пока в каждой группе останутся лишь неразличимые объекты.

Описанный алгоритм использовался нами в задачах таксономии описаний яйцеклеток рыб (см. § 5.5). Ниже приводится один из вариантов позиционного дихотомического ключа для определения 10 видов пресноводных рыб по морфоструктуре ооцитов на фазе развития E_1.

S_1	S_2	S_3	S_4	
1	1	1	*	горбушка
1	1	0	*	уклей
1	0	1	*	корейская востробрюшка
1	0	0	1	конь пестрый
1	0	0	0	угай
0	1	1	*	верхогляд
0	1	0	1	серебряный карась
0	1	0	0	краснопер
0	0	0	1	сазан
0	0	0	0	ханкайская востробрюшка

С помощью ЭВМ отобраны четыре признака:

S_1 — наименьший диаметр наружных вакуолей равен 8—12 мкм;

S_2 — диаметр вакуолей по окружности среза ооцита, проходящего через внешний ряд вакуолей, 27—75 мкм;

S_3 — диаметр ооцитов 350—450 *мкм*;

S_4 — максимальный диаметр наружных вакуолей более 28 мкм.

Часть II

СТАТИСТИЧЕСКИЕ МЕТОДЫ РАСПОЗНАВАНИЯ

Глава 7

ДОСТОВЕРНОСТЬ РАЗЛИЧИЙ И «РАССТОЯНИЕ» МЕЖДУ ВЫБОРКАМИ

Требование попарного непересечения классов ограничивает применимость детерминистского подхода, и когда это условие не отвечает существу дела, используются статистические методы. Достоинства их заключаются в том, что они допускают наличие ошибок и неполноту знаний о распознаваемых объектах.

Главная цель настоящей главы — попытаться показать на конкретных примерах использование статистического подхода в решении практических задач экологии и зоологической систематики. Как и ранее, мы не будем стремиться к обзору всех существующих в кибернетике и математике методов и направлений, но подробно рассмотрим те из них, которые, по нашему мнению, имеют наибольшую прикладную ценность для биогеографов, морфологов и систематиков.

7.1. Об ошибочном использовании одномерных критериев различия

Специалистам, изучающим биологические объекты, приходится иметь дело с выборками, поэтому после обработки экспериментального материала одной из самых частых процедур является проверка гипотезы о принадлежности выборок к одной и той же генеральной совокупности (ГС). Это и понятно: при статистической однородности выборок задачи классификации теряют практический смысл.

Для проверки гипотезы обычно используются одномерные параметрические критерии, из которых самым популярным в экологии можно назвать критерий Стьюдента. Как это ни удивительно, но ни легкость вычисления, ни популярность этого по-

казателя не спасли многих исследователей от ошибок в его использовании и интерпретации получаемых результатов.

Как пример можно привести рекомендации «Руководства по изучению рыб» [42], в котором для различения средних арифметических значений двух выборок предлагается использовать «дифференцию рядов»:

$$\mathrm{Diff} = \pm \frac{\hat{x}^{(1)} - \hat{x}^{(2)}}{\sqrt{\sigma^2_{\hat{x}_1} + \sigma^2_{\hat{x}_2}}} > 3, \qquad (7.1)$$

где $\hat{x}^{(k)}$ — оценки средних арифметических значений изучаемого признака; $\sigma_{x_k} = \frac{\sigma^{(k)}}{\sqrt{N_k}}$; $\sigma^{(k)}$ — оценка среднеквадратического отклонения; N_k — объем выборки.

Легко заметить аналогию (7.1) с приближенным критерием Стьюдента

$$t = \frac{|\hat{x}^{(1)} - \hat{x}^{(2)}|}{\sqrt{\sigma^2_{\hat{x}_1} + \sigma^2_{\hat{x}_2}}} \geqslant t_\alpha(u), \qquad (7.2)$$

где $t_\alpha(u)$ — критическое значение для доверительной вероятности α и числа степеней свободы u.

Разница между (7.1) и (7.2) состоит в том, что при использовании (7.1) вопросы об учете степеней свободы, а также о формулировке нуль-гипотезы полностью игнорируются. Допустим, например, что при $u = 100$ значение критерия Стьюдента равно 2,7. Тогда согласно (7.1) разность средних значений следует считать недостоверной, хотя в пользу такого решения приходится менее одного случая из 1000. С другой стороны, если бы число степеней свободы было бы небольшим, скажем $u = 5$, и значение критерия — 3,1, то согласно (7.1) разность средних следует считать доказанной, хотя на самом деле она оказывается недостоверной даже при 95% доверительной вероятности.

Другая ошибка, как указано, заключается в игнорировании характера нуль-гипотезы, для проверки которой используется критерий (7.2). Как известно, нормальное распределение признака в ГС характеризуется параметрами $\hat{x}'$ и σ' (выборочными оценками которых являются $\hat{x}$, σ). Относительно $\hat{x}$ и σ могут быть выдвинуты разные гипотезы, в частности $H_0 : \hat{x}^{(1)} = \hat{x}^{(2)}$ безотносительно к тому, имеет место равенство $\sigma^{(1)} = \sigma^{(2)}$ или нет. Для проверки именно этой гипотезы и применяется критерий (7.2). Его использование целесообразно в тех случаях, когда заранее известно, что обе выборки принадлежат к различным ГС и исследователя интересует вопрос: одинаковы или нет средние арифметические значения признака в обеих ГС?

Гипотеза проверяется сравнением расчетного значения с критическими (табличными) при числе степеней свободы

$$u = (N_1 + N_2 - 2)\left[\frac{1}{2} + \frac{(\sigma^{(1)}\sigma^{(2)})^2}{(\sigma^{(1)})^4 + (\sigma^{(2)})^4}\right].$$

Если t оказывается не меньше табличного, то различия считаются достоверными с вероятностью α, где α обычно берется равной 95, 99 или 99,9%.

Типичная ситуация, где возможно применение критерия (7.2), описывается в работе [28]. Карликовые особи кунджи визуально не отличаются от молоди. Принадлежность особи к той или иной совокупности устанавливается после вскрытия: зрелые гонады указывают на принадлежность к карликам, незрелые — к молоди. Таким образом, заранее известно, что выборки карликов и молоди относятся к разным ГС, а существо задачи сводится к тому, чтобы установить, различаются ли особи двух совокупностей морфологически.

На практике гораздо чаще встречается ситуация, когда происхождение выборок неизвестно, и возникает необходимость проверки гипотезы о принадлежности их к одной и той же ГС: $H_0'' : \hat{x}^{(1)} = \hat{x}^{(2)}$, $\sigma^{(1)} = \sigma^{(2)}$. Для проверки H_0'' используется критерий

$$t = \frac{|\hat{x}^{(1)} - \hat{x}^{(2)}|}{\sqrt{\frac{(N_1 - 1)(\sigma^{(1)})^2 + (N_2 - 1)(\sigma^{(2)})^2}{N_1 + N_2 - 2} \cdot \frac{N_1 + N_2}{N_1 N_2}}} \geqslant t_\alpha(u) \qquad (7.3)$$

при степенях свободы $u = N_1 + N_2 - 2$.

Гипотеза отвергается, если (нестрогое) неравенство (7.3) выполняется, в противном случае она остается недоказанной. Последнее обстоятельство часто игнорируется, и многие исследователи ошибочно полагают, что недостоверные различия есть доказательство неразличимости выборочных показателей.

Приведенный перечень ошибок не является исчерпывающим: кроме перечисленных аспектов, формула (7.1) ошибочно рекомендуется и для количественной характеристики величины различий двух выборок. Появились даже инструкции о приписывании таксономического ранга на основе расчетных значений (7.1). Например, в книге [31] рекомендуется приписывать различиям ранг вида, если $\mathrm{Diff} > 7$. Между тем уже поверхностный анализ формулы (7.1) показывает, что при любом малом, но фиксированном различии средних (числитель) увеличением численности выборок (знаменатель) можно довести значение Diff до сколь угодно большой величины. Следовательно, показатель Diff, а также критерии (7.2) и (7.3) не могут характеризовать величину различий выборок, они предназначены исключительно для проверки гипотез о достоверности этих различий.

Ввиду особой важности затронутого вопроса рассмотрим не-олько подробнее способы математически корректной характе-зации величины различий двух выборок.

Самым простым способом было бы вычисление абсолютных ачений разности двух средних арифметических значений

$$\delta = |\hat{x}^{(1)} - \hat{x}^{(2)}|,$$

нако для многих случаев такая характеристика оказывается лоприемлемой. Во-первых, δ — величина именованная и имеет же размерность, что и средние арифметические, поэтому сопоавление различий по разным признакам оказывается неудобным.)-вторых, величина различий выборок из нормальной ГС, естевенно, должна включать не только различия средних значений, и дисперсий. Этим требованиям вполне отвечает «дивергенция» ульбака [3, 5]

$$D^2_{1,2} = \frac{(\hat{x}^{(1)} - \hat{x}^{(2)})^2 [(\sigma^{(1)})^2 + (\sigma^{(2)})^2] + [(\sigma^{(1)})^2 - (\sigma^{(2)})^2]^2}{2(\sigma^{(1)}\sigma^{(2)})^2}. \qquad (7.4)$$

:ли дисперсии равны, $\sigma^{(1)} = \sigma^{(2)} = \sigma$, то

$$D^2_{1,2} = \frac{(\hat{x}^{(1)} - \hat{x}^{(2)})^2}{\sigma^2}. \qquad (7.5)$$

:ли же равны средние значения, $\hat{x}^{(1)} = \hat{x}^{(2)}$, то

$$D^2_{1,2} = \frac{[(\sigma^{(1)})^2 - (\sigma^{(2)})^2]^2}{2(\sigma^{(1)}\sigma^{(2)})^2}.$$

Для иллюстрации вычислений рассмотрим пример. Пусть даны ıраметры двух выборок: $N_1 = 50$, $N_2 = 100$, $(\sigma^{(1)})^2 = 0{,}9$, $^{(2)})^2 = 1{,}0$, $\hat{x}^{(1)} = 2$, $\hat{x}^{(2)} = 3$, требуется охарактеризовать ве-ичину их различий. Согласно (7.4) находим

$$D^2_{1,2} = \frac{(2-3)^2(0{,}9 - 1{,}0) + (0{,}9 - 1{,}0)^2}{2 \cdot 0{,}9} = 1{,}61. \qquad (7.6)$$

Может оказаться, однако, что дисперсии достоверно не разли-аются, и тогда величина $D^2_{1,2}$ оказывается несколько завышенной. роверим гипотезу $H_0 : \sigma^{(1)} = \sigma^{(2)}$, используя критерий Фишера:

$$F = \left[\frac{\sigma^{(1)}}{\sigma^{(2)}}\right]^2 \geqslant F_\alpha(u_1; u_2) \qquad (7.7)$$

ри $u_1 = N_1 - 1$, $u_2 = N_2 - 1$ и $\sigma^{(1)} \geqslant \sigma^{(2)}$.
ля данных этого примера

$$F = \frac{1{,}0}{0{,}9} = 1{,}1,$$

го оказывается ниже табличных значений. Поскольку дис-ерсии значимо не различаются, найдем усредненное значение

$$\sigma^2 = \frac{(N_1 - 1)(\sigma^{(1)})^2 + (N_2 - 1)(\sigma^{(2)})^2}{N_1 + N_2 - 2} = 0{,}97.$$

Следовательно, расстояние между выборками согласно (7.5)

$$D^2_{1,2} = \frac{(2-3)^2}{0{,}97} = 1{,}03.$$

Связь между критерием Стьюдента и «расстоянием» легко заметить, если формулу (7.3) записать в виде

$$t = \sqrt{\frac{N_1 N_2}{N_1 - N_2}} \frac{|\hat{x}^{(1)} - \hat{x}^{(2)}|}{\sigma} = \sqrt{\frac{N_1 N_2}{N_1 + N_2}} D_{1,2}. \qquad (7.8)$$

Величина

$$t^2 = \frac{N_1 N_2}{N_1 + N_2} D^2_{1,2} \qquad (7.9)$$

имеет распределение Фишера с $u_1 = 1$ и $u_2 = N_1 + N_2 - 2$ степенями свободы.

Среди отечественных биологов все более популярным становится CD-коэффициент, предложенный Э. Майром [35]:

$$CD = \frac{\hat{x}^{(1)} - \hat{x}^{(2)}}{\sigma^{(1)} + \sigma^{(2)}}. \qquad (7.10)$$

Этот показатель «расстояния» не имеет математического обоснования (модели), и, следовательно, последствия его использования неизвестны.

7.2. Многомерные показатели

До сих пор рассматривался «одномерный» случай — ситуация, когда на изучаемых объектах измеряется всего лишь один признак. Однако на биологических объектах измеряются, как правило, сразу несколько признаков, поэтому для критериев «различия» и показателей «расстояния» желательно найти многомерные аналоги.

В отличие от нормального одномерного распределения в многомерном, кроме параметров $\hat{x}$ и σ, необходимо учитывать также взаимосвязь между любой парой измеряемых признаков. Количественно эта связь характеризуется коэффициентом парной корреляции

$$\rho_{ik} = \frac{\sum_m^N (x_{im} - \hat{x}_i)(x_{km} - \hat{x}_k)}{\sqrt{\sum_m (x_{im} - \hat{x}_i)^2 \sum_m (x_{km} - \hat{x}_k)^2}}, \qquad (7.11)$$

где x_{im} — значения i-го признака у m-го объекта. Сумму, находящуюся в числителе (7.11), называют суммой смешанных произведений (СП), а СП, деленную на число степеней свободы $(N-1)$, ковариацией. Знаменатель формулы (7.11) — нормирующий множитель, представляющий собой среднее геометрическое СП i-го и k-го признаков в отдельности.

Многомерное нормальное распределение характеризуется следующими параметрами: $\hat{\mathbf{R}}$ — вектор средних арифметических значений всех признаков (он называется также «центром» выборки), $\mathbf{W}$ — ковариационная матрица порядка $p \times p$. Так что роль дисперсии в данном случае играет не отдельное число, а симметричная матрица, содержащая в общем случае $p(p+1)/2$ различающихся значений, причем элементы главной диагонали суть дисперсии каждого признака.

В связи с этим при многомерных распределениях обычные алгебраические операции заменяются операциями над матрицами. Ранее (см. гл. 5) были введены понятия транспонирования, сложения и умножения матриц, теперь введем новую операцию — обращение. Матрица A^{-1} — называется обратной матрице A, если $A \cdot A^{-1} = 1$, где 1 — единичная матрица, у которой все диагональные элементы равны единце, а прочие — нулю. В обычной алгебре этой операции соответствует $a \cdot a^{-1} = 1$. Обращение матрицы достаточно большого порядка — трудоемкая и утомительная процедура, но, к счастью, настолько общеупотребительная, что для любой ЭВМ входит в комплект типовых программ. Методы обращения матриц для ручного счета можно найти в [12, 34].

Пусть даны параметры двух выборок

$$N_1, \hat{\mathbf{R}}_1, \mathbf{W}_1; \qquad N_2, \hat{\mathbf{R}}_2, \mathbf{W}_2.$$

Расстояние между центрами выборок можно найти по аналогии с (7.5):

$$D^2_{1,2} = (\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2)\,\mathbf{W}^{-1}(\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2)^T, \tag{7.12}$$

где $\mathbf{W}$ — обобщенная ковариационная матрица

$$(N_1 + N_2 - 2)\,W = \sum_{m=1}^{N_1} (x_{im} - \hat{x}_i^{(1)})(x_{km} - \hat{x}_k^{(1)}) + \sum_{m=N_1+1}^{N_1+N_2} (x_{im} - \hat{x}_i^{(2)})(x_{km} - \hat{x}_k^{(2)}).$$

Для получения элементов матрицы $\mathbf{W}$ нужно сложить соответствующие элементы СП-матриц и каждую сумму разделить на $N_1 + N_2 - 2$. Расстояние (7.12) называется обобщенным расстоянием Махаланобиса.

По аналогии с (7.9) можно записать

$$T^2_{1,2} = \frac{N_1 N_2}{N_1 + N_2} D^2_{1,2}, \tag{7.13}$$

где T^2 — многомерный аналог критерия Стьюдента (7.3), который называется критерием Хотеллинга. Критерий (7.13) служит для проверки гипотезы о принадлежности двух выборок к одной и той же ГС. В частности, величина

$$F = \frac{N_1 + N_2 - p - 1}{p} \frac{T^2_{1,2}}{N_1 + N_2 - 2} \tag{7.14}$$

имеет F-распределение Фишера с p и $N_1 + N_2 - p - 1$ степенями свободы. При $p = 1$ (одномерный случай) выражение (7.13) становится в точности равным (7.9).

