8

Кластерный анализ: Основные понятия и Алгоритмы

Кластерный анализ делит данные на группы (кластеры), которые являются значимыми, полезными или и тем, и другим. Если целью являются значимые группы, то кластеры должны отражать естественную структуру данных. Однако в некоторых случаях кластерный анализ является лишь полезной отправной точкой для других целей, таких как обобщение данных. Будь то понимание или полезность, кластерный анализ уже давно играет важную роль в самых разных областях: психологии и других социальных науках, биологии, статистике, распознавании образов, поиске информации, машинном обучении и интеллектуальном анализе данных.

Было много применений кластерного анализа для решения практических задач. Мы приводим несколько конкретных примеров, упорядоченных по тому, является ли целью кластеризации понимание или полезность.

Кластеризация для понимания Классы или концептуально значимые группы объектов, которые имеют общие характеристики, играют важную роль в том, как люди анализируют и описывают мир. Действительно, люди умеют делить объекты на группы (кластеризация) и относить к этим группам отдельные объекты (классификация). Например, даже относительно маленькие дети могут быстро обозначить объекты на фотографии как здания, транспортные средства, людей, животных, растения и т. д. В контексте понимания данных кластеры являются потенциальными классами, а кластерный анализ — это изучение методов автоматический поиск классов. Ниже приведены некоторые примеры:

- Биология. Биологи потратили много лет на создание таксономии (иерархической классификации) всех живых существ: царства, типа, класса, отряда, семейства, рода и вида. Таким образом, возможно, неудивительно, что большая часть ранних работ по кластерному анализу была направлена на создание дисциплины математической таксономии, которая могла бы автоматически находить такие классификационные структуры. Совсем недавно биологи применили кластеризацию для анализа больших объемов генетической информации, которая сейчас доступна. Например, кластеризация использовалась для поиска групп генов, выполняющих схожие функции.

- Поиск информации. Всемирная паутина состоит из миллиардов

  Веб-страницы и результаты запроса к поисковой системе могут возвращать тысячи страниц. Кластеризацию можно использовать для группировки результатов поиска в небольшое количество кластеров, каждый из которых фиксирует определенный аспект запроса. Например, запрос «фильм» может вернуть веб-страницы, сгруппированные по таким категориям, как обзоры, трейлеры, звезды и кинотеатры. Каждую категорию (кластер) можно разбить на подкатегории (подкластеры), создавая иерархическую структуру, которая дополнительно помогает пользователю исследовать результаты запроса.

- Климат. Понимание климата Земли требует поиска закономерностей в атмосфере и океане. С этой целью был применен кластерный анализ для выявления закономерностей атмосферного давления в полярных регионах и районах океана, которые оказывают существенное влияние на климат суши.

- Психология и медицина. Заболевание или состояние часто имеет ряд вариаций, и для выявления этих различных подкатегорий можно использовать кластерный анализ. Например, кластеризация использовалась для выявления различных типов депрессии. Кластерный анализ также можно использовать для выявления закономерностей в пространственном или временном распределении заболевания.

- Бизнес. Предприятия собирают большие объемы информации о текущих и потенциальных клиентах. Кластеризация может использоваться для разделения клиентов на небольшое количество групп для дополнительного анализа и маркетинговой деятельности.

Кластеризация для утилиты. Кластерный анализ обеспечивает абстракцию отдельных объектов данных на кластеры, в которых находятся эти объекты данных. Кроме того, некоторые методы кластеризации характеризуют каждый кластер с точки зрения прототипа кластера; т. е. объект данных, который представляет другие объекты в кластере. Эти прототипы кластеров могут быть использованы в качестве основы для

количество методов анализа или обработки данных. Таким образом, с точки зрения полезности кластерный анализ — это исследование методов поиска наиболее репрезентативных прототипов кластеров.

- Подведение итогов. Многие методы анализа данных, такие как регрессия или PCA, имеют временную или пространственную сложность O(m2) или выше (где m — количество объектов) и, следовательно, непрактичны для больших наборов данных. Однако вместо применения алгоритма ко всему набору данных его можно применить к сокращенному набору данных, состоящему только из прототипов кластеров. В зависимости от типа анализа, количества прототипов и точности, с которой прототипы представляют данные, результаты могут быть сопоставимы с теми, которые были бы получены, если бы можно было использовать все данные.

- Сжатие. Прототипы кластеров также можно использовать для сжатия данных. В частности, создается таблица, состоящая из прототипов для каждого кластера; т. е. каждому прототипу присваивается целочисленное значение, которое является его позицией (индексом) в таблице. Каждый объект представлен индексом прототипа, связанного с его кластером. Этот тип сжатия известен как векторное квантование и часто применяется к данным изображения, звука и видео, где (1) многие объекты данных очень похожи друг на друга, (2) допустима некоторая потеря информации и (3) а желательно существенное уменьшение размера данных.

- Эффективный поиск ближайших соседей. Для поиска ближайших соседей может потребоваться вычисление попарного расстояния между всеми точками. Часто кластеры и их кластерные прототипы можно найти гораздо эффективнее. Если объекты относительно близки к прототипу своего кластера, то мы можем использовать прототипы, чтобы уменьшить количество вычислений расстояния, необходимых для поиска ближайших соседей объекта. Интуитивно понятно, что если два прототипа кластера находятся далеко друг от друга, то объекты в соответствующих кластерах не могут быть ближайшими соседями друг друга. Следовательно, для нахождения ближайших соседей объекта необходимо лишь вычислить расстояние до объектов в соседних кластерах, причем близость двух кластеров измеряется расстоянием между их прототипами. Эта идея конкретизируется в упражнении 25 на стр. 94.

В этой главе представлено введение в кластерный анализ. Мы начнем с общего обзора кластеризации, включая обсуждение различных подходов к разделению объектов на наборы кластеров и различные типы кластеров. Затем мы опишем три конкретных метода кластеризации, которые представляют

широкие категории алгоритмов и иллюстрируют различные концепции: K-средние, агломеративная иерархическая кластеризация и DBSCAN. Последний раздел этой главы посвящен валидности кластеров — методам оценки качества кластеров, создаваемых алгоритмом кластеризации. Более продвинутые концепции и алгоритмы кластеризации будут обсуждаться в главе 9. По возможности мы обсуждаем сильные и слабые стороны различных схем. Кроме того, в библиографических примечаниях содержатся ссылки на соответствующие книги и статьи, в которых кластерный анализ рассматривается более глубоко.

8.1 Обзор

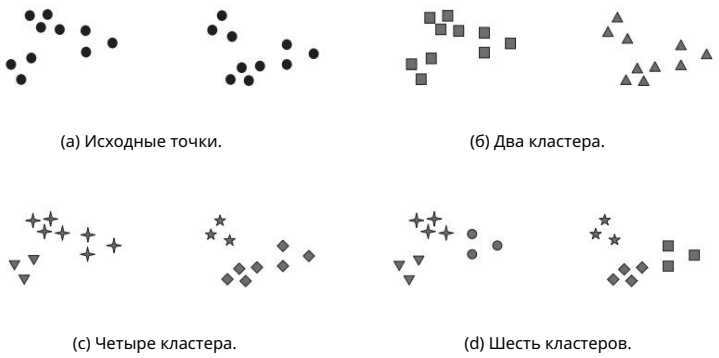
Прежде чем обсуждать конкретные методы кластеризации, мы предоставим некоторую необходимую информацию. Во-первых, мы даем определение кластерному анализу, показывая, почему он сложен, и объясняя его связь с другими методами группировки данных. Затем мы исследуем две важные темы: (1) различные способы группировки набора объектов в набор кластеров и (2) типы кластеров.

8.1.1 Что такое кластерный анализ?

Кластерный анализ группирует объекты данных только на основе информации, содержащейся в данных, которая описывает объекты и их отношения. Цель состоит в том, чтобы объекты внутри группы были похожи (или связаны) друг с другом и отличались (или не были связаны) с объектами в других группах. Чем больше сходство (или однородность) внутри группы и чем больше разница между группами, тем лучше или отчетливее кластеризация.

Во многих приложениях понятие кластера не определено четко. Чтобы лучше понять сложность принятия решения о том, что представляет собой кластер, рассмотрим рисунок 8.1, на котором показаны двадцать точек и три различных способа их разделения на кластеры. Форма маркеров указывает на принадлежность к кластеру. На рисунках 8.1(b) и 8.1(d) данные разделены на две и шесть частей соответственно. Однако кажущееся разделение каждого из двух более крупных кластеров на три подкластера может быть просто артефактом зрительной системы человека. Кроме того, возможно, не будет неразумным сказать, что точки образуют четыре кластера, как показано на рисунке 8.1(c). Этот рисунок показывает, что определение кластера неточно и что лучшее определение зависит от характера данных и желаемых результатов.

Кластерный анализ связан с другими методами, которые используются для разделения объектов данных на группы. Например, кластеризацию можно рассматривать как форму классификации, поскольку она создает маркировку объектов метками класса (кластера). Однако он извлекает эти метки только из данных. Напротив, классификация



(а) Исходные точки. (б) Два кластера.

(c) Четыре кластера. (d) Шесть кластеров.

Рисунок 8.1. Различные способы кластеризации одного и того же набора точек.

в смысле Главы 4 – контролируемая классификация; т. е. новым немаркированным объектам присваивается метка класса с использованием модели, разработанной на основе объектов с известными метками классов. По этой причине кластерный анализ иногда называют неконтролируемой классификацией. Когда термин «классификация» используется без каких-либо уточнений в рамках интеллектуального анализа данных, он обычно относится к контролируемой классификации.

Кроме того, хотя термины «сегментация» и «разделение» иногда используются как синонимы кластеризации, эти термины часто используются для подходов, выходящих за традиционные рамки кластерного анализа. Например, термин секционирование часто используется в связи с методами, которые делят графы на подграфы и не тесно связаны с кластеризацией. Сегментация часто означает разделение данных на группы с использованием простых методов; например, изображение можно разделить на сегменты только на основе интенсивности и цвета пикселей, или людей можно разделить на группы в зависимости от их дохода. Тем не менее, некоторая работа по разбиению графов, а также по сегментации изображений и рынков связана с кластерным анализом.

8.1.2 Различные типы кластеризации

Всю совокупность кластеров обычно называют кластеризацией, и в этом разделе мы различаем различные типы кластеров: иерархические (вложенные) и секционированные (невложенные), исключающие, перекрывающиеся и нечеткие, а также полные и частичные.

Иерархическая и партиционная. Наиболее часто обсуждаемое различие между различными типами кластеризации заключается в том, является ли набор кластеров вложенным.

или невложенные, или, в более традиционной терминологии, иерархические или секционированные. Секциональная кластеризация — это просто разделение набора объектов данных на непересекающиеся подмножества (кластеры), так что каждый объект данных находится ровно в одном подмножестве. В отдельности каждая совокупность кластеров на рисунках 8.1 (b–d) представляет собой секционную кластеризацию.

Если мы разрешим кластерам иметь подкластеры, то мы получим иерархическую кластеризацию, которая представляет собой набор вложенных кластеров, организованных в виде дерева. Каждый узел (кластер) дерева (кроме листовых узлов) представляет собой объединение его дочерних элементов (подкластеров), а корнем дерева является кластер, содержащий все объекты. Часто, но не всегда, листья дерева представляют собой одноэлементные кластеры отдельных объектов данных. Если мы позволим кластерам быть вложенными, то одна из интерпретаций рисунка 8.1(a) состоит в том, что он состоит из двух подкластеров (рис. 8.1(b)), каждый из которых, в свою очередь, имеет три подкластера (рис. 8.1(d)). Кластеры, показанные на рисунках 8.1 (a–d), если рассматривать их в таком порядке, также образуют иерархическую (вложенную) кластеризацию с соответственно 1, 2, 4 и 6 кластерами на каждом уровне. Наконец, обратите внимание, что иерархическую кластеризацию можно рассматривать как последовательность секционированных кластеров, и секционную кластеризацию можно получить, взяв любой член этой последовательности; т.е. путем разрезания иерархического дерева на определенном уровне.

Эксклюзивное, перекрывающееся и нечеткое Все кластеризации, показанные на рис. 8.1, являются исключительными, поскольку они относят каждый объект к одному кластеру.
Существует множество ситуаций, в которых точку можно разумно поместить более чем в один кластер, и эти ситуации лучше решать с помощью немонопольной кластеризации. В самом общем смысле перекрывающаяся или неисключительная кластеризация используется для отражения того факта, что объект может одновременно принадлежать более чем одной группе (классу). Например, человек в университете может быть как зачисленным студентом, так и сотрудником университета. Неэксклюзивная кластеризация также часто используется, когда, например, объект находится «между» двумя или более кластерами и может быть обоснованно отнесен к любому из этих кластеров.
Представьте себе точку на полпути между двумя кластерами на рис. 8.1. Вместо того, чтобы произвольно относить объект к одному кластеру, он помещается во все «одинаково хорошие» кластеры.

В нечеткой кластеризации каждый объект принадлежит каждому кластеру с весом членства от 0 (абсолютно не принадлежит) до 1 (абсолютно принадлежит). Другими словами, кластеры рассматриваются как нечеткие множества. (Математически нечеткое множество — это такое множество, в котором объект принадлежит любому множеству с весом от 0 до 1. При нечеткой кластеризации мы часто налагаем дополнительное ограничение, согласно которому сумма весов каждого объекта должна равняться 1. .) Аналогично, методы вероятностной кластеризации вычисляют вероятность того, что каждый

точка принадлежит каждому кластеру, и эти вероятности также должны быть в сумме равны 1. Поскольку веса членства или вероятности для любого объекта в сумме равны 1, нечеткая или вероятностная кластеризация не подходит для истинных многоклассовых ситуаций, таких как случай сотрудника-студента. , где объект принадлежит нескольким классам. Вместо этого эти подходы наиболее подходят для того, чтобы избежать произвольного отнесения объекта только к одному кластеру, когда их может быть близко к нескольким. На практике нечеткая или вероятностная кластеризация часто преобразуется в исключительную кластеризацию путем отнесения каждого объекта к кластеру, в котором вес или вероятность его членства являются наибольшими.

Полная и частичная. Полная кластеризация присваивает каждый объект кластеру, а частичная — нет. Мотивом для частичной кластеризации является то, что некоторые объекты в наборе данных могут не принадлежать к четко определенным группам. Зачастую объекты в наборе данных могут представлять собой шум, выбросы или «неинтересный фон». Например, некоторые газетные статьи могут иметь общую тему, например, глобальное потепление, в то время как другие статьи носят более общий или единственный в своем роде характер. Таким образом, чтобы найти важные темы в статьях за прошлый месяц, нам может потребоваться искать только группы документов, тесно связанных общей темой. В других случаях желательна полная кластеризация объектов.
Например, приложение, которое использует кластеризацию для организации документов для просмотра, должно гарантировать возможность просмотра всех документов.

8.1.3 Различные типы кластеров

Кластеризация направлена на поиск полезных групп объектов (кластеров), полезность которых определяется целями анализа данных. Неудивительно, что существует несколько различных понятий кластера, которые оказываются полезными на практике. Чтобы наглядно проиллюстрировать различия между этими типами кластеров, мы используем двумерные точки, как показано на рисунке 8.2, в качестве объектов данных. Однако мы подчеркиваем, что описанные здесь типы кластеров в равной степени применимы и к другим типам данных.

Хорошо разделенный Кластер — это набор объектов, в котором каждый объект ближе (или более похож) к любому другому объекту в кластере, чем к любому объекту, не входящему в кластер. Иногда порог используется, чтобы указать, что все объекты в кластере должны быть достаточно близки (или похожи) друг к другу. Это идеалистическое определение кластера удовлетворяется только тогда, когда данные содержат естественные кластеры, находящиеся довольно далеко друг от друга. На рисунке 8.2(a) показан пример хорошо разделенных кластеров, состоящих из двух групп точек в двумерном пространстве. Расстояние между любыми двумя точками в разных группах больше, чем

расстояние между любыми двумя точками внутри группы. Хорошо разделенные кластеры не обязательно должны быть шаровидными, но могут иметь любую форму.

Основанный на прототипе Кластер — это набор объектов, в котором каждый объект ближе (более похож) к прототипу, определяющему кластер, чем к прототипу любого другого кластера. Для данных с непрерывными атрибутами прототипом кластера часто является центроид, т.е. среднее значение всех точек в кластере. Когда центроид не имеет смысла, например, когда данные имеют категориальные атрибуты, прототипом часто является медоид, т. е. наиболее репрезентативная точка кластера. Для многих типов данных прототип можно рассматривать как наиболее центральную точку, и в таких случаях мы обычно называем кластеры на основе прототипов кластерами на основе центра. Неудивительно, что такие скопления имеют тенденцию быть шаровидными. На рисунке 8.2(b) показан пример кластеров с центром.

На основе графа. Если данные представлены в виде графа, где узлы являются объектами, а связи представляют связи между объектами (см. раздел 2.1.2), то кластер можно определить как связный компонент; т. е. группа объектов, которые связаны друг с другом, но не имеют связи с объектами вне группы. Важным примером кластеров на основе графов являются кластеры на основе смежности, в которых два объекта соединяются только в том случае, если они находятся на определенном расстоянии друг от друга. Это означает, что каждый объект в кластере на основе смежности находится ближе к какому-либо другому объекту в кластере, чем к любой точке другого кластера. На рисунке 8.2(c) показан пример таких кластеров для двумерных точек. Такое определение кластера полезно, когда кластеры нерегулярны или переплетены, но может возникнуть проблема при наличии шума, поскольку, как показано на двух сферических кластерах на рисунке 8.2(c), небольшой мост из точек может объединить два отдельных кластера.

Возможны и другие типы кластеров на основе графов. Один из таких подходов (раздел 8.3.2) определяет кластер как клику; т. е. набор узлов графа, которые полностью связаны друг с другом. В частности, если мы добавим связи между объектами в порядке их расстояния друг от друга, кластер образуется, когда набор объектов образует клику. Подобно кластерам на основе прототипов, такие кластеры имеют тенденцию быть шаровидными.

