понент: мы можем рассматривать число векторов со значениями главных компонент как гиперпараметр, подлежащий нахождению при помощи перекрестной проверки или какого–либо другого родственного ей метода. Относительная простота выбора числа главных компонент для обучения с учителем — это одно из проявлений того факта, что обучение с учителем обычно является более четко определенной и объективно решаемой задачей, чем обучение без учителя.

10.2.4 Другие приложения PCA

В подразделе 6.3.1 мы видели, что регрессионный анализ можно выполнить, используя векторы значений главных компонент в качестве предикторов. Более того, многие статистические методы регрессии, классификации и кластеризации можно легко применять не к исходной матрице данных размером $n \times p$, а к матрице $n \times M$, столбцы в которой представляют собой первые $M \ll p$ векторов со значениями главных компонент. Это может приводить к *менее зашумленным* результатам, поскольку часто случается так, что сигнал в данных (в отличие от шума) сконцентрирован в первых нескольких главных компонентах.

10.3 Методы кластеризации

Под кластеризацией понимают очень широкий круг методов, предназначенных для обнаружения *групп*, или *кластеров*, в данных. Выполняя кластеризацию наблюдений из некоторого набора данных, мы пытаемся разбить их на отдельные группы таким образом, чтобы наблюдения внутри каждой группы были похожи друг на друга, а наблюдения из разных групп заметно отличались друг от друга. Конечно, чтобы конкретизировать эту задачу, мы должны дать определение тому, что подразумевается под *сходством* или *различием* двух или более наблюдений. Безусловно, часто этот выбор будет определяться предметной областью и знанием исследуемых данных.

Предположим, например, что у нас есть набор из n наблюдений, каждое из которых описано по p признакам. Эти n наблюдений могли бы соответствовать образцам тканей, взятых у пациенток с раком груди, а p признаков могли бы соответствовать измерениям, выполненным на каждом образце (например, измерения клинических параметров, таких как стадия рака или уровень экспрессии генов). Возможно, у нас есть основания считать, что эти n наблюдений в некоторой степени неоднородны; например, могут существовать несколько *неизвестных* типов рака груди. Для обнаружения этих типов можно применить кластеризацию. Здесь мы имеем дело с задачей обучения без учителя, поскольку мы пытаемся выявить структуру (в данном случае различные кластеры) на основе некоторого набора данных. Цель же задач обучения с учителем заключается в предсказании вектора значений некоторой зависимой переменной, такой как время выживания или реакция на введение лекарства.

И кластерный анализ, и PCA пытаются упростить данные, вычисляя небольшой набор сводных величин, однако механизмы этих методов различаются:

- цель PCA состоит в представлении данных в подпространстве малой размерности, объясняющем достаточную долю общей дисперсии;
- цель кластеризации состоит в нахождении однородных групп наблюдений.

С еще одним приложением кластеризации можно столкнуться в маркетинге. У нас может иметься доступ к большому числу измерений (например, медианного уровня дохода в расчете на домохозяйство, профессионального положения, расстояния до ближайшего городского поселения и т. п.) для большого числа людей. Наша задача заключается в *сегментировании рынка* путем обнаружения групп людей, которые могли бы с большей частотой реагировать на рекламу определенного вида или с большей вероятностью покупать определенный продукт. Эта задача сводится к кластеризации людей, входящих в наш набор данных.

Поскольку кластерный анализ популярен во многих областях, существует большое количество методов для его выполнения. В этом разделе мы сосредоточимся на двух, возможно, наиболее популярных подходах: *кластеризации по методу K средних* и *иерархической кластеризации*. В случае с кластеризацией по методу K средних мы пытаемся разбить наблюдения на некоторое заранее заданное число кластеров. В случае же с иерархической кластеризацией желаемое число кластеров нам заранее неизвестно; более того, в результате такого анализа мы получаем древовидное представление наблюдений — *дендрограмму*, которая позволяет нам одновременно увидеть все возможные кластеры — от 1 до n. Каждый из этих методов обладает своими преимуществами и недостатками, которые мы опишем в этой главе.

метод K средних

иерархическая кластеризация

дендрограмма

В целом мы можем выполнить кластеризацию наблюдений по их признакам для обнаружения групп среди этих наблюдений или кластеризацию признаков по наблюдениям для обнаружения групп среди этих признаков. Ниже для простоты мы будем обсуждать кластеризацию наблюдений на основе признаков, хотя обратную задачу можно выполнить посредством простой транспозиции матрицы данных.

10.3.1 Кластеризация по методу K средних

Кластеризация по методу K средних представляет собой простой и элегантный подход для разбиения некоторого набора данных на K отдельных, непересекающихся кластеров. Для выполнения кластеризации по этому методу мы должны сначала указать желаемое число кластеров K; затем соответствующий алгоритм отнесет каждое наблюдение в точности к одному из K кластеров. На рис. 10.5 показаны результаты, полученные в результате применения кластеризации по методу K средних с разными значениями K к набору имитированных данных, содержащему 150 наблюдений в двумерном пространстве.

Процедура кластеризации на основе K средних вытекает из простой и интуитивно понятной математической проблемы. Мы начнем с введения некоторых обозначений. Пусть $C_1, \ldots, C_K$ обозначают наборы индексов наблюдений из каждого кластера. Эти наборы обладают двумя свойствами:

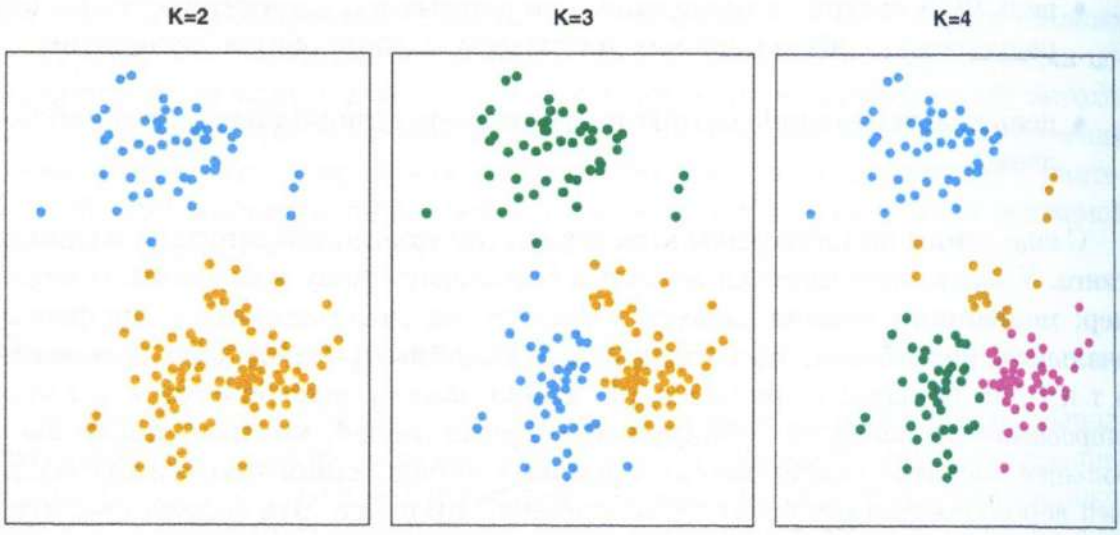

РИСУНОК 10.5. *Набор имитированных данных, содержащий 150 наблюдений и два признака. Графики показывают результаты применения метода K средних с разными значениями K (число кластеров). Цвет каждого наблюдения обозначает кластер, к которому оно было отнесено алгоритмом кластеризации. Заметьте, что кластеры никак не упорядочены, в связи с чем выбор цвета точек произволен. Данные метки кластеров не были использованы в ходе кластеризации — они являются результатом работы самой этой процедуры*

1. $C_1 \cup C_2 \cup \cdots \cup C_k = \{1, \ldots, n\}$. Другими словами, каждое наблюдение принадлежит по меньшей мере к одному из K кластеров.