Имеются и другие критерии, основанные на $T^2_{1,2}$, например статистика

$$-a \ln \Lambda_{1,2}, \tag{7.15}$$

где $\Lambda_{1,2} = \dfrac{1}{1 + \dfrac{T^2_{1,2}}{N_1 + N_2 - 2}}$, $a = N_1 + N_2 - 1 - \dfrac{p+2}{2}$,

имеет асимптотическое распределение χ^2 с p степенями свободы. Кроме того, статистика

$$\frac{1 - \Lambda_{1,2}}{\Lambda_{1,2}} \frac{N_1 + N_2 - p}{p} \tag{7.16}$$

имеет F-распределение Фишера с p и $N_1 + N_2 - p$ степенями свободы.

Многомерным аналогом критерия Стьюдента (7.2) при неравных дисперсиях является критерий

$$T^2_{1,2} = \frac{N_1 N_2}{N_1 + N_2} D^2_{1,2}. \tag{7.17}$$

где

$$D^2_{1,2} = (\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2)(\mathbf{W}_1^{-1} + \mathbf{W}_2^{-1})(\hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2)^T. \tag{7.18}$$

Критерий (7.17) имеет асимптотическое распределение χ^2 с p степенями свободы.

В многомерных распределениях в отличие от одномерных проверяется равенство не дисперсий, а ковариационных матриц сравниваемых выборок. Делается это с помощью критерия

$$N_1 \ln \frac{|\mathbf{V}|}{|\mathbf{V}_1|} + N_2 \ln \frac{|\mathbf{V}|}{|\mathbf{V}_2|}, \tag{7.19}$$

где элементы матрицы $\mathbf{V}$ подсчитываются по формуле

$$\mathbf{V} = \frac{N_1 \mathbf{V}_1 + N_2 \mathbf{V}_2}{N_1 + N_2},$$

$|\mathbf{V}|$ — детерминант $\mathbf{V}$ (вычисление детерминантов см. [34]). Величина (7.19) имеет асимптотическое распределение χ^2 с p $(p + 1)/2$ степенями свободы.

Вычисление большей части показателей несложно, если известна величина $D^2_{1,2}$ — обобщенное расстояние Махаланобиса. Поэтому прежде всего опишем алгоритм подсчета $D^2_{1,2}$. Если задан набор исходных измерений, то для получения $D^2_{1,2}$ необходимо вычислить следующее.

1. Средние арифметические значения всех признаков отдельно для каждой выборки, в результате получим $\hat{\mathbf{R}}_1$, $\hat{\mathbf{R}}_2$.

2. СП-матрицы $\mathbf{V}_1$ и $\mathbf{V}_2$ для каждой выборки, где каждый элемент подсчитывается по формуле

$$v_{ik} = \sum_{m=1}^{N_1} (x_{im} - \hat{x}_i^{(1)})(x_{km} - \hat{x}_k^{(1)}), \tag{7.20}$$

а матрицы $\mathbf{V}_2$ по формуле

$$v_{ik} = \sum_{m=N_1+1}^{N_1+N_2} (x_{im} - \hat{x}_i^{(2)})(x_{km} - \hat{x}_k^{(2)}). \tag{7.21}$$

3. Обобщенную СП-матрицу $\mathbf{V} = \mathbf{V}_1 + \mathbf{V}_2$, где сложение ведется поэлементно.

4. Матрицу $\mathbf{V}^{-1}$, причем при $p > 3$ вычисление $\mathbf{V}^{-1}$ вручную нецелесообразно и следует обратиться в вычислительный центр.

5. Матрицу $\mathbf{W}^{-1}$, где

$$\mathbf{W}^{-1} = (N_1 + N_2 - 2)\,\mathbf{V}^{-1}. \tag{7.22}$$

6. Вектор разностей средних арифметических значений

$$\boldsymbol{\delta} = \hat{\mathbf{R}}_1 - \hat{\mathbf{R}}_2.$$

Таблица 7.1

Результаты измерений трех признаков в двух выборках окуней

Номер особи	Угличское водохранилище			Номер особи	Сям-озеро		
	H	h	CD		H	h	CD
1	0,261	0,080	0,261	1	0,275	0,070	0,218
2	0,256	0,078	0,267	2	0,269	0,083	0,200
3	0,261	0,076	0,261	3	0,262	0,083	0,228
4	0,223	0,085	0,255	4	0,236	0,074	0,209
5	0,295	0,084	0,242	5	0,262	0,079	0,211
6	0,258	0,072	0,258	6	0,281	0,085	0,222
7	0,257	0,082	0,247	7	0,281	0,085	0,209
8	0,235	0,082	0,245	8	0,253	0,071	0,221
9	0,250	0,080	0,240	9	0,262	0,081	0,219
10	0,245	0,078	0,265	10	0,269	0,075	0,219
11	0,245	0,078	0,255	11	0,273	0,075	0,218
12	0,235	0,078	0,255	12	0,256	0,073	0,201
13	0,248	0,076	0,238	13	0,281	0,066	0,204
14	0,243	0,075	0,243	14	0,260	0,075	0,214
15	0,231	0,074	0,231	15	0,303	0,085	0,229
16	0,279	0,081	0,225				
17	0,252	0,081	0,252				
18	0,268	0,071	0,241				
19	0,250	0,071	0,250				
20	0,274	0,080	0,204				
$\hat{x}_i^{(1)}$	0,253	0,078	0,247	$\hat{x}_i^{(2)}$	0,269	0,077	0,215

7. Коэффициенты дискриминантной функции (см. гл. 8)

$$\boldsymbol{\omega} = \boldsymbol{\delta}\mathbf{W}^{-1}. \tag{7.23}$$

8. Расстояние Махаланобиса

$$D^2_{1,2} = \boldsymbol{\omega}\boldsymbol{\delta}^T. \tag{7.24}$$

Рассмотрим упрощенный пример из практики морфологических исследований. В двух выборках окуней из Угличского водохранилища и Сям-озера (Карелия) измерялись три признака, которые взяты как отношение к длине тела и условно названы H, h, CD (табл. 7.1) *.

Для подсчета $D^2_{1,2}$ используем все пункты вычислительной схемы.

1. Средние арифметические значения каждого признака:

$$\hat{x}_i^{(1)} = \frac{1}{N_1}\sum_{m=1}^{N_1} x_{im}, \qquad \hat{x}^{(2)} = \frac{1}{N_2}\sum_{m=N_1+1}^{N_1+N_2} x_{im}.$$

Элементы векторов $\hat{\mathbf{R}}_1$ и $\hat{\mathbf{R}}_2$ приведены в последней строке табл. 7.1.

2. Для расчета СП-матриц удобно использовать следующие приемы:

а) все значения каждого признака убавляем на величину соответствующего среднего арифметического значения и вместо исходной табл. 7.1 получаем таблицу со значениями

Номер особи	Угличское водохранилище			Номер особи	Сям-озеро		
	H	h	CD		H	h	CD
1	0,008	0,002	0,014	1	0,006	−0,007	$0,00^3$
2	0,003	0,000	0,020	2	0,000	0,006	$-0,01^5$
3	0,008	−0,002	0,014	3	−0,007	0,006	0,013
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
20	0,021	0,002	−0,043	15	−0,034	0,008	0,014

б) подсчитываем элементы матрицы $\mathbf{V}_1$:

$$\mathbf{V}_1 = \begin{Vmatrix} v_{11} & - & - \\ v_{21} & v_{22} & - \\ v_{31} & v_{32} & v_{33} \end{Vmatrix},$$

где для получения v_{11} перемножаем значения первой строки со значениями первой же (возведение в квадрат):

$$v_{11} = 0{,}008 \cdot 0{,}008 + 0{,}003 \cdot 0{,}003 + \ldots$$
$$\ldots + 0{,}021 \cdot 0{,}021 = 0{,}005817.$$

* Измерения О. А. Поповой.

алее соответственно

$$v_{21} = 0{,}002 \cdot 0{,}008 + 0 \cdot 0{,}003 + \ldots + 0{,}002 \cdot 0{,}021 = 0{,}000120,$$
$$v_{22} = 0{,}002 \cdot 0{,}002 + 0 + \ldots + 0{,}002 \cdot 0{,}002 = 0{,}000314$$

т. д.

в) аналогично подсчитываются элементы другой СП-матрицы; результате получим

$$\mathbf{V}_1 = \begin{Vmatrix} 5{,}8169 & — & — \\ 0{,}1199 & 0{,}3138 & — \\ -0{,}2781 & -0{,}0661 & 4{,}3909 \end{Vmatrix} \cdot 10^{-3},$$

$$\mathbf{V}_2 = \begin{Vmatrix} 3{,}3723 & — & — \\ 0{,}4670 & 0{,}5253 & — \\ 0{,}5856 & 0{,}2430 & 1{,}1304 \end{Vmatrix} \cdot 10^{-3}.$$

оскольку матрицы симметричны, заполняем только нижний реугольник.

3. Обобщенная СП-матрица

$$\mathbf{V} = \mathbf{V}_1 + \mathbf{V}_2 = \begin{Vmatrix} 9{,}1892 & — & — \\ 0{,}5869 & 0{,}8391 & — \\ 0{,}6925 & 0{,}1769 & 5{,}5213 \end{Vmatrix} \cdot 10^{-3}.$$

4. Обращение V можно провести согласно рекомендациям [12], пределив сначала детерминант $|\mathbf{V}|$ по правилу Саррюса, а за-ем — алгебраические дополнения каждого элемента. Нами же спользован метод Гаусса, реализованный на ЭВМ «Мир-2» по тандартной программе:

$$\mathbf{V}^{-1} = \begin{Vmatrix} 115{,}505 & — & — \\ -84{,}406 & 1261{,}460 & — \\ 17{,}190 & -50{,}999 & 184{,}900 \end{Vmatrix}.$$

5. До сих пор вместо ковариаций использовались СП-матрицы, оэтому для получения результата по формуле (7.11) нужно каж-ый элемент $\mathbf{V}^{-1}$ умножить на $N_1 + N_2 - 2 = 33$, получим

$$\mathbf{W}^{-1} = \begin{Vmatrix} 3811{,}66 & -2785{,}38 & 567{,}28 \\ -2785{,}38 & 41628{,}20 & -1682{,}98 \\ 567{,}28 & -1682{,}98 & 6101{,}70 \end{Vmatrix}.$$

Теперь найдем вектор

$$\delta = \begin{Vmatrix} 0{,}253 - 0{,}269 \\ 0{,}078 - 0{,}077 \\ 0{,}247 - 0{,}215 \end{Vmatrix} = \begin{Vmatrix} -0{,}014 \\ 0{,}001 \\ 0{,}032 \end{Vmatrix}.$$

Коэффициенты дискриминантной функции:

$$\omega_1 = -0{,}014 \cdot 3811{,}66 + 0{,}001\,(-2785{,}38) + 0{,}032 \cdot 567{,}28 = \\ = -35{,}49,$$

$$\omega_2 = -0{,}014\,(-2785{,}38) + 0{,}001 \cdot 41628{,}20 + \\ + 0{,}032\,(-1682{,}98) = -10{,}70,$$

$$\omega_3 = -0{,}014 \cdot 567{,}28 + 0{,}001\,(-1682{,}98) + \\ + 0{,}032 \cdot 6101{,}70 = 187{,}14.$$

8. Наконец, расстояние Махаланобиса равно

$$D^2_{1,2} = -0{,}014\,(-35{,}49) + 0{,}001 \cdot (-10{,}70) + \\ + 0{,}032 \cdot 187{,}14 = 6{,}49.$$

Величина $D^2_{1,2}$ — неименованная и не зависит от единиц измерения признаков, но зависит от их числа.

Достоверность морфологических расхождений окуней из двух сравниваемых выборок определим на основе (7.14):

$$T^2_{1.2} = \frac{20 \cdot 15}{20 + 15} \cdot 6{,}49 = 55{,}63,$$

$$F = \frac{20 + 15 - 3 - 1}{3} \cdot \frac{55{,}63}{33} = 17{,}42.$$

При степенях свободы 3 и 31 критическое (табличное) значение $F_{95}\,(3;\ 31) = 4{,}5$. Расчетное значение выше табличного, значит, различия между выборками статистически достоверны.

Поскольку число измерений $N_1 + N_2 = 35$ мало, то применение асимптотически распределенных критериев нецелесообразно. Вместо этого можно использовать критерий (7.16), который имеет точное дисперсионное отношение:

$$\Lambda_{1,2} = \left(1 + \frac{55{,}63}{33}\right)^{-1} = 0{,}37,$$

$$\frac{1 - \Lambda_{1,2}}{\Lambda_{1,2}} \cdot \frac{N_1 + N_2 - p}{p} = 18{,}16,$$

что для 3 и 32 степеней свободы существенно выше критического.

Глава 8

ДИСКРИМИНАНТНЫЙ АНАЛИЗ

В дискриминантном анализе должен быть задан алфавит классов $\mathcal{H}$, указаны признаки S, в пространстве которых эти классы следует отличать друг от друга, и необходимо найти такое решающее правило, которое позволяет с наименьшей ошибкой относить новые объекты к одному из заданных классов.

Для примера можно использовать задачу, поставленную в цитированной ранее работе [28]. Алфавит состоит из двух классов:

H_1 — карликовые особи и H_2 — молодь кунджи, и требуется любую вновь встретившуюся особь на основе учета морфологических признаков с наименьшей вероятностью ошибки отнести к одному из классов.

Допустим, что из всех рыб малого размера карлики в природных условиях встречаются в 9 раз реже, чем некарлики. Тогда любую взятую для распознавания особь на основе только этой информации мы должны считать карликом с вероятностью $P(H_1)=0{,}1$ и молодью — с вероятностью $P(H_2)=0{,}9$. Эти априорные вероятности отражают исходные знания о том, с какой степенью уверенности можно предсказать карликовую или неполовозрелую особь до их действительного появления. Если решение необходимо принять на основе столь малой информации, то разумно воспользоваться решающим правилом: принять H_1, если $P(H_1) > P(H_2)$, и принять H_2 в противном случае. В таких условиях никакое другое правило не дает меньшей вероятности ошибки.

Разумеется, в практической деятельности мы редко ограничиваемся столь малой информацией и для ее увеличения используем какие-либо определяющие признаки. Рассмотрим некоторые случаи, в которых учитывается либо один, либо сразу несколько признаков.

8.1. Одномерные распределения

Чтобы увеличить эффективность распознавания, кроме априорной вероятности классов, необходимо задать условные плотности распределения признака. В самом деле, пусть $p(x \mid H_1)$ — плотность распределения величины x при условии, что она принадлежит первому классу. В этом случае $p(x \mid H_1)$ и $p(x \mid H_2)$ отражают различия в распределении признака у карликов и молоди. Произведя измерения и получив значение x, можно определить апостериорную вероятность $P(H_k \mid x)$ по правилу Байеса:

$$P(H_k \mid x) = \frac{p(x \mid H_k)\, P(H_k)}{\sum p(x \mid H_k)\, P(H_k)}. \tag{8.1}$$

Если $P(H_1 \mid x) > P(H_2 \mid x)$, то нужно выбрать H_1, а если $P(H_1 \mid x) < P(H_2 \mid x)$, то принять решение H_2. Вероятность ошибки при этом

$$P(\text{ош} \mid x) = \begin{cases} P(H_1 \mid x), & \text{если выбрать } H_1, \\ P(H_2 \mid x), & \text{если выбрать } H_2. \end{cases} \tag{8.2}$$

Для каждого нового значения x можно минимизировать ошибку, все время выбирая H_1, если $P(H_1 \mid x), > P(H_2 \mid x)$ и H_2, если $P(H_1 \mid x) < P(H_2 \mid x)$.

Пользуясь правилом Байеса, можно заменить апостериорные вероятности априорными условными плотностями, в результате

чего получаем эквивалентное правило: принять решение H_1, если

$$p(x \mid H_1)\, P(H_1) > p(x \mid H_2)\, P(H_2)$$

или

$$\frac{p(x \mid H_1)}{p(x \mid H_2)} > \frac{P(H_2)}{P(H_1)}. \tag{8.3}$$

Величина $p(x \mid H_1)$ показывает, насколько правдоподобно при данном x наличие H_k. Поэтому отношение

$$p(x \mid H_1)/p(x \mid H_2), \tag{8.4}$$

называется отношением правдоподобия.

Действие решающего правила состоит в том, чтобы разбить пространство признаков на области решения Ω_1 и Ω_2, которые разделяются границей. Рассмотрим несколько случаев построения таких границ для нормального распределения.

Случай 1 : $\sigma^{(1)} = \sigma^{(2)} = \sigma$, т. е. дисперсии признака в обоих классах равны.