На основе плотности Кластер — это плотная область объектов, окруженная областью с низкой плотностью. На рисунке 8.2(d) показаны некоторые кластеры на основе плотности для данных, созданных путем добавления шума к данным рисунка 8.2(c). Два круговых кластера не объединяются, как на рис. 8.2(c), потому что мост между ними растворяется в шуме. Аналогично, кривая, представленная на рис. 8.2(c), также

исчезает в шуме и не образует кластер на рисунке 8.2(d). Определение кластера на основе плотности часто используется, когда кластеры нерегулярны или переплетены, а также когда присутствуют шум и выбросы. Напротив, определение кластера на основе смежности не будет хорошо работать для данных рисунка 8.2(d), поскольку шум будет иметь тенденцию образовывать мосты между кластерами.

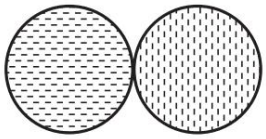
Совместное свойство (концептуальные кластеры) В более общем смысле мы можем определить кластер как набор объектов, имеющих некоторое общее свойство. Это определение охватывает все предыдущие определения кластера; например, объекты в кластере на основе центра имеют общее свойство: все они расположены ближе всего к одному и тому же центроиду или медоиду. Однако подход с общей собственностью также включает новые типы кластеров. Рассмотрим кластеры, показанные на рисунке 8.2(e). Треугольная область (кластер) соседствует с прямоугольной, и имеются два переплетенных круга (кластера). В обоих случаях алгоритму кластеризации потребуется очень специфическая концепция кластера для успешного обнаружения этих кластеров. Процесс поиска таких кластеров называется концептуальной кластеризацией. Однако слишком сложное понятие кластера увело бы нас в область распознавания образов, и поэтому в этой книге мы рассматриваем только более простые типы кластеров.

Дорожная карта

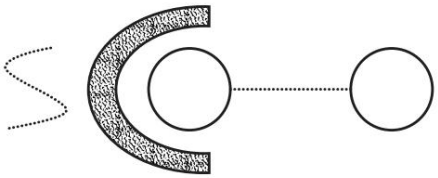
В этой главе мы используем следующие три простых, но важных метода, чтобы представить многие концепции кластерного анализа.

- K-средство. Это метод секционной кластеризации на основе прототипа, который пытается найти указанное пользователем количество кластеров (K), которые представлены их центроидами.
- Агломеративная иерархическая кластеризация. Этот подход к кластеризации относится к набору тесно связанных методов кластеризации, которые создают иерархическую кластеризацию, начиная с каждой точки как одноэлементного кластера, а затем многократно объединяя два ближайших кластера, пока не останется один всеобъемлющий кластер. Некоторые из этих методов имеют естественную интерпретацию с точки зрения кластеризации на основе графов, тогда как другие интерпретируются с точки зрения подхода, основанного на прототипах.
- ДБСКАН. Это алгоритм кластеризации на основе плотности, который создает секционную кластеризацию, при которой количество кластеров автоматически определяется алгоритмом. Точки в регионах с низкой плотностью классифицируются как шум и опускаются; таким образом, DBSCAN не обеспечивает полную кластеризацию.

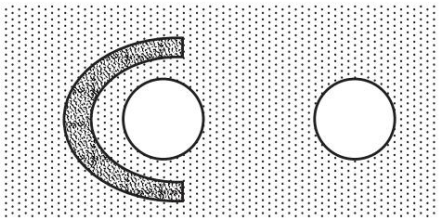
(a) Хорошо разделенные кластеры. Каждая точка находится ближе ко всем точкам своего кластера, чем к любой точке другого кластера.



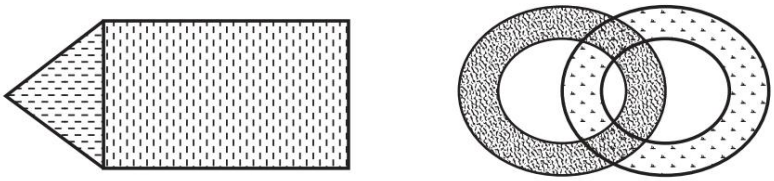
(b) Центрированные кластеры. Каждая точка находится ближе к центру своего кластера, чем к центру любого другого кластера.



(c) Кластеры на основе смежности. Каждая точка находится ближе хотя бы к одной точке своего кластера, чем к любой точке другого кластера.



(d) Кластеры на основе плотности. Кластеры — это области высокой плотности, разделенные областями низкой плотности.



(e) Концептуальные кластеры. Точки в кластере имеют некоторые общие свойства, вытекающие из всего набора точек. (Точки на пересечении окружностей принадлежат обоим.)

Рисунок 8.2. Различные типы кластеров, проиллюстрированные наборами двумерных точек.

8.2 K-средние

Методы кластеризации на основе прототипов создают одноуровневое секционирование объектов данных. Существует несколько таких методов, но два из наиболее известных — это K-means и K-medoid. K-means определяет прототип в терминах центроида, который обычно является средним значением группы точек и обычно

применяется к объектам в непрерывном n-мерном пространстве. K-medoid определяет прототип в терминах медоида, который является наиболее репрезентативной точкой для группы точек и может применяться к широкому диапазону данных, поскольку требует только меры близости для пары объектов. Хотя центроид почти никогда не соответствует фактической точке данных, медоид по своему определению должен быть фактической точкой данных. В этом разделе мы сосредоточимся исключительно на K-средних, который является одним из старейших и наиболее широко используемых алгоритмов кластеризации.

8.2.1 Базовый алгоритм K-средних

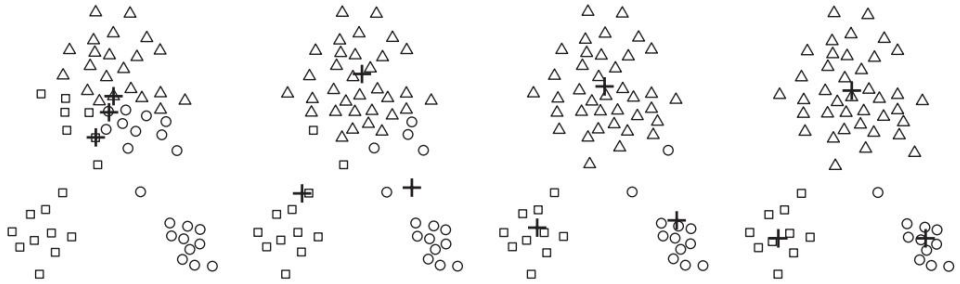
Методика кластеризации K-средних проста, и мы начнем с описания основного алгоритма. Сначала мы выбираем K начальных центроидов, где K — заданный пользователем параметр, а именно количество желаемых кластеров. Затем каждая точка присваивается ближайшему центроиду, и каждый набор точек, назначенный центроиду, представляет собой кластер. Затем центроид каждого кластера обновляется на основе точек, назначенных кластеру. Мы повторяем шаги назначения и обновления до тех пор, пока ни одна точка не изменит кластеры, или, что то же самое, до тех пор, пока центроиды не останутся такой же.

K-means формально описывается алгоритмом 8.1. Работа K-средних проиллюстрирована на рисунке 8.3, где показано, как, начиная с трех центроидов, конечные кластеры находятся за четыре шага обновления назначения. На этих и других рисунках, отображающих кластеризацию K-средних, каждый подрисунок показывает (1) центроиды в начале итерации и (2) присвоение точек этим центроидам. Центроиды обозначаются знаком «+»; все точки, принадлежащие одному кластеру, имеют одинаковую форму маркера.

Алгоритм 8.1. Базовый алгоритм K-средних.

1: выберите точки K в качестве начальных центроидов.
2: повторить 3: Сформировать K кластеров, назначив каждую точку ближайшему центроиду.
4: Пересчитать центроид каждого кластера. 5: до тех пор, пока центроиды не изменятся.

На первом этапе, показанном на рисунке 8.3(a), точкам присваиваются начальные центроиды, которые все входят в большую группу точек. В этом примере мы используем среднее значение в качестве центроида. После присвоения точек центроиду центроид обновляется. Опять же, на рисунке для каждого шага показан центр тяжести в начале шага и присвоение точек этим центроидам. На втором этапе обновленным центроидам присваиваются точки, а центроиды



(а) Итерация 1. (б) Итерация 2. (в) Итерация 3. (г) Итерация 4.

Рисунок 8.3. Использование алгоритма K-средних для поиска трех кластеров в выборочных данных.

обновляются снова. На шагах 2, 3 и 4, показанных на рисунках 8.3 (b), (c) и (d) соответственно, два центроида перемещаются в две небольшие группы точек внизу рисунков. Когда алгоритм K-средних завершается на рис. 8.3(d), поскольку изменений больше не происходит, центроиды идентифицируют естественные группы точек.

Для некоторых комбинаций функций близости и типов центроидов K-средние всегда сходятся к решению; т. е. K-средние достигают состояния, в котором никакие точки не перемещаются из одного кластера в другой, и, следовательно, центроиды не изменяются. Однако, поскольку большая часть сходимости происходит на ранних этапах, условие в строке 5 алгоритма 8.1 часто заменяется более слабым условием, например, повторять до тех пор, пока только 1% точек не изменят кластеры.

Рассмотрим каждый из шагов базового алгоритма K-средних более подробно. а затем провести анализ пространственной и временной сложности алгоритма.

Назначение точек ближайшему центроиду

Чтобы назначить точку ближайшему центроиду, нам нужна мера близости, которая количественно определяет понятие «ближайший» для конкретных рассматриваемых данных. Евклидово расстояние (L2) часто используется для точек данных в евклидовом пространстве, тогда как косинусное подобие больше подходит для документов. Однако может существовать несколько типов мер близости, подходящих для данного типа данных. Например, для евклидовых данных можно использовать манхэттенское расстояние (L1) , а для документов часто применяют меру Жаккара.

Обычно меры сходства, используемые для K-средних, относительно просты, поскольку алгоритм неоднократно вычисляет сходство каждой точки с каждым центроидом. Однако в некоторых случаях, например, когда данные находятся в низкоразмерном формате

Таблица 8.1. Таблица обозначений.

Символ	Описание
x	Объект.
Кластер C_i	i-й кластер.
c_i	Центр тяжести кластера C_i.
c	Центр тяжести всех точек.
m_i	Количество объектов в i-м кластере.
m	Количество объектов в наборе данных.
K	Количество кластеров.

В евклидовом пространстве можно избежать многих сходств,
тем самым значительно ускоряя алгоритм K-средних. Биссектриса K-средних
(описанный в разделе 8.2.3) — это еще один подход, который ускоряет K-средние за счет
уменьшение количества вычисляемых сходств.

Центроиды и целевые функции

Шаг 4 алгоритма K-средних был сформулирован в общем как «пересчитать
центроид каждого кластера», поскольку центроид может меняться в зависимости от
мера близости данных и цель кластеризации. Цель
кластеризация обычно выражается целевой функцией, которая зависит от
близость точек друг к другу или к центроидам кластера; например,
минимизируйте квадрат расстояния каждой точки до ближайшего к ней центроида. Проиллюстрируем
это двумя примерами. Однако ключевой момент заключается в следующем: как только мы получим
определили меру близости и целевую функцию — центроид, который мы
выбор, часто можно определить математически. Мы приводим математические
подробности в разделе 8.2.6 и даем нематематическое обсуждение
это наблюдение здесь.

Данные в евклидовом пространстве Рассмотрим данные, мерой близости которых является
евклидово расстояние. Для нашей целевой функции, измеряющей качество
При кластеризации мы используем сумму квадратов ошибок (SSE), которая также известна
как разброс. Другими словами, мы вычисляем ошибку каждой точки данных, т. е. ее
Евклидово расстояние до ближайшего центроида, а затем вычислить общую сумму
квадратов ошибок. Учитывая два разных набора кластеров, которые производятся
двумя разными сериями K-средних, мы предпочитаем тот, у которого наименьший квадрат
ошибка, так как это означает, что прототипы (центроиды) этой кластеризации
лучшее представление точек в их кластере. Используя обозначения в
В таблице 8.1 SSE формально определяется следующим образом:

$$\text{SSE} = \sum_{i=1}^{K} \sum_{x \in C_i} \text{расстояние}(c_i, x)^2 \quad (8.1)$$

где dist — стандартное евклидово (L2) расстояние между двумя объектами в евклидовом пространстве.

Учитывая эти предположения, можно показать (см. раздел 8.2.6), что центроид, который минимизирует SSE кластера, является средним. Используя обозначения таблицы 8.1, центроид (среднее значение) i-й кластер определяется уравнением 8.2.

$$c_i = \frac{1}{m_i} \sum_{x \in C_i} x \quad (8.2)$$

Для иллюстрации: центроид кластера, содержащего три двумерные точки (1,1), (2,3) и (6,2), равен ((1 + 2 + 6)/3,(1 + 3 + 2)/3) = (3, 2).

Шаги 3 и 4 алгоритма K-средних напрямую пытаются минимизировать SSE (или, в более общем смысле, целевую функцию). Шаг 3 формирует кластеры путем присвоения точек их ближайшему центроиду, что минимизирует SSE для данного набора центроидов. Шаг 4 пересчитывает центроиды, чтобы еще больше минимизировать SSE. Однако действия K-средних на шагах 3 и 4 гарантированно находят только локальный минимум по отношению к SSE, поскольку они основаны на оптимизации SSE для конкретного выбора центроидов и кластеров, а не для всех возможных вариантов. Позже мы увидим пример, в котором это приводит к неоптимальной кластеризации.

Данные документа Чтобы проиллюстрировать, что K-средние не ограничиваются данными в евклидовом пространстве, мы рассматриваем данные документа и косинусную меру подобия. Здесь мы предполагаем, что данные документа представлены в виде матрицы терминов документа, как описано на стр. 31. Наша цель — максимизировать сходство документов в кластере с центроидом кластера; эта величина известна как сплоченность кластера. Для этой цели можно показать, что центроид кластера, как и для евклидовых данных, является средним. Аналогичной величиной общего SSE является общее сцепление, которое определяется уравнением 8.3.

$$\text{Общая сплоченность} = \sum_{i=1}^{K} \sum_{x \in C_i} \text{косинус}(x, c_i) \quad (8,3)$$

Общий случай Существует несколько вариантов функции близости, центроида и целевой функции, которые можно использовать в базовых K-средних.

Таблица 8.2. K-средние: распространенный выбор близости, центроидов и целевых функций.

Функция близости	Центроидная	целевая функция
Манхэттен (L1)	медиана	Минимизировать сумму расстояний L1 объекта наблюдения. спроектировать его центроид кластера
Квадрат Евклидова (L2 2)	среднее	Минимизировать сумму квадрата расстояния L2 объекта к центроиду его кластера
косинус	среднее	Максимизировать сумму косинусного подобия объект в его центроиде кластера
Дивергенция Брегмана	среднее значение	Минимизировать сумму расхождения Брегмана объекта к центроиду его кластера

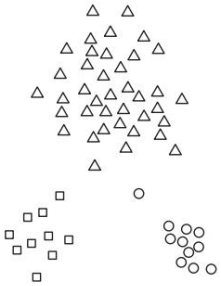
алгоритм и которые гарантированно сходятся. В Таблице 8.2 показаны некоторые возможные вариантов, включая те два, которые мы только что обсудили. Обратите внимание, что для расстояния Манхэттен (L1) и цели минимизации суммы расстояний:
соответствующий центроид — это медиана точек в кластере.

Последняя запись в таблице, Дивергенция Брегмана (раздел 2.4.5), на самом деле класс мер близости, включающий квадрат евклидова расстояния L2 2, расстояние Махаланобиса и косинусное подобие. Важность Брегмана Дивергенция заключается в том, что любая такая функция может быть использована в качестве основы алгоритма кластеризации в стиле K-средних со средним значением в качестве центроида. Конкретно, если мы используем дивергенцию Брегмана в качестве функции близости, то полученный алгоритм кластеризации будет иметь обычные свойства K-средних по отношению к сходимость, локальные минимумы и т. д. Более того, свойства такого алгоритма кластеризации можно развить для всех возможных расходимостей Брегмана. Действительно, Алгоритмы K-средних, использующие косинусное подобие или квадрат евклидова расстояния. являются частными примерами общего алгоритма кластеризации, основанного на Брегмане. расхождения.

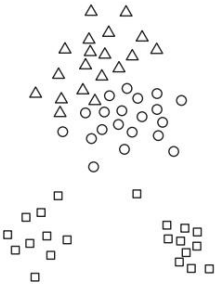
В остальной части нашего обсуждения K-средних мы используем двумерные данные, поскольку легко объяснить K-средние и его свойства для этого типа данных. Но, как следует из последних нескольких абзацев, K-средние — это очень общая кластеризация. алгоритм и может использоваться с самыми разными типами данных, такими как документы и временной ряд.

Выбор начальных центроидов

Когда используется случайная инициализация центроидов, различные прогоны K-средних обычно производят разные общие SSE. Мы проиллюстрируем это на примере набора двумерных точек, показанного на рисунке 8.3, который имеет три естественных кластера. точки. На рис. 8.4(a) показано решение кластеризации, которое представляет собой глобальный минимум



(а) Оптимальная кластеризация.



(б) Субоптимальная кластеризация.

Рисунок 8.4. Три оптимальных и неоптимальных кластера.

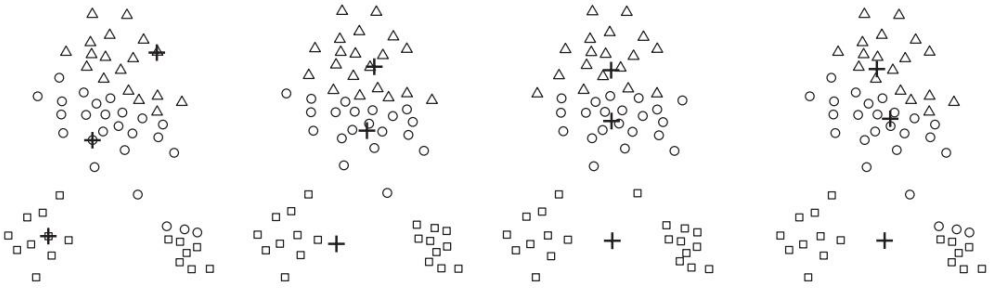
SSE для трех кластеров, а на рис. 8.4(b) показана субоптимальная кластеризация, которая является лишь локальным минимумом.