2. $C_k \cap C_{k'} = \emptyset$ для всех $k \neq k'$. Другими словами, кластеры не пересекаются: ни одно из наблюдений не может принадлежать к более чем одному кластеру.

Например, если i–е наблюдение входит в состав k–го кластера, то $i \in C_k$. Идея, лежащая в основе метода K средних, заключается в том, что *хорошей* является кластеризация, при которой *внутрикластерный разброс*[6] минимален. Внутрикластерный разброс для кластера C_k — это величина $W(C_k)$, отражающая степень отличий наблюдений из данного кластера друг от друга. Таким образом, мы хотим решить следующую задачу:

$$\underset{C_1,\ldots,C_K}{\text{минимизировать}} \left\{ \sum_{k=1}^{K} W(C_k) \right\}. \tag{10.9}$$

Другими словами, согласно этой формуле, мы хотим разбить наблюдения на K кластеров таким образом, чтобы общий внутрикластерный разброс, полученный путем суммирования по всем K кластерам, был минимальным.

Попытка найти решение для (10.9) выглядит как разумная идея, однако для ее практической реализации нам необходимо дать определение

[6] В оригинале используется термин «within–cluster variation». — *Прим. пер.*

внутрикластерного разброса. Для этого существует большое число способов, но чаще всего это определение базируется на возведенном в квадрат *евклидовом расстоянии*:

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \tag{10.10}$$

где $|C_k|$ обозначает число наблюдений в k–м кластере. Другими словами, внутрикластерный разброс в кластере k — это сумма квадратов евклидовых расстояний между всеми парами наблюдений в этом кластере, разделенная на общее число входящих в него наблюдений. Объединение (10.9) и (10.10) дает оптимизационную задачу кластеризации по методу K средних:

$$\underset{C_1,\ldots,C_K}{\text{минимизировать}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}. \tag{10.11}$$

Теперь мы хотели бы найти алгоритм для решения (10.11), т. е. такой метод разбиения наблюдений на K кластеров, который минимизирует целевую функцию из (10.11). Оказывается, что получить точное решение этой задачи очень трудно, поскольку существует почти K^n способов разбиения n наблюдений на K кластеров. Это огромное число (если только K не является очень малым)! К счастью, можно показать, что один очень простой алгоритм позволяет найти локальный оптимум, т. е. *очень хорошее приблизительное решение* оптимизационной задачи (10.11). Этот подход представлен в алгоритме 10.1.

Алгоритм 10.1 Кластеризация по методу K средних

1. Каждому наблюдению присвойте случайно выбранное число из интервала от 1 до K. Эти числа будут служить в качестве исходных меток кластеров.

2. Повторите следующие шаги несколько раз до тех пор, пока метки классов не перестанут изменяться:

   (a) вычислите *центроид* для каждого из K кластеров. Центроид k–го класса представляет собой вектор из p средних значений признаков, описывающих наблюдения из этого кластера;

   (b) присвойте каждому наблюдению метку того кластера, чей центроид находится ближе всего к этому наблюдению (здесь удаленность выражается в виде евклидова расстояния).

Алгоритм 10.1 на каждом шаге будет гарантированно снижать значение целевой функции (10.11). Следующий пример помогает понять, почему это так:

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2, \qquad (10.12)$$

где $\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$ — это среднее значение признака j в кластере C_k. На шаге 2(a) средние значения признаков в каждом кластере являются константами, которые минимизируют сумму квадратов отклонений, а на шаге 2(b) перераспределение меток кластеров может только улучшать (10.12). Это означает, что по мере выполнения алгоритма получаемая кластеризация будет постепенно улучшаться до тех пор, пока результат не перестанет изменяться; целевая функция (10.11) никогда не будет возрастать. Стабилизация результата кластеризации означает, что *локальный оптимум* достигнут. На рис. 10.6 приведена динамика выполнения алгоритма на примере имитированных данных из рис. 10.5. Название метода K средних обусловлено тем обстоятельством, что на шаге 2(a) центроиды вычисляются как средние значения наблюдений, отнесенных к каждому кластеру.

Поскольку метод K средних находит локальный, а не глобальный оптимум, то полученные результаты будут зависеть от исходного (случайного) разбиения наблюдений на кластеры на первом шаге алгоритма 10.1. По этой причине важно запускать алгоритм несколько раз, используя разные исходные случайные конфигурации. После этого выбирают *оптимальное* решение, т. е. решение, обеспечивающее наименьшее значение целевой функции (10.11). На рис. 10.7 показаны локальные оптимумы, полученные в результате шестикратного запуска алгоритма K средних на данных из рис. 10.5 с использованием шести разных исходных разбиений на кластеры. В данном случае оптимальной является кластеризация со значением целевой функции, равным 235.8.

Итак, для выполнения кластеризации по методу K средних мы должны решить, сколько кластеров мы ожидаем обнаружить в данных. Проблема выбора K далеко непроста. Эта проблема, а также некоторые другие практические аспекты, возникающие при выполнении кластеризации по методу K средних, будут рассмотрены в подразделе 10.3.3.

10.3.2 Иерархическая кластеризация

Одним из потенциальных недостатков кластеризации по методу K средних является то, что она требует от нас предварительного указания числа кластеров. *Иерархическая кластеризация* представляет собой альтернативный подход, который не требует, чтобы мы придерживались какого-то конкретного выбора K. Дополнительное преимущество иерархической кластеризации в сравнении с методом K средних заключается в том, что она приводит к привлекательному представлению данных в виде древовидной структуры, которую называют *дендрограммой*.

агломеративная кластеризация

В этом подразделе мы опишем кластеризацию типа «*снизу вверх*», или *агломеративную* кластеризацию[7]. Это наиболее распространенный тип иерархической кластеризации, который заключается в том, что построение дендрограммы (обычно изображается в виде перевернутого дерева —

[7] В оригинале используются термины «bottom–up clustering» и «agglomerative clustering» соответственно. — *Прим. пер.*