Плотность вероятности в точке x для H_1

$$p(x \mid H_1) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(x-\hat{x}^{(1)})^2}{\sigma^2}\right],$$

а для H_2

$$p(x \mid H_2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\frac{(x-\hat{x}^{(2)})^2}{\sigma^2}\right].$$

Отношение правдоподобия согласно (8.3)

$$\frac{p(x \mid H_1)}{p(x \mid H_2)} = \exp\left[\frac{1}{2}(Q_2 - Q_1)\right] > \frac{P(H_2)}{P(H_1)}, \tag{8.5}$$

где

$$Q_k = \frac{(x-\hat{x}^{(k)})^2}{\sigma^2}.$$

Удобно использовать логарифм отношения (8.5)

$$\frac{1}{2}(Q_2 - Q_1) > \ln\frac{P(H_2)}{P(H_1)}$$

или же

$$\frac{\hat{x}^{(1)} - \hat{x}^{(2)}}{\sigma^2}\, x - \frac{(\hat{x}^{(1)} + \hat{x}^{(2)})(\hat{x}^{(1)} - \hat{x}^{(2)})}{2\sigma^2} > \ln\frac{P(H_2)}{P(H_1)}.$$

Обозначая

$$\omega = \frac{\hat{x}^{(1)} - \hat{x}^{(2)}}{\sigma^2}, \qquad \omega_0 = \frac{(\hat{x}^{(1)} + \hat{x}^{(2)})(\hat{x}^{(1)} - x^{(2)})}{2\sigma^2},$$

получим

$$L(x) = \omega x - \omega_0 - \ln\frac{P(H_2)}{P(H_1)} > 0, \tag{8.6}$$

где $L(x)$ называется разделяющей или дискриминантной функцией. Подставив конкретное значение x в дискриминантную функцию, получим величину $L(x)$, которая укажет на принадлежность x к одному из классов: если $L(x) > 0$, то H_1, если $L(x) < 0$, то H_2, если $L(x) = 0$, то точка x находится на разделяющей границе.

Пример. Пусть имеются две обучающие выборки: $N_1 = 100$, $N_2 = 200$, $\hat{x}^{(1)} = 20$, $\hat{x}^{(2)} = 25$, $(\sigma^{(1)})^2 = (\sigma^{(2)})^2 = 4$. Требуется рассортировать по классам три неизвестных объекта: $x = 21$, $y = 22$, $z = 18$.

В предположении, что численности выборок пропорциональны объемам ГС, найдем

$$P(H_1) = \frac{100}{100+200} = 0{,}33, \quad P(H_2) = 1 - P(H_1) = 0{,}67,$$

$$\ln \frac{P(H_2)}{P(H_1)} = 0{,}71.$$

Коэффициенты разделяющей функции

$$\omega = \frac{20-25}{4} = -1{,}25; \qquad \omega_0 = \frac{20^2 - 25^2}{2 \cdot 4} = -28{,}13.$$

Следовательно, условия принадлежности к классу H_1

$$L(x) = 27{,}42 - 1{,}25\,x > 0.$$

Подставляя в эту формулу значения признака трех неизвестных объектов, получим

$$L(x) = 1{,}17; \; L(y) = -0{,}08; \; L(z) = 4{,}92.$$

Первый и третий объекты следует отнести к классу H_1, а второй — к классу H_2.

Если предположить, что объемы ГС одинаковы, то

$$\ln \frac{P(H_2)}{P(H_1)} = 0$$

и разделяющая функция будет иметь вид

$$L(x) = 28{,}13 - 1{,}25\,x > 0.$$

При этих условиях все три распознаваемых объекта относятся к первому классу (H_1).

Случай 2 : $\sigma^{(1)} \neq \sigma^{(2)}$. Отношение правдоподобия в данном случае

$$\ln \frac{\sigma^{(1)}}{\sigma^{(2)}} + \frac{1}{2}(Q_2 - Q_1) > \ln \frac{P(H_2)}{P(H_1)}. \tag{8.7}$$

Раскрывая скобки в левой части неравенства, получим

$$L(x) = \frac{1}{2}\left[\frac{(\sigma^{(1)})^2 (x - \hat{x}^{(2)}) - (\sigma^{(2)})^2 (x - \hat{x}^{(1)})^2}{(\sigma^{(1)}\sigma^{(2)})^2}\right] - \ln \frac{P(H_2)(\sigma^{(1)})^2}{P(H_1)(\sigma^{(2)})^2} > 0, \tag{8.8}$$

где $L(x)$ — область принятия гипотезы о принадлежнс классу H_1.

Пример. В условиях предыдущего примера положим, ч $(\sigma^{(1)})^2 = 4$, $(\sigma^{(2)})^2 = 5$.

Тогда

$$\ln \frac{P(H_2)}{P(H_1)} = 0{,}71.$$

Для неизвестных объектов получим

Q_1	Q_2
$\frac{(21-20)^2}{4} = 0{,}25,$	$\frac{(21-25)^2}{5} = 3{,}2,$
$\frac{(22-20)^2}{4} = 1{,}00,$	$\frac{(22-25)^2}{5} = 1{,}8,$
$\frac{(18-20)^2}{4} = 1{,}00,$	$\frac{(18-20)^2}{5} = 9{,}8,$

Условия принадлежности к классу H_1

$$L(x) = \frac{1}{2}(Q_2 - Q_1) > 0{,}71 + 0{,}11 = 0{,}82.$$

Следовательно,

$$L(x) = 1{,}48 > 0{,}82, \qquad L(y) = 0{,}40 < 0{,}82, \quad L(z) = 4{,}4 > 0{,}82,$$

т. е. два объекта следует отнести к классу H_1.

Обобщения рассмотренных случаев можно осуществлять в следующих направлениях:

1) использование более одного признака;
2) распознавание при числе классов более двух;
3) использование более общего, чем вероятность ошибки, понятия функции потерь.

Несмотря на усложнение формул, связанное с этими обобщениями, принципиальная сторона распознавания не изменяется.

8.2. Байесовская теория решений при многомерных распределениях

При использовании большего числа признаков апостериорная вероятность вычисляется по правилу Байеса:

$$P(H_k \mid \mathbf{R}) = \frac{p(\mathbf{R} \mid H_k) P(H_k)}{\sum p(\mathbf{R} \mid H_k) P(H_k)}, \tag{8.9}$$

где $\mathbf{R}$ — вектор значений признаков объекта $\mathbf{R}'$.

Общий вид нормального распределения задается формулой

$$p(\mathbf{R} \mid H_k) = (2\pi)^{-p/2} |\mathbf{W}_k|^{-1/2} \exp\left(-\frac{1}{2}\mathbf{Q}_k\right), \tag{8.10}$$

где

$$Q_k = (\mathbf{R} - \hat{\mathbf{R}}_k)\,\mathbf{W}_k^{-1}(\mathbf{R} - \hat{\mathbf{R}}_k)^T. \qquad (8.11)$$

Случай 3 : $\mathbf{W}_1 = \mathbf{W}_2 = \mathbf{W} = \sigma_i^2\mathbf{I}$, т. е. когда признаки статистически независимы и имеют одинаковую дисперсию в обоих классах. Ковариационная матрица при независимых признаках становится диагональной, превращаясь в произведение σ_i^2 на единичную матрицу I.

$$\mathbf{W}_k = \begin{Vmatrix} \sigma_{11}^2 & 0 \ldots 0 \\ \cdots\cdots\cdots \\ 0 \quad 0 \ldots \sigma_{pp}^2 \end{Vmatrix}, \qquad \mathbf{W}_k^{-1} = \begin{Vmatrix} 1/\sigma_{11}^2 & 0 \ldots 0 \\ \cdots\cdots\cdots \\ 0 \quad 0 \ldots 1/\sigma_{pp}^2 \end{Vmatrix}.$$

Логарифм отношения правдоподобия равен

$$\frac{1}{2}(Q_2 - Q_1),$$

где Q_k определяется формулой (8.11). Уравнение дискриминантной функции

$$L(\mathbf{X}) = \sum_{i=1}^{p}\left(\frac{\hat{x}_i^{(1)} - \hat{x}_i^{(2)}}{\sigma_i^2}\right)x_i - \frac{1}{2}\sum_{i}^{p}\frac{(\hat{x}_i^{(1)} + \hat{x}_i^{(2)})(\hat{x}_i^{(1)} - \hat{x}_i^{(2)})}{\sigma_i^2} - \ln\frac{P(H_2)}{P(H_1)} > 0 \qquad (8.12)$$

или по аналогии с (8.6)

$$L(\mathbf{X}) = \sum \omega_i x_i - \omega_0 - \ln\frac{P(H_2)}{P(H_1)} > 0, \qquad (8.13)$$

где $L(\mathbf{X})$ — линейная дискриминантная функция.

Имея некоторый объект X, заданный как вектор значений признаков, и подставляя эти значения в (8.13), получим ответ о принадлежности X к классу H_1.

При одинаковых вероятностях классов

$$\ln\frac{P(H_2)}{P(H_1)} = 0,$$

поэтому решающее правило достаточно просто: чтобы определить класс, к которому принадлежит описание X, следует подсчитать расстояния до каждой выборки

$$Q_k = (\mathbf{X} - \hat{\mathbf{R}}_k)\,\mathbf{W}^{-1}(\mathbf{X} - \hat{\mathbf{R}}_k)^T$$

и отнести объект к тому классу, до которого расстояние наименьшее. Граница между двумя областями выражается в виде гиперплоскости, ортогональной («перпендикулярной») отрезку, соединяющему центры классов в пространстве признаков.

Пример. Для прогнозирования уловов Y горбуши в году $t+2$ использовались данные о размерах уловов S_1 и о пропуске на

Таблица 8.1

Данные учета уловов в году t (S_1) и $t+2$ (Y), а также численности производителей на нерестилищах в году t (S_2)

H_1				H_2			
t	S_1	S_2	Y	t	S_1	S_2	Y
1	6	3	29	1	22	10	19
2	8	4	31	2	30	14	9
3	8	3	30	3	28	12	14
4	11	5	29	4	26	10	17
5	10	4	39	5	30	12	12
6	9	4	38	6	35	14	9
7	13	5	28	7	29	12	13
8	14	6	25	8	32	14	10
9	9	4	32	9	27	11	14
10	12	5	30				
11	10	5	31				
12	11	5	31				

естественные нерестилища производителей S_2 в году t. Результаты учетов приведены в табл. 8.1.

Все значения уловов от возврата разбиты на два класса: H_1 — «большие» и H_2 — «маленькие».

Допустим, что имеются данные о родительском стаде: значение признака S_1, равное $x_1^{(\cdot)} = 15$, и признака S_2, равное $x_2^{(\cdot)} = 9$, т. е. $\mathbf{X} = \|\,15\;\;9\,\|$. Требуется узнать, какой улов, «большой» или «маленький», следует ожидать при возврате потомства через два года.

Средние значения и дисперсии признаков равны

$$\|\,\hat{x}_i^{(1)}\,\| = \|\,10{,}1\;\;4{,}42\,\|, \qquad \|\,\hat{x}_i^{(2)}\,\| = \|\,28{,}8\;\;12{,}1\,\|,$$
$$\|\,(\sigma_i^{(2)})^2\,\| = \|\,12{,}5\;\;2{,}46\,\|, \qquad \|\,(\sigma^{(1)})^2\,\| = \|\,5{,}24\;\;0{,}82\,\|.$$

В первом приближении взаимосвязь признаков не учитываем и, кроме того, полагаем, что соответствующие дисперсии в обучающих выборках равны. Усреднение последних проводим по формуле

$$\frac{(N_1-1)(\sigma_i^{(1)})^2 + (N_2-1)(\sigma^{(2)})^2}{N_1+N_2-2},$$

что дает $\|\,\sigma_i^2\,\| = \|\,8{,}3\;\;1{,}5\,\|$. Считаем, что $P(H_1) = P(H_2)$, и по формуле (8.12) определяем

$$L(\mathbf{X}) = \frac{10{,}1-28{,}8}{8{,}3}S_1 + \frac{4{,}4-12{,}1}{1{,}5}S_2 + \frac{1}{2}\left(\frac{10{,}1^2-28{,}8^2}{8{,}3} + \frac{4{,}4^2-12{,}1^2}{1{,}5}\right) =$$
$$2{,}25\,S_1 - 5{,}13\,S_2 + 86{,}2 > 0. \qquad (8.14)$$

Подставляя в (8.14) значения $X = \| 15\ 9 \|$, получим

$$L(X) = -79{,}9 + 86{,}2 = 6{,}28 > 0,$$

т. е. улов следует ожидать «большим».

Формулу (8.14) в данном примере следует расценивать как прогнозирующую систему для данной популяции. Подставляя каждый раз новые значения S_1 и S_2, можно прогнозировать улов с наименьшей возможной в заданных условиях ошибкой.

Разумеется, на практике прогнозирующая система должна учитывать больший материал обучения, большее число признаков, большее число классов и разные наборы решающих правил.

Случай 4 : $W_1 = W_2 = W$ также является сравнительно простым: ковариационные матрицы для классов одинаковы

$$L(X) = \frac{1}{2}(Q_2 - Q_1) = XW^{-1}(\hat{R}_1 - \hat{R}_2)^T - \frac{1}{2}(\hat{R}_1 + \hat{R}_2)W^{-1}(\hat{R}_1 - \hat{R}_2)^T > \frac{P(H_2)}{P(H_1)}. \qquad (8.15)$$

Если вероятности классов одинаковы, то правая часть неравенства равна нулю, и для классификации вектора признаков достаточно определить махаланобисово расстояние до центра каждого класса

$$(X - \hat{R}_k)W^{-1}(X - \hat{R}_k)^T$$

и отнести вектор к тому классу, до которого это расстояние минимально. Дискриминантная функция в данном случае линейна, а граница между двумя областями не обязательно ортогональна отрезку, соединяющему центры классов.

Пример. СП-матрицы для обучающих выборок предыдущего примера

$$V_1 = \begin{Vmatrix} 56{,}72 & 20{,}48 \\ 20{,}48 & 8{,}96 \end{Vmatrix}, \qquad V_2 = \begin{Vmatrix} 109{,}56 & 42{,}22 \\ 42{,}22 & 26{,}89 \end{Vmatrix}.$$

Усредненную СП-матрицу

$$V = \begin{Vmatrix} 79{,}37 & 29{,}80 \\ 29{,}80 & 16{,}64 \end{Vmatrix}$$

используем для построения прогнозирующей системы по формуле (8.15). Детерминант V равен

$$|V| = 79{,}37 \cdot 16{,}64 - 29{,}80 \cdot 29{,}80 = 432{,}68,$$

что дает возможность определить

$$V^{-1} = \begin{Vmatrix} 0{,}038 & -0{,}069 \\ -0{,}069 & 0{,}183 \end{Vmatrix}, \qquad W^{-1} = \begin{Vmatrix} 0{,}722 & -1{,}311 \\ -1{,}311 & 3{,}477 \end{Vmatrix},$$

$$W^{-1}\delta^T = \begin{Vmatrix} 0{,}722 & -1{,}311 \\ -1{,}311 & 3{,}477 \end{Vmatrix} \times \begin{Vmatrix} -18{,}7 \\ -7{,}7 \end{Vmatrix} = \begin{Vmatrix} -3{,}42 \\ -2{,}28 \end{Vmatrix}.$$

Используя полученные коэффициенты как фрагмент решающего правила $-3{,}42\,S_1 - 2{,}28\,S_2$ и подставляя вместо переменных S_1 и S_2 средние значения каждой выборки, получим

$$M_1 = -3{,}42 \cdot 10{,}1 - 2{,}28 \cdot 4{,}4 = -44{,}57,$$

$$M_2 = -3{,}42 \cdot 28{,}8 - 2{,}28 \cdot 12{,}1 = -126{,}16,$$

$$\Gamma = 1/2\,(M_1 + M_2) = -85{,}37.$$

Таким образом, прогнозирующая система при сделанных предположениях имеет вид

$$-3{,}42\,S_1 - 2{,}28\,S_2 + 85{,}37. \qquad (8.16)$$

Задавая ситуацию $X = \| 19\ 5 \|$, получим

$$L(X) = 12{,}65 > 0,$$

т. е. прогноз совпадает с полученным ранее.

Случай 5 : $W_1 \neq W_2$. Это наиболее общий случай для нормальных распределений: ковариационные матрицы различны.

Дискриминантная функция

$$L(X) = \ln \sqrt{\frac{|W_2|}{|W_1|}} + \frac{1}{2}(Q_2 - Q_1) > \ln \frac{P(H_2)}{P(H_1)}. \qquad (8.17)$$

Как и ранее, если априорные вероятности равны, то область принятия решения H_1 определяется неравенством $L(X) > 0$. При $L(X) < 0$ принимается H_2. Нелинейная решающая граница (гиперповерхность второго порядка) проходит через точки, для которых $L(X) = 0$.