Выбор правильных начальных центроидов является ключевым шагом базовой процедуры K-средних. Распространенный подход заключается в случайном выборе начальных центроидов, но полученные кластеры часто бывают плохими.

Пример 8.1 (плохие начальные центроиды). Случайно выбранные начальные центроиды могут быть плохими. Мы приводим пример этого, используя тот же набор данных, что и на рисунках 8.3 и 8.4. На рисунках 8.3 и 8.5 показаны кластеры, возникающие в результате двух конкретных вариантов выбора начальных центроидов. (На обоих рисунках положения центроидов кластера в различных итерациях обозначены крестиками.) На рисунке 8.3, хотя все начальные центроиды взяты из одного естественного кластера, минимальная кластеризация SSE все же обнаружена. Однако на рисунке 8.5, хотя исходные центроиды кажутся более распределенными, мы получаем неоптимальную кластеризацию с более высоким квадратом ошибки.

■

Пример 8.2 (Пределы случайной инициализации). Один из методов, который обычно используется для решения проблемы выбора начальных центроидов, заключается в выполнении нескольких прогонов, каждый со своим набором случайно выбранных начальных центроидов, а затем выборе набора кластеров с минимальным SSE. Несмотря на простоту, эта стратегия может работать не очень хорошо, в зависимости от набора данных и количества искомых кластеров. Мы продемонстрируем это, используя пример набора данных, показанный на рисунке 8.6 (а). Данные состоят из двух пар кластеров, причем кластеры в каждой паре (верх-низ) расположены ближе друг к другу, чем к кластерам в другой паре. Рисунок 8.6 (б–г) показывает, что если мы начнем с двух начальных центроидов на пару кластеров, то даже когда оба центроида находятся в одном



(а) Итерация 1. (б) Итерация 2. (в) Итерация 3. (г) Итерация 4.

Рисунок 8.5. Плохие начальные центроиды для K-средних.

В кластере центроиды перераспределятся так, что будут найдены «истинные» кластеры. Однако на рис. 8.7 показано, что если пара кластеров имеет только один начальный центроид, а другая пара — три, то два истинных кластера будут объединены, а один истинный кластер будет разделен.

Обратите внимание, что оптимальная кластеризация будет получена до тех пор, пока два начальных центроида попадают в любое место пары кластеров, поскольку центроиды будут перераспределяться по одному в каждый кластер. К сожалению, по мере того, как количество кластеров становится больше, возрастает вероятность того, что хотя бы одна пара кластеров будет иметь только один начальный центроид. (См. упражнение 4 на стр. 559.) В этом случае, поскольку пары кластеров расположены дальше друг от друга, чем кластеры внутри пары, алгоритм K-средних не будет перераспределять центроиды между парами кластеров, и, таким образом, будет использоваться только локальный минимум. быть достигнуто. ■

Из-за проблем с использованием случайно выбранных начальных центроидов, которые невозможно преодолеть даже при повторных запусках, для инициализации часто используются другие методы. Один из эффективных подходов — взять выборку точек и сгруппировать их с использованием метода иерархической кластеризации. K-кластеры извлекаются из иерархической кластеризации, а центроиды этих кластеров используются в качестве начальных центроидов. Этот подход часто работает хорошо, но практичен только в том случае, если (1) выборка относительно мала, например, от нескольких сотен до нескольких тысяч (иерархическая кластеризация обходится дорого), и (2) K относительно мал по сравнению с размером выборки.

Следующая процедура представляет собой еще один подход к выбору начальных центроидов. Выберите первую точку случайным образом или возьмите центр тяжести всех точек. Затем для каждого последующего начального центроида выберите точку, которая находится дальше всего от любого из уже выбранных начальных центроидов. Таким образом, мы получаем набор исходных

(а) Начальные точки.

(б) Итерация 1.

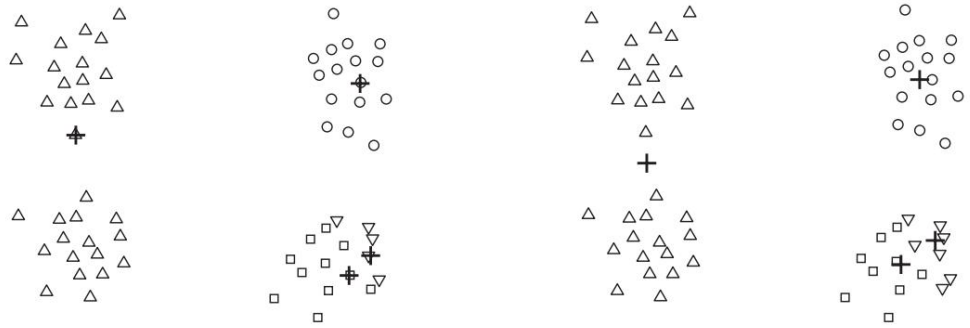
(в) Итерация 2.

(г) Итерация 3.

Рисунок 8.6. Две пары кластеров с парой начальных центроидов внутри каждой пары кластеров.

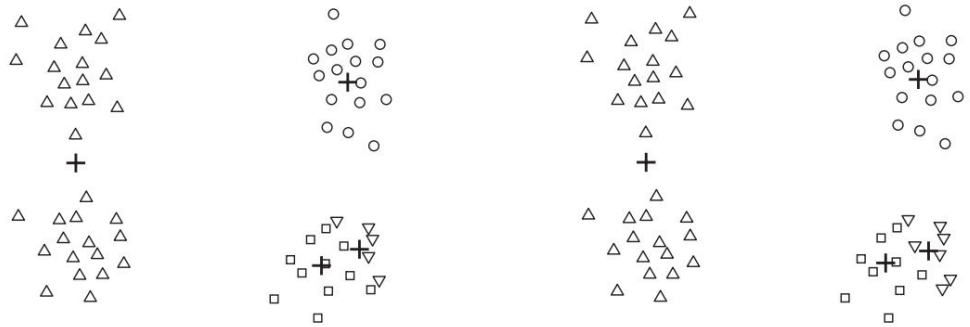
центроиды, которые гарантированно будут не только случайно выбраны, но и хорошо разделены. К сожалению, такой подход позволяет выбирать выбросы, а не точки в плотных регионах (кластерах). Кроме того, вычисление самой дальней точки от текущего набора начальных центроидов требует больших затрат. Чтобы преодолеть эти проблемы, этот подход часто применяется к выборке точек. Поскольку выбросы редки, они, как правило, не появляются в случайной выборке. Напротив, точки из каждой плотной области, скорее всего, будут включены, если только размер выборки не очень мал. Кроме того, значительно сокращаются вычисления, необходимые для поиска начальных центроидов, поскольку размер выборки обычно намного меньше количества точек.

Позже мы обсудим два других подхода, которые полезны для создания кластеров лучшего качества (с более низким SSE): использование варианта K-средних, который



(а) Итерация 1.

(б) Итерация 2.



(в) Итерация 3.

(г) Итерация 4.

Рисунок 8.7. Две пары кластеров с более или менее двумя начальными центроидами внутри пары кластеров.

менее подвержен проблемам инициализации (деление K-средних пополам) и использованию постобработки для «исправления» набора созданных кластеров.

Временная и пространственная сложность

Требования к пространству для K-средних скромны, поскольку сохраняются только точки данных и центроиды. В частности, требуемый объем памяти равен O((m + K)n), где m — количество точек, а n — количество атрибутов. Требования ко времени для K-средних также скромны — в основном линейны по количеству точек данных. В частности, требуемое время равно O(I K m n), где I — количество итераций, необходимых для сходимости. Как уже упоминалось, I часто невелик и обычно может быть безопасно ограничен, поскольку большинство изменений обычно происходит в

первые несколько итераций. Таким образом, K-средние линейны по m (количество точек) и являются эффективными и простыми при условии, что K (количество кластеров) значительно меньше m.

8.2.2 K-средние: дополнительные вопросы

Обработка пустых кластеров

Одна из проблем базового алгоритма K-средних, приведенного ранее, заключается в том, что пустые кластеры могут быть получены, если на этапе присваивания кластеру не выделяются никакие точки. Если это произойдет, то необходима стратегия выбора замещающего центроида, поскольку в противном случае квадрат ошибки будет больше, чем необходимо. Один из подходов — выбрать точку, которая находится дальше всего от текущего центроида. По крайней мере, это устраняет точку, которая в настоящее время вносит наибольший вклад в общую квадратичную ошибку. Другой подход — выбрать замещающий центроид из кластера с самым высоким SSE. Обычно это приводит к разделению кластера и уменьшению общего SSE кластеризации. Если пустых кластеров несколько, то этот процесс можно повторить несколько раз.

Выбросы

При использовании критерия квадратичной ошибки выбросы могут оказывать чрезмерное влияние на найденные кластеры. В частности, при наличии выбросов результирующие центроиды кластера (прототипы) могут быть не такими репрезентативными, как в противном случае, и, следовательно, SSE также будет выше. По этой причине часто бывает полезно обнаружить выбросы и устранить их заранее. Однако важно понимать, что существуют определенные приложения кластеризации, для которых выбросы не следует исключать. Когда для сжатия данных используется кластеризация, каждая точка должна быть кластеризована, и в некоторых случаях, например при финансовом анализе, наиболее интересными точками могут быть явные выбросы, например, необычно прибыльные клиенты.

Очевидная проблема заключается в том, как идентифицировать выбросы. Ряд методов выявления выбросов будет обсуждаться в главе 10. Если мы используем подходы, которые удаляют выбросы перед кластеризацией, мы избегаем точек кластеризации, которые не кластеризуются должным образом. Альтернативно, выбросы также могут быть идентифицированы на этапе постобработки. Например, мы можем отслеживать SSE, вносимый каждой точкой, и исключать те точки с необычно высоким вкладом, особенно при нескольких прогонах. Кроме того, мы можем захотеть исключить небольшие кластеры, поскольку они часто представляют собой группы выбросов.

Уменьшение SSE с помощью постобработки

Очевидный способ уменьшить SSE — найти больше кластеров, т. е. использовать большее значение K. Однако во многих случаях мы хотели бы улучшить SSE, но не хотим увеличивать количество кластеров. Это часто возможно, поскольку K-средние обычно сходятся к локальному минимуму. Для «исправления» полученных кластеров используются различные методы, чтобы создать кластеризацию с более низким SSE. Стратегия состоит в том, чтобы сосредоточиться на отдельных кластерах, поскольку общий SSE представляет собой просто сумму SSE, внесенную каждым кластером. (Мы будем использовать терминологию «общее SSE» и «кластерное SSE» соответственно, чтобы избежать возможной путаницы.) Мы можем изменить общий SSE, выполняя различные операции над кластерами, такие как разделение или слияние кластеров. Одним из часто используемых подходов является использование альтернативных фаз разделения и слияния кластеров. На этапе разделения кластеры разделяются, а на этапе слияния кластеры объединяются. Таким образом, часто можно избежать локальных минимумов SSE и при этом создать решение для кластеризации с желаемым количеством кластеров. Ниже приведены некоторые методы, используемые на этапах разделения и слияния.

Две стратегии, которые уменьшают общую SSE за счет увеличения количества кластеры следующие:

Разделение кластера. Обычно выбирается кластер с наибольшим SSE, но мы также можем разделить кластер с наибольшим стандартным отклонением для одного конкретного атрибута.

Введение нового центроида кластера. Часто выбирается точка, наиболее удаленная от любого центра кластера. Мы можем легко это определить, если будем отслеживать SSE, вносимый каждой точкой. Другой подход заключается в случайном выборе из всех точек или из точек с самым высоким SSE.

Две стратегии, которые уменьшают количество кластеров, пытаясь минимизировать Чтобы увеличить общий объем SSE, можно выделить следующие:

Распределить кластер. Это достигается путем удаления центроида, соответствующего кластеру, и переназначения точек другим кластерам. В идеале рассредоточенным кластером должен быть тот, который в наименьшей степени увеличивает общий SSE.

Объединение двух кластеров. Обычно выбираются кластеры с ближайшими центроидами, хотя другой, возможно, лучший подход заключается в объединении двух кластеров, что приводит к наименьшему увеличению общего SSE. Эти две стратегии слияния аналогичны тем, которые используются в иерархической структуре.

методы кластеризации, известные как метод центроида и метод Уорда соответственно. Оба метода обсуждаются в разделе 8.3.

Постепенное обновление центроидов

Вместо обновления центроидов кластера после того, как все точки были назначены кластеру, центроиды можно обновлять постепенно после каждого назначения точки кластеру. Обратите внимание, что для этого требуется либо ноль, либо два обновления центроидов кластера на каждом этапе, поскольку точка либо перемещается в новый кластер (два обновления), либо остается в своем текущем кластере (ноль обновлений). Использование стратегии инкрементного обновления гарантирует, что пустые кластеры не будут создаваться, поскольку все кластеры начинаются с одной точки, и если в кластере когда-либо имеется только одна точка, то эта точка всегда будет переназначена тому же кластеру.

Кроме того, если используется постепенное обновление, относительный вес добавляемой точки может быть скорректирован; например, вес точек часто уменьшается по мере кластеризации. Хотя это может привести к повышению точности и более быстрой сходимости, может быть сложно сделать правильный выбор относительного веса, особенно в самых разных ситуациях. Эти проблемы обновления аналогичны тем, которые возникают при обновлении весов искусственных нейронных сетей.

Еще одно преимущество дополнительных обновлений связано с использованием целей, отличных от «минимизации SSE». Предположим, что нам дана произвольная целевая функция для измерения качества набора кластеров. Когда мы обрабатываем отдельную точку, мы можем вычислить значение целевой функции для каждого возможного назначения кластера, а затем выбрать то, которое оптимизирует целевую функцию. Конкретные примеры альтернативных целевых функций приведены в разделе 8.5.2.

С другой стороны, постепенное обновление центроидов приводит к зависимости от порядка. Другими словами, создаваемые кластеры могут зависеть от порядка обработки точек. Хотя эту проблему можно решить путем рандомизации порядка обработки точек, базовый подход K-средних, заключающийся в обновлении центроидов после того, как все точки были присвоены кластерам, не имеет зависимости от порядка. Кроме того, дополнительные обновления немного дороже. Однако K-средние сходятся довольно быстро, и поэтому количество точек, переключающих кластеры, быстро становится относительно небольшим.

8.2.3 Биссектриса K-средних

Алгоритм деления K-средних пополам является прямым расширением базового алгоритма. Алгоритм K-средних, основанный на простой идее: чтобы получить K кластеров, разделите набор всех точек на два кластера, выберите один из этих кластеров для разделения и

и так далее, пока не будет создано K кластеров. Детали разделения K-средних пополам даны алгоритмом 8.2.

Алгоритм 8.2 Алгоритм деления K-средних пополам.

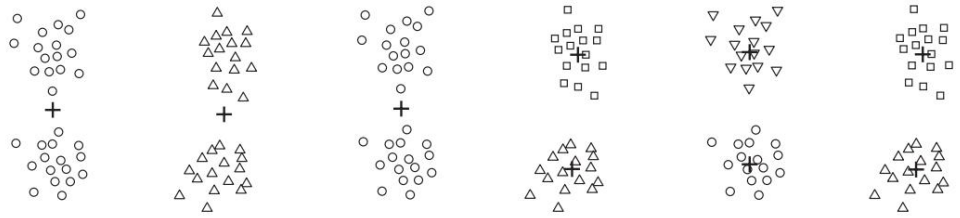
```
1: Инициализировать список кластеров, чтобы он содержал кластер, состоящий из всех
точек. 2:
повторить 3: удалить кластер из списка кластеров.
4: {Выполните несколько «пробных» делений выбранного кластера
пополам.} 5: для i = 1 к количеству попыток
выполните 6: Разделите выбранный кластер пополам, используя
базовые K-
средние. 7: конец для 8: выберите два кластера из биссектрисы с наименьшим общим SSE.
9: Добавьте эти два кластера в список кластеров. 10: до тех
пор, пока список кластеров не будет содержать K кластеров.
```

Существует несколько различных способов выбора кластера для разделения. Мы можем выбирать самый большой кластер на каждом этапе, выбирать кластер с самым большим SSE или использовать критерий, основанный как на размере, так и на SSE. Разные варианты выбора приводят к созданию разных кластеров.

Мы часто уточняем полученные кластеры, используя их центроиды в качестве начальных центроидов для базового алгоритма K-средних. Это необходимо, потому что, хотя алгоритм K-средних гарантированно находит кластеризацию, которая представляет собой локальный минимум по отношению к SSE, при разделении K-средних пополам мы используем алгоритм K-средних «локально», т. е. делим пополам отдельные кластеры. Следовательно, окончательный набор кластеров не представляет собой кластеризацию, которая является локальным минимумом по отношению к общему SSE.

Пример 8.3 (Разделение K-средних пополам и инициализация). Чтобы проиллюстрировать, что разделение K-средних пополам менее подвержено проблемам инициализации, мы покажем на рисунке 8.8, как разделение K-средних пополам находит четыре кластера в наборе данных, первоначально показанном на рисунке 8.6 (a). На итерации 1 находятся две пары кластеров; на итерации 2 разбивается самая правая пара кластеров; и на итерации 3 самая левая пара кластеров расщепляется. При разделении K-средних пополам возникает меньше проблем с инициализацией, поскольку оно выполняет несколько пробных делений пополам и выбирает то, у которого SSE наименьшее, а также потому, что на каждом шаге имеется только два центроида. ■

Наконец, записывая последовательность кластеризаций, созданных как кластеры K-средних пополам, мы также можем использовать разделение K-средних пополам для создания иерархической кластеризации.



(а) Итерация 1. (б) Итерация 2. (в) Итерация 3.

Рисунок 8.8. Разделение K-средних пополам на примере четырех кластеров.