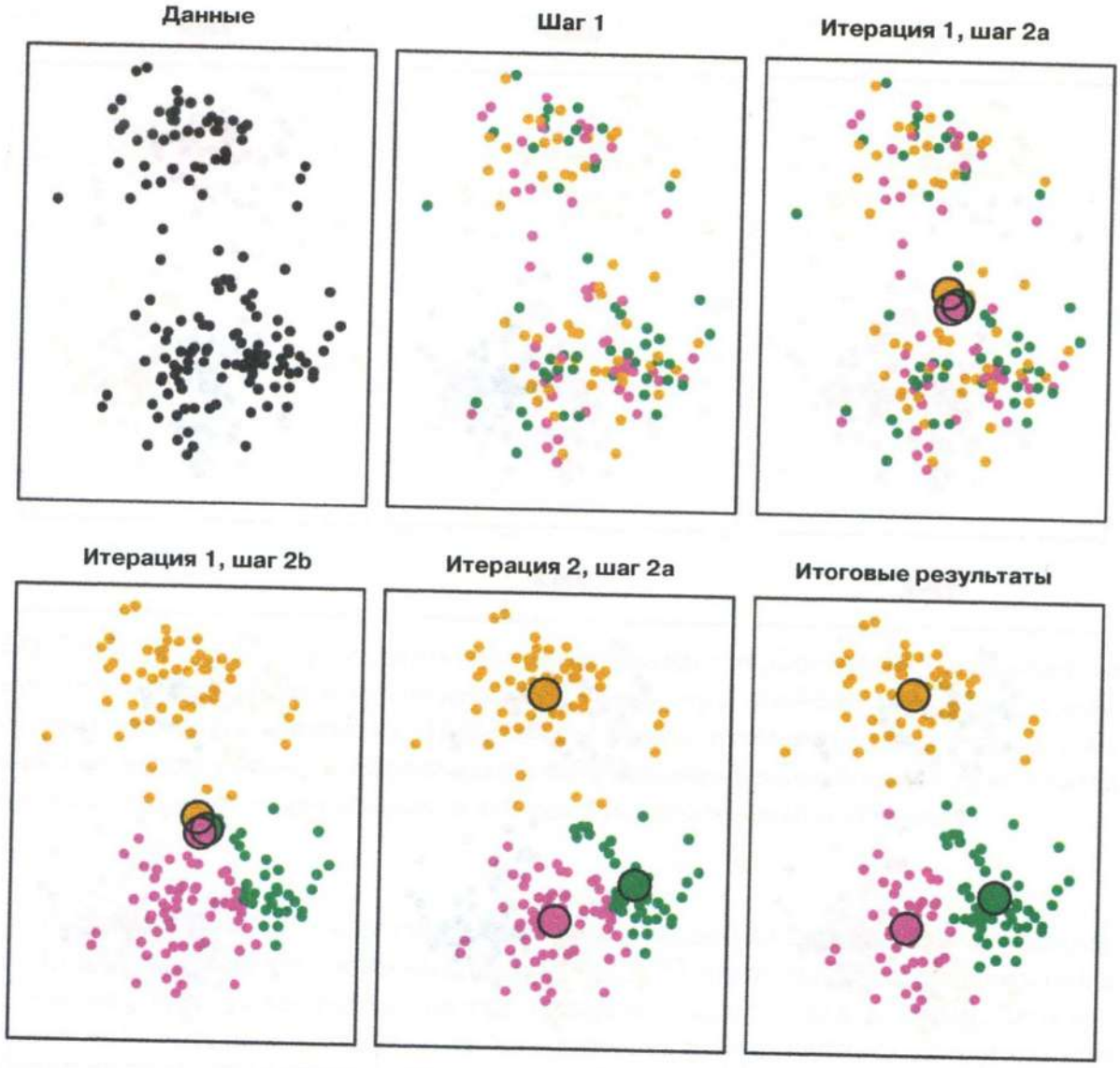

РИСУНОК 10.6. *Динамика выполнения алгоритма K средних на примере данных из рис. 10.5 с $K = 3$.* Слева вверху: *представлены исходные данные.* В центре вверху: *на первом шаге алгоритма наблюдения разбиваются на кластеры случайным образом.* Справа вверху: *на шаге 2(a) вычисляются центроиды кластеров. Они показаны в виде больших цветных кружков. Изначально центроиды почти полностью пересекаются, поскольку исходное разбиение на кластеры выполняется случайным образом.* Слева внизу: *на шаге 2(b) метка кластера присваивается каждому наблюдению в соответствии с ближайшим центроидом.* Справа внизу: *результаты, полученные после шести итераций*

см. рис. 10.9) начинается с листьев и продолжается путем их последовательного объединения в кластеры вплоть до самого «ствола». Мы начнем с обсуждения интерпретации дендрограммы, а затем обсудим собственно механизм выполнения иерархической кластеризации, т. е. способ построения дендрограммы.

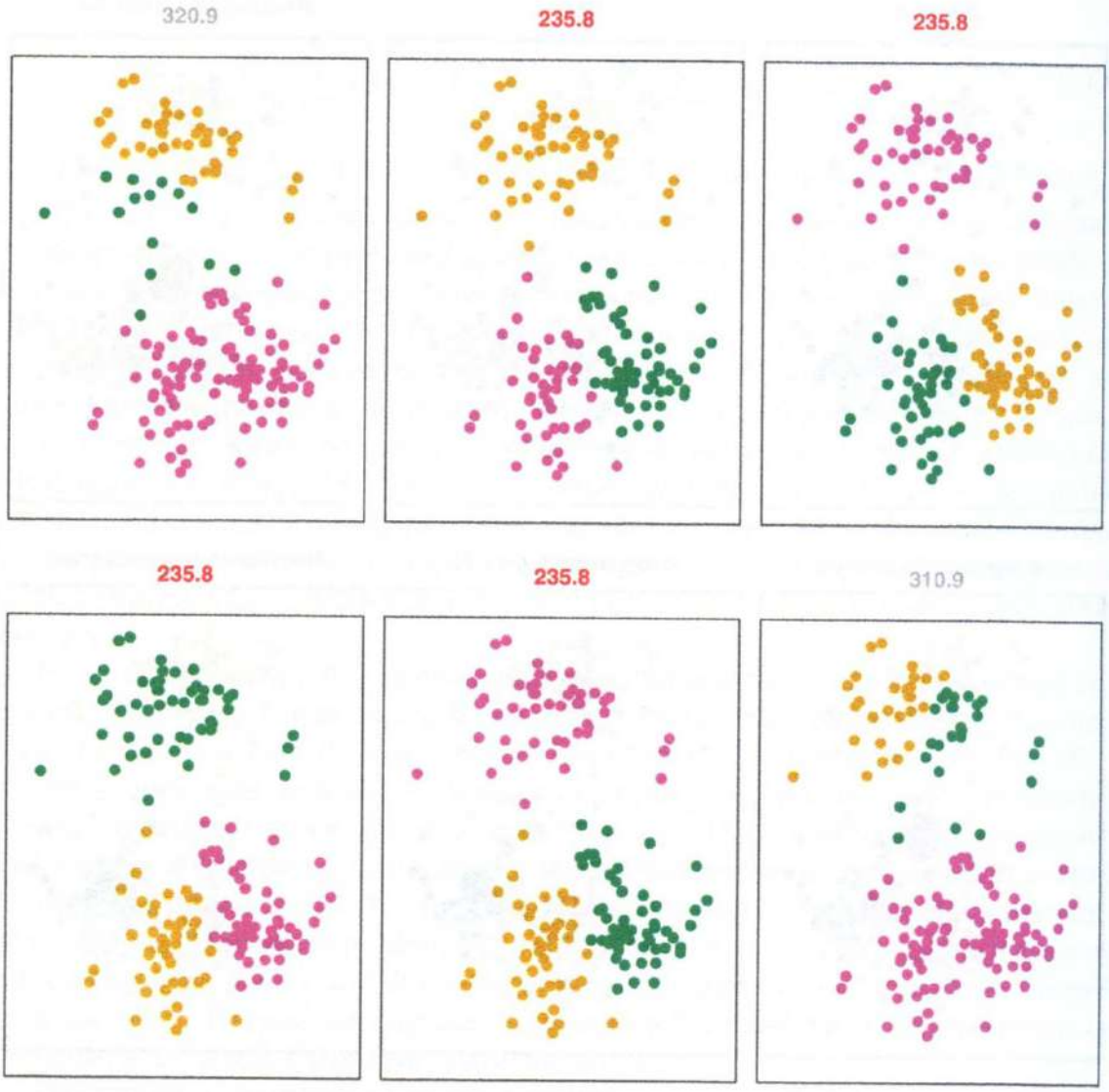

РИСУНОК 10.7. *Кластеризация по методу K средних ($K = 3$), выполненная шесть раз на данных из рис. 10.5 с разными случайными разбиениями наблюдений на кластеры на первом шаге алгоритма. Над каждым графиком показано значение целевой функции (10.11). Получены три локальных оптимума, один из которых привел к более низкому значению целевой функции и дал более качественное разбиение на кластеры. Все итерации, у которых значения целевой функции выделены красным цветом (235.8), дали одинаково оптимальные решения*

Интерпретация дендрограммы

Сначала рассмотрим представленный на рис. 10.8 набор имитированных данных, включающий 45 наблюдений и две переменные. Эти данные были сгенерированы на основе модели с тремя классами; истинные метки классов показаны разными цветами. Предположим, однако, что данные были получены без информации о метках классов и что мы хотели бы выполнить иерархическую кластеризацию этих наблюдений. Иерархическая кластеризация (с «полным присоединением», которое обсуждается ниже) дает результат, показанный слева на рис. 10.9. Как нам интерпретировать эту дендрограмму?