Границы в этом случае называются гиперквадриками, поскольку они могут принимать любую из общих форм гиперплоскостей: гиперсфер, гиперэллипсоидов, гиперпараболоидов и гипергиперболоидов.

Пример. Продолжим изучение возможностей построения прогнозирующей системы уловов. На этот раз не будем усреднять ковариационные матрицы, которые для каждой выборки равны соответственно

$$W_1 = \begin{Vmatrix} 5{,}16 & 1{,}86 \\ 1{,}86 & 0{,}81 \end{Vmatrix}, \qquad W_2 = \begin{Vmatrix} 13{,}70 & 5{,}28 \\ 5{,}28 & 3{,}36 \end{Vmatrix}.$$

Обратные матрицы равны

$$W_1^{-1} = \begin{Vmatrix} 0{,}253 & -0{,}396 \\ -0{,}396 & 1{,}034 \end{Vmatrix}, \qquad W_2^{-1} = \begin{Vmatrix} 0{,}81 & -1{,}85 \\ -1{,}85 & 5{,}07 \end{Vmatrix}.$$

Расстояния объекта $X = \| 15\ 9 \|$ до центров обучающих выборок

$$Q_1 = \| -4{,}9 \quad -4{,}6 \| \times \begin{Vmatrix} 0{,}253 & -0{,}396 \\ -0{,}396 & 1{,}034 \end{Vmatrix} \times \begin{Vmatrix} -4{,}9 \\ -4{,}6 \end{Vmatrix} = 10{,}1,$$

$$Q_2 = \| -13{,}8 \quad -7{,}7 \| \times \begin{Vmatrix} 0{,}81 & -1{,}85 \\ -1{,}85 & 5{,}07 \end{Vmatrix} \times \begin{Vmatrix} -13{,}8 \\ -7{,}7 \end{Vmatrix} = 61{,}6,$$

где Q_k определяются формулой (8.11).

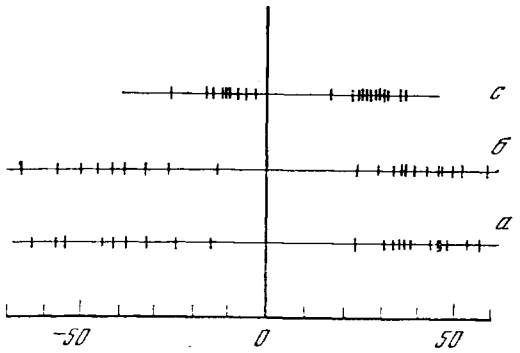


Рис. 8.1. Дискриминация обучающих выборок для прогноза уловов на основе различных дискриминантных функций

а — случай 3; *б* — случай 4; *в* — случай 5

Детерминанты ковариационных матриц равны

$$|W_1| = 0{,}73, \quad |W_2| = 18{,}18.$$

Следовательно, согласно (8.17)

$$L(X) = \frac{1}{2}\left(\ln\frac{18{,}18}{0{,}73} + 61{,}6 - 10{,}1\right) = 27{,}4 > 0.$$

Таким образом, прогноз остается прежним: в заданной ситуации следует ожидать «больших» уловов.

Для сравнения эффективности прогнозирующих систем (8.14), (8.16) и (8.17) проведем дискриминацию представителей обучающих выборок с помощью соответствующих дискриминантных функций. Результаты расчетов приведены на рис. 8.1. Сравнивая визуально соотношения внутривыборочного и межвыборочного разброса для трех вариантов, можно убедиться в том, что в данном случае усложнение прогнозирующего аппарата дает весьма небольшое улучшение дискриминации. В общем случае это не так, и качество дискриминации существенно улучшается при последовательном переходе от случая 3 к случаю 5.

8.3. Другие приложения решающего правила Байеса. Независимые бинарные признаки

Приведенные случаи не исчерпывают всех возможностей применения байесовского решающего правила для распознавания образов. В частности, это может касаться понятия «функции потерь», как более общего по сравнению с вероятностью ошибки распознавания и рассмотрения ситуаций с числом классов более двух.

Пусть

$$\mathcal{H} = \{H_1, \ldots, H_l\}$$

и $A = \{\alpha_1, \ldots, \alpha_s\}$ — конечное множество из возможных действий; $\lambda(\alpha_i \mid H_k)$ — потери, связанные с принятием действия для образа H_k.

Как и ранее,

$$P(H_k \mid X) = \frac{p(X \mid H_k)\, P(H_k)}{\sum_l p(X \mid H_k)\, P(H_k)}. \tag{8.18}$$

Допустим, что мы наблюдаем некоторое описание X и собираемся произвести действие α_i. Если распознаваемый образ есть H_k, то мы понесем определенные потери $\lambda(\alpha_i \mid H_k)$. Так как $P(H_k \mid X)$ есть вероятность того, что распознаваемый образ действительно H_k, то ожидаемые потери, связанные с действием α_i,

$$F(\alpha_i \mid X) = \sum \lambda(\alpha_i \mid H_k)\, P(H_k \mid X). \tag{8.19}$$

Ожидаемые потери называют также риском, а $F(\alpha_i \mid X)$ — условным риском. Всякий раз при наблюдении конкретного значения X ожидаемые потери сводятся к минимуму выбором действия, минимизирующего условный риск. Тогда и для всех X в длинном ряду испытаний общий условный риск будет минимальным.

Обозначим $\lambda_{uk} = \lambda(\alpha_u \mid H_k)$ — потери вследствие принятия решения H_u, когда на самом деле распознаваемый образ есть H_k. Для случая двух классов

$$\begin{aligned} F(\alpha_1 \mid X) &= \lambda_{11} P(H_2 \mid X) + \lambda_{12} P(H_2 \mid X), \\ F(\alpha_2 \mid X) &= \lambda_{21} P(H_1 \mid X) + \lambda_{22} P(H_1 \mid X). \end{aligned} \tag{8.20}$$

Решающее правило с минимальным риском заключается в выборе H_1, если

$$F(\alpha_1 \mid X) < F(\alpha_2 \mid X)$$

или

$$(\lambda_{21} - \lambda_{11})\, P(H_1 \mid X) > (\lambda_{12} - \lambda_{22})\, P(H_2 \mid X).$$

Пользуясь правилом Байеса для априорных вероятностей и полагая, что потери в случае ошибки больше, чем при правильном ответе

$$(\lambda_{21} - \lambda_{11})\, p(X \mid H_1)\, P(H_1) > (\lambda_{12} - \lambda_{22})\, p(X \mid H_2)\, P(H_2),$$

получим отношение правдоподобия

$$\frac{p(X \mid H_1)}{p(X \mid H_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(H_2)}{P(H_1)} \tag{8.21}$$

Байесовское правило может быть распространено на случай дискретных переменных, в частности на бинарные признаки.

Пусть компоненты векторов равны либо 0, либо 1; классу H_1 принадлежат N_1 и классу H_2 — N_2 объектов. Обозначим

$$p_i = \frac{n_i^{(1)}}{N_1}, \qquad q_i = \frac{n_i^{(2)}}{N_2},$$

где $n_i^{(k)}$ — число векторов (описаний) первого класса, для которых i-я компонента равна 1. Далее естественно предположить, что

$$P(H_1) = \frac{N_1}{N_1+N_2}, \qquad P(H_2) = 1 - P(H_1).$$

Как и ранее, $p(X \mid H_k)$ — вероятность появления описания X при условии, что оно принадлежит классу H_k, а $P(H_k \mid X)$ — вероятность того, что X принадлежит классу H_k.

Для двух классов согласно (8.1) естественно выбрать

$$\ln \frac{p(X \mid H_1)}{p(X \mid H_2)} > \frac{1-P(H_1)}{P(H_2)} \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}. \qquad (8.22)$$

Если признаки независимы, то плотность вероятностей можно записать как произведения вероятностей для компонент вектора X

$$p(X \mid H_1) = \prod_{i=1}^{p} p_i^{x_i}(1-p_i)^{1-x_i}, \qquad p(X \mid H_2) = \prod_{i=1}^{p} q_i^{x_i}(1-q_i)^{1-x_i}.$$

Подставляя эти выражения в (8.22), получим уравнение разделяющей функции

$$L(X) = \sum_i^p \left[x_i \ln \frac{p_i}{q_i} + (1-x_i) \ln \frac{1-p_i}{1-q_i}\right] - \frac{1-P(H_1)}{P(H_1)} \frac{\lambda_{12}-\lambda_2^2}{\lambda_{21}-\lambda_{11}} > 0.$$

Преобразуем выражение в квадратных скобках:

$$x_i \ln \frac{p_i}{q_i} + (1-x_i) \ln \frac{1-p_i}{1-q_i} = x_i \ln \frac{p_i}{q_i} + \ln \frac{1-p_i}{1-q_i} + x_i \ln \frac{1-p_i}{1-q_i}.$$

Получим

$$L(X) = \sum_i^p x_i \ln \frac{p_i(1-q_i)}{q_i(1-p_i)} + \sum_i^p \ln \frac{1-p_i}{1-p_i} - \ln \frac{1-P(H_1)}{P(H_1)} \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} > 0.$$

Обозначим

$$\omega_i = \ln \frac{p_i(1-q_i)}{q_i(1-p_i)}, \qquad \omega_0 = \sum \ln \frac{1-p_i}{1-q_i}.$$

Если положить

$$P(H_1) = P(H_2), \qquad \lambda_{11} = \lambda_{22} = 0, \qquad \lambda_{12} = \lambda_{21} = 1,$$

то область решения H_1 определяется неравенством

$L(X) = \sum \omega_i x_i + \omega_0 > 0$, т. е. дискриминантная функция линейна.

Пример. Пусть имеются две выборки: $N_1 = 4$, $N_2 = 5$, число признаков $p = 6$, $P(H_1) = \frac{4}{4+5} = 0{,}44$, $P(H_2) = 0{,}56$; описание объектов

H_1						H_2						
R_1	R_2	R_3	R_4	p_i	$1-p_i$	R_1	R_2	R_3	R_4	R_5	q_i	$1-q_i$
1	1	1	0	0,75	0,25	0	1	0	0	0	0,20	0,80
0	1	0	0	0,25	0,75	0	0	0	0	1	0,20	0,80
0	0	0	1	0,25	0,75	0	0	0	1	0	0,20	0,80
0	0	1	0	0,25	0,75	1	1	1	1	0	0,80	0,20
0	1	0	1	0,50	0,50	1	1	0	1	0	0,40	0,60
0	0	1	1	0,50	0,50	0	0	1	0	1	0,40	0,60

Находим коэффициенты дискриминантной функции

i	ω_i	$\ln \frac{1-p_i}{1-q_i}$	i	ω_i	$\ln \frac{1-p_i}{1-q_i}$	
1	2,48	−1,16	4	2,48	1,32	
2	0,29	−0,06	5	0,41	0,18	$\omega_0 = -0{,}32$
3	0,29	−0,06	6	0,41	0,18	

Вектор $X = \| 1\ 0\ 0\ 0\ 1\ 1 \|$ необходимо с наименьшей вероятностью ошибки отнести к одному из классов: H_1 или H_2. Подставляя X в выражение

$$L(x) = \sum \omega_i x_i + \omega_0 - \ln \frac{P(H_2)}{P(H_1)} > 0,$$

получим

$$L(x) = 3{,}30 - 0{,}32 = 2{,}98 > 0,$$

т. е. распознаваемый объект относится к классу H_1.

Заметим, что если $p_i = q_i$, то величина x_i не несет информации о принадлежности к классам и $\omega_i = 0$. В случае $p_i > q_i$ имеем $1 - p_i < 1 - q_i$, так что ω_i положителен и для x_i имеем ω_i «голосов» в пользу H_1. Кроме того, при любом постоянном $q_i < 1$ чем больше p_i, тем больше ω_i. С другой стороны, при $p_i < q_i$ величина ω_i становится отрицательной и мы имеем $|\omega_i|$ «голосов» в пользу H_2.

8.4. Метод наименьших квадратов. Отбор информативных признаков

Дискриминация представителей различных ГС может формально рассматриваться как регрессионный анализ, основанный на методе наименьших квадратов. Этот подход удобно иллюстрировать примером прогноза уловов, рассмотренным в параграфе 8.3.

Каждое значение прогнозируемой переменной $\hat{Y}$ определяется как функция независимых переменных

$$\hat{Y} = f(S_1, S_2),$$

причем самым простым видом функции можно считать линейную зависимость

$$\hat{Y} = \omega_0 + \omega_1 S_1 + \omega_2 S_2$$

или при p переменных

$$\hat{Y} = \omega_0 + \sum_i^p \omega_i S_i. \tag{8.23}$$

В более общем случае можно вводить нелинейные члены: S_i^2, $S_i \cdot S_j$ и т. п., например, для двух переменных полином второго порядка

$$\hat{Y} = \omega_0 + \omega_1 S_1 + \omega_2 S_2 + \omega_3 S_1^2 + \omega_4 S_2^2 + \omega_5 S_1 S_2, \tag{8.24}$$

где значения $S_1 S_2 = \{x_{1j} \cdot x_{2j} \mid j \in J\}$ можно считать «новыми» переменными, т. е. значения «новых» признаков вводятся посредством попарного перемножения значений заданных. Уравнение (8.24) нелинейно относительно S_i, но линейно относительно $\hat{Y}$.

Ситуация примера примечательна тем, что значения прогнозируемой переменной $\hat{Y}_j$ являются числовыми и все $\hat{Y}_j$ могут быть упорядочены. Если же значения зависимой переменной не могут быть упорядочены, то регрессионный анализ эффективен при распознавании только двух классов. Соответственно этому всем $y_j \in H_1$ приписывают значения N_1/N, а $y_j \in H_2$ — значения $(-N_2/N)$, где $N = N_1 + N_2$.

Коэффициенты ω_i находятся по формулам

$$\omega_i = \sum v^{ij} v_{0i}, \tag{8.25}$$

где v^{ij} — элементы матрицы $\mathbf{V}^{-1}$,

$$v_{ij} = \sum_k^N (x_{ik} - \hat{x}_i)(x_{jk} - \hat{x}_j), \tag{8.26}$$

$$\hat{x}_i = \frac{1}{N} \sum_k^N x_{ik}, \tag{8.27}$$

$$v_{0i} = \sum_k^N (x_{ik} - \hat{x}_i)(y_j - \hat{Y}_0),$$

$$\hat{Y}_0 = \frac{1}{N} \sum_j^N y_j,$$

$$v_{00} = \sum (y_j - \hat{Y}_0)^2,$$

$$\omega_0 = \hat{Y}_0 - \sum \omega_i \hat{x}_i, \tag{8.28}$$

т. е. значения матриц $\mathbf{V}$ и $\mathbf{V}_0$ получаются при объединении всех выборок в одну.

Множественная корреляция, которая служит мерой точности предсказания или мерой связи между зависимой и независимыми переменными, оценивается величиной

$$R^2 = \frac{1}{v_{00}} \sum \omega_i v_{0i}. \tag{8.29}$$

Используя технику дисперсионного анализа, можно проверить гипотезу о значимости связи с помощью критерия

$$F = \frac{R^2}{1 - R^2} \frac{N - p - 1}{p}, \tag{8.30}$$

который имеет F-распределение Фишера с p и $N - p - 1$ степенями свободы.

При незначимой связи задача прогнозирования (распознавания) теряет практический смысл.