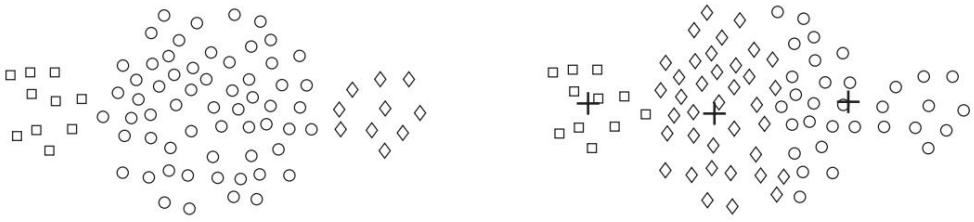
8.2.4 K-средние и различные типы кластеров

K-средние и его варианты имеют ряд ограничений в отношении поиска различных типов кластеров. В частности, K-средние испытывают трудности с обнаружением «естественных» кластеров, когда кластеры имеют несферическую форму или сильно различаются по размеру или плотности. Это иллюстрируется рисунками 8.9, 8.10 и 8.11. На рисунке 8.9 K-средние не могут найти три естественных кластера, поскольку один из кластеров намного больше двух других, и, следовательно, больший кластер разрывается, а один из меньших кластеров объединяется с частью большего кластера. . На рисунке 8.10 метод K-средних не может найти три естественных кластера, поскольку два меньших кластера намного плотнее, чем больший кластер. Наконец, на рисунке 8.11 метод K-means находит два кластера, в которых смешаны части двух естественных кластеров, поскольку форма естественных кластеров не является шаровидной.

Трудность в этих трех ситуациях заключается в том, что целевая функция K-средних не соответствует типам кластеров, которые мы пытаемся найти, поскольку она минимизируется шаровыми кластерами одинакового размера и плотности или кластерами, которые хорошо разделены. Однако эти ограничения в некотором смысле можно преодолеть, если пользователь готов принять кластеризацию, которая разбивает естественные кластеры на несколько подкластеров. На рис. 8.12 показано, что произойдет с тремя предыдущими наборами данных, если мы обнаружим шесть кластеров вместо двух или трех. Каждый меньший кластер является чистым в том смысле, что он содержит только точки одного из естественных кластеров.

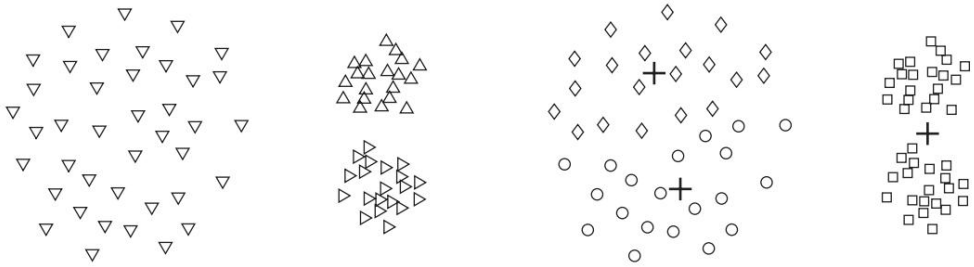
8.2.5 Сильные и слабые стороны

K-средние просты и могут использоваться для самых разных типов данных. Это также весьма эффективно, хотя часто выполняется несколько прогонов. Некоторые варианты, включая разделение K-средних пополам, еще более эффективны и менее подвержены проблемам инициализации. K-means подходит не для всех типов данных.



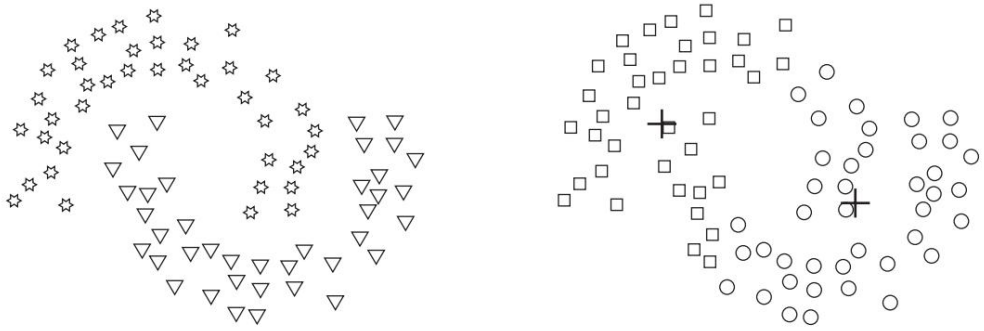
(а) Исходные точки. (б) Три кластера K-средних.

Рисунок 8.9. K-означает с кластерами разного размера.



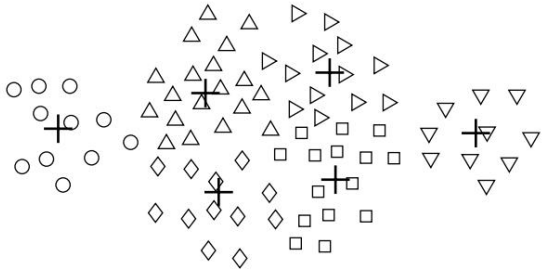
(а) Исходные точки. (б) Три кластера K-средних.

Рисунок 8.10. K-средства с кластерами разной плотности.

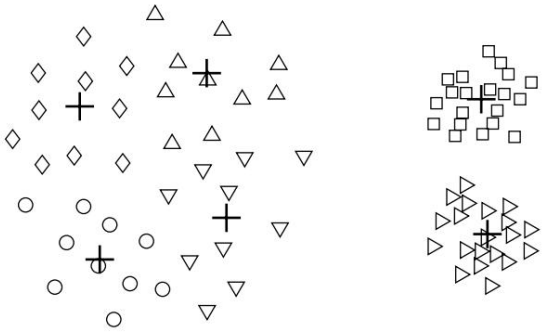


(а) Исходные точки. (б) Два кластера K-средних.

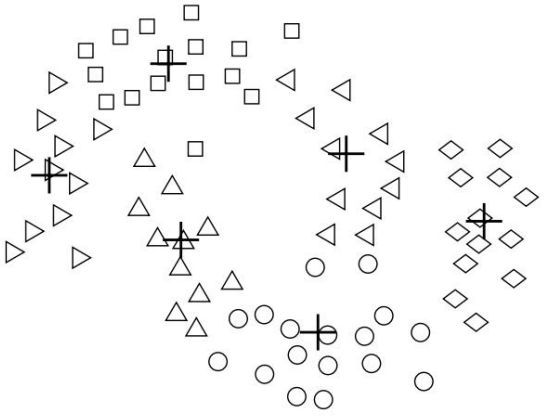
Рисунок 8.11. K-средства с неглобулярными кластерами.



а) Неравные размеры.



(б) Неравные плотности.



(с) Несферические формы.

Рисунок 8.12. Использование K-средних для поиска кластеров, которые являются подкластерами естественных кластеров.

однако. Он не может обрабатывать нешаровые кластеры или кластеры разных размеров и плотности, хотя обычно может найти чистые подкластеры, если указано достаточно большое количество кластеров. K-means также имеет проблемы с кластеризацией данных, содержащих выбросы. Обнаружение и удаление выбросов может существенно помочь в таких ситуациях. Наконец, K-средние ограничены данными, для которых существует понятие центра (центроида). Родственный метод, кластеризация K-медоидов, не имеет этого ограничения, но является более дорогим.

8.2.6 K-средние как задача оптимизации

Здесь мы углубимся в математику, лежащую в основе K-средних. Этот раздел, который можно пропустить без потери непрерывности, требует знания исчисления через частные производные. Также может оказаться полезным знакомство с методами оптимизации, особенно основанными на градиентном спуске.

Как упоминалось ранее, при наличии такой целевой функции, как «минимизация SSE», кластеризацию можно рассматривать как задачу оптимизации. Один из способов решения этой проблемы — найти глобальный оптимум — состоит в том, чтобы перечислить все возможные способы разделения точек на кластеры, а затем выбрать набор кластеров, который лучше всего удовлетворяет целевой функции, например, который минимизирует общее SSE. Конечно, эта исчерпывающая стратегия вычислительно неосуществима, и в результате необходим более практичный подход, даже если такой подход находит решения, которые не гарантированно будут оптимальными. Один из методов, известный как градиентный спуск, основан на выборе начального решения и последующем повторении следующих двух шагов: вычисление изменения решения, которое лучше всего оптимизирует целевую функцию, а затем обновление решения.

Мы предполагаем, что данные одномерны, т. е. dist(x, y)=(x y)2.
Ничего существенного это не меняет, но существенно упрощает обозначения.

Вывод K-средних как алгоритм минимизации SSE

В этом разделе мы покажем, как можно математически вывести центроид для алгоритма K-средних, когда функция близости представляет собой евклидово расстояние, а цель состоит в том, чтобы минимизировать SSE. В частности, мы исследуем, как лучше всего обновить центроид кластера, чтобы минимизировать SSE кластера. С математической точки зрения мы стремимся минимизировать уравнение 8.1, которое мы повторяем здесь, специально для одномерных данных.

$$CCE = \sum_{i=1}^{K} \sum_{x \varepsilon C_i} (ци - x)^2 \qquad (8,4)$$

Здесь Ci — это i й кластер, x — точка в Ci, а ci — среднее значение i й кластер. Полный список обозначений см. в Таблице 8.1.

Мы можем найти k-й центроид ck, который минимизирует уравнение 8.4, путем дифференцирования SSE, установки его равным 0 и решения, как указано ниже.

$$\frac{\partial}{\partial c_k} SSE = \frac{\partial}{\partial c_k} \sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2$$

$$= \sum_{i=1}^{K} \sum_{x \in C_i} \frac{\partial}{\partial c_k} (c_i - x)^2$$

$$= \sum_{x \in C_k} 2 (c_k - x_k) = 0$$

$$\sum_{x \in C_k} 2 (c_k - x_k) = 0 \quad m_k c_k = \sum_{x \in C_k} x_k \quad c_k = m_k \frac{1}{\sum_{x \in C_k}} x_к$$

Таким образом, как указывалось ранее, лучший центроид для минимизации SSE кластер — это среднее значение точек в кластере.

Вывод K-средних для SAE

Чтобы продемонстрировать, что алгоритм K-средних можно применять к множеству различных целевых функций, мы рассмотрим, как разделить данные на K кластеры так, чтобы сумма манхэттенских (L1) расстояний точек от центра их кластеров была минимизирована. . Мы стремимся минимизировать сумму абсолютных ошибок L1 (SAE), как указано в следующем уравнении, где distL1 — расстояние L1 . Опять же, для простоты обозначений, мы используем одномерные данные, т. е. distL1 = |ci x|.

$$САЭ = \sum_{я=1}^{K} \sum_{x \in Ci} distL1 (ci, x) \qquad (8,5)$$

Мы можем найти k-й центроид ck, который минимизирует уравнение 8.5, дифференцируя SAE, устанавливая его равным 0 и решая.

$$\frac{\partial SAE}{\partial c_k} = \frac{\partial}{\partial c_k} \sum_{i=1}^{K} \sum_{x \varepsilon C_i} |c_i - x|$$

$$= \sum_{i=1}^{K} \sum_{x \varepsilon C_i} \frac{\partial}{\partial c_k} |c_i - x|$$

$$= \sum_{x \varepsilon C_k} \frac{\partial}{\partial c_k} |c_k - x| = 0$$

$$\sum_{x \varepsilon C_k} \frac{\partial}{\partial c_k} |c_к - x| = 0 \qquad \text{знак}(x - c_к) = 0$$

$$x \varepsilon C_k$$

Если мы найдем ck, мы обнаружим, что ck = median{x Ck}, медиана точек в кластере. Медиану группы точек легко вычислить, и она менее подвержена искажениям из-за выбросов.

8.3 Агломеративная иерархическая кластеризация

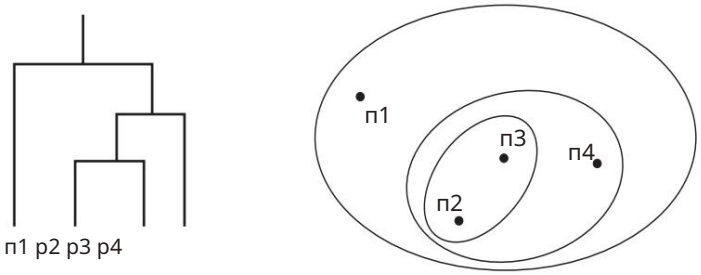
Методы иерархической кластеризации являются второй важной категорией методов кластеризации. Как и в случае с K-средними, эти подходы относительно старые по сравнению со многими алгоритмами кластеризации, но они до сих пор широко используются. Существует два основных подхода к созданию иерархической кластеризации:

Агломерация: начните с точек как отдельных кластеров и на каждом этапе объединяйте ближайшую пару кластеров. Для этого необходимо определить понятие близости кластера.

Разделение: начните с одного комплексного кластера и на каждом этапе разбивайте кластер до тех пор, пока не останутся только одноэлементные кластеры из отдельных точек. В этом случае нам нужно решить, какой кластер разбивать на каждом шаге и как это делать.

Методы агломеративной иерархической кластеризации являются наиболее распространенными, и в этом разделе мы сосредоточимся исключительно на этих методах. Метод разделительной иерархической кластеризации описан в разделе 9.4.2.

Иерархическая кластеризация часто отображается графически с использованием древовидной диаграммы, называемой дендрограммой, которая отображает как кластер-подкластер, так и кластер-подкластер.



(а) Дендрограмма. (б) Диаграмма вложенных кластеров.

Рисунок 8.13. Иерархическая кластеризация из четырех точек, показанная в виде дендрограммы и вложенных кластеров.

отношения и порядок, в котором кластеры были объединены (агломеративный вид) или разделены (разделенный вид). Для наборов двумерных точек, таких как те, которые мы будем использовать в качестве примеров, иерархическую кластеризацию также можно представить графически с помощью диаграммы вложенных кластеров. На рис. 8.13 показан пример этих двух типов фигур для набора из четырех двумерных точек. Эти точки были сгруппированы с использованием метода одиночной связи, описанного в разделе 8.3.2.

8.3.1 Базовый алгоритм агломеративной иерархической кластеризации

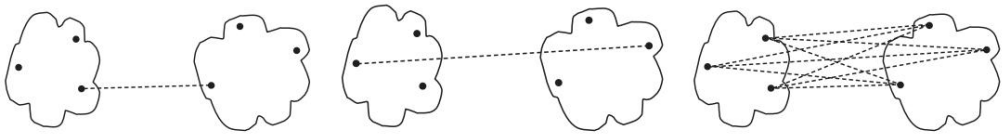
Многие методы агломеративной иерархической кластеризации представляют собой вариации единого подхода: начиная с отдельных точек в виде кластеров, последовательно объединяя два ближайших кластера, пока не останется только один кластер. Более формально этот подход выражен в алгоритме 8.3.

Алгоритм 8.3 Базовый алгоритм агломеративной иерархической кластеризации.

1: При необходимости вычислите матрицу близости.
2: повторить
3: объединить два ближайших кластера.
4. Обновите матрицу близости, чтобы отразить близость между новым кластером и исходными кластерами.
5: до тех пор, пока не останется только один кластер.

Определение близости между кластерами

Ключевой операцией алгоритма 8.3 является вычисление близости между двумя кластерами, и именно определение близости кластеров отличает различные методы агломеративной иерархии, которые мы будем обсуждать. Близость кластера обычно определяется для определенного типа кластера.
иметь в виду — см. раздел 8.1.2. Например, многие агломеративные иерархические методы кластеризации, такие как MIN, MAX и Group Average, взяты из графическое представление кластеров. MIN определяет близость кластера как близость между двумя ближайшими точками, которые находятся в разных кластерах, или используя термины графа — кратчайшее ребро между двумя узлами в разных подмножествах узлов. В результате получаются кластеры на основе смежности, как показано на рисунке 8.2(c). Альтернативно, MAX принимает близость между двумя самыми дальними точками в разных кластерах. быть близостью кластера или, используя термины графа, самое длинное ребро между два узла в разных подмножествах узлов. (Если наши близости — это расстояния, то имена MIN и MAX короткие и наводящие на размышления. Однако из-за сходства где более высокие значения указывают на более близкие точки, имена кажутся перевернутыми. Для этого По этой причине мы обычно предпочитаем использовать альтернативные имена, одиночную ссылку и полную ссылку соответственно.) Другой подход, основанный на графике, — среднее значение по группе. Метод определяет близость кластера как среднюю попарную близость (среднюю длину ребер) всех пар точек из разных кластеров. Рисунок 8.14
иллюстрирует эти три подхода.



(а) MIN (одиночное звено.) (б) MAKC (полная ссылка.) (c) Среднее значение по группе.

Рисунок 8.14. Определения близости кластеров на основе графов

Если вместо этого мы возьмем представление, основанное на прототипах, в котором каждый кластер представлен центроидом, различные определения близости кластеров будут более естественными. При использовании центроидов близость кластера обычно определяется как близость между центроидами кластера. Альтернативный метод, метод Уорда.
также предполагает, что кластер представлен его центроидом, но измеряет близость между двумя кластерами с точки зрения увеличения SSE, что

результат слияния двух кластеров. Как и K-средние, метод Уорда пытается минимизировать сумму квадратов расстояний точек от центроидов их кластеров.

Временная и пространственная сложность

Только что представленный базовый алгоритм агломеративной иерархической кластеризации м2 (при условии, что матрица близости. Для этого требуется, 1 использует близости площадью 2 чтобы матрица близости была симметричной), где m — количество точек данных. Пространство, необходимое для отслеживания кластеров, пропорционально количеству кластеров, которое равно m-1, исключая одноэлементные кластеры. Следовательно, общая сложность пространства равна O(m2).