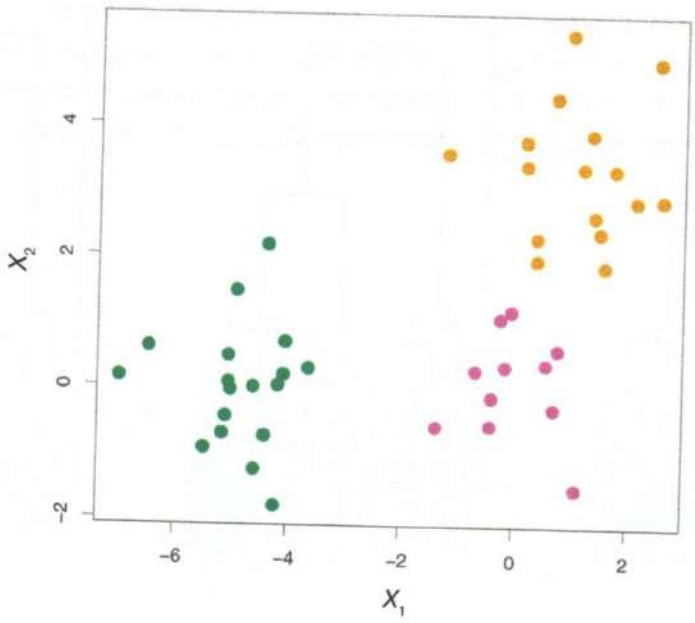

РИСУНОК 10.8. *Сорок пять имитированных наблюдений в двумерном пространстве. В действительности есть три отдельных класса, показанных разными цветами. Однако мы будем считать, что эти метки классов неизвестны, и попытаемся объединить наблюдения в группы для восстановления информации о классах по имеющимся данным*

Слева на рис. 10.9 каждый *лист* дендрограммы соответствует одному из 45 наблюдений, показанных на рис. 10.8. Однако по мере продвижения вверх по дереву некоторые листья начинают *сливаться* в ветви. Это наблюдения, которые похожи друг на друга. При дальнейшем продвижении вверх ветви также начинают сливаться — либо с листьями, либо с другими ветвями. Чем раньше (т. е. чем ниже в структуре дерева) происходит слияние, тем выше степень сходства между группами наблюдений. В то же время наблюдения, которые сливаются позже (т. е. недалеко от ствола дерева), могут быть довольно непохожими друг на друга. Более того, это утверждение можно сделать строгим: для любых двух наблюдений мы можем найти точку в структуре дерева, где происходит первое слияние ветвей, содержащих эти два наблюдения. Высота, на которой происходит слияние (показано на вертикальной оси), отражает степень отличий между двумя наблюдениями. Таким образом, наблюдения, которые сливаются у самого основания дерева, являются очень похожими друг на друга, тогда как наблюдения, которые сливаются у ствола дерева, обычно довольно разные.

Это подчеркивает одно очень важное обстоятельство, возникающее при интерпретации дендрограмм и часто понимаемое неправильно. Рассмотрим простую дендрограмму, приведенную слева на рис. 10.10, которая была получена в результате иерархической кластеризации девяти наблюдений. Можно увидеть, что наблюдения 5 и 7 довольно похожи друг на друга, поскольку они сливаются в наиболее низко расположенной точке дендрограммы. В то же время неверным (хотя и заманчивым) было бы заключение о том, что в силу своего близкого расположения друг к другу на этой дендрограмме похожими являются также наблюдения 9 и 2. В дей-

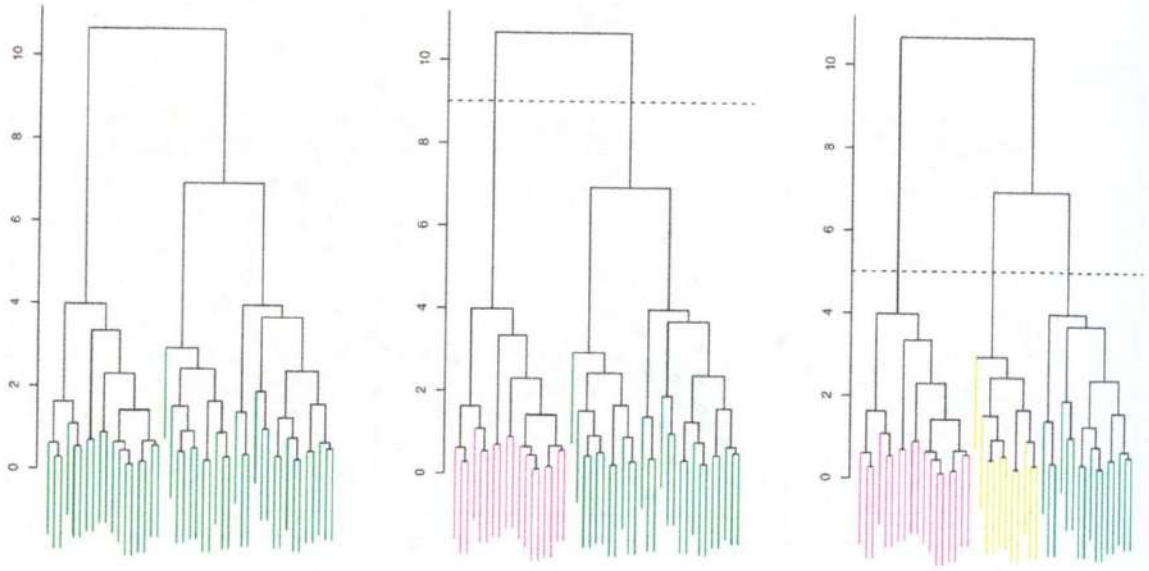

РИСУНОК 10.9. Слева: *дендрограмма, полученная путем иерархической кластеризации данных из рис. 10.8 с использованием полного присоединения и евклидова расстояния.* В центре: *дендрограмма, приведенная слева, рассечена на высоте 9 (отмечена прерывистой линией). Это приводит к двум кластерам, показанным при помощи разных цветов.* Справа: *дендрограмма, приведенная слева, рассечена на высоте 5. Это приводит к трем отдельным кластерам, показанным при помощи разных цветов. Обратите внимание, что цветовые метки не были задействованы в ходе кластеризации и используются здесь просто для наглядности*

ствительности, как следует из представленной на дендрограмме информации, наблюдение 9 похоже на наблюдение 2 не больше, чем оно похоже на наблюдения 8, 5 и 7. (Это можно увидеть справа на рис. 10.10, где приведены исходные данные.) С математической точки зрения, существует 2^{n-1} возможных вариантов упорядочения ветвей диаграммы, где n — это число листьев. Это обусловлено тем, что в каждой из $n-1$ точек слияния позиции двух слившихся ветвей можно поменять местами, не изменяя смысла дендрограммы. Следовательно, мы не можем делать выводы о сходстве двух наблюдений на основе их взаимной близости по *горизонтальной оси*. Вместо этого мы делаем выводы о сходстве двух наблюдений, исходя из положения на вертикальной оси той точки, где происходит слияние этих двух наблюдений.