Продолжим изучение задачи прогноза уловов. Согласно формулам (8.25) — (8.27) необходимо вычислить элементы матрицы $\mathbf{V}$ при объединении выборок в одну. В данном случае нет необходимости все расчеты делать заново, так как элементы $\mathbf{V}$ определяются через известные величины

$$v_{ij} = v_{ij}^{(1)} + v_{ij}^{(2)} + N_1 (\hat{x}_i^{(1)} - \hat{x}_i)(x_j^{(1)} - \hat{x}_j) + N_2 (\hat{x}_i^{(2)} - \hat{x}_i)(\hat{x}_j^{(2)} - \hat{x}_j), \tag{8.31}$$

где

$$\hat{x}_i = \frac{N_1 \hat{x}_i^{(1)} + N_2 \hat{x}_i^{(2)}}{N}.$$

Расчеты показывают, что

$$\mathbf{V} = \begin{Vmatrix} 1964{,}7 & 796{,}3 \\ 796{,}3 & 339{,}2 \end{Vmatrix}; \quad \mathbf{V}^{-1} = \begin{Vmatrix} 0{,}0104 & -0{,}0246 \\ -0{,}0246 & 0{,}0608 \end{Vmatrix};$$

$$\mathbf{V}_0 = \| -1740{,}7 \quad -714{,}9 \|; \quad v_{00} = 1949{,}8;$$

$$\hat{x}_1 = 18{,}1; \quad \hat{x}_2 = 7{,}7; \quad \hat{y}_0 = 23{,}3.$$

Согласно (8.25) коэффициенты ω_i определяются как

$$\begin{Vmatrix} 0{,}0104 & -0{,}0246 \\ -0{,}0246 & 0{,}0608 \end{Vmatrix} \times \begin{Vmatrix} -1740{,}7 \\ -714{,}9 \end{Vmatrix} = \begin{Vmatrix} -0{,}406 \\ -1{,}062 \end{Vmatrix}.$$

Свободный член в (8.28) равен

$$\omega_0 = 23{,}3 + 18{,}1 \cdot 0{,}406 + 7{,}7 \cdot 1{,}062 = 38{,}9.$$

В результате получаем эмпирическое уравнение

$$\hat{Y} = 38{,}9 - 0{,}406\, S_1 - 1{,}062\, S_2, \tag{8.32}$$

которое дает возможность по известным значениям S_1 и S_2 определять возможные значения $\hat{Y}$. Однако, прежде чем это делать, необходимо убедиться, что по крайней мере один ω_i значимо отличается от нуля. На основе (8.29) и (8.30) получаем

$$R^2 = \frac{1581{,}9}{1949{,}8} \approx 0{,}81, \qquad F = \frac{0{,}81}{0{,}19}\,\frac{18}{2} = 38{,}4.$$

При 2 и 18 степенях свободы табличное значение $F_{95}\,(2;\,18) = 19{,}4$ ниже расчетного, следовательно, связь между уловами и учитываемыми факторами следует признать высокодостоверной.

Теперь используем для прогноза вектор $\mathbf{X} = \| 19 \;\; 5 \|$:

$$\hat{Y} = 38{,}9 - 0{,}406 \cdot 19 - 1{,}062 \cdot 5 = 23{,}2, \tag{8.33}$$

т. е. ожидаемый улов явно относится к классу «больших».

Регрессионный анализ дает возможность построить доверительный интервал для средних значений прогнозируемой переменной Y. Величина

$$t = \frac{|y - \hat{Y}|}{\sigma_{\hat{y}}} \tag{8.34}$$

имеет t-распределение Стьюдента с $N - p - 1$ степенями свободы. Значение дисперсии в знаменателе (8.34) оценивается как

$$\sigma^2_{\hat{y}} = \sigma^2_{00}\left[\frac{1}{N} + \sum_i \sum_j (x_i - \hat{x}_i)(x_j - \hat{x}_j)\, v^{ij}\right], \tag{8.35}$$

где

$$\sigma^2_{00} = \frac{v_{00} - \sum \omega_i v_{0i}}{N - p - 1}$$

называют остаточной дисперсией $\hat{Y}$.

Для условий примера $\sigma^2_{00} = 20{,}5$, а выражение в квадратных скобках (8.35) равно

$$\frac{1}{25} + \begin{Vmatrix} -3{,}1 \\ 1{,}3 \end{Vmatrix} \times \begin{Vmatrix} 0{,}0104 & -0{,}0246 \\ -0{,}0246 & 0{,}0608 \end{Vmatrix} = 0{,}15.$$

Табличное значение $t_{95}\,(18) = 2{,}1$. Тогда из (8.34) находим

$$y = \hat{Y} \pm 2{,}1 \cdot \sqrt{20{,}5 \cdot 0{,}15} = 23{,}2 \pm 3{,}7,$$

т. е. среднее значение прогнозируемого улова находится в пределах от 19,5 до 26,9 тыс. ц.

В задаче регрессионного анализа можно поставить вопрос: увеличивает ли точность прогноза каждая переменная в отдельности? Для рассматриваемых условий, например, интересно выяснить: нужно ли учитывать переменную S_2 в дополнение к S_1 (измерения S_2 значительно труднее, чем S_1)? Такой вопрос эквивалентен проверке нуль-гипотезы : $\omega_2 = 0$. Для ее проверки используется величина

$$F = \frac{\omega_i^2}{v_{ii}\sigma^2_{00}}, \tag{8.36}$$

которая имеет F-распределение Фишера с 1 и $N - p - 1$ степенями свободы. Подставляя в (8.36) ранее вычисленные значения, получаем

$$F = \frac{1{,}06^2}{0{,}0608 \cdot 20{,}5} = 0{,}91,$$

что при 1 и 18 степенях свободы ниже табличного значения. Следовательно, для построения прогнозирующей системы уловов переменную S_2 (численность производителей на нерестилищах) можно было бы и не учитывать.

Попробуем проверить гипотезу: $\omega_1 = 0$. Согласно (8.36)

$$F = \frac{0{,}406^2}{0{,}0104 \cdot 20{,}5} = 0{,}77,$$

что указывает на незначимость переменной S_1 как дополнительной к S_2. Тот факт, что каждая в отдельности переменная оказывается незначимой, а совместный их учет приводит к значимому прогнозу, является следствием большой корреляции значений S_1 и S_2. Так что при построении прогнозирующей системы какой-либо одной из этих переменных можно пренебречь.

Если одна из независимых переменных вычеркивается из исходных данных, то все коэффициенты ω_i пересчитываются заново. Однако не нужно делать все с самого начала. При отбрасывании u-го признака элементы новой (редуцированной) матрицы находятся по правилам

$$v^{ij}_{\text{ред}} = v^{ij} - v^{iu}v^{ju}/v^{uu}. \tag{8.37}$$

Таким образом, если имеются сомнения в целесообразности использования в прогнозирующей системе какой-либо переменной S_u, то с помощью критерия (8.36) следует установить значимость S_u и отбросить ее, если она оказалась незначимой. Заметим, что проверка гипотезы $\omega_u = 0$ осуществляется при допущении, что оставшиеся ω_i имеют фиксированные значения. Поэтому при p переменных последовательное удаление из них наименее значимых не обязательно приводит к наилучшей системе распознавания. Ведь каждая из отброшенных переменных незначима в отдельности, но при совместном учете они могут оказаться высокозначимыми.

В тех случаях, когда по каким-либо содержательным условиям задачи под сомнение ставится целесообразность измерения сразу m признаков, следует ориентироваться на применение критерия (8.30): если оставшиеся $p - m$ признаков оказались значимыми, то сомнительные m могут быть отброшены.

Если же с содержательных позиций оценить целесообразность изъятия каких-либо конкретных m признаков невозможно, то следует ориентироваться на формальное правило, согласно которому отбрасываются те признаки, которые в наименьшей степени влияют на величину множественной корреляции R^2. Допустим, мы отбрасываем m признаков из p. Тогда вся процедура оценки R^2 выполняется C_p^m раз. В сравнительно несложном варианте выбора 20 признаков из 50 число $C_{50}^{20} \approx 10^{13}$, что невозможно выполнить в разумное время даже с помощью современных ЭВМ. Чаще всего для этих целей используются приближенные методы. Например, первым в распознающую систему включается признак, который дает наибольший R^2 при индивидуальном учете, вторым включается тот, который совместно с первым отобранным дает наибольший R^2, и т. д.

При другом, более грубом подходе первым удаляется наименее значимый признак, вторым — наименее значимый из оставшихся и т. д. до тех пор, пока R^2 остается значимым. После каждого отбрасывания все необходимые параметры пересчитываются заново.

В распознающих системах, основанных на дискриминантных функциях и расстоянии Махаланобиса, можно использовать нуль-гипотезу о том, что при отбрасывании m признаков оставшиеся $p - m$ признаков не несут добавочной информации относительно разделения. Для проверки гипотезы используется дисперсионное отношение

$$F = \frac{N - p - 1}{p - m} \frac{N_1 N_2 (D_p^2 - D_m^2)}{(N_1 + N_2)(N - 2) + N_1 N_2 D_m^2}, \tag{8.38}$$

где D_m^2 — махаланобисово расстояние, основанное на m признаках. Величина (8.38) имеет F-распределение Фишера с $p - m$ и $N - p - 1$ степенями свободы.

Все трудности, связанные с отбором m признаков из p, имеют такой же характер, как и в регрессионном анализе. Различие заключается лишь в том, что в данном случае наименее значимыми считаются признаки, которые в наименьшей степени влияют на величину D_{p-m}^2.

Последовательное отбрасывание наименее значимых переменных совпадает с процедурой оптимального отбора в одном частном случае, когда учитываемые признаки независимы (СП-матрицы диагональны). В самом деле, расстояние Махаланобиса в этом

случае равно

$$D_{1,2}^2 = \sum_i^p \left(\frac{\hat{x}_i^{(1)} - \hat{x}_i^{(2)}}{\sigma_i} \right)^2, \tag{8.39}$$

откуда видно, что вклад i-го признака в величину $D_{1,2}^2$ есть i-е слагаемое в (8.39). Чем больше эта величина по отношению к другим слагаемым, тем больше доля участия i-го признака в различении классов.

В. Ю. Урбахом [64] предложен способ оценки уменьшения разделительной мощности дискриминантной функции после отбрасывания одного признака. Если удаляется i-й признак, то величина махаланобисова расстояния уменьшается на ω_i^2/σ^{ii}, где ω_i — коэффициенты дискриминантной функции, σ^{ii} — элементы матрицы $\mathbf{W}^{-1}$. С помощью критерия

$$\hat{D}^2 = \frac{D_{1,2}^2 + (N - 2^p - 2) N}{(N - p - 2)(N_1 + N_2)}$$

устанавливается, что при справедливости неравенства

$$\frac{\omega_i^2}{\sigma^{ii}} < \hat{D}^2$$

i-й признак является «вредным» и должен быть отброшен.

Глава 9

МЕТОД ГЛАВНЫХ КОМПОНЕНТ И ОБУЧЕНИЕ БЕЗ УЧИТЕЛЯ

9.1. Линейные комбинации признаков и автоматическая сортировка объектов по классам

В практике экологических исследований часто возникает необходимость выявления неоднородности некоторой группы объектов, составляющих случайную выборку из ГС. Одни из этих объектов даже при беглом осмотре могут быть отнесены к разным классам, другие — только после тщательных измерений и количественного сравнения каких-либо характерных признаков, и, наконец, встречаются промежуточные формы, которые создают существенные трудности для однозначного выбора.

Ясно, что при непосредственном разбиении «трудной» выборки на классы результаты будут в сильной степени зависеть от интуиции исследователя или, другими словами, от накопленного и подсознательно закрепленного опыта. Имея в виду этот недостаток

и учитывая ограниченные возможности человека оперировать измерениями многих признаков, можно всегда в таких случаях испытывать некоторые сомнения по поводу правильности и полноты проведенных построений.

В данной ситуации полезно использовать некоторые статистические модели и получить результаты, не зависящие от указанных субъективных причин. Для объяснения существа подобных приемов обратимся к геометрическим интерпретациям.

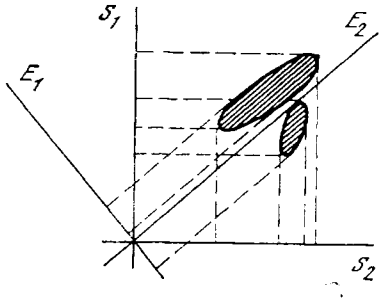


Рис. 9.1. Выбор подпространства E_s для разделения неоднородной совокупности объектов (пояснение см. в тексте)

Анализируемая выборка может быть представлена как выборочная «гроздь» из N точек в p-мерном пространстве признаков. Эту гроздь необходимо спроектировать в подпространство E_s ($s \leqslant p$) так, чтобы получить максимально возможное рассеивание спроектированных точек.

Сказанное можно пояснить простым примером. На рис. 9.1 в пространстве координат двух признаков S_1 и S_2 (двухмерное пространство) со значениями x_1. и x_2. объекты выборки образуют две грозди точек, оконтуренные эллипсами рассеивания. Сразу видно, что выборка неоднородна и состоит из двух подсовокупностей (на рисунке заштриховано). Проектируя все точки выборки на любую из осей — S_1 или S_2, можно убедиться, что в каждом случае обнаруживается сильная трансгрессия обеих подсовокупностей. Следовательно, четкое разделение возможно только при одновременном учете S_1 и S_2, в то время как ни S_1 ни S_2, взятые в отдельности, такой возможности не предоставляют.

Попробуем, однако, использовать следующий прием. Повернем оси координат на такой угол, чтобы одна из осей оказалась расположенной в направлении наибольшего разброса всех N точек (рис. 9.1). Если теперь спроектировать точки на новые направления, то на одном из них (E_1) получим две четко различающиеся группы. Вторую ось для разделения можно и не учитывать. Итак, вместо исходного двухмерного пространства используется подпространство E_1, в котором и решается поставленная задача разделения.

Рассмотрим алгебраические операции, требуемые для описания подпространства E_1.. Из аналитической геометрии известно, что при повороте осей координат S_i на некоторый угол α, новые координаты точек выражаются через старые с помощью формул

$$e_{1.} = \cos\alpha \cdot x_{1.} - \sin\alpha x_{2.},$$
$$e_{2.} = \sin\alpha x_{1.} + \cos\alpha x_{2.}. \qquad (9.1)$$

Обозначив ω_i коэффициенты при $x_{i.}$, получим

$$e_k = \sum_i \omega_{ik} x_{ik}, \qquad (9.2)$$

т. е. новые координаты представляют собой линейные комбинации исходных измерений. Набор коэффициентов ω_{ik} для каждой e_k есть вектор, при проектировании на который исходных точек достигается их максимальный разброс. Значения e_k составляют k-ю главную компоненту дисперсии. Аналитически элементы вектора ω_{ik} находятся как решение в системе уравнений

$$\sum_i^p (v_{ji} - \lambda_k \delta_{ji})\,\omega_{ik} = 0, \qquad (9.3)$$

где $\| v_{ji} \|$ — ковариационная матрица исходных измерений, рассматриваемых как единая выборка; $\| \delta_{ji} \|$ — единичная матрица.

Решение системы (9.3) хорошо известно и сводится к поиску так называемых собственных векторов (ω_k) и собственных значений ковариационной матрицы (λ_k). Нет необходимости осуществлять этот поиск расчетами вручную, так как для любой ЭВМ имеются типовые программы для его автоматической реализации.

Если значения ω_{ik} известны, то, подставив в (9.2) значения признаков конкретных объектов, получим значения e_k, которые принадлежат k-й главной компоненте. Отметим, что дисперсия k-й компоненты равна λ_k и все компоненты попарно некоррелированы. Кроме того, сумма всех собственных чисел факторизуемой матрицы равна сумме диагональных элементов этой же матрицы, а сумма квадратов элементов собственного вектора равна единице.

Наибольший интерес, конечно, представляют такие направления изменчивости, для которых λ_k достаточно большое. При сильной коррелированности исходных измерений вся изменчивость распределяется по s направлениям, где $s \ll p$. Это иногда дает возможность графически исследовать ситуацию, рассматривая вместо исходного p-мерного пространства s-мерное подпространство «новых» признаков.

Из способа нахождения всех собственных векторов и собственных значений ковариационной матрицы следует, что элементы ω_{ik} зависят от масштаба исходных измерений. Поскольку учитываемые признаки часто имеют различную природу (и соответственно разные единицы измерения), все значения их следует подходящим образом нормировать. Вопрос выбора способа нормировки достаточно сложный и в каждом конкретном случае зависит от содержательных аспектов задачи, но в большинстве случаев он решается делением всех значений каждого i-го признака x_{ik} на сред

нее арифметическое значение $\hat{x}_i$, или на среднее квадратическое отклонение.

Допустим, мы определили главные направления изменчивости и число направлений оказалось небольшим. Однако, распределив исходные точки в новых координатах, мы получили картину разброса с неясно выраженными плеядами, и нет уверенности в том, что полученные плеяды не есть результат случайности. Чтобы избежать напрасной работы по определению главных компонент, можно воспользоваться некоторыми критериями проверки однородности выборки [33]. В частности, полезна проверка следующей нуль-гипотезы.

Пусть выборка, включающая N независимых наблюдений, взята из нормальной ГС и имеет параметры $\| \hat{x}_i \|$, $\| \sigma_{ij} \|$, где $i, j = 1, 2, \ldots p$, p — число признаков. Нуль-гипотеза H_2^0 утверждает, что наблюдения в выборке принадлежат к одной и той же p-мерной ГС с заданной ковариационной матрицей. Альтернативой H_2^0 является гипотеза H_1^0 о том, что наблюдения принадлежат p-мерным ГС с различными средними, но одной и той же ковариационной матрицей. Критерием проверки H_2^0 служит величина

$$U = \sum_k^N D_k^2, \tag{9.4}$$

где

$$D_k^2 = \sum_i^p \sum_j^p \sigma^{ij} (x_{ik} - \hat{x}_i)(x_{jk} - \hat{x}_j), \tag{9.5}$$

т. е. D_k^2 — махаланобисово расстояние каждой варианты до центра выборки, а U является суммой расстояний всех точек до этого центра $\| \hat{x}_i \|$.