Анализ базового алгоритма агломеративной иерархической кластеризации также прост с точки зрения вычислительной сложности. Для вычисления матрицы близости требуется время O(m2) . После этого шага выполняется m - 1 итераций, включающих шаги 3 и 4, поскольку в начале имеется m кластеров, и во время каждой итерации два кластера объединяются. Если выполняется как линейный поиск итерации, шаг 3 требует O((m i+ 1)2) матрицы близости, затем в течение времени кластеров в квадрате. Шаг 4 требует всего й i, которое пропорционально текущему числу O(m - i + 1) времени для обновления матрицы близости после слияния двух кластеров. (Слияние кластеров влияет только на O(m i+1) близости для рассматриваемых нами методов.) Без изменений это привело бы к временной сложности O(m3). Если расстояния от каждого кластера до всех остальных кластеров храниться в виде отсортированного списка (или кучи), можно снизить стоимость поиска двух ближайших кластеров до O(m i + 1). Однако из-за дополнительной сложности хранения данных в отсортированном списке или куче общее время, необходимое для иерархической кластеризации на основе алгоритма 8.3, составляет O(m2 log m).

Пространственная и временная сложность иерархической кластеризации серьезно ограничивает размер наборов данных, которые можно обработать. Мы обсуждаем подходы к масштабируемости алгоритмов кластеризации, включая методы иерархической кластеризации, в разделе 9.5.

8.3.2 Специальные методы

Образец данных

Чтобы проиллюстрировать поведение различных алгоритмов иерархической кластеризации, мы будем использовать выборочные данные, состоящие из 6 двумерных точек, которые показаны на рисунке 8.15. Координаты x и y точек и евклидовы расстояния между ними показаны в таблицах 8.3 и 8.4 соответственно.

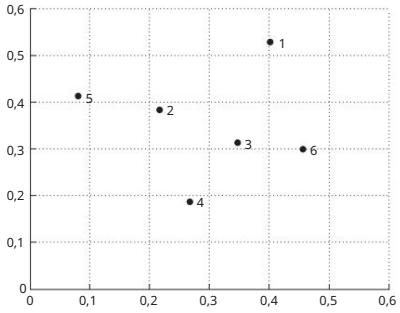


Рисунок 8.15. Набор из 6 двумерных точек.

Точка x	Координата y	Координата
p1	0,40	0,53
p2	0,22	0,38
p3	0,35	0,32
p4	0,26	0,19
p5	0,08	0,41
p6	0,45	0,30

Таблица 8.3. координаты xy 6 точек.

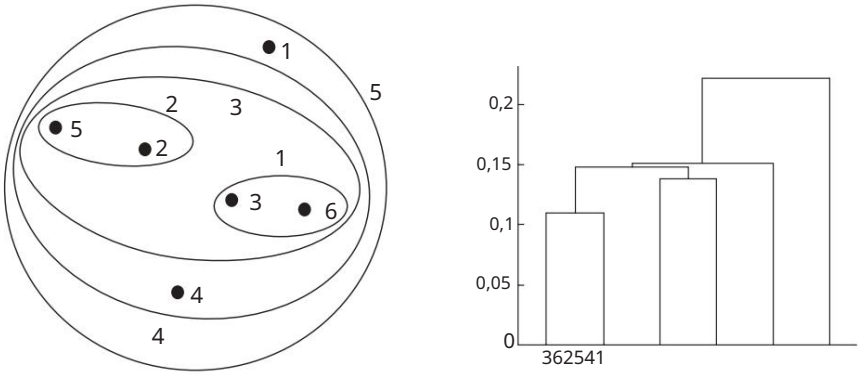
	п1	p2	p3	p4	p5	p6		
p1	0,00	0,24	0,22	0,37	0,34	0,23		
p2	0,24	0,00	0,15	0,20	0,14	0,25		
p3	0,22	0,15	0,00	0,15	0,28	0,11		
p4	0,37	0,20	0,15	0,00	0,29	0,22		
p5	0,34	0,14	0,28	0,29	0,00	0,39		
p6	0,23	0,25	0,11	0,22	0,39	0,00		

Таблица 8.4. Матрица евклидовых расстояний для 6 точек.

Одиночное соединение или МИН.

Для одноканальной или MIN-версии иерархической кластеризации близость двух кластеров определяется как минимум расстояния (максимум сходство) между любыми двумя точками в двух разных кластерах. Использование графика терминологии, если вы начнете со всех точек как одноэлементных кластеров и добавите ссылки между точками по одной, сначала кратчайшие связи, затем эти одиночные связи объединяют точки в кластеры. Метод одной ссылки хорошо справляется с неэллиптической формы, но чувствителен к шуму и выбросам.

Пример 8.4 (одиночная ссылка). На рис. 8.16 показан результат применения метод одной связи с нашим примером набора данных из шести точек. Рисунок 8.16(а) показывает вложенные кластеры как последовательность вложенных эллипсов, где числа связанные с эллипсами, указывают порядок кластеризации. Рисунок 8.16(б) показывает ту же информацию, но в виде дендрограммы. Высота, на которой двое кластеры сливаются в дендрограмме, что отражает расстояние между двумя кластерами. Например, из таблицы 8.4 мы видим, что расстояние между точками 3 и 6



(а) Однозвенная кластеризация. (б) Дендрограмма одной связи.

Рисунок 8.16. Одноканальная кластеризация шести точек, показанная на рисунке 8.15.

составляет 0,11, и это высота, на которой они объединяются в один кластер на дендрограмме. В качестве другого примера расстояние между кластерами {3, 6} и {2, 5} определяется выражением

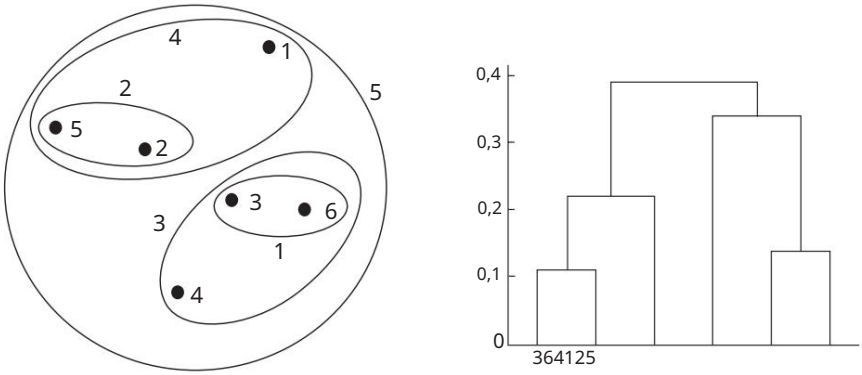
$$\begin{aligned}\text{dist}(\{3,6\},\{2,5\}) &= \min(\text{dist}(3,2),\text{dist}(6,2),\text{dist}(3,5),\text{dist}(6,5)) = \min(0{,}15,0{,}25,0{,}28,\\ &\qquad 0{,}39)\\ &= 0{,}15.\end{aligned}$$

■

Полная ссылка или MAX или CLIQUE

Для полной ссылки или версии MAX иерархической кластеризации близость двух кластеров определяется как максимальное расстояние (минимум сходства) между любыми двумя точками в двух разных кластерах. Используя терминологию графа, если вы начинаете со всех точек как одноэлементных кластеров и добавляете связи между точками по одной, сначала самые короткие связи, то группа точек не является кластером до тех пор, пока все точки в ней не будут полностью связаны, т. е. не образуют клика.
Полная связь менее восприимчива к шуму и выбросам, но может разбивать большие кластеры и предпочитает шаровидную форму.

Пример 8.5 (Полная ссылка). На рисунке 8.17 показаны результаты применения MAX к набору выборочных данных из шести точек. Как и в случае с одиночной ссылкой, пункты 3 и 6.



(а) Полная кластеризация ссылок. (б) Полная дендрограмма связей.

Рисунок 8.17. Полная кластеризация каналов из шести точек показана на рисунке 8.15.

объединяются в первую очередь. Однако {3, 6} объединяется с {4} вместо {2, 5} или {1}, поскольку

$$\text{dist}(\{3, 6\}, \{4\}) = \max(\text{dist}(3, 4), \text{dist}(6, 4)) = \max(0{,}15, 0{,}22) = 0{,}22.$$

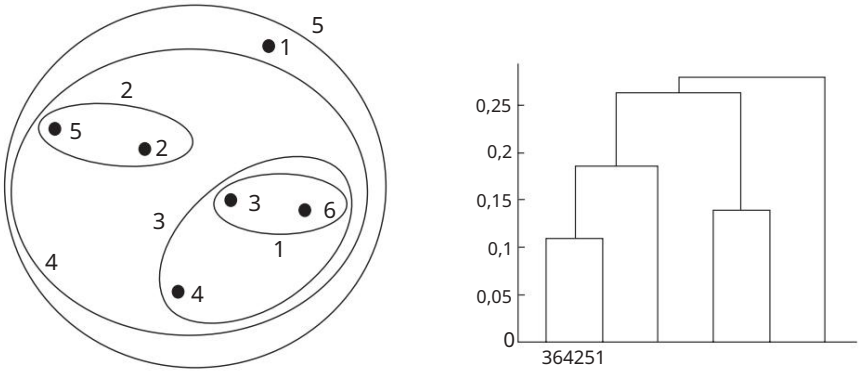
$$\text{dist}(\{3, 6\}, \{2, 5\}) = \max(\text{dist}(3, 2), \text{dist}(6, 2), \text{dist}(3, 5), \text{dist}(6, 5)) = \max(0.15, 0{,}25, 0{,}28, 0{,}39) = 0{,}39.$$

$$\text{dist}(\{3, 6\}, \{1\}) = \max(\text{dist}(3, 1), \text{dist}(6, 1)) = \max(0{,}22, 0{,}23) = 0{,}23.$$

■

Группа Средний

Для среднегрупповой версии иерархической кластеризации близость двух кластеров определяется как средняя попарная близость среди всех пар точек в разных кластерах. Это промежуточный подход между подходами с одним и полным соединением. Таким образом, для среднего по группе кластерная близость



(а) Групповая средняя кластеризация. (б) Средняя дендрограмма группы.

Рисунок 8.18. Групповая средняя кластеризация из шести точек показана на рисунке 8.15.

Близость(Ci, Cj) кластеров Ci и Cj , имеющих размеры mi и mj соответственно, выражается следующим уравнением:

$$\text{близость}(C_i, C_j) = \frac{\sum_{x \in C_i, y \in C_j} \text{близость}(x, y)}{m_i * m_j}. \quad (8,6)$$

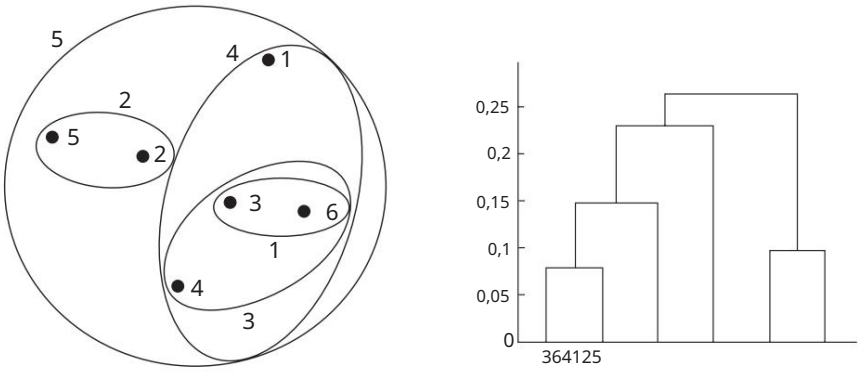
Пример 8.6 (Среднее по группе). На рисунке 8.18 показаны результаты применения метода группового среднего к выборочному набору данных из шести точек. Чтобы проиллюстрировать, как работает среднее значение по группе, мы рассчитаем расстояние между некоторыми кластерами.

dist({3, 6, 4}, {1}) = (0,22 + 0,37 + 0,23)/(3 1) = 0,28

dist({2, 5}, {1}) = (0,2357 + 0,3421)/(2 1) = 0,2889

dist({3, 6, 4}, {2, 5}) = (0,15 + 0,28 + 0,25 + 0,39 + 0,20 + 0,29)/(6 2)
= 0,26

Поскольку dist({3, 6, 4}, {2, 5}) меньше, чем dist({3, 6, 4}, {1}) и dist({2, 5}, {1}), кластеры { 3, 6, 4} и {2, 5} объединяются на четвертом этапе. ■



(а) Кластеризация Уорда. (б) Дендрограмма Уорда.

Рисунок 8.19. Кластеризация Уорда из шести точек показана на рис. 8.15.

Метод Уорда и методы центроида

Для метода Уорда близость между двумя кластерами определяется как увеличение квадрата ошибки, возникающее при слиянии двух кластеров. Таким образом, этот метод использует ту же целевую функцию, что и кластеризация K-средних. Хотя может показаться, что эта особенность делает метод Уорда несколько отличным от других иерархических методов, математически можно показать, что метод Уорда очень похож на метод группового среднего, когда близость между двумя точками принимается за квадрат расстояния. между ними.

Пример 8.7 (метод Уорда). На рисунке 8.19 показаны результаты применения метода Уорда к набору выборочных данных из шести точек. Создаваемая кластеризация отличается от кластеризации, создаваемой одиночной ссылкой, полной ссылкой и средним значением группы. ■

Методы центроидов вычисляют близость между двумя кластерами, вычисляя расстояние между центроидами кластеров. Эти методы могут показаться похожими на K-средние, но, как мы уже отмечали, метод Уорда является правильным иерархическим аналогом.

У методов центроидов также есть особенность (часто считающаяся плохой), которой нет у других методов иерархической кластеризации, которые мы обсуждали: возможность инверсий. В частности, два объединенных кластера могут быть более похожими (менее удаленными), чем пара кластеров, которые были объединены на предыдущем этапе. Для других методов расстояние между

Таблица 8.5. Таблица коэффициентов Лэнса-Вильямса для распространенных подходов иерархической кластеризации.

Метод кластеризации	αA	αБ	β	γ
Одиночная ссылка	1/2	1/2	0	1/2
Полная ссылка	1/2	1/2	0	1/2
Группа Средний	$\frac{м_A}{м_A+м_Б}$	$\frac{м_Б}{м_A+м_Б}$	0	0
центроид	$\frac{м_A}{м_A+м_Б}$	$\frac{м_Б}{м_A+м_Б}$	$\frac{м_A м_Б}{(м_A+м_Б)2}$	0
Уорд	$\frac{м_A+м_К}{м_A+м_В+м_К}$	$\frac{м_Б+м_К}{м_A+м_В+м_К}$	$\frac{-м_Q}{м_A+м_В+м_К}$	0

объединенных кластеров монотонно возрастает (или, в худшем случае, не возрастает) по мере мы переходим от одноэлементных кластеров к одному комплексному кластеру.

8.3.3 Формула Лэнса-Вильямса для определения близости кластеров

Любая из близостей кластера, которые мы обсуждали в этом разделе, может быть рассматривается как выбор различных параметров (в формуле Лэнса-Вильямса показано ниже в уравнении 8.7) для близости между кластерами Q и R, где R образуется путем слияния кластеров A и B. В этом уравнении p(., .) равно функция близости, а mA, mB и mQ — количество точек в кластеры A, B и Q соответственно. Другими словами, после объединения кластеров A и B для формирования кластера R, близость нового кластера R к существующему кластер Q является линейной функцией близости Q по отношению к исходные кластеры A и B. В таблице 8.5 приведены значения этих коэффициентов для методы, которые мы обсуждали.

$$p(R, Q) = \alpha A\, p(A, Q) + \alpha B\, p(B,Q) + \beta\, p(A, B) + \gamma\, |p(A, Q) - p(B,Q)| \quad (8,7)$$

Любой метод иерархической кластеризации, который можно выразить с помощью Формула Лэнса-Вильямса не требует сохранения исходных данных. Вместо этого матрица близости обновляется по мере возникновения кластеризации. В то время как генерал формула привлекательна, особенно для реализации, ее легче понять различные иерархические методы, непосредственно рассматривая определение близости кластеров, которое использует каждый метод.

8.3.4 Ключевые проблемы иерархической кластеризации

Отсутствие глобальной целевой функции

Ранее мы упоминали, что агломеративную иерархическую кластеризацию невозможно рассматривается как глобальная оптимизация целевой функции. Вместо этого агломеративный методы иерархической кластеризации используют различные критерии для принятия решений локально, на каждом этапе.

шаг, какие кластеры следует объединить (или разделить в случае разногласий). Этот подход дает алгоритмы кластеризации, которые позволяют избежать трудностей, связанных с решением сложной задачи комбинаторной оптимизации. (Можно показать, что общая задача кластеризации для такой целевой функции, как «минимизировать SSE», вычислительно неосуществима.) Кроме того, такие подходы не имеют проблем с локальными минимумами или трудностями с выбором начальных точек. Конечно, временная сложность O(m2 log m) и пространственная сложность O(m2) во многих случаях непомерно высоки.

Возможность обработки кластеров разных размеров

Один из аспектов агломеративной иерархической кластеризации, который мы еще не обсуждали, — это то, как обращаться с относительными размерами пар объединяемых кластеров. (Это обсуждение применимо только к схемам близости кластеров, которые включают суммы, такие как центроид, Уорд и среднее значение группы.) Существует два подхода: взвешенный, который рассматривает все кластеры одинаково, и невзвешенный, который учитывает количество точек в каждом кластере. счет. Обратите внимание, что терминология «взвешенные» или «невзвешенные» относится к точкам данных, а не к кластерам. Другими словами, обработка кластеров неравного размера одинаково дает разный вес точкам в разных кластерах, а учет размера кластера дает точкам в разных кластерах одинаковый вес.

Мы проиллюстрируем это, используя метод группового среднего, обсуждавшийся в разделе 8.3.2, который представляет собой невзвешенную версию метода группового среднего. В литературе по кластеризации полное название этого подхода — «Метод невзвешенной пары пар с использованием средних арифметических значений» (UPGMA). В таблице 8.5, где приведена формула для обновления сходства кластеров, коэффициенты для UPGMA включают размер каждого из объединенных кластеров: $\alpha A = \frac{mA}{mA+mB}$, $\alpha B = \frac{mB}{mA+mB}$, $\beta = 0$, $\gamma = 0$. Для взвешенной версии группового среднего значения — известный как WPGMA — коэффициенты являются константами: $\alpha A = 1/2$, $\alpha B = 1/2$, $\beta = 0$, $\gamma = 0$.