Теперь, когда мы узнали, как интерпретировать дендрограмму, представленную на рис. 10.9 слева, мы можем продолжить и определить кластеры на ее основе. Для этого мы проводим горизонтальную линию, рассекающую диаграмму, как показано на рис. 10.9 в центре и справа. Отдельные наборы наблюдений, лежащие ниже этой линии, можно интерпретировать как кластеры. В центре на рис. 10.9 рассечение дендрограммы на высоте 9 приводит к получению двух кластеров, выделенных разными цветами. На графике справа рассечение дендрограммы на высоте 5 приводит к решению с тремя кластерами. Спускаясь ниже, можно выполнить дополнительные рассечения для получения любого числа кластеров — от 1 (что эквивалентно отсутствию рассечения) до n (что соответствует

рассечению на высоте 0, в результате которого каждое наблюдение оказывается в своем собственном кластере). Другими словами, высота, на которой выполняется рассечение дендрограммы, играет ту же роль, что и K в кластеризации по методу K средних: она задает число получаемых кластеров.

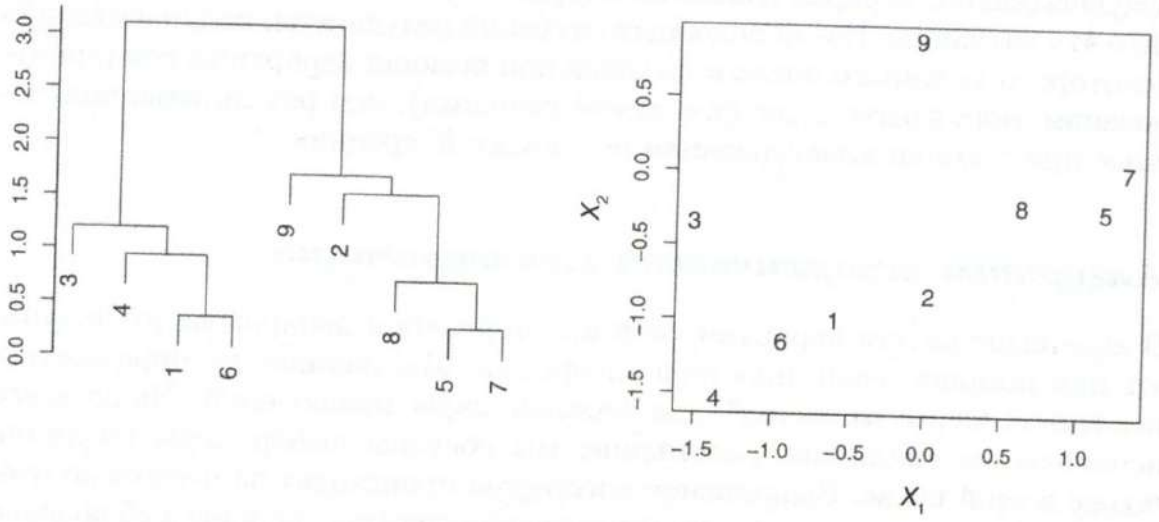

РИСУНОК 10.10. *Иллюстрация того, как правильно интерпретировать дендрограмму, построенную по девяти наблюдениям в двумерном пространстве.* Слева: *дендрограмма, созданная с использованием евклидова расстояния и полного присоединения. Наблюдения 5 и 7 довольно похожи друг на друга, равно как и наблюдения 1 и 6. Однако наблюдение 9 похоже на наблюдение 2 не больше, чем на наблюдения 8, 5 и 7, даже несмотря на то, что наблюдения 9 и 2 находятся близко друг к другу по горизонтали. Это обусловлено тем, что наблюдения 2, 8, 5 и 7 одновременно объединяются с наблюдением 9 на одинаковой высоте, равной примерно 1.8.* Справа: *исходные данные, по которым была построена дендрограмма, можно использовать для подтверждения того, что наблюдение 9 действительно похоже на наблюдение 2 не больше, чем на наблюдения 8, 5 и 7*

Таким образом, рис. 10.9 подчеркивает очень привлекательный аспект иерархической кластеризации: одну и ту же дендрограмму можно использовать для получения любого числа кластеров. На практике подходящее число кластеров выбирают путем визуального изучения дендрограммы, учитывая высоту, на которой происходит слияние отдельных ветвей, и желаемое число кластеров. В случае с рис. 10.9 можно было бы выбрать два или три кластера. Однако часто место рассечения дендрограммы не будет таким очевидным.

Термин «*иерархическая* кластеризация» предполагает, что кластеры, полученные в результате рассечения дендрограммы на некоторой высоте, обязательно являются «вложенными» в кластеры, полученные при рассечении этой дендрограммы на любой другой большей высоте. Однако для некоторых наборов данных это условие может не выполняться. Представьте, например, что наши наблюдения соответствуют группе людей, в которой в равных пропорциях представлены женщины и мужчины, а также жители Америки, Японии и Франции. Возможен сценарий, в кото-

ром оптимальное разбиение на две группы могло бы разделить этих людей по полу, а оптимальное разбиение на три группы могло бы разделить их по национальности. В таком случае кластеры не являются вложенными (в том смысле, что оптимальное разбиение на три группы не является естественным результатом предварительного разбиения на две группы). Следовательно, иерархическая кластеризация не смогла бы хорошо описать эту ситуацию. Из-за подобных ситуаций результаты, получаемые для некоторого заданного числа кластеров при помощи иерархической кластеризации, могут быть хуже (т. е. менее точными), чем результаты, получаемые при помощи кластеризации по методу K средних.

Алгоритм иерархической кластеризации

В ходе выполнения иерархической кластеризации дендрограмму получают при помощи очень простого алгоритма. Мы начинаем с определения некоторой меры *различия*[8] для каждой пары наблюдений. Чаще всего используется евклидово расстояние; мы обсудим выбор меры различия позже в этой главе. Выполнение алгоритма происходит за несколько итераций. Начиная с самого основания дендрограммы, каждое наблюдение рассматривается как самостоятельный кластер. Далее два наиболее похожих друг на друга кластера *сливаются*, в результате чего образуется $n-1$ кластеров. Затем два наиболее похожих друг на друга кластера снова сливаются, что приводит к формированию $n-2$ кластеров. Выполнение алгоритма продолжается аналогичным образом до тех пор, пока все наблюдения не становятся частью одного большого кластера, завершая формирование дендрограммы. На рис. 10.11 показаны первые несколько шагов алгоритма для данных из рис. 10.9. В обобщенном виде реализация иерархической кластеризации описана в алгоритме 10.2.

Алгоритм 10.2 Иерархическая кластеризация

1. Начните с n наблюдений и вычислите значения меры различия (например, евклидова расстояния) для всех $\binom{n}{2} = n(n-1)/2$ пар наблюдений. Рассматривайте каждое наблюдение как самостоятельный кластер.
2. Для $i = n, n-1, \ldots, 2$:
   (a) ранжируйте значения меры межкластерных различий для всех i кластеров и найдите пару наименее различающихся (т. е. наиболее похожих) кластеров. Объедините эти два кластера. Различие между этими двумя кластерами соответствует высоте, на которой должно происходить их слияние в дендрограмме;
   (b) вычислите новые значения меры различия для всех пар оставшихся $i-1$ кластеров.