Величина U распределена как χ^2 с $n=(N-1)\,p$ степенями свободы, а при $n > 30$ величина $\sqrt{2U}$ имеет приближенно нормальное распределение со средним значением $\sqrt{2n-1}$ и стандартным отклонением, равным единице. Следовательно, при $n > 30$ и $t = \sqrt{2U} - \sqrt{2n-1} > 2$ нуль-гипотеза отвергается и можно утверждать неоднородность выборки с вероятностью, большей 95%.

Рассмотрим пример. В исследованиях дрейфующей станции СП-6 и на судах АтлантНИРО были собраны пробы рачков-калянусов в Норвежском и Гренландском морях, а также в Девисовом проливе. При обработке проб было замечено, что пробы из Норвежского и Гренландского морей сравнительно однородны, а в пробах из Девисова пролива встречалось большое количество исключительно крупных особей. В связи с этим было выдвинуто предположение о том, что эти особи относятся к другому виду, хотя оба вида морфологически трудно различимы. Для проверки предположения были измерены некоторые морфологические признаки у 61 особи и проведен компонентный анализ выборки.

Таблица 9.1

Данные промеров трех признаков у 47 особей калянуса

Признак	Номер особи									
	1	2	3	4	5	6	7	8	9	10
LC	3,20	3,00	3,10	3,15	2,85	2,70	2,60	2,80	3,00	3,10
LG	0,35	0,30	0,30	0,33	0,25	0,28	0,25	0,30	0,30	0,33
HG	0,34	0,28	0,28	0,28	0,25	0,25	0.25	0,28	0,28	0,28
Признак	11	12	13	14	15	16	17	18	19	20
LC	2,85	2,90	2,85	2,95	2,90	2,75	2,80	3,75	4,00	4,20
LG	0,28	0,28	0,28	0,33	0,28	0,28	0,28	0,35	0,48	0,42
HG	0,25	0,25	0,25	0,25	0,25	0,25	0,25	0,34	0,38	0,38
Признак	21	22	23	24	25	26	27	28	29	30
LC	4,00	4,10	3,90	4,15	3,95	3,95	4,00	3,95	3,75	4,15
LG	0,45	0,45	0,40	0,40	0,38	0,40	0,43	0,40	0,35	0,48
HG	0,35	0,38	0,35	0,38	0,35	0,34	0,35	0,35	0,33	0,40
Признак	31	32	33	34	35	36	37	38	39	40
LC	4,00	3,70	4,05	4,05	3,90	3,60	3,70	3,50	4,20	4,00
LG	0,40	0,38	0,38	0,38	0,40	0,35	0,35	0,35	0,45	0,43
HG	0,38	0,34	0,39	0,35	0,36	0,33	0,33	0,30	0,38	0,38
Признак	41	42	43	44	45	46	47			
LC	4,20	4,05	4,15	4,00	3,70	3,80	4,10			
LG	0,43	0,40	0,43	0,38	0,43	0,40	0,40			
HG	0,38	0,38	0,38	0,38	0,35	0,38	0,38			

Чтобы не загромождать изложение метода, в данном примере мы ограничимся анализом трех признаков у 47 особей. Признаки условно обозначим как LC — длина цефалоторакса, LG — длина и HG — ширина генитального сегмента (табл. 9.1).

По данным измерений 35 представителей из Норвежского и 55 — из Гренландского морей определены внутривыборочные ковариации и получена усредненная матрица, имеющая $90 - 2 = 88$ степеней свободы:

$$\begin{Vmatrix} 2{,}715 & & \\ 0{,}352 & 0{,}097 & \\ 0{,}273 & 0{,}045 & 0{,}047 \end{Vmatrix}. \tag{9.6}$$

Эти значения затем рассматривались как соответствующие значения в ГС.

Средние значения и ковариационная матрица выборки из Девисова пролива равны

$$\|\hat{x}_i\| = \begin{Vmatrix} 3{,}576 \\ 0{,}364 \\ 0{,}326 \end{Vmatrix}, \quad \|\sigma_{ij}\| = \begin{Vmatrix} 13{,}184 & — & — \\ 1{,}432 & 0{,}183 & — \\ 1{,}215 & 0{,}137 & 0{,}122 \end{Vmatrix}. \qquad (9.7)$$

Расчеты на ЭВМ «Мир-2» по формулам (9.4) и (9.5) дали результат: $U = 534{,}5$. Для степеней свободы $(N - 1)\ p = 138$ величина $t = \sqrt{2U} - \sqrt{2 \cdot 138 - 1} = 16{,}1$, что значительно больше критических значений, следовательно, выборку из Девисова пролива с высокой вероятностью можно считать неоднородной.

Для выполнения компонентного анализа каждое значение ковариационной матрицы выборки нормализуем, получая

$\sigma_{ij}/(\hat{x}_i\hat{x}_j)$.

Собственные значения и собственные векторы этой матрицы равны

$$\|\lambda_k\| = \begin{Vmatrix} 3{,}394 \\ 0{,}119 \\ 0{,}046 \end{Vmatrix}, \quad \|\omega_{ik}\| = \begin{Vmatrix} 0{,}539 & 0{,}621 & 0{,}569 \\ 0{,}385 & -0{,}782 & 0{,}489 \\ -0{,}749 & 0{,}044 & 0{,}061 \end{Vmatrix}.$$

Соответственно значениям λ_k первая компонента учитывает 95,4% всей дисперсии, первые две компоненты — 98,7%. Следовательно, третью компоненту, на долю которой приходится всего лишь 1,3%, без существенной потери информации можно опустить. Таким образом, исходные измерения рачков из Девисова пролива можно представить графически, используя в качестве осей координат первые две компоненты и откладывая в этих координатах e_1. и e_2. для каждой особи (рис. 9.2).

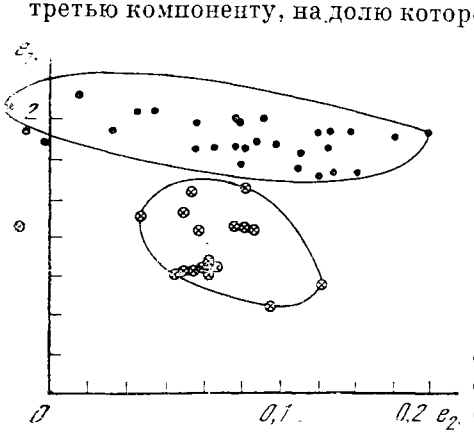


Рис. 9.2. Главные компоненты выборки калянусов из Девисова пролива

Координаты точек для графика находятся по следующей схеме. Берем первое по списку описание из табл. 9.1 и делим значения всех признаков на соответствующее среднее арифметическое. Получившийся вектор умножаем на первый собственный вектор и получаем первую координату: $e_{11} = 3{,}7/3{,}576 \cdot 0{,}539 + 0{,}35/0{,}364 \cdot 0{,}621 + 0{,}33/0{,}326 \cdot 0{,}569 = 1{,}73$. Вторая координата находится аналогично: $e_{21} = 3{,}7/3{,}576 \cdot 0{,}385 - 0{,}35/0{,}364 \cdot 0{,}782 + 0{,}33/0{,}326 \cdot 0{,}489 = 0{,}14$. Далее берется второе по списку описание и т. д. до тех пор, пока для каждой особи не будут получены координаты соответствующих им точек.

На рис. 9.2 можно видеть, что координата e_2. практически не несет информации о разделении и что вся задача могла быть решена с использованием одной компоненты e_1., на долю которой приходится 95,4% всей дисперсии. Граничное значение e_1., разбивающее всю выборку на две совокупности, можно принять равным 1,7: выше него все точки соответствуют крупным особям, ниже — мелким.

В заключение этого параграфа обратим внимание, что наибольший по абсолютной величине элемент первого собственного вектора (0,621) соответствует второму признаку— длине генитального сегмента: «вес» этого элемента превышает «веса» других признаков, однако превышение не слишком большое. Так что участие всех признаков в разделении довольно равномерное.

9.2. Анализ фенетической изменчивости симпатрических и аллопатрических популяций одного вида

Рассмотрим еще один пример из практики ихтиологических исследований. В р. Анадырь обитают две симпатрические формы сигов: горбун и востряк. Многие из особей, попадающих в выборку, имеют часть признаков, присущих востряку, а другую часть — горбуну, но имеются и такие, которые по всей совокупности учитываемых признаков примерно одинаково похожи на обе формы. Визуально свежих рыб могут различать только опытные специалисты, местные рыбаки и жители в большинстве случаев затрудняются провести такое разделение.

В связи с этим была предпринята попытка применить формальную процедуру для сортировки неоднородного материала, которая обладала бы достаточной эффективностью и не зависела от субъективных причин.

Материалом послужили морфологические измерения 200 рыб, выловленных в июле—августе 1974 г. Все измерения проводились одним и тем же исследователем (Ю. С. Решетниковым) на свежих (не фиксированных) особях. В данном примере использованы только 12 признаков, хотя измерялись 31. На основании этого материала предстояло найти в 12-мерном пространстве такие направления, при проектировании на которые точки обладали бы наибольшим разбросом.

Используя в качестве нормирующих делителей средние арифметические значения и применяя метод Якоби для решения системы (9.3), мы получили 12 собственных значений и соответствующих им собственных векторов.

Оказалось, что сумма двух самых больших собственных значений составляет около 80% от суммы всех собственных значений. Следовательно, соответствующие им главные компоненты учитывают около 80% от изменчивости всех 12 признаков. На рис. 9.3 указаны проекции 200 точек в новое двухмерное подпространство

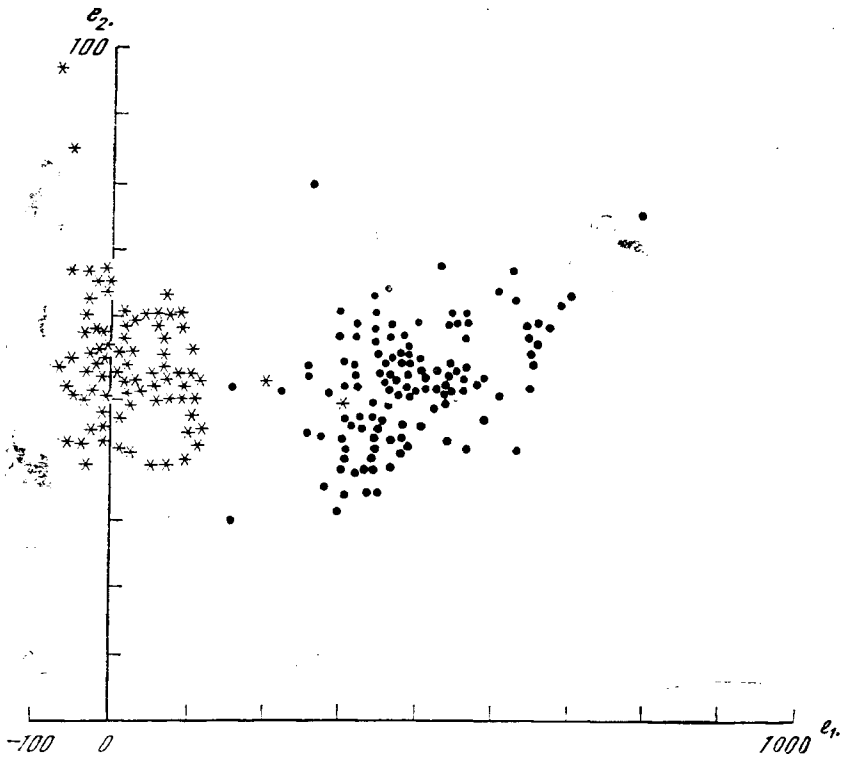


Рис. 9.3. Главные компоненты выборки сигов из р. Анадырь

и таким образом выделены две четко различающиеся плеяды, которые, по предположению, соответствуют представителям двух разных форм сига. На рис. 9.3 звездочкам соответствуют особи, которые при полевых сборах материалов были определены как горбуны, а точкам — особи востряков. Как можно видеть, расхождения между определениями ЭВМ и опытного специалиста составляют две точки, которые морфолог определил как соответствующие горбунам.

Полученные результаты позволяют утверждать следующее: во-первых, в р. Анадырь обитают две морфологически четко различимые формы сига и, во-вторых, различение их с помощью ЭВМ вполне сравнимо с результатами достаточно опытного морфолога.

Параметрическая характеристика выделенных групп приведена в табл. 9.2. Пожалуй, самой примечательной особенностью ее данных является сильная трансгрессия обеих форм по всем 12 признакам. В отдельных случаях перекрытие составляет 100%, в то время как при использовании новой системы признаков (координат) трансгрессия отсутствует. Отсюда понятно, что ни один из учитываемых признаков в отдельности не может обеспечить

Таблица 9.2

Параметрическая характеристика выделенных групп сигов

Признак	Востряк			Горбун		
	min	max	$\hat{x}$	min	max	$\hat{x}$
AC	0,930	0,973	0,949	0,926	0,968	0,947
h_{max}	0,052	0,096	0,065	0,066	0,107	0,075
P_3D	0,348	0,467	0,418	0,402	0,474	0,433
l_D	0,093	0,140	0,118	0,107	0,187	0,128
h_D	0,111	0,162	0,143	0,123	0,177	0,160
h_A	0,089	0,116	0,103	0,098	0,125	0,111
l_v	0,134	0,170	0,148	0,144	0,175	0,160
C	0,182	0,230	0,211	0,170	0,225	0,192
r	0,044	0,099	0,069	0,042	0,060	0,054
l_{max}	0,035	0,068	0,056	0,039	0,047	0,043
ll	72	89	72,92	74	90	81,46
Sp.br.	23	31	26,93	21	30	24,74

четкую классификацию особей, но если использовать линейные комбинации этих признаков, то классификация оказывается вполне возможной.

Следует обратить внимание еще на одну особенность рис. 9.3. Описываемые плеяды выделяются только по первой координате — e_1., а вторая координата (e_2.) не несет информации о разделении. На световом экране ЭВМ были дополнительно просмотрены 66 графиков проекций 200 точек на все парные сочетания из 12 новых осей координат, но указанные группировки выделялись только на рисунках, в которых участвовала координата e_1.. Этот результат свидетельствует о том, что только одно главное направление e_1. несет информацию о генетически обусловленной неоднородности, другие же направления, по-видимому, учитывают возрастные, половые и прочие различия. Такое предположение представляет специальный интерес, поскольку является основой для дальнейшей реификации [10], однако подобный анализ выходит за рамки настоящей темы.

То обстоятельство, что симпатрические формы сига р. Анадырь эффективно разделяются с помощью единственной переменной (линейной комбинации), еще не освобождает от необходимости выполнять трудоемкие измерения многих признаков. Поэтому естественным продолжением выполняемого анализа следует считать отбор информативных признаков, которые позволяют осуществлять классификацию с заданной точностью. В первом приближении для этой цели используем прием, основанный на расчете обобщенных расстояний между выделенными группами особей. Линейная дискриминантная функция из-за достоверных различий

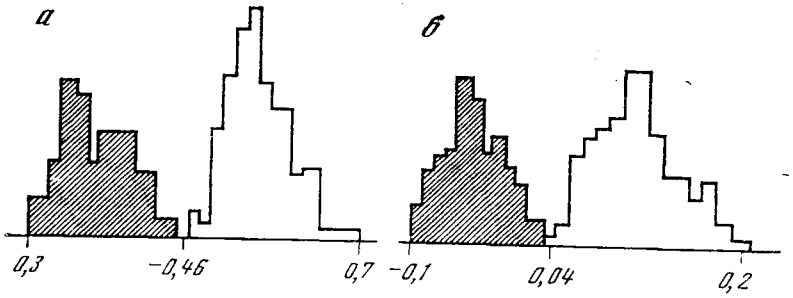


Рис. 9.4. Дискриминация горбунов (заштриховано) и востряков на основе 12 признаков (*а*) и шести (*б*)

ковариационных матриц определялась в данном случае как решение системы

$$\omega\delta = W_1 + W_2, \tag{9.8}$$

где W_1, W_2 — внутреннее рассеивание выборок; δ — вектор разности средних арифметических. Здесь нет необходимости в нормировании исходных данных, поэтому система признаков несколько изменилась.

Соответственно элементам первого собственного вектора из рассмотрения были удалены шесть признаков, имеющих малые «веса», а оставшиеся шесть (размеры плавников и висцерального скелета) были использованы для подсчета коэффициентов линейной дискриминантной функции. Коэффициенты ее для каждого признака равны

— 1,092	— 0,382	— 2,104,	1,355	0,892	4,874
hD	hA	lV	C	r	l_{max}

(9.9)

Оказалось, что проведенное сокращение числа переменных практически не отразилось на величине ошибки распознавания (рис. 9.4), поэтому была предпринята попытка удалить также все измерения плавников, однако использование всего лишь трех признаков — C, r, l_{max} — привело к резкому увеличению трансгрессии, и этот вариант был отвергнут.