В целом, невзвешенные подходы предпочтительнее, если нет оснований полагать, что отдельные точки должны иметь разные веса; например, возможно, классы объектов были выбраны неравномерно.

Решения по слиянию являются окончательными

Алгоритмы агломеративной иерархической кластеризации имеют тенденцию принимать хорошие локальные решения об объединении двух кластеров, поскольку они могут использовать информацию о попарном сходстве всех точек. Однако если принято решение об объединении двух кластеров, его нельзя будет отменить позднее. Такой подход предотвращает превращение локального критерия оптимизации в глобальный критерий оптимизации.

Например, хотя критерий «минимизации квадратичной ошибки» из K-средних используется при принятии решения о том, какие кластеры объединять в методе Уорда, кластеры на каждом уровне не представляют собой локальные минимумы по отношению к общему SSE. Действительно, кластеры даже не стабильны в том смысле, что точка в одном кластере может быть ближе к центроиду какого-то другого кластера, чем к центроиду текущего кластера. Тем не менее, метод Уорда часто используется в качестве надежного метода инициализации кластеризации K-средних, указывая на то, что локальная целевая функция «минимизировать квадратичную ошибку» действительно имеет связь с глобальной целевой функцией «минимизировать квадратичную ошибку».

Существуют некоторые методы, которые пытаются преодолеть ограничение, заключающееся в том, что слияния являются окончательными. Один из подходов пытается исправить иерархическую кластеризацию путем перемещения ветвей дерева, чтобы улучшить глобальную целевую функцию. Другой подход использует метод секционной кластеризации, такой как K-средние, для создания множества небольших кластеров, а затем выполняет иерархическую кластеризацию, используя эти небольшие кластеры в качестве отправной точки.

8.3.5 Сильные и слабые стороны

Сильные и слабые стороны конкретных алгоритмов агломеративной иерархической кластеризации обсуждались выше. В более общем плане такие алгоритмы обычно используются потому, что базовое приложение, например создание таксономии, требует иерархии. Кроме того, были проведены некоторые исследования, которые предполагают, что эти алгоритмы могут создавать кластеры более высокого качества. Однако алгоритмы агломеративной иерархической кластеризации являются дорогостоящими с точки зрения требований к вычислениям и хранению. Тот факт, что все слияния являются окончательными, также может вызвать проблемы с зашумленными многомерными данными, такими как данные документа. В свою очередь, эти две проблемы можно в некоторой степени решить, сначала частично кластеризовав данные с использованием другого метода, такого как K-средние.

8.4 ДБСКАН

Кластеризация на основе плотности позволяет обнаружить области с высокой плотностью, которые отделены друг от друга областями с низкой плотностью. DBSCAN — это простой и эффективный алгоритм кластеризации на основе плотности, который иллюстрирует ряд важных концепций, важных для любого подхода к кластеризации на основе плотности. В этом разделе мы сосредоточимся исключительно на DBSCAN, предварительно рассмотрев ключевое понятие плотности. Другие алгоритмы поиска кластеров на основе плотности описаны в следующей главе.

8.4.1 Традиционная плотность: центральный подход

Хотя подходов к определению плотности не так много, как к определению сходства, существует несколько различных методов. В этом разделе мы обсуждаем центральный подход, на котором основан DBSCAN. Другие определения плотности будут представлены в главе 9.

В подходе на основе центра плотность оценивается для конкретной точки набора данных путем подсчета количества точек в пределах заданного радиуса Eps этой точки. Это включает в себя саму точку. Этот метод графически иллюстрируется рисунком 8.20. Число точек в радиусе Eps точки A равно 7, включая саму A.

Этот метод прост в реализации, но плотность любой точки будет зависеть от заданного радиуса. Например, если радиус достаточно велик, то все точки будут иметь плотность m — количество точек в наборе данных.
Аналогично, если радиус слишком мал, все точки будут иметь плотность 1.
Подход к выбору подходящего радиуса для низкоразмерных данных представлен в следующем разделе в контексте нашего обсуждения DBSCAN.

Классификация точек по центральной плотности

Подход к плотности на основе центра позволяет нам классифицировать точку как находящуюся (1) внутри плотной области (центральная точка), (2) на краю плотной области (граничная точка) или (3) в малонаселенной области (шумовая или фоновая точка). Рисунок 8.21 графически иллюстрирует концепции ядра, границы и шумовых точек с использованием набора двумерных точек. Следующий текст дает более точное описание.

Основные точки: эти точки находятся внутри кластера на основе плотности. Точка является базовой точкой, если количество точек в данной окрестности вокруг точки, определяемое функцией расстояния и заданным пользователем параметром расстояния Eps, превышает определенный порог M inP ts, который также является заданным пользователем параметром расстояния. параметр. На рисунке 8.21 точка A является центральной точкой для указанного радиуса (Eps), если M inP ts 7.

Пограничные точки: Пограничная точка не является основной точкой, но находится в пределах окрестности основной точки. На рисунке 8.21 точка B является границей. Пограничная точка может находиться в пределах нескольких основных точек.

Точки шума: Точка шума — это любая точка, которая не является ни основной, ни граничной точкой. На рисунке 8.21 точка C является точкой шума.

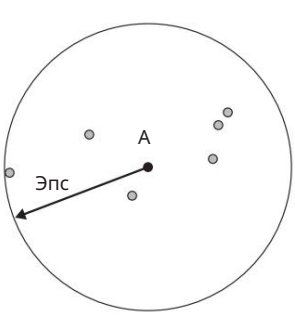


Рисунок 8.20. Плотность по центру.

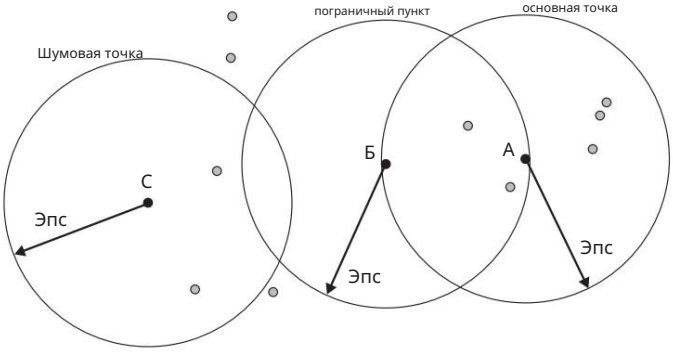


Рисунок 8.21. Ядро, граница и точки шума.

8.4.2 Алгоритм DBSCAN

Учитывая предыдущие определения основных точек, граничных точек и точек шума, алгоритм DBSCAN можно неформально описать следующим образом. Любые две основные точки, находящиеся достаточно близко (на расстоянии Eps друг от друга), помещаются в один кластер. Аналогично, любая пограничная точка, расположенная достаточно близко к основной точке, помещается в тот же кластер, что и основная точка. (Возможно, потребуется устранить связи, если граничная точка находится близко к основным точкам из разных кластеров.) Точки шума отбрасываются. Формальные детали приведены в алгоритме 8.4. Этот алгоритм использует те же концепции и находит те же кластеры, что и исходный DBSCAN, но оптимизирован для простоты, а не эффективности.

Алгоритм 8.4 Алгоритм DBSCAN.

1: обозначьте все точки как точки ядра, границы или точки шума.
2: Устранить шумовые точки.
3: Поместите ребро между всеми основными точками, которые находятся в пределах Eps друг от друга.
4. Сделайте каждую группу связанных основных точек отдельным кластером.
5. Назначьте каждую пограничную точку одному из кластеров связанных с ней основных точек.

Временная и пространственная сложность

Базовая временная сложность алгоритма DBSCAN равна O(m × время поиска точек в Eps-окрестности), где m — количество точек. В худшем случае эта сложность равна O(m2). Однако в пространствах низкой размерности существуют структуры данных, такие как kd-деревья, которые позволяют эффективно извлекать все данные.

точек на заданном расстоянии от указанной точки, а временная сложность может составлять всего O(m log m). Требуемое пространство DBSCAN, даже для многомерных данных, составляет O(m), поскольку необходимо хранить лишь небольшой объем данных для каждой точки, т. е. метку кластера и идентификацию каждой точки как ядра, границы. , или шумовая точка.

Выбор параметров DBSCAN

Возникает, конечно, вопрос, как определить параметры Eps и M inP ts. Основной подход состоит в том, чтобы посмотреть на поведение расстояния от точки до ее k-го ближайшего соседа, которое мы назовем k-расстоянием. Для точек, принадлежащих некоторому кластеру, значение k-dist будет небольшим, если k не превышает размер кластера. Обратите внимание, что будут некоторые различия в зависимости от плотности кластера и случайного распределения точек, но в среднем диапазон изменений не будет огромным, если плотности кластеров не различаются радикально. Однако для точек, которые не входят в кластер, например точек шума, k-dist будет относительно большим. Следовательно, если мы вычислим k-dist для всех точек данных для некоторого k, отсортируем их в порядке возрастания, а затем построим график отсортированных значений, мы ожидаем увидеть резкое изменение значения k-dist, которое соответствует подходящему значению. значение Эпс. Если мы выберем это расстояние в качестве параметра Eps и возьмем значение k в качестве параметра M inP ts, то точки, для которых k-dist меньше Eps, будут помечены как основные точки, а другие точки будут помечены как шум или граница. точки.

На рисунке 8.22 показан пример набора данных, а график k-dist для данных показан на рисунке 8.23. Определяемое таким образом значение Eps зависит от k, но при изменении k существенно не меняется. Если значение k слишком мало, то даже небольшое количество близко расположенных точек, которые являются шумом или выбросами, будут неправильно помечены как кластеры. Если значение k слишком велико, то небольшие кластеры (размером меньше k), скорее всего, будут помечены как шум. Исходный алгоритм DBSCAN использовал значение k = 4, которое кажется разумным значением для большинства наборов двумерных данных.

Кластеры различной плотности

У DBSCAN могут возникнуть проблемы с плотностью, если плотность кластеров сильно варьируется. Рассмотрим рисунок 8.24, на котором показаны четыре кластера, погруженные в шум. Плотность кластеров и областей шума обозначена их темнотой. Шум вокруг пары более плотных кластеров A и B имеет ту же плотность, что и кластеры C и D. Если порог Eps достаточно низок, чтобы DBSCAN обнаружил C и D как кластеры, тогда A и B и окружающие их точки станут один

Рисунок 8.22. Образец данных.

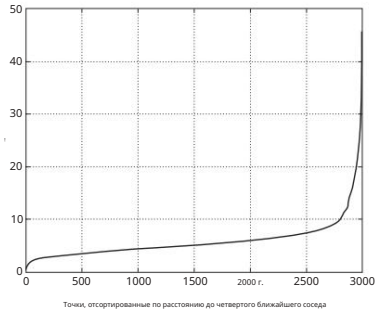


Рисунок 8.23. График K-dist для выборочных данных.

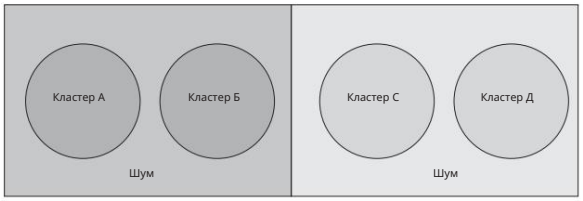


Рисунок 8.24. Четыре кластера, погруженные в шум.

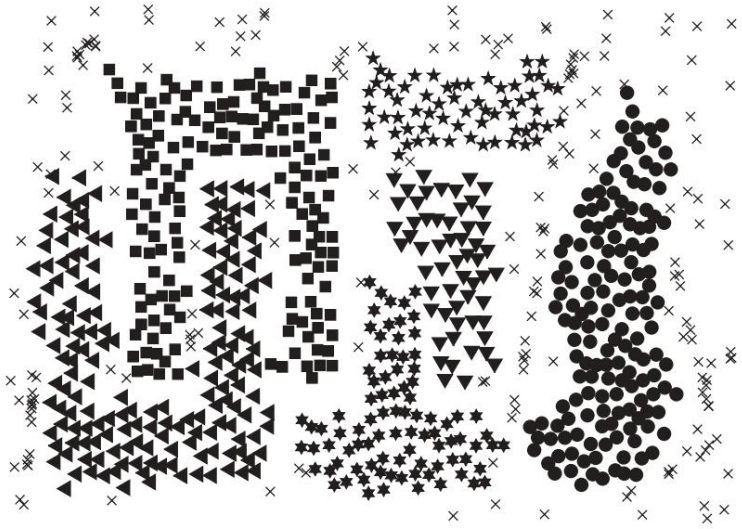
кластер. Если порог Eps достаточно высок, чтобы DBSCAN обнаружил A и B как отдельные кластеры, а окружающие их точки отмечаются как шум, то C и D и окружающие их точки также будут отмечены как шум.

Пример

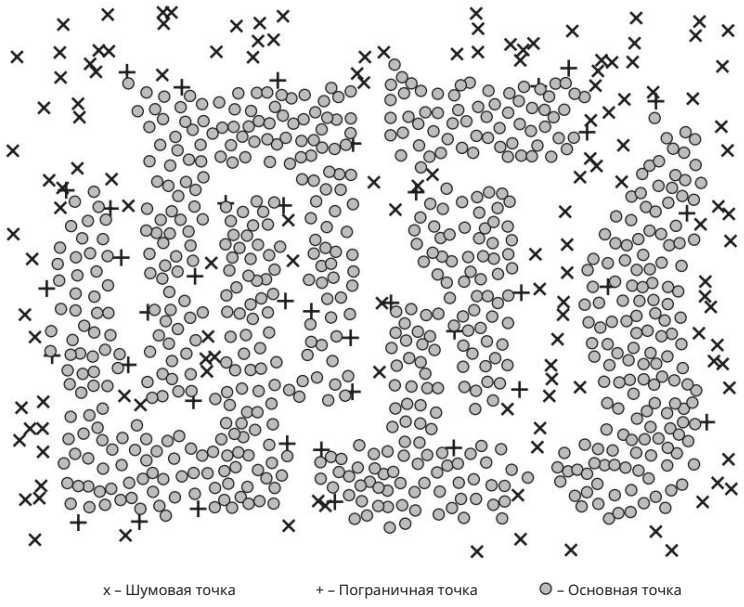
Чтобы проиллюстрировать использование DBSCAN, мы покажем кластеры, которые он находит в относительно сложный двумерный набор данных, показанный на рисунке 8.22. Этот набор данных состоит из 3000 двумерных точек. Порог Eps для этого данные были найдены путем построения отсортированных расстояний четвертого ближайшего соседа каждой точки (рис. 8.23) и выявление значения, при котором наблюдается резкий увеличивать. Мы выбрали Eps = 10, что соответствует перегибу кривой. Кластеры, найденные DBSCAN с использованием этих параметров, т.е. M inP ts = 4 и Eps = 10 показаны на рисунке 8.25(a). Основные точки, пограничные точки и точки шума показаны на рисунке 8.25(b).

8.4.3 Сильные и слабые стороны

Поскольку DBSCAN использует определение кластера на основе плотности, оно относительно устойчив к шуму и может обрабатывать кластеры произвольных форм и размеров. Таким образом,



(а) Кластеры, обнаруженные с помощью DBSCAN.



(б) Ядро, граница и точки шума.

Рисунок 8.25. DBSCAN-кластеризация 3000 двумерных точек.

DBSCAN может найти множество кластеров, которые невозможно найти с помощью K-средних, например, показанных на рисунке 8.22. Однако, как указывалось ранее, у DBSCAN возникают проблемы, когда плотность кластеров сильно различается. У него также возникают проблемы с многомерными данными, поскольку для таких данных сложнее определить плотность. Один из возможных подходов к решению таких проблем приведен в разделе 9.4.8. Наконец, DBSCAN может быть дорогостоящим, когда вычисление ближайших соседей требует вычисления всех попарных близостей, как это обычно бывает с многомерными данными.

8.5 Оценка кластера

В контролируемой классификации оценка полученной классификационной модели является неотъемлемой частью процесса разработки классификационной модели, и существуют общепринятые меры и процедуры оценки, например, точность и перекрестная проверка соответственно. Однако по своей природе кластерная оценка не является хорошо разработанной и широко используемой частью кластерного анализа.
Тем не менее, оценка кластера или валидация кластера, как ее традиционно называют, важна, и в этом разделе будут рассмотрены некоторые из наиболее распространенных и легко применяемых подходов.

Может возникнуть некоторая путаница относительно необходимости оценки кластера. Часто кластерный анализ проводится как часть исследовательского анализа данных. Таким образом, оценка кажется излишне сложным дополнением к тому, что должно быть неформальным процессом. Более того, поскольку существует множество различных типов кластеров (в некотором смысле каждый алгоритм кластеризации определяет свой собственный тип кластера), может показаться, что каждая ситуация может потребовать разных мер оценки. Например, кластеры K-средних можно оценивать с точки зрения SSE, но для кластеров на основе плотности, которые не обязательно должны быть шаровидными, SSE вообще не будет работать хорошо.

Тем не менее, оценка кластера должна быть частью любого кластерного анализа. Ключевой мотивацией является то, что почти каждый алгоритм кластеризации находит кластеры в наборе данных, даже если этот набор данных не имеет естественной кластерной структуры. Например, рассмотрим рисунок 8.26, на котором показан результат кластеризации 100 точек, случайно (равномерно) распределенных на единичном квадрате. Исходные точки показаны на рисунке 8.26(a), а кластеры, найденные с помощью DBSCAN, K-средних и полной ссылки, показаны на рисунках 8.26(b), 8.26(c) и 8.26(d) соответственно. Поскольку DBSCAN нашел три кластера (после того, как мы установили Eps, просматривая расстояния до четвертых ближайших соседей), мы установили K-средние и полную ссылку, чтобы также найти три кластера. (На рис. 8.26(b) шум показан маленькими маркерами.) Однако кластеры не выглядят убедительно ни для одного из них.

три метода. В более высоких измерениях такие проблемы не так легко обнаружить.