[8] В оригинале используется термин «dissimilarity». — *Прим. пер.*

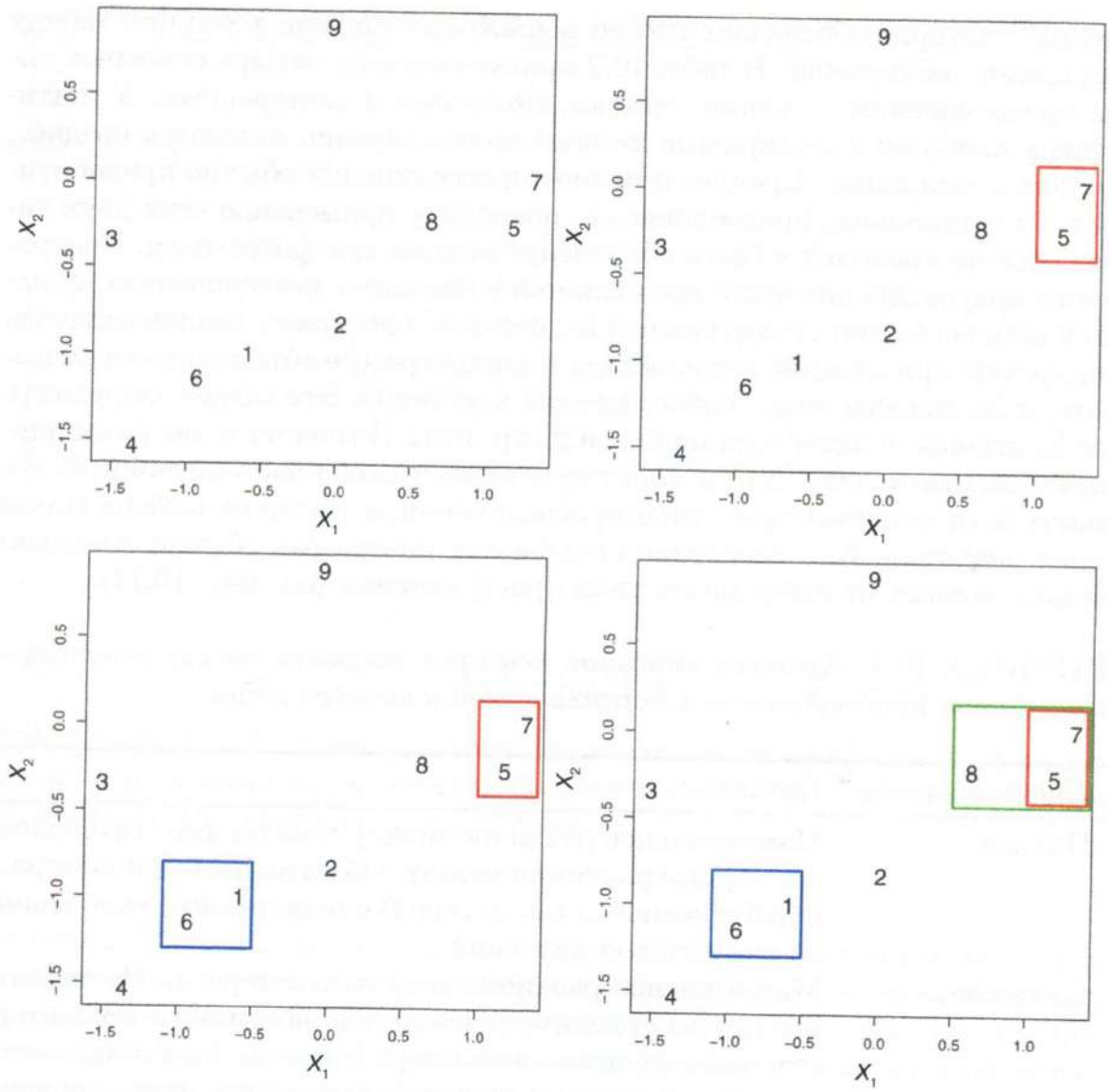

РИСУНОК 10.11. *Иллюстрация первых нескольких шагов алгоритма иерархической кластеризации с использованием полного присоединения и евклидова расстояния на примере данных из рис. 10.10.* Слева вверху: *сначала есть девять кластеров:* $\{1\}, \{2\}, \ldots, \{9\}$. Справа вверху: *два наиболее близких кластера* — $\{5\}$ *и* $\{7\}$ — *объединены в один кластер.* Слева внизу: *два наиболее близких кластера* — $\{6\}$ *и* $\{1\}$ — *объединены в один кластер.* Справа внизу: *два наиболее близких кластера* — $\{8\}$ *и* $\{5, 7\}$ — *объединены в один кластер по методу полного присоединения*

Этот алгоритм довольно прост, однако нерешенной остается одна проблема. Рассмотрим график, представленный справа внизу на рис. 10.11. Как мы определили, что кластер $\{5, 7\}$ должен быть объединен с кластером $\{8\}$? У нас есть понятие о различии между парой наблюдений, но как нам определить различие между двумя кластерами, если один или оба из них содержат несколько наблюдений? Идею о различии между парой наблюдений необходимо обобщить на случай пары *групп, включающих несколько наблюдений.* Это обобщение получают, вводя понятие *присоеди-*

присоединение

нения[9], которое обозначает способ выражения степени различий между группами наблюдений. В табл. 10.2 кратко описаны четыре основных типа присоединения — *полное*, *среднее*, *единичное* и *центроидное*. У статистиков наиболее популярными типами присоединения являются среднее, полное и единичное. Среднее и полное присоединения обычно предпочтительны единичному присоединению, поскольку применение этих двух типов обычно приводит к более сбалансированным дендрограммам. Центроидное присоединение часто применяется в геномных исследованиях, однако у него есть один существенный недостаток: оно может сопровождаться *инверсией*, при которой два кластера в дендрограмме объединяются на высоте, находящейся ниже любого из этих кластеров. Это может затруднять визуализацию и интерпретацию дендрограммы. Значения меры различия, вычисляемые на шаге 2(b) алгоритма иерархической кластеризации, будут зависеть от используемого типа присоединения, а также от выбора самой меры различия. Как следствие получаемая диаграмма обычно довольно сильно зависит от выбранного типа присоединения (см. рис. 10.12).

ТАБЛИЦА 10.2. *Краткое описание четырех наиболее часто используемых типов присоединения в иерархической кластеризации*

Присоединение	*Описание*
Полное	Максимальное различие между кластерами. Вычислите все парные различия между наблюдениями в кластере A и наблюдениями в кластере B и используйте *наибольшее* из полученных значений
Одиночное	Минимальное различие между кластерами. Вычислите все парные различия между наблюдениями в кластере A и наблюдениями в кластере B и используйте *наименьшее* из полученных значений. Одиночное присоединение может привести к «растянутым» кластерам, в которых отдельные наблюдения входят в кластер по одному за раз
Среднее	Среднее различие между кластерами. Вычислите все парные различия между наблюдениями в кластере A и наблюдениями в кластере B и используйте *среднее* из полученных значений
Центроидное	Различие между центроидом кластера A (средний вектор длиной p) и центроидом кластера B. Центроидное присоединение может приводить к нежелательному явлению *инверсии*

Выбор меры различия

До сих в примерах из этой главы в качестве меры различия применялось евклидово расстояние. Однако иногда предпочтительными могут оказаться и другие меры. Например, согласно *расстоянию, основанному на кор-*