Таким образом, в качестве практических приложений данного исследования рекомендуется простая формула, с помощью которой любой неспециалист может практически безошибочно различать две симпатрические формы сигов р. Анадырь.

Если при подстановке значений признаков в формулу

$$L(x_k) = \frac{1}{AC}\sum_i \omega_i x_{i_k} + 0{,}07 \tag{9.10}$$

где x_{ik} — непосредственно замеряемое значение признака у k-й особи) значения $L(x_k)$ будут отрицательными, то классифицируемые особи относятся к горбунам, в противном случае — к востря-кам.

В практике рыболовства недифференцированный вылов ведет к подрыву численности менее устойчивой популяции. Определители типа (9.10) дают возможность дифференцировать улов и тем самым способствуют рациональному распределению промыслового усилия, обеспечивающему максимальную добычу при устойчивом запасе.

Теоретическая ценность развиваемого подхода очевидна: определители типа (9.10) открывают возможность быстрым путем познать многие интимные стороны экологии генетически разобщенных, но морфологически трудно различимых таксонов. Смешанный материал обедняет выводы, так как вносит в изучаемые характеристики «информационный шум». Дифференциация материала дает способ борьбы с вечным врагом экспериментатора — «шумом», который в статистике называют также остаточной дисперсией, а в биологии — долей неопознанного или случайностью.

Рассмотрим более общий пример анализа морфологической изменчивости в совокупности, состоящей из аллопатрических и симпатрических популяций одного вида. В основу этого исследования положены материалы по сигам Coregonus lavaretus, собранные по единой методике одним и тем же исследователем (Ю. С. Решетниковым). Всего учитывалось 22 пластических и четыре меристических признаков, наиболее часто применяемых в морфологических исследованиях по сиговым рыбам. Выборки произведены из 11 популяций, которым ради удобства изложения присвоим следующую нумерацию: 1 — Чун-озеро (154 экз.); 2 — малотычинковая форма Охт-озера (84 экз.); 3 — многотычинковая форма Охт-озера (12 экз.); 4 — многотычинковая форма Чингльс-явра (15 экз.); 5 — малотычинковая форма Чингльс-явра (17 экз.); 6 — многотычинковая форма Кензис-явра (12 экз.); 7 — малотычинковая форма Кензис-явра (51 экз.); 8 — оз. Подпахтинское (31 экз.); 9 — р. Воронья (36 экз.); 10—востряк р. Анадырь (120 экз.); 11 — горбун р. Анадырь (80 экз.). Первые девять выборок относятся к сигам Кольского полуострова.

Предварительные исследования показали, что используемые признаки слабо коррелированы, так что недиагональными элементами ковариационных матриц можно пренебречь и расчет обобщенных расстояний вести по формуле (7.4). Эти расчеты, отдельно по меристическим и пластическим признакам, позволили всю цифровую информацию, содержащую около 16 тысяч измерений, свести к 55 числовым значениям расстояний для каждой системы признаков.

Объединенная матрица расстояний, рассчитанных по всем 26 признакам, была использована для поиска «лидера» — наиболее «типичной» популяции. Оказалось (табл. 9.3), что наибольшее сходство со всеми выборками имеет популяция Чун-озера, которая

Таблица 9.3

Матрица дивергенций 11 популяций сига

Популяция	Популяция										
	1	2	3	4	5	6	7	8	9	10	11
1	0	11,0	82,3	70,3	11,3	91,1	10,0	33,0	33,7	81,4	62,4
2	—	0	50,6	90,2	24,4	121,9	16,1	40,9	41,7	65,5	62,2
3	—	—	0	108,0	75,6	139,9	135,3	36,3	143,0	123,3	174,1
4	—	—	—	0	35,1	24,6	111,3	99,1	102,8	186,2	179,1
5	—	—	—	—	0	52,4	18,4	40,1	43,0	96,2	84,8
6	—	—	—	—	—	0	116,4	89,8	140,4	245,0	202,4
7	—	—	—	—	—	—	0	39,2	36,3	87,8	68,8
8	—	—	—	—	—	—	—	0	50,3	118,2	48,1
9	—	—	—	—	—	—	—	—	0	92,0	77,1
10	—	—	—	—	—	—	—	—	—	0	51,5
11	—	—	—	—	—	—	—	—	—	—	0

в дальнейших, чисто морфологических исследованиях использовалась как «модельная». Наименьшим сходством со всеми малотычинковыми формами характеризуются сиги р. Анадырь, и этот факт также представляет научный интерес, так как указывает популяции, удобные для изучения внутривидовой дивергенции под влиянием внешних условий.

В дальнейшем все три матрицы расстояний использовались для построения дендрограмм, но в отличие от предыдущих примеров объединение групп проводилось с подсчетом фактических параметров объединения. Такой способ привел к характерным особенностям, заключающимся в том, что образование групп более высокого уровня происходило при меньших различиях, чем групп предыдущего уровня (рис. 9.5). Это обстоятельство, однако, не мешает поставленной цели наглядно представить сходство и различие популяций.

На основании анализа исследуемых признаков первыми объединились малотычинковые сиги популяций 1, 2, 5, 7. Все они обитают в сравнительно близких географических районах бассейна р. Имандра. Несколько ниже к этой группе присоединяются малотычинковые сиги из бассейна Баренцева моря (популяции 8 и 9). Довольно четко выделяются крупные малотычинковые сиги р. Анадырь (популяции 10 и 11), и, наконец, особняком стоят многотычинковые формы (популяции 3, 4 и 6).

Природа образуемых группировок становится более очевидной, если рассматривать сходство популяций только в системе меристических или только пластических признаков (рис. 9.5).

При анализе только счетных признаков более четко выделяются многотычинковые сиги, малотычинковые сиги р. Анадырь, а все

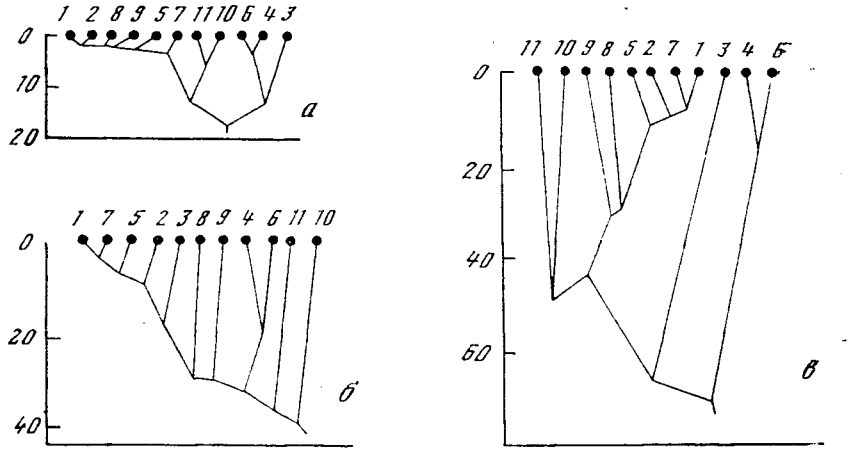


Рис. 9.5. Дендрограммы различий 11 популяций сигов

а — меристические; *б* — пластические; *в* — все учитываемые признаки

прочие малотычинковые популяции образуют довольно тесно сплоченную группу.

Дендрограмма, построенная только по пластическим признакам, позволяет отметить зависимость «близости» популяций от темпа роста: например, четко обособлена группа схожих между собой популяций мелких тугорослых сигов (1, 5 и 7) от группы крупных быстрорастущих озерных и полупроходных сигов Кольского полуострова (популяции 2, 3, 8 и 9). Самостоятельную плеяду образуют мелкие многотычинковые сиги-планктофаги (4 и 6). Интересен факт, что много- и малотычинковые формы Охт-озера (2 и 3) мало различаются по темпу роста, достигаемым размерам и форме тела, что согласуется с их близким расположением на дендрограмме. Крайнее положение занимают крупные сиги р. Анадырь, из которых востряк мигрирует даже в лиман. Востряк и горбун отличаются преимущественно пластическими признаками.

Заметим, что в данном примере использовались всего лишь две системы переменных. Допустим, что таких систем имелось бы несколько, например, дополнительно учитывались бы физиологические и этологические признаки. Для свертки информации в этом случае можно было бы произвести поиск «лидера» среди дендрограмм, соответствующих каждой системе. Полученные «веса» для каждой системы позволили бы не только определить наиболее «типичную» классификацию, но и в дальнейших исследованиях учитывать признаки каждой системы с соответствующими «весами».

9.3. Корреляционные ансамбли признаков и оценки направления изменчивости под действием отбора

Существует немало примеров того, что изменение образа жизни, способа питания или размножения животных влечет за собой комплекс морфологических изменений. Известно, например, что среди рыб хорошие пловцы характеризуются торпедообразной формой тела: мощными мышцами в передней части тела, тонким хвостовым стеблем и т. п. Эти особенности характерны для всех хороших пловцов, независимо от их видовой принадлежности.

Можно говорить, таким образом, о целом комплексе, или ансамбле, признаков, которые взаимно обусловлены и, следовательно, сильно коррелированы друг с другом. Если под действием отбора происходит дивергенция популяций, то ее внешними проявлениями будут изменения обсуждаемых корреляционных ансамблей, а не разрозненных признаков.

В процессе разделения смешанной выборки симпатрических форм сигов установлено, что внутривидовая дивергенция охватывает совокупность переменных, которую в целом можно назвать характеристикой охотничьих способностей. По крайней мере, строение висцерального скелета и величина плавников имеют непосредственное отношение к этим способностям. Если другие, географически отдаленные популяции сигов обнаруживают изменчивость признаков, совпадающую с указанным направлением, то мы получаем сильный довод в пользу предположения, что весь «ансамбль» контролируется организмом как единое целое.

В данном исследовании мы используем ранее описанный материал по малотычинковой форме Coregonus lavaretus, образуя смешанные выборки следующих популяций: I — популяции бассейна р. Имандра (Чун-озеро и Охт-озеро); II — популяции баренцевоморских бассейнов (оз. Подпахтинское и р. Воронья); III — симпатрические формы сигов р. Анадырь (горбун и востряк). По чисто техническим причинам (малый объем оперативной памяти используемой ЭВМ) из всех 26 признаков учитывались только 17: 1 — h_{max} — максимальная высота тела; 2 — h_{min} — минимальная высота тела; 3 — ped — толщина хвостового стебля; 4 — AnD — антедорсальное расстояние; 5 — PsD — постдорсальное расстояние; 6 — l_D — длина спинного плавника; 7 — h_D — высота его; 8 — l_A — длина анального плавника; 9 — h_A — высота его; 10 — l_P — длина грудного плавника; 11 — l_V — длина вентрального плавника; 12 — C — длина головы; 13 — r — длина рыла; 14 — l_{max} — длина верхней челюсти; 15 — l_{md} — длина нижней челюсти; 16 — ll — число чешуй в боковой линии; 17 — Sp. br. — число жаберных тычинок.

Признаки 1—12 брались как отношение к длине тела AC, 13—15 — как отношение к длине головы; 16 и 17 — в штуках. Самцов и самок в пробах не разделяли, поскольку половые различия не выражены.

Расчеты показали, что выборки из бассейна р. Имандра образуют корреляционную матрицу со сравнительно слабыми связями и неясно выраженными направлениями изменчивости (табл. 9.4). Первые три наибольшие собственные значения объединенной корреляционной матрицы составляют всего лишь 38,8% всей дисперсии, причем на долю главного направления приходится около 22%. Все это указывает на относительную однородность объединенной выборки со слабо выраженной дивергенцией. Наибольшие нагрузки в главном направлении имеют признаки h_A, l_P, h_D, l_V, h_{min}. Иначе говоря, основная дивергенция у сигов бассейна р. Имандра происходит в направлении изменчивости плавников, высоты тела и толщины хвостового стебля. Другое направление имеет наибольшие нагрузки у признаков ll, AnD, l_{md}, C, Sp. br., r, т. е. захватывает счетные признаки и признаки, связанные с размерами головы.

Выборки из баренцевоморских бассейнов значительно более неоднородны: первые три наибольшие собственные значения объе-

Таблица 9.4

Собственные значения и собственные векторы корреляционных матриц объединенных выборок популяций сига

Номер признака	Охт-озеро и Чун-озеро			Оз. Подпахтинское и р. Воронья			р. Анадырь		
	Собственные значения								
	3,72	1,53	1,33	4,59	3,52	1,47	7,26	3,02	1,48
1	33	—05	—15	26	—32	—08	23	—13	—28
2	35	00	00	31	10	—23	25	—15	—23
3	—11	11	31	24	29	—19	04	—09	60
4	10	—41	—44	38	06	00	—19	18	—26
5	—18	14	23	25	—22	—33	12	—19	43
6	23	15	41	23	23	27	38	60	40
7	37	08	15	31	24	06	29	—12	—05
8	24	04	33	17	—21	21	15	—02	—32
9	39	03	11	28	—07	25	42	54	10
10	38	05	—18	35	—05	18	23	—11	—18
11	35	21	—06	38	—10	00	28	—15	—11
12	12	—37	10	17	41	—04	—26	19	—01
13	11	27	—30	—11	08	32	—26	22	—12
14	05	15	—35	—09	43	—05	—25	22	—12
15	—04	41	—03	—06	45	—13	—09	12	—16
16	—12	42	—19	00	12	50	09	—10	—14
17	—02	37	—16	—02	05	—45	—24	17	06

Примечание. При записи коэффициентов собственных векторов нуль и запятая для краткости опущены. Каждое собственное значение расположено над вектором-столбцом, которому оно соответствует.

диненной корреляционной матрицы составляют уже 56,4% всей дисперсии, причем на долю главного направления приходится 26,7%. Как и в предыдущем случае, здесь наибольшие нагрузки имеют относительные размеры плавников, высота тела и толщина хвостового стебля. На долю другого направления приходится 20,7% всей дисперсии, т. е. одного порядка с главным, причем наибольшие нагрузки приходятся на признаки головы; счетные же признаки входят с наибольшими нагрузками в третье направление.

Наконец, в выборках из р. Анадырь неоднородность выражена совершенно отчетливо: на долю первых трех собственных значений приходится 79% всей дисперсии, а на главное направление — 42,5%. Интересно, что главное направление имеет наибольшие нагрузки на все те признаки, которые захватываются первыми двумя векторами в предыдущих случаях (см. табл. 9.4). Второй вектор не несет дополнительной информации о направлениях изменчивости, поскольку признаки с наибольшими «весами» у первого и второго векторов совпадают.

Итак, дивергенции внутри популяций сигов в разных частях ареала существенно различаются, хотя и имеют много общего.

К сожалению, даже в таком простом примере из-за обилия цифровых данных довольно трудно провести более четкое обобщение различий и сходства дивергенций в разных популяциях. Поэтому ниже предлагается простой метод графического представления полученных результатов.

Для сравнения направлений дивергенции в двух районах отложим в системе прямоугольных координат значения коэффициентов собственных векторов, соответствующих наибольшим собственным значениям (рис. 9.6). Если дивергенции внутри разных совокупностей абсолютно одинаковы, то точки, соответствующие значениям элементов собственных векторов, лягут на прямую $O - O'$, расположенную под углом 45° к осям координат. Точки будут тем дальше расположены от начала координат, чем большую нагрузку имеет соответствующий признак. Так что вблизи начала координат концентрируются те точки, которые соответствуют наиболее «стабильным» признакам.

На рис. 9.6, *а* можно видеть, что вся совокупность точек, представляющих коэффициенты первых собственных векторов для выборок из водоемов Кольского полуострова, разбивается на три группы.

Первая из них располагается вдоль упомянутой прямой $O - O'$ и на достаточно большом расстоянии от начала координат. Это и есть те признаки, по которым дивергенция сравниваемых совокупностей параллельна.