8.5.1 Обзор

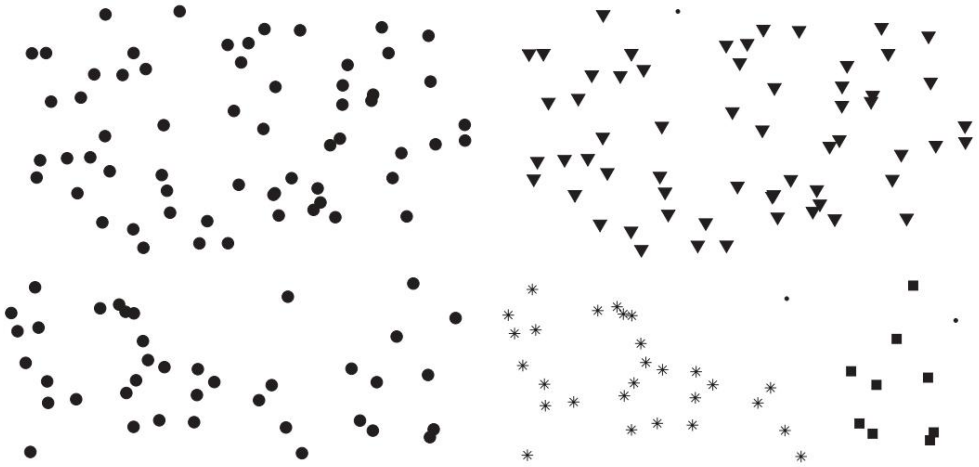
Возможность определить, есть ли в данных неслучайная структура, — это лишь один важный аспект проверки кластера. Ниже приведен список нескольких важных вопросов для проверки кластера.

1. Определение тенденции к кластеризации набора данных, т. е. определение того, действительно ли в данных существует неслучайная структура.
2. Определение правильного количества кластеров.
3. Оценка того, насколько хорошо результаты кластерного анализа соответствуют данным без ссылка на внешнюю информацию.
4. Сравнение результатов кластерного анализа с известными извне результатами, например, предоставленные извне метки классов.
5. Сравнение двух наборов кластеров, чтобы определить, какой из них лучше.

Обратите внимание, что пункты 1, 2 и 3 не используют никакой внешней информации (это неконтролируемые методы), тогда как пункт 4 требует внешней информации.
Пункт 5 может выполняться как контролируемым, так и неконтролируемым способом. В отношении пунктов 3, 4 и 5 можно провести дальнейшее различие: хотим ли мы оценить всю кластеризацию или только отдельные кластеры?

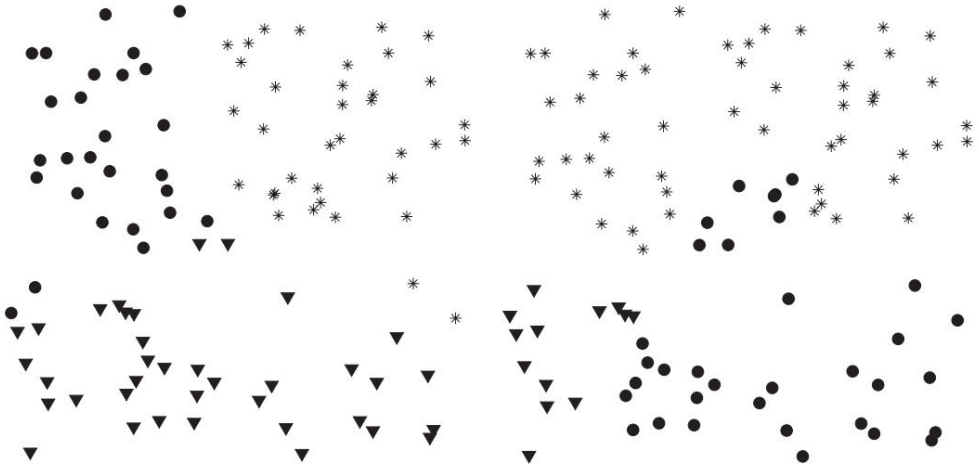
Хотя можно разработать различные числовые показатели для оценки различных аспектов валидности кластера, упомянутых выше, существует ряд проблем. Во-первых, мера валидности кластера может быть весьма ограничена в сфере своей применимости. Например, большая часть работы по измерению тенденции к кластеризации была проделана для двух- или трехмерных пространственных данных. Во-вторых, нам нужна основа для интерпретации любой меры. Если мы получим значение 10 для показателя, который оценивает, насколько хорошо метки кластера соответствуют меткам классов, предоставленным извне, представляет ли это значение хорошее, удовлетворительное или плохое соответствие? Качество совпадения часто можно измерить, рассматривая статистическое распределение этого значения, т. е. насколько вероятно, что такое значение возникнет случайно. Наконец, если мера слишком сложна для применения или понимания, мало кто будет ее использовать.

Меры оценки или индексы, которые применяются для оценки различных аспектов валидности кластера, традиционно подразделяются на следующие три типа.



(а) Исходные точки.

(б) Три кластера, обнаруженные DBSCAN.



(с) Три кластера, найденные с помощью K-средних.

(г) Три кластера, обнаруженные методом полного связь.

Рисунок 8.26. Кластеризация 100 равномерно распределенных точек.

Без присмотра. Измеряет качество структуры кластеризации без учета внешней информации. Примером этого является SSE. Неконтролируемые меры валидности кластера часто подразделяются на два класса: меры сплоченности кластера (компактность, плотность), которые определяют, насколько тесно связаны объекты в кластере, и меры разделения кластера (изоляции), которые определяют, насколько различны или хорошо отделен кластер от других кластеров. Неконтролируемые меры часто называют внутренними индексами, поскольку они используют только информацию, присутствующую в наборе данных.

Под присмотром. Измеряет степень, в которой структура кластеризации, обнаруженная алгоритмом кластеризации, соответствует некоторой внешней структуре. Примером контролируемого индекса является энтропия, которая измеряет, насколько хорошо метки кластера соответствуют меткам классов, предоставленным извне. Контролируемые показатели часто называют внешними индексами, поскольку они используют информацию, отсутствующую в наборе данных.

Родственник. Сравнивает различные кластеризации или кластеры. Относительная мера оценки кластера — это контролируемая или неконтролируемая мера оценки, которая используется в целях сравнения. Таким образом, относительные меры на самом деле не являются отдельным типом меры оценки кластера, а представляют собой конкретное использование таких мер. Например, две кластеризации K-средних можно сравнить, используя либо SSE, либо энтропию.

В оставшейся части этого раздела мы приводим конкретные подробности, касающиеся валидности кластера. Сначала мы опишем темы, связанные с неконтролируемой оценкой кластеров, начиная с (1) мер, основанных на сплоченности и разделении, и (2) двух методов, основанных на матрице близости. Поскольку эти подходы полезны только для разделенных наборов кластеров, мы также описываем популярный коэффициент кофенетической корреляции, который можно использовать для неконтролируемой оценки иерархической кластеризации. Мы заканчиваем наше обсуждение неконтролируемой оценки краткими обсуждениями поиска правильного количества кластеров и оценки тенденции кластеризации. Затем мы рассмотрим контролируемые подходы к валидности кластера, такие как энтропия, чистота и мера Жаккара. Мы завершаем этот раздел кратким обсуждением того, как интерпретировать значения показателей достоверности (неконтролируемых или контролируемых).

8.5.2 Неконтролируемая оценка кластера с использованием сплоченности и разделения

Многие внутренние меры валидности кластера для схем секционной кластеризации основаны на понятиях сплоченности или разделения. В этом разделе мы используем меры кластерной достоверности для методов кластеризации на основе прототипов и графов, чтобы более подробно изучить эти понятия. В процессе мы также увидим некоторые интересные взаимосвязи между кластеризацией на основе прототипов и графов.

В общем, мы можем рассмотреть возможность выражения общей валидности кластера для набора K кластеров как взвешенная сумма достоверности отдельных кластеров,

$$\text{общая достоверность} = \sum_{i=1}^{K} \text{с действительностью } (C_i). \tag{8,8}$$

Функцией достоверности может быть сцепление, разделение или некоторая комбинация этих величин. Веса будут варьироваться в зависимости от меры достоверности кластера.
В некоторых случаях веса просто равны 1 или размеру кластера, тогда как в других случаях они отражают более сложное свойство, такое как квадратный корень из связности. См. Таблицу 8.6.
Если функция достоверности является связностью, то чем выше значение, тем лучше. Если это разделение, то лучше использовать более низкие значения.

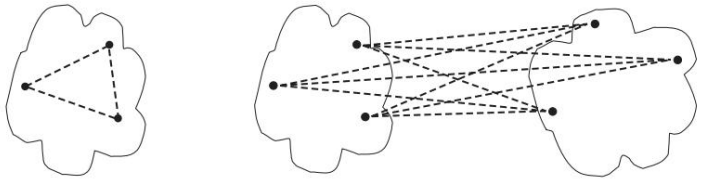
Графический взгляд на сплоченность и разделение

Для кластеров на основе графов связность кластера можно определить как сумму весов связей в графе близости, которые соединяют точки внутри кластера. См. рисунок 8.27(a). (Напомним, что граф близости имеет объекты данных в качестве узлов, связь между каждой парой объектов данных и вес, присвоенный каждой ссылке, который представляет собой близость между двумя объектами данных, соединенными ссылкой.) Аналогично, разделение между двумя кластерами может быть измерена суммой весов связей от точек одного кластера к точкам другого кластера.
Это показано на рисунке 8.27(b).

Математически связность и разделение кластера на основе графа можно выразить с помощью уравнений 8.9 и 8.10 соответственно. Функция близости может быть сходством, несходством или простой функцией этих величин.

$$\text{сплоченность}(C_i) = \sum_{\substack{x \in C_i \\ y \in C_i}} \text{близость}(x, y) \tag{8,9}$$

$$\text{разделение}(C_i, C_j) = \sum_{\substack{x \in C_i \\ y \in C_j}} \text{близость}(x, y) \tag{8.10}$$



(а) Сплоченность. (б) Разделение.

Рисунок 8.27. Графическое представление сплоченности и разделения кластеров.

Взгляд на сплоченность и разделение на основе прототипа

Для кластеров на основе прототипов сплоченность кластера можно определить как сумма близостей относительно прототипа (центроида или медоида) кластер. Аналогичным образом можно измерить расстояние между двумя кластерами. близостью двух прототипов кластера. Это показано на рисунке 8.28, где центр тяжести кластера обозначен знаком «+».

Сплоченность кластера на основе прототипа представлена уравнением 8.11, а две меры разделения даны в уравнениях 8.12 и 8.13 соответственно, где ci — прототип (центроид) кластера Ci , а c — общий прототип (центроид). Есть две меры разделения, потому что, как мы вскоре мы увидим отделение прототипов кластера от общего прототипа иногда напрямую связано с отделением прототипов кластеров от одного другой. Обратите внимание, что уравнение 8.11 представляет собой SSE кластера, если мы позволяем близости быть квадрат евклидова расстояния.

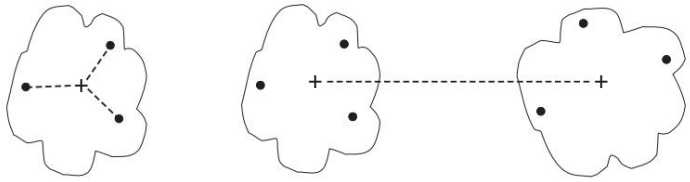
$$\text{сплоченность}(C_i) = \sum_{x \in C_i} \text{близость}(x, c_i) \tag{8.11}$$

$$\text{разделение}(C_i, C_j) = \text{близость}(c_i, c_j) \tag{8.12}$$

$$\text{разделение}(C_i) = \text{близость}(c_i, c) \tag{8.13}$$

Общие меры сплоченности и разделения

Предыдущие определения сплоченности и разделения кластеров дали нам несколько простых и четко определенных показателей валидности кластера, которые можно объединить в общая мера достоверности кластера с использованием взвешенной суммы, как указано



(а) Сплоченность. (б) Разделение.

Рисунок 8.28. Представление о сплоченности и разделении кластеров на основе прототипов.

в уравнении 8.8. Однако нам нужно решить, какие веса использовать. Неудивительно, что используемые веса могут сильно различаться, хотя обычно они немного различаются. мера размера кластера.

В Таблице 8.6 приведены примеры показателей достоверности, основанных на сплоченности и разлука. I1 является мерой сцепления с точки зрения попарной близости объектов в кластере, разделенных на размер кластера. I2 — мера сплоченности на основе суммы близостей объектов в кластере к кластеру центроид. E1 — это мера разделения, определяемая как близость кластера. центроида к общему центроиду, умноженному на количество объектов в кластер. G1, который является мерой, основанной как на сплоченности, так и на разделении, сумма попарной близости всех объектов в кластере со всеми объектами вне кластера — общий вес ребер графа близости, которые необходимо разрезать, чтобы отделить кластер от всех других кластеров, разделенных сумма попарной близости объектов в кластере.

Таблица 8.6. Таблица показателей оценки кластера на основе графиков.

Название кластера	Мера	Вес кластера	Тип
я1	$\sum_{x \varepsilon Ci} \sum_{y \varepsilon Ci}$ близость (x, y)	$\frac{1}{\text{ми}}$	основанный на графе сплоченность
я2	$\sum_{x \varepsilon Ci}$ близость (x, ci)	1	на основе прототипа сплоченность
Е1	близость (ci, c)	ми	на основе прототипа разделение
Г1	$\sum_{j=1}^{K} \sum_{j=я} \sum_{x \varepsilon Ci} \sum_{y \varepsilon Cj}$ близость (x, y)	$\frac{1}{\sum_{x \varepsilon Ci} \sum_{y \varepsilon Ci} \text{близость (x, y)}}$	основанный на графе разделение и сплоченность

Обратите внимание, что любая неконтролируемая мера достоверности кластера потенциально может использоваться в качестве целевой функции для алгоритма кластеризации и наоборот. Инструмент CLUstering TOolkit (CLUTO) (см. библиографические примечания) использует меры оценки кластеров, описанные в таблице 8.6, а также некоторые другие меры оценки, не упомянутые здесь, для управления процессом кластеризации. Для этого используется алгоритм, аналогичный алгоритму дополнительных K-средних, обсуждаемому в разделе 8.2.2. В частности, каждая точка присваивается кластеру, который дает наилучшее значение для функции оценки кластера. Показатель оценки кластера I2 соответствует традиционным K-средним и создает кластеры с хорошими значениями SSE. Другие меры создают кластеры, которые не так хороши с точки зрения SSE, но более оптимальны с точки зрения указанной меры валидности кластера.

Связь между связностью на основе прототипов и связностью на основе графов

Сплоченность

Хотя подходы к измерению сплоченности и разделения кластера на основе графов и прототипов кажутся разными, для некоторых показателей близости они эквивалентны. Например, для SSE и точек евклидова пространства можно показать (уравнение 8.14), что среднее попарное расстояние между точками в кластере эквивалентно SSE кластера. См. упражнение 27 на стр. 566.

$$\text{Кластер SSE} = \sum_{x \in C_i} \text{расстояние}(c_i, x)^2 = \frac{1}{2m_i} \sum_{x \in C_i} \sum_{y \in C_i} \text{расстояние}(x, y)^2 \qquad (8.14)$$

Два подхода к разделению на основе прототипов

Когда близость измеряется евклидовым расстоянием, традиционной мерой разделения между кластерами является сумма квадратов между группами (SSB), которая представляет собой сумму квадрата расстояния центроида кластера c_i к общему среднему значению c всех точки данных. Суммируя SSB по всем кластерам, мы получаем общий SSB, который определяется уравнением 8.15, где c_i — среднее значение i-го кластера, а c — общее среднее значение. Чем выше общий SSB кластеризации, тем более отделены кластеры друг от друга.

$$\text{Всего SSB} = \sum_{i=1}^{K} m_i \, \text{расстояние}(c_i, c)^2 \qquad (8.15)$$

Несложно показать, что общий SSB напрямую связан с попарными расстояниями между центроидами. В частности, если размеры кластеров

равны, т. е. mi = m/K, то это соотношение принимает простую форму, заданную уравнением 8.16. (См. упражнение 28 на стр. 566.) Именно этот тип эквивалентности мотивирует определение разделения прототипов в терминах уравнений 8.12 и 8.13.

$$\text{Всего SSB} = \frac{1}{2K}\sum_{i=1}^{K}\sum_{j=1}^{K} \frac{m}{K}\,\text{расстояние}(c_i, c_j)^2 \qquad (8,16)$$

Связь между сплоченностью и разделением

В некоторых случаях также существует тесная взаимосвязь между сплоченностью и разделением. В частности, можно показать, что сумма общего SSE и общего SSB является константой; т.е. оно равно общей сумме квадратов (TSS), которая представляет собой сумму квадратов расстояния каждой точки до общего среднего значения данных. Важность этого результата заключается в том, что минимизация SSE (сплоченности) эквивалентна максимизации SSB (разделения).

Ниже мы приводим доказательство этого факта, поскольку этот подход иллюстрирует методы, которые также применимы для доказательства отношений, изложенных в последних двух разделах. Чтобы упростить обозначения, мы предполагаем, что данные одномерны, т. е. dist(x, y)=(x y)2 . Кроме того, мы используем тот факт, что перекрестный $\sum_{i=1}^{K}\sum_{x\varepsilon Ci}$ член (x ci)(c ci) равен 0. (См. упражнение 29 на стр. 566.)

$$\begin{aligned}
\text{TCC} &= \sum_{i=1}^{K}\sum_{x\varepsilon Ci} (x - c)^2 \\
&= \sum_{i=1}^{K}\sum_{x\varepsilon Ci} ((x - ci) - (c - ci))^2 \\
&= \sum_{i=1}^{K}\sum_{x\varepsilon Ci} (x - ци)^2 - 2\sum_{i=1}^{K}\sum_{x\varepsilon Ci} (x - ci)(c - ci) + \sum_{i=1}^{K}\sum_{x\varepsilon Ci} (c - ци)^2 \\
&= \sum_{i=1}^{K}\sum_{x\varepsilon Ci} (x - ци)^2 + \sum_{i=1}^{K}\sum_{x\varepsilon Ci} (c - ци)^2 \\
&= \sum_{i=1}^{K}\sum_{x\varepsilon Ci} (x - ци)^2 + \sum_{i=1}^{K} |Ci|(c - ci)^2 \\
&= \text{CCE} + \text{CCC}
\end{aligned}$$

Оценка отдельных кластеров и объектов

До сих пор мы фокусировались на использовании сплоченности и разделения при общей оценке группы кластеров. Многие из этих показателей достоверности кластера также можно использовать для оценки отдельных кластеров и объектов. Например, мы можем ранжировать отдельные кластеры в соответствии с их конкретным значением валидности кластера, т. е. сплоченности или разделения кластера. Кластер с высоким значением сплоченности можно считать лучшим, чем кластер с более низким значением. Эту информацию часто можно использовать для улучшения качества кластеризации. Если, например, кластер не очень сплочен, мы можем захотеть разделить его на несколько подкластеров. С другой стороны, если два кластера относительно сплочены, но плохо разделены, мы можем захотеть объединить их в один кластер.