[9] В оригинале используется термин «linkage». — *Прим. пер.*

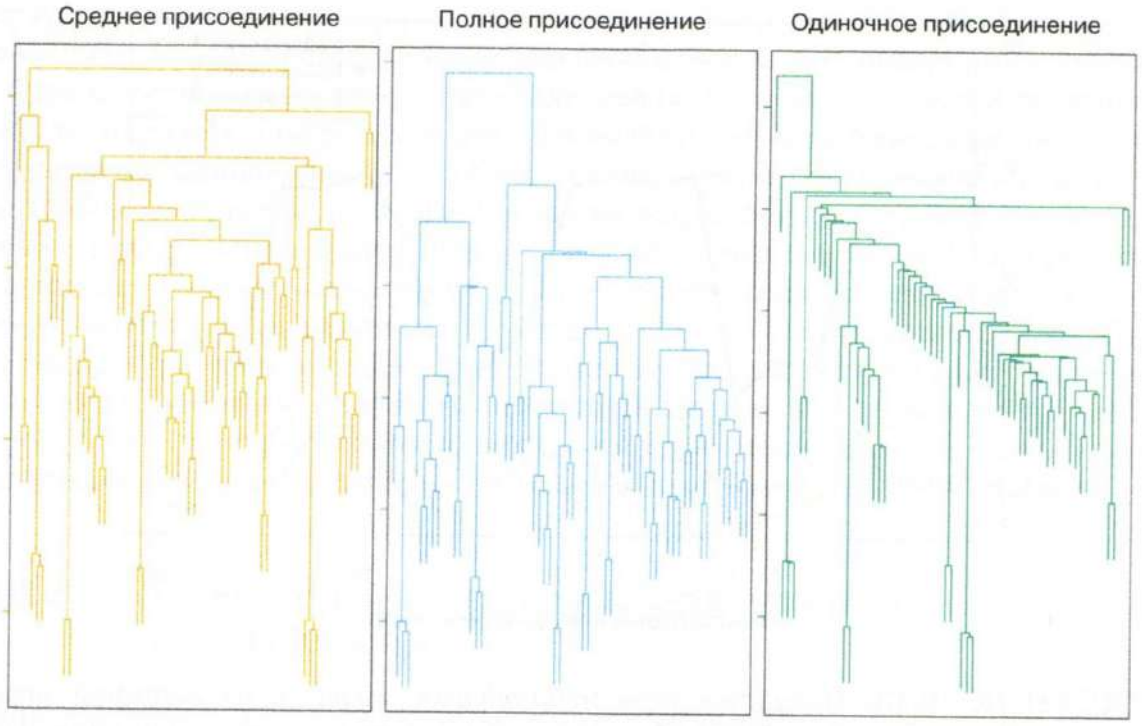

РИСУНОК 10.12. *Среднее, полное и одиночное присоединения применены к некоторому набору данных. Среднее и полное присоединения обычно дают более сбалансированные кластеры*

реляции, два наблюдения являются похожими, если их признаки тесно коррелируют, даже несмотря на то, что наблюдаемые значения могут при этом находиться далеко друг от друга в терминах евклидова расстояния. Это необычный пример использования корреляции, которая в большинстве случаев рассчитывается для переменных; здесь же она рассчитываться для профилей наблюдений. Рисунок 10.13 демонстрирует разницу между евклидовым расстоянием и расстоянием, основанным на корреляции. Расстояние, основанное на корреляции, уделяет основное внимание очертанию «профилей» наблюдений, а не относительной удаленности наблюдений друг от друга.

Выбор меры различия очень важен, поскольку он оказывает существенное влияние на итоговую дендрограмму. В целом необходимо очень внимательно относиться к типу кластеризуемых данных, а также к стоящему научному вопросу. Эти аспекты должны определять выбор меры различия для выполнения иерархической кластеризации.

Рассмотрим, например, интернет–магазин, который хотел бы сгруппировать своих покупателей на основе их прошлых покупок. Цель заключается в нахождении групп *похожих* покупателей, чтобы в последующем показывать каждой группе такие товары и рекламные объявления, которые с большой вероятностью будут интересны членам соответствующих групп. Предположим, что данные представлены в виде матрицы, в которой строки соответствуют покупателям, а столбцы — всем доступным для покупки товарам; в ячейках матрицы представлены значения, показывающие, сколько раз тот или иной товар был куплен тем или иным покупателем (0, если товар не был куплен ни разу, 1, если товар был

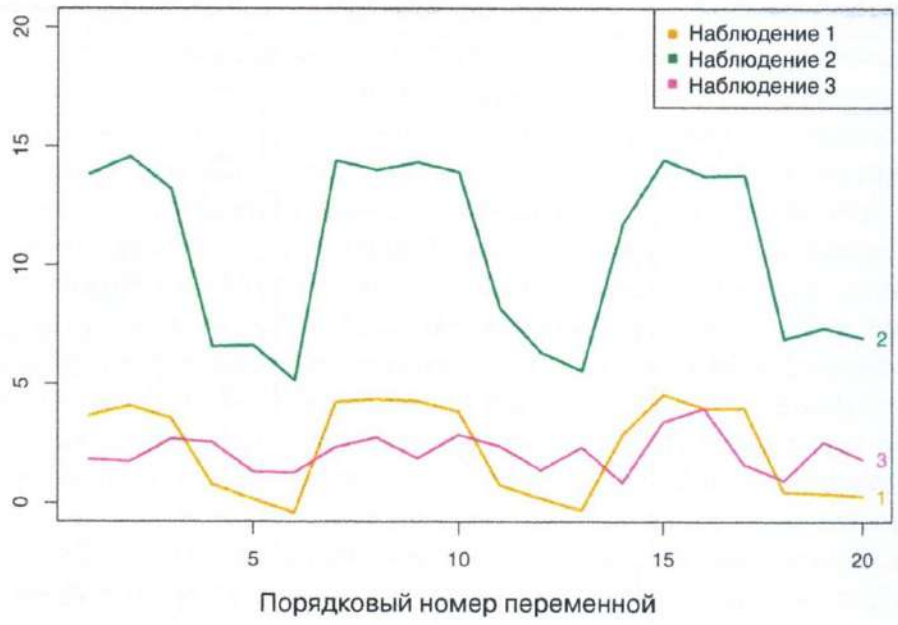

РИСУНОК 10.13. *Показаны три наблюдения, каждое из которых описано по 20 признакам. Наблюдения 1 и 3 имеют похожие значения для всех признаков, в связи с чем евклидово расстояние между ними невелико. Однако они очень слабо коррелируют, и поэтому с точки зрения корреляции они находятся далеко друг от друга. С другой стороны, значения большинства признаков у наблюдений 1 и 2 заметно различаются, в связи с чем евклидово расстояние между ними велико. В то же время эти два наблюдения сильно коррелируют, и поэтому с точки зрения корреляции расстояние между ними невелико*

куплен один раз, и т. д.). Какую меру различия следует применить для кластеризации покупателей? Если использовать евклидово расстояние, то покупатели, которые купили лишь небольшое количество разных товаров (т. е. нечастые посетители сайта магазина), попадут в один кластер. Такой результат может оказаться нежелательным. С другой стороны, если использовать расстояние, основанное на корреляции, то в одном кластере окажутся покупатели с одинаковыми предпочтениями (т. е. те, кто купил товары A и B, но никогда не покупал C и D), даже если некоторые покупатели с такими предпочтениями совершают очень много покупок в сравнении с другим покупателями. Следовательно, для этой конкретной задачи расстояние, основанное на корреляции, может оказаться более подходящим.

Помимо тщательного выбора меры различия, необходимо также подумать о том, следует ли перед ее вычислением масштабировать все переменные таким образом, чтобы их стандартные отклонения стали равны 1. Для иллюстрации этого аспекта мы продолжим рассматривать только что описанный пример с интернет–магазином. Некоторые товары могут быть популярнее других; например, покупатель мог бы приобретать десять пар носков в год, но при этом очень редко заказывать компьютеры. В связи с этим часто покупаемые товары вроде носков будут оказывать большее влияние на различия между товарами, а следовательно, и на результат кластеризации, чем редко покупаемые товары вроде компьютеров. Такой

результат может оказаться нежелательным. Если же перед вычислением различий между наблюдениями масштабировать все переменные таким образом, чтобы их стандартные стали равны 1, то вклад каждой переменной в получаемое решение задачи иерархической кластеризации, по сути, становится одинаковым. Подобное масштабирование переменных может потребоваться также в случае, когда они выражаются в разных единицах, иначе выбор шкалы измерения для той или иной переменной (например, сантиметры вместо километров) окажет значительное влияние на получаемое значение меры различия. Конечно, решение о выполнении масштабирования переменных перед вычислением меры различия зависит также от характера решаемой практической задачи. На рис. 10.14 приведен пример. Как видим, решение о масштабировании переменных перед выполнением кластерного анализа является важным также и для метода K средних.