Другую группу составляет точка *4* — это номер признака, по которому дивергенции не совпадают, а именно она находится сравнительно близко от оси ординат, но имеет большое значение по оси абсцисс. Другими словами, в выборках из бассейнов р. Иманд-

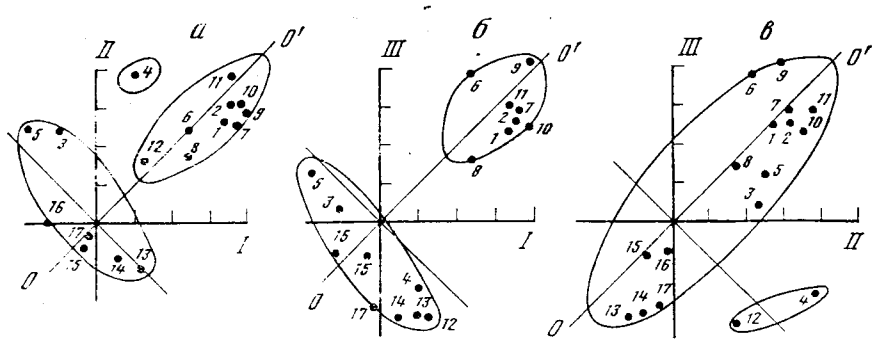


Рис. 9.6. Соотношение направлений изменчивости в выборках сигов

Объединение выборок: *а* — I и II; *б* — I и III; *в* — II и III. Римские цифры — номера объединений; арабские цифры — номера признаков

ра этот признак «стабилен», а в выборках из баренцевоморских бассейнов он участвует в дивергенции.

Наконец, третью группу составляют точки, лежащие вдоль направления, перпендикулярного направлению первой группы. Здесь указаны номера признаков, по которым дивергенции в сравниваемых выборках противоположны. Правда, эти точки расположены вблизи начала координат и, следовательно, расхождение по этим признакам находится на начальном этапе.

На рис. 9.6, *б* аналогичные сведения представлены по дивергенциям внутри популяций из бассейнов рек Имандра и Анадырь. Это самые многочисленные выборки, и для построения данного рисунка использовано около 7,5 тыс. измерений 17 признаков (не считая АС) у 438 особей.

Здесь вся совокупность точек разбивается на две группы, каждая из которых расположена вдоль направлений, перпендикулярных друг другу. Примечательно, что в группу с высокими значениями нагрузок попадают все те же признаки (за исключением 12), что и на рис. 9.6, *а*. Не совпадающие по расхождениям и «стабильные» признаки также почти одинаковы по составу, но образуют здесь единую плеяду.

Наконец, на рис. 9.6, *в* можно заметить, что дивергенции внутри популяций из баренцевоморских бассейнов и р. Анадырь практически параллельны по всем признакам (за исключением 4 и 12).

Таким образом, можно констатировать, что наибольшее сходство по направлениям дивергенции наблюдается между популяциями сигов из баренцевоморских бассейнов и р. Анадырь, в то время как сиги из бассейнов р. Имандра дивергируют по собственному направлению, лишь частично совпадающему с направлениями прочих анализируемых популяций.

Одним из главных результатов такого анализа следует считать установление факта дивергенции самых разных популяций не по

случайному набору признаков, а по их устойчивым корреляционным ансамблям. Так, во всех рассматриваемых частях ареала относительные размеры плавников имеют наибольшие корреляции и соответственно наибольшие «веса» во всех главных собственных векторах корреляционных матриц объединенных выборок. Эти связи сохраняются даже тогда, когда факторизуются матрицы единичных выборок.

Рассмотрим возможные причины, приводящие к такому эффекту. Еще во времена Кювье было известно, что признаки в организме могут быть сильно коррелированы, так что по значениям одного из них достаточно точно прогнозируются значения других. Несколько позднее было установлено, что коррелированные признаки не составляют единого целого, а распадаются на отдельные плеяды, такие, когда внутри одной плеяды они оказываются тесно связанными, но слабо коррелированными с признаками других плеяд. Р. Э. Блэкит [10] в связи с этим полагает, что, несмотря на многочисленность измеримых признаков, их взаимные отношения могут быть описаны в рамках значительно меньшего числа фундаментально различных «схем роста». Причиной этого, по его мнению, является то, что организм животного располагает сравнительно небольшим числом гормональных и иных регулирующих рост механизмов.

Блэкит ссылается на работы Б. С. Краус и С. Чои (см. [10]), которые прямыми данными показали, что 12 скелетных признаков у человеческого плода сводятся к 4 «схемам роста», и привели некоторые доказательства того, что каждая схема может быть блокирована в результате одной генной мутации, контролирующей ее как целое. Эти и многие другие данные свидетельствуют о том, что генетические механизмы действуют не на отдельные признаки, а на «схемы роста», приводящие к изменению корреляционных ансамблей фенетических признаков.

В свете сказанного можно полагать, что генетические расхождения двух симпатрических форм сига неизбежно приводят к дивергенции признаков, принадлежащих, вероятно, нескольким «схемам роста». Другими словами, наблюдаемые в анализе расхождения не обусловлены изменениями одной такой «схемы», однако, чтобы выяснить их истинное число, нужна более подробная информация об изменчивости популяций на всем ареале вида.

Использованные материалы показывают, что комплекс измеряемых признаков распадается в разных популяциях на несколько ансамблей. В частности, различия в дивергенциях позволяют отметить, что в анализе их участвовало не менее трех: первый — относительные размеры головы, второй — относительные размеры плавников и третий — меристические признаки. В то же время факты устойчивого проявления второго ансамбля во всех трех сравнениях (рис. 9.6, *а, б, в*) позволяют предполагать его неделимость.

Значение поиска и выделения обсуждаемых ансамблей признаков трудно переоценить, так как именно на этом пути появляются возможности прогнозировать направления морфологической изменчивости разных популяций при действии естественного и искусственного отбора. Можно ожидать, например, что в силу одинаковой направленности дивергенций популяции сигов р. Анадырь и сигов баренцевоморских бассейнов у последних с течением времени образуются формы, аналогичные симпатрическим формам горбуна и востряка. Как отмечалось, визуально современные горбун и востряк различаются только опытными специалистами или с помощью изощренных математических методов, реализуемых на ЭВМ. Ясно, что в случае сигов из баренцевоморских бассейнов, у которых подобные различия выражены значительно слабее, единственный путь их выявления — это применение косвенных методов, основанных на точных измерениях. То, что недоступно умозрительному анализу, становится легко обозримым и интерпретируемым при использовании обсуждаемой методики.

Если действительно в процессе отбора или генных мутаций дивергируют не отдельные признаки, а их ансамбли, то данная методика позволяет предусмотреть многие нежелательные и неожиданные последствия для случая, когда отбор является искусственным. Может оказаться, например, что, добиваясь от какой-либо генетической линии изменения полезных для потребителя признаков, можно с неизбежностью получать нежелательные изменения других признаков, находящихся в одном ансамбле с первыми.

Добавим к этому, что в свете предлагаемого подхода возможности традиционного морфологического анализа, который порой считается одиозным, оказываются по существу не до конца используемыми. Это совсем не означает, что обсуждаемый анализ пригоден только для случая морфологических измерений. Видимо, он мог бы оказаться намного более эффективным, если бы измеряемые признаки имели различную физическую природу. Использование анатомических, физиологических, экологических или этологических признаков позволило бы разрешить сформулированные выше проблемы гораздо полнее.

ЛИТЕРАТУРА

1. *Александров П. С.* Введение в теорию множеств и общую топологию. М.: Наука, 1977, с. 367.
2. *Андреев В. Л., Никулин О. А.* О различении внутрипопуляционных группировок анадырской кеты на основе анализа рисунков чешуи.— В кн.: Динамика вязкой жидкости: Измерение параметров состояния сложных систем. Владивосток: ДВНЦ, 1977, с. 64—71.
3. *Андреев В. Л., Решетников Ю. С.* Исследование внутривидовой морфологической изменчивости сига методами многомерного статистического анализа.— Вопр. ихт., 1977, т. 14, вып. 5(106), с. 862—878.
4. *Андреев В. Л., Решетников Ю. С.* Анализ состова пресноводной ихтиофауны северо-восточной части СССР на основе методов теории множеств.— Зоол. ж., 1978, т. 58, вып. 2, с. 165—175.
5. *Андреев В. Л., Решетников Ю. С.* Использование ЭВМ для распознавания симпатрических форм сига р. Анадырь.— В кн.: Систематика и биология пресноводных организмов северо-восточной Азии. Владивосток: ДВНЦ, 1978, с. 112—122.
6. *Барабаш Ю. Л., Варский Б. В., Зиновьев В. Г., Кириченко В. С., Сапегин В. Ф.* Вопросы статистической теории распознавания. М.: Сов. радио, 1967, с. 400.
7. *Бейли Н.* Математика в биологии и медицине. М.: Мир, 1970, с. 326.
8. *Берж К.* Теория графов и ее применения. М.: ИЛ, 1962, с. 300.
9. *Берзтисс А. Т.* Структуры данных. М.: Статистика, с. 407.
10. *Блэкит Р. Э.* Морфометрический анализ: Теоретическая и математическая биология. М.: Мир, 1968, с. 247—273.
11. *Бонгард М. М.* Проблема узнавания. М.: Наука, 1967, с. 320.
12. *Бронштейн И. Н., Семендяев К. А.* Справочник по математике. М.: Наука, 1967, с. 668.
13. *Василевич В. И.* Статистические методы в геоботанике. Л.: Наука, 1969, с. 230.
14. *Васильев В. И.* Распознающие системы (справочник). Киев.: Наукова думка, 1968, с. 292.
15. *Вентцель Е. С.* Теория вероятностей. М.: Наука, 1969, с. 576.
16. Геология и математика/Под руководством Ю. А. Воронина/Новосибирск: Наука, 1965, с. 107.
17. *Гжегорчик А.* Популярная логика. М.: Наука, 1967, с. 107.
18. *Гильманов Т. Г.* Математическое моделирование биогеохимических циклов в травяных экосистемах. М.: Изд-во МГУ, 1978, с. 168.
19. *Гладких Г. Н.* Характеристика фитопланктона юго-восточного побережья острова Хонсю в сезонном аспекте.— Изв. ТИНРО, 1975, т. 96, с. 46—56.
20. *Горелик А. Л., Скрипкин В. А.* Методы распознавания. М.: Высшая школа, 1977, с. 221.
21. *Дмитриев А. Я., Журавлев Ю. И., Кренделев Ф. П.* О математических принципах классификации предметов и явлений.— В кн.: Дискретный анализ. Новосибирск: СО АН СССР, 1966, вып. 7, с. 3—15.
22. *Дружинин В. В., Конторов Д. С.* Проблемы системологии. М.: Сов. радио, 1976, с. 296.

23. *Дуда Р., Харт П.* Распознавание образов и анализ сцен. М.: Мир, с. 510.
24. *Дюран Б., Оделл П.* Кластерный анализ. М.: Статистика, 1977, с. 128.
25. *Елисеева И. И., Рукавишников В. О.* Группировка, корреляции, распознавание образов. М.: Статистика, 1977, с. 143.
26. *Ежов И. И., Скороход А. В., Ядренко М. И.* Элементы комбинаторики. М.: Наука, 1977, с. 80.
27. *Загоруйко Н. Г.* Методы распознавания и их применение. М.: Сов. радио, 1972 с. 206.
28. *Иванков В. Н., Броневский А. М.* Неотения у лососевых.— В кн.: Лососевидные рыбы. Л.: Наука, 1976, с. 39—40.
29. *Иванков В. Н., Свирский В. Г.* Таксономические различия и экологическая обусловленность особенностей оогенеза и половых циклов трескообразных, размножающихся на шельфе.— В кн.: Биология шельфа. Владивосток: ДВНЦ, 1975, с. 60—61.
30. *Калужнин Л. А.* Что такое математическая логика. М.: Наука, 1964, с. 150.
31. *Кириллов Ф. Н.* Рыбы Якутии. М.: Наука, 1972, с. 300.
32. *Константинов А. С.* Использование теории множеств в биогеографическом и экологическом анализе.— Усп. соврем. биол., 1969, т. 67, вып. 1, с. 99—108.
33. *Кульбак С.* Теория информации и статистика. М.: Наука, 1967, с. 408.
34. *Курош А. Г.* Курс высшей алгебры. М.: Наука, 1965, с. 431.
35. *Майр Э.* Принципы зоологической систематики. М.: Мир, 1971, с. 454.
36. *Макфедьен Э.* Экология животных. М.: Мир, 1965, с. 540.
37. *Мельник Г. П.* Азбука математической логики. М.: Знание, с. 103.
38. *Мендельсон Э.* Введение в математическую логику. М.: Наука, 1976, с. 320.
39. *Миркин Б. Г.* Проблема группового выбора. М.: Наука, 1974, с. 255.
40. *Нечипоренко В. И.* Структурный анализ систем.: М.: Сов. радио, 1977, с. 214.
41. *Олейников А. П.* Цифровое кодирование систематических признаков древних организмов. М.: Наука, 1972, с. 188.
42. *Правдин И. Ф.* Руководство по изучению рыб. М.: Пищевая промышленность, 1966, с. 376.
43. *Пузаченко Ю. Г., Мошкин А. В.* Информационно-логический анализ в медико-географических исследованиях.— Медицинская география, 1969, вып. 3, с. 5—74.
44. *Райзер Г. Дж.* Комбинаторная математика. М.: Мир, 1966, с. 154.
45. *Рао С. Р.* Линейные статистические методы и их приложения. М.: Наука, 1968, с. 547.
46. *Семкин Б. И.* Дескриптивные множества и их приложения.— В кн.: Исследование систем. 1. Сложные системы. Владивосток: ДВНЦ, 1973, с. 83—94.
47. *Семкин Б. И., Двойченков В. И.* Об эквивалентности мер сходства и различия.— В кн.: Исследование систем. 1. Сложные системы. Владивосток: ДВНЦ, 1973, с. 95—104.
48. *Смирнов Е. С.* Таксономический анализ. М.: Изд-во МГУ, с.186.
49. *Столл Р. Р.* Множества. Логика. Аксиоматические теории. М.: Просвещение, 1968, с. 232.
50. *Уилкс С.* Математическая статистика. М.: Наука, 1967, с. 632.
51. *Уилсон Р.* Введение в теорию графов. М.: Мир, 1977, с. 207.
52. *Харари Ф.* Теория графов. М.: Мир, 1973, с. 300.
53. *Цыпкин Я. З.* Основы теории обучающихся систем. М.: Наука, 1970, с. 300.
54. *Чернов Ю. И.* О некоторых индексах, используемых при анализе структуры животного населения суши.— Зоол. ж., 1971, т. 50, вып. 7, с. 1079—1091.
55. *Шаракшанэ А. С., Железнов И. Г., Ивницкий В. А.* Сложные системы. М.: Высшая школа, 1977, с. 247.
56. *Шрейдер Ю. А.* Равенство, сходство, порядок. М.: Наука, 1971, с. 254

57. *Эдельман С. Л.* Математическая логика. М.: Высшая школа, 1975, с. 176.
58. *Юрцев Б. А.* Флора Сунтар-Хаята.— В кн.: Проблемы истории высокогорных ландшафтов северо-востока Сибири, 1968, 1.
59. *Juchasz-Nagy P.* On association among plant populations.— J. Acta biologica Debricina Y., 1967, vol. V, p. 59—66.
60. *Kussakin O. G.* A list of the macrofauna in the intertidal zone of the Kuril islands, with remarcs on zoogeographical structure of the region.— Pabl. of the Seto Marine Lab., 1975, vol. 32, № 114, p 47—74.
61. *Sispson G. G.* Mammals and Nature of continents.— Amer. J. Sci., 1943, vol. 241.
62. *Sneath P. H. A., Sokal R. R.* Numerical taxonomy. San Francisco, 1973, p. 573.
63. *Sokal R. R., Sneath P. H. A.* Principles of numerical taxonomy. San Francisco, 1963, p 359.
64. *Urbach V. Y.* Linear discriminant analysis: loss of the power when a variate is omitted.— Biometrics, 1971, vol. 27, N 3, p. 531—534.

ПРЕДМЕТНЫЙ УКАЗАТЕЛЬ

ОГЛАВЛЕНИЕ

Часть II

СТАТИСТИЧЕСКИЕ МЕТОДЫ РАСПОЗНАВАНИЯ

Валентин Леонидович Андреев

КЛАССИФИКАЦИОННЫЕ ПОСТРОЕНИЯ В ЭКОЛОГИИ И СИСТЕМАТИКЕ

Утверждено к печати
Тихоокеанским институтом географии
ДВНЦ АН СССР

Редактор издательства К. Ф. Пашковская
Художник Н. Н. Якубовская
Художественный редактор Н. Н. Власик
Технические редакторы Л. Н. Золотухина, А. М. Сатарова
Корректоры Н. И. Казарина, Г. Н. Лащ

ИБ № 17423

Сдано в набор 9.10.79. Подписано к печати 3.03.80
Т-02493. Формат 60×90¹/₁₆. Бумага типографская № 2
Гарнитура обыкновенная. Печать высокая
Усл. печ. л. 9. Уч.-изд. л. 9,1. Тираж 1550 экз.
Тип. зак. 2418. Цена 1 р. 40 к.

Издательство «Наука»
117864 ГСП-7, Москва, В-485, Профсоюзная ул., 90

2-я типография издательства «Наука»
121099, Москва, Г-99, Шубинский пер., 10