Мы также можем оценить объекты внутри кластера с точки зрения их вклада в общую сплоченность или разделение кластера. Объекты, которые больше способствуют сплочению и разделению, находятся вблизи «внутренней части» кластера. Те объекты, для которых верно обратное, вероятно, находятся вблизи «края» скопления. В следующем разделе мы рассмотрим меру оценки кластеров, которая использует основанный на этих идеях подход для оценки точек, кластеров и всего набора кластеров.

Коэффициент силуэта

Популярный метод коэффициентов силуэта сочетает в себе как сплоченность, так и разделение. Следующие шаги объясняют, как вычислить коэффициент силуэта для отдельной точки. Этот процесс состоит из следующих трех шагов.
Мы используем расстояния, но аналогичный подход можно использовать и для сходства.

1. Для меня объекта, вычислите его среднее расстояние до всех остальных объектов в его кластер. Назовите это значение ai.
2. Для меня объект и любой кластер, не содержащий объект, вычисляет среднее расстояние объекта до всех объектов в данном кластере. Найдите минимальное такое значение по отношению ко всем кластерам; назовите это значение bi.
3. Для меня Для объекта коэффициент силуэта равен si = (bi ai)/ max(ai, bi).

Значение коэффициента силуэта может варьироваться от 1 до 1. Отрицательное значение нежелательно, поскольку оно соответствует случаю, когда ai, среднее расстояние до точек в кластере, больше, чем bi, минимальное среднее расстояние до точек в кластере. еще один кластер. Мы хотим, чтобы коэффициент силуэта был положительным (ai < bi), а значение ai было как можно ближе к 0, поскольку коэффициент принимает максимальное значение 1, когда ai = 0.

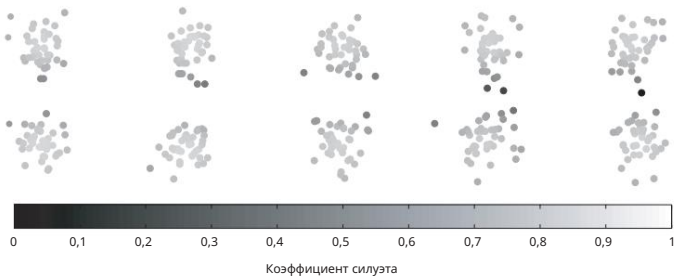


Рисунок 8.29. Коэффициенты силуэта для точек в десяти кластерах.

Мы можем вычислить средний коэффициент силуэта кластера, просто взяв среднее значение коэффициентов силуэта точек, принадлежащих кластер. Общую меру качества кластеризации можно получить с помощью вычисление среднего коэффициента силуэта всех точек.

Пример 8.8 (Коэффициент силуэта). На рис. 8.29 показан график коэффициенты силуэта для точек в 10 кластерах. Более темные оттенки указывают на более низкую коэффициенты силуэта. ■

8.5.3 Неконтролируемая оценка кластера с использованием близости Матрица

В этом разделе мы рассмотрим несколько неконтролируемых подходов к оценке. валидность кластера, основанная на матрице близости. Первый сравнивает реальная и идеализированная матрица близости, тогда как вторая использует визуализацию.

Измерение достоверности кластера посредством корреляции

Если нам дана матрица сходства для набора данных и метки кластера из кластерный анализ набора данных, тогда мы сможем оценить «качественность» кластеризацию, рассматривая корреляцию между матрицей сходства и идеальная версия матрицы сходства, основанная на метках кластеров. (С незначительные изменения, следующее относится к матрицам близости, но для простоты: мы обсуждаем только матрицы сходства.) Более конкретно, идеальный кластер — это один чьи точки имеют сходство 1 со всеми точками в кластере, а сходство 0 для всех точек в других кластерах. Таким образом, если мы отсортируем строки и столбцы матрицы сходства так, чтобы все объекты, принадлежащие одному классу, были вместе, то идеальная матрица подобия имеет блочно-диагональную структуру. В другими словами, сходство не равно нулю, т. е. 1, внутри блоков сходства

матрица, элементы которой представляют сходство внутри кластера, и 0 в других местах. Идеальная матрица сходства строится путем создания матрицы, которая имеет одну строку и один столбец для каждой точки данных (точно так же, как реальная матрица сходства) и присвоения 1 записи, если соответствующая пара точек принадлежит к одному и тому же кластеру. Все остальные записи равны 0.

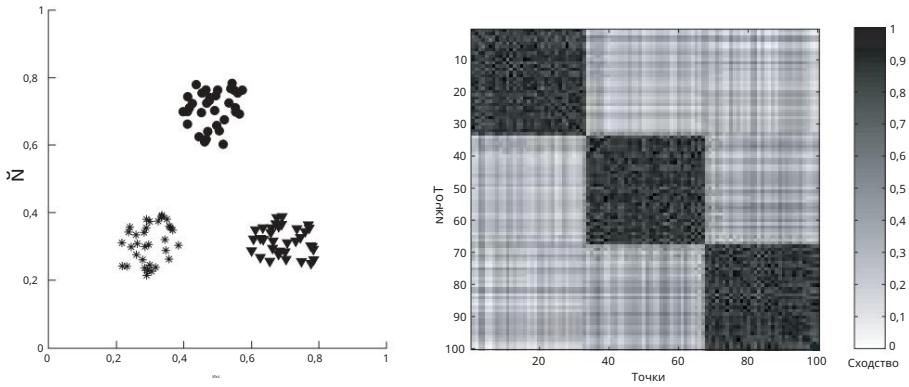
Высокая корреляция между идеальной и фактической матрицами сходства указывает на то, что точки, принадлежащие одному кластеру, расположены близко друг к другу, тогда как низкая корреляция свидетельствует об обратном. (Поскольку фактическая и идеальная матрицы сходства симметричны, корреляция вычисляется только среди n(n 1)/2 элементов ниже или выше диагонали матриц.) Следовательно, это не является хорошей мерой для многих плотностей или значений. кластеры на основе смежности, поскольку они не являются шаровидными и могут быть тесно переплетены с другими кластерами.

Пример 8.9 (Соотношение фактической и идеальной матриц подобия).
Чтобы проиллюстрировать эту меру, мы рассчитали корреляцию между идеальной и фактической матрицами сходства для кластеров K-средних, показанных на рисунке 8.26(c) (случайные данные) и рисунке 8.30(a) (данные с тремя хорошо разделенными кластерами). Корреляции составили 0,5810 и 0,9235 соответственно, что отражает ожидаемый результат: кластеры, найденные с помощью K-средних в случайных данных, хуже, чем кластеры, найденные с помощью K-средних в данных с хорошо разделенными кластерами. ■

Визуальная оценка кластеризации по ее матрице сходства

Предыдущий метод предлагает более общий, качественный подход к оценке набора кластеров: упорядочите матрицу сходства по меткам кластеров, а затем постройте ее график. Теоретически, если у нас есть хорошо разделенные кластеры, то матрица сходства должна быть примерно блочно-диагональной. Если нет, то закономерности, отображаемые в матрице сходства, могут выявить связи между кластерами.
Опять же, все это можно применить и к матрицам различий, но для простоты мы будем обсуждать только матрицы сходства.

Пример 8.10 (Визуализация матрицы подобия). Рассмотрим точки на рисунке 8.30(a), которые образуют три хорошо разделенных кластера. Если мы используем K-средние для группировки этих точек в три кластера, то у нас не должно возникнуть проблем с поиском этих кластеров, поскольку они хорошо разделены. Разделение этих кластеров иллюстрируется переупорядоченной матрицей сходства, показанной на рисунке 8.30(b). (Для единообразия мы преобразовали расстояния в сходства, используя формулу s = 1 (d min d)/(max d min d).) На рисунке 8.31 показаны переупорядоченные матрицы сходства для кластеров, найденных в случайных данных. набор рис. 8.26 с помощью DBSCAN, K-средних и полной ссылки.



(а) Хорошо разделенные кластеры.

(б) Матрица сходства, отсортированная по меткам кластера K-средних.

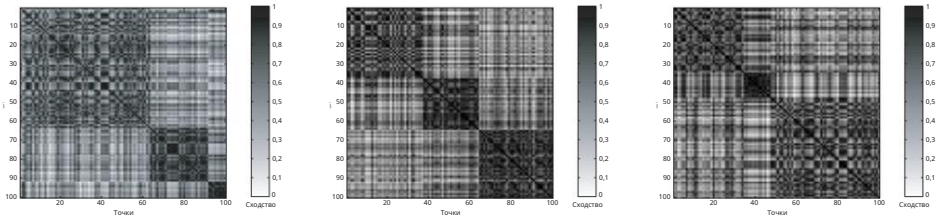
Рисунок 8.30. Матрица сходства для хорошо разделенных кластеров.

Хорошо разделенные кластеры на рис. 8.30 демонстрируют очень четкую блочно-диагональную структуру в переупорядоченной матрице сходства. Однако существуют также слабые диагональные шаблоны блоков (см. рисунок 8.31) в переупорядоченных матрицах сходства кластеризаций, найденных с помощью K-средних, DBSCAN и полной связи в случайных данных. Точно так же, как люди могут находить закономерности в облаках, алгоритмы интеллектуального анализа данных могут находить кластеры в случайных данных. Находить закономерности в облаках интересно, но бессмысленно и, возможно, стыдно находить скопления в шуме. ■

Этот подход может показаться безнадежно дорогим для больших наборов данных, поскольку вычисление матрицы близости занимает время O(m2), где m — количество объектов, но при выборке этот метод все же можно использовать. Мы можем взять выборку точек данных из каждого кластера, вычислить сходство между этими точками и построить график результата. Может потребоваться избыточная выборка небольших кластеров и недостаточная выборка больших, чтобы получить адекватное представление всех кластеров.

8.5.4 Неконтролируемая оценка иерархической кластеризации

Предыдущие подходы к оценке кластеров предназначены для секционной кластеризации. Здесь мы обсуждаем кофенетическую корреляцию, популярную меру оценки иерархических кластеров. Кофенетическое расстояние между двумя объектами — это близость, при которой технология агломеративной иерархической кластеризации



(а) Матрица сходства отсортировано по DBSCAN метки кластера.

(б) Матрица подобия отсортировано по K-средним метки кластера.

(в) Матрица сходства отсортировано по полной ссылке метки кластера.

Рисунок 8.31. Матрицы сходства для кластеров из случайных данных.

nique впервые помещает объекты в один и тот же кластер. Например, если в какой-то момент в процессе агломеративной иерархической кластеризации, наименьший расстояние между двумя объединенными кластерами равно 0,1, тогда все точки в один кластер имеет кофенетическое расстояние 0,1 по отношению к точкам в другой кластер. В матрице кофенетических расстояний элементами являются кофенетические расстояния. расстояния между каждой парой объектов. Кофенетическое расстояние разное. для каждой иерархической кластеризации набора точек.

Пример 8.11 (Матрица кофенетических расстояний). В таблице 8.7 показана матрица кофентических расстояний для однозвенной кластеризации, показанной на рисунке 8.16. (данные для этого рисунка состоят из 6 двумерных точек, приведенных в таблице. 8.3.)

Таблица 8.7. Матрица кофенетических расстояний для одиночной линии связи и данные в таблице 8.3.

Точка P1	П2	П3	П4	П5	П6
П1	0,222 0,222 0,222 0,222 0,222				
0 П2 0,222 П3	0	0,148 0,151 0,139 0,148			
0,222 0,148		0	0,151 0,148 0,110		
П4 0,222 0,151 0,151 П5 0,222			0	0,151 0,151	
0,139 0,148 0,151 P6 0,222 0,148 0,110				0 0,148	
0,151 0,148					0

■

Кофенетический коэффициент корреляции (CPCC) – это корреляция между элементами этой матрицы и исходной матрицей несходства и

стандартная мера того, насколько хорошо иерархическая кластеризация (определенного типа) соответствует данным. Одним из наиболее распространенных применений этой меры является оценка того, какой тип иерархической кластеризации лучше всего подходит для определенного типа данных.

Пример 8.12 (Кофенетический коэффициент корреляции). Мы рассчитали CPCC для иерархических кластеров, показанных на рисунках 8.16–8.19. Эти значения показаны в таблице 8.8. Иерархическая кластеризация, полученная с помощью метода одиночной связи, по-видимому, хуже соответствует данным, чем кластеризация, полученная с помощью полной ссылки, среднего по группе и метода Уорда.

Таблица 8.8. Кофенетический коэффициент корреляции для данных таблицы 8.3 и четырех методов агломеративной иерархической кластеризации.

Техника CPCC Одиночное звено	
0,44 Полное звено 0,63 Среднее по группе 0,66 Уорд 0,64	

■

8.5.5 Определение правильного количества кластеров

Для приблизительного определения правильного или натурального числа кластеров можно использовать различные меры оценки кластеров без присмотра.

Пример 8.13 (Количество кластеров). Набор данных на рисунке 8.29 содержит 10 естественных кластеров. На рисунке 8.32 показан график SSE в зависимости от количества кластеров для кластеризации набора данных (по биссектрисе) K-средними, а на рисунке 8.33 показан средний коэффициент силуэта в зависимости от количества кластеров для тех же данных. Имеется отчетливый излом на SSE и отчетливый пик коэффициента силуэта при числе кластеров, равном 10.

■

Таким образом, мы можем попытаться найти натуральное количество кластеров в наборе данных, ища количество кластеров, в которых имеется перегиб, пик или провал на графике оценочной меры, когда она отображается в зависимости от количества кластеров. . Конечно, такой подход не всегда работает хорошо. Кластеры могут быть значительно более переплетенными или перекрывающимися, чем те, что показаны на рисунке 8.29. Также данные могут состоять из вложенных кластеров. На самом деле кластеры на рис. 8.29 являются в некоторой степени вложенными; т. е. имеется 5 пар кластеров, поскольку кластеры расположены ближе сверху вниз, чем слева направо. На кривой SSE есть излом, который указывает на это, но на кривой коэффициента силуэта это не так.

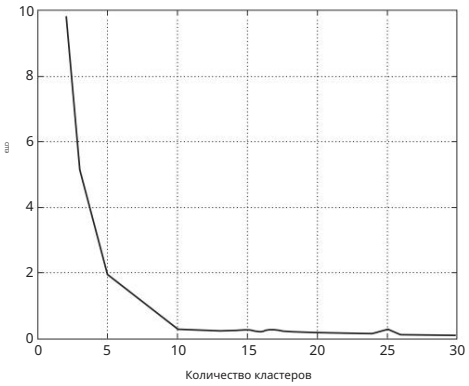


Рисунок 8.32. SSE в зависимости от количества кластеров для данные рисунка 8.29.

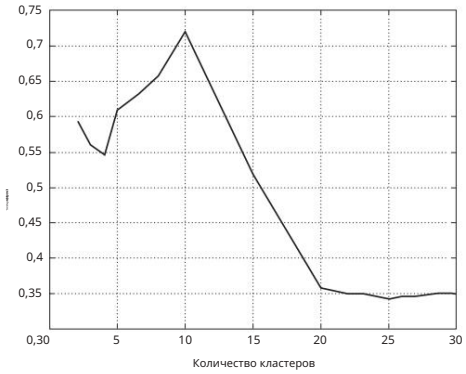


Рисунок 8.33. Средний коэффициент силуэта в зависимости от количества кластеров для данных, представленных на рисунке. 8.29.

как ясно. Подводя итог, можно сказать, что, хотя необходима осторожность, техника, которую мы только что рассмотрели, Описанное может дать представление о количестве кластеров в данных.

8.5.6 Тенденция к кластеризации

Один очевидный способ определить, есть ли в наборе данных кластеры, — попытаться сгруппировать их. это. Однако почти все алгоритмы кластеризации исправно находят кластеры, когда данные данные. Чтобы решить эту проблему, мы могли бы оценить полученные кластеры и утверждать, что набор данных содержит кластеры, только в том случае, если хотя бы некоторые из кластеров имеют хорошее качество. качество. Однако этот подход не учитывает тот факт, что кластеры в данные могут быть другого типа, чем те, которые ищет наш алгоритм кластеризации. Чтобы справиться с этой дополнительной проблемой, мы могли бы использовать несколько алгоритмов и снова оценить качество полученных кластеров. Если кластеры равномерно бедны, то это действительно может указывать на то, что в данных нет кластеров.

В качестве альтернативы, и именно в этом суть показателей тенденции к кластеризации, мы можно попытаться оценить, есть ли в наборе данных кластеры без кластеризации. наиболее распространенным подходом, особенно для данных в евклидовом пространстве, было используйте статистические тесты на пространственную случайность. К сожалению, выбор правильной модели, оценка параметров и оценка статистической значимости гипотезы о неслучайности данных могут оказаться весьма сложной задачей. Тем не менее, было разработано множество подходов, большинство из которых рассчитаны на баллы. в низкомерном евклидовом пространстве.

Пример 8.14 (статистика Хопкинса). Для этого подхода мы генерируем p точек которые случайным образом распределены по пространству данных, а также выборку p