10.3.3 Практические аспекты применения кластеризации

Кластеризация может оказаться очень полезным инструментом при решении задач обучения без учителя. Однако при выполнении кластеризации возникает ряд затруднений. Рассмотрим эти затруднения.

Небольшие решения с большими последствиями

При выполнении кластерного анализа необходимо принять ряд решений:

- Следует ли выполнить определенную стандартизацию переменных или наблюдений? Возможно, например, что переменные следует центрировать и масштабировать таким образом, чтобы их средние значения стали равны 0, а стандартные отклонения — 1.
- В случае с иерархической кластеризацией:
  - Какую меру различия выбрать?
  - Какой тип присоединения выбрать?
  - На какой высоте рассекать дендрограмму для получения кластеров?
- В случае с кластеризацией по методу K средних сколько кластеров нам следует искать в данных?

Каждое из этих решений может оказать большое влияние на получаемые результаты. На практике мы обычно проверяем несколько разных вариантов, пытаясь найти наиболее полезное или интерпретируемое решение. Простого ответа не существует ни для одного из этих методов — рассматривать следует любое решение, которое обнажает интересные свойства данных.

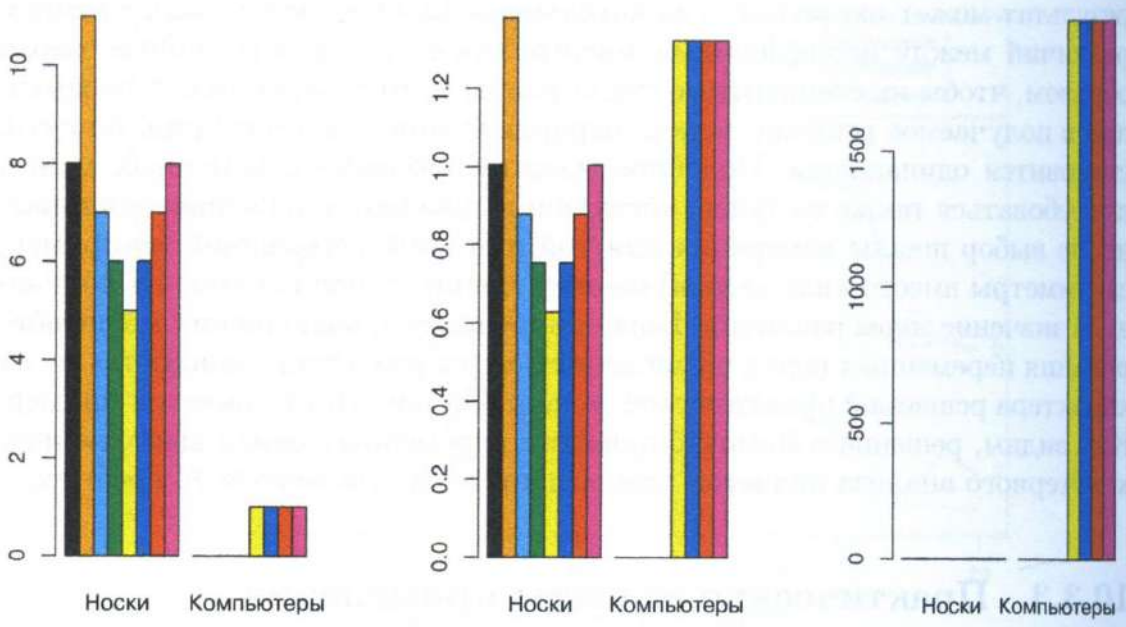

РИСУНОК 10.14. *Необычный интернет–магазин продает два вида товаров: носки и компьютеры.* Слева: *показано количество покупок носков и компьютеров, совершенных восемью покупателями. Каждый покупатель выделен отдельным цветом. Если выразить различия между наблюдениями при помощи евклидова расстояния, рассчитанного по исходным значениям переменных, то полученные различия в основном будут определяться количеством покупок носков, а количество покупок компьютеров будет оказывать очень слабый эффект. Такая ситуация может быть нежелательной, поскольку: (1) компьютеры являются более дорогими, чем носки, в связи с чем продавец может быть больше заинтересован в подталкивании посетителей магазина к покупке компьютеров, а не носков; (2) большая разница в количестве носков, купленных двумя покупателями, может оказаться менее информативной в отношении покупательских предпочтений этих людей, чем небольшое различие в количестве купленных компьютеров.* В центре: *показаны те же данные, но после деления каждой переменной на ее стандартное отклонение. Теперь количество приобретенных компьютеров будет оказывать гораздо большее влияние на уровень различий между наблюдениями.* Справа: *показаны те же данные, но теперь по оси Y отложены общие суммы (в долларах), потраченные каждым покупателем на носки и компьютеры. Поскольку компьютеры гораздо дороже носков, то полученные различия между наблюдениями будут определяться историей покупок компьютеров*

Проверка качества полученных кластеров

Мы всегда обнаружим какие–то кластеры, выполняя кластеризацию на том или ином наборе данных. Поэтому важно знать, соответствуют ли найденные кластеры реальным группам в данных или это просто результат *кластеризации содержащегося в данных шума.* Например, если у нас в распоряжении окажется независимый набор наблюдений, образуют ли эти наблюдения тот же набор кластеров? Ответить на этот вопрос нелег-

ко. Существует целый ряд методов, позволяющих присвоить кластеру p–значение для описания вероятности того, что этот кластер получен неслучайно. Однако мнения в отношении наиболее оптимального метода для решения этой задачи расходятся. Более подробное обсуждение данной проблемы можно найти в работе Hastie et al. (2009).

Другие аспекты кластерного анализа

И метод K средних, и иерархическая кластеризация отнесут каждое наблюдение к определенному кластеру. Однако иногда такой результат может оказаться неприемлемым. Предположим, например, что большинство наблюдений в действительности принадлежит к небольшому (и неизвестному) количеству групп, а остальная часть наблюдений существенно отличается как друг от друга, так и от большинства других наблюдений. Поскольку метод K средних и иерархическая кластеризация заставляют *каждое* наблюдение быть частью того или иного кластера, то найденные кластеры могут оказаться очень неоднородными из–за наличия выбросов, которые на самом деле не принадлежат ни к одному из этих кластеров. Привлекательным альтернативным подходом, позволяющим учесть наличие подобных выбросов, является применение смешанных моделей[10]. По сути, эти модели эквивалентны *мягкой* версии метода K средних, описанной в книге Hastie et al. (2009).

Кроме того, методы кластеризации обычно не очень устойчивы к возмущениям в данных. Предположим, например, что мы выполняем кластеризацию n наблюдений, а затем повторяем эту процедуру после удаления нескольких случайно выбранных наблюдений. Хотелось бы надеяться, что полученные два набора кластеров будут довольно похожими, однако часто это совсем не так!

Здравый подход к интерпретации результатов кластеризации

Мы описали несколько трудностей, связанных с выполнением кластерного анализа. Однако при правильном использовании кластеризация может быть очень полезным и эффективным инструментом. Мы уже отметили, что небольшие решения, определяющие ход выполнения кластеризации (например, необходимость стандартизации и выбор типа присоединения), могут оказать большое влияние на результаты. Поэтому мы рекомендуем выполнять кластеризацию с разными сочетаниями этих параметров и рассматривать всю совокупность результатов с целью обнаружения повторяющихся закономерностей. Поскольку кластеризация может давать неустойчивые решения, мы рекомендуем выполнять анализ на нескольких частях данных, чтобы получить представление об устойчивости получаемых кластеров. И самое главное: мы должны быть осторожны с представлением результатов кластерного анализа. Эти результаты не должны восприниматься как абсолютная истина в отношении данных. Вместо этого они должны служить начальной точкой для разработки той или иной научной гипотезы и последующего исследования, которое в идеале должно выполняться на независимом наборе данных.

[10] В оригинале используется термин «mixture models». — *Прим. пер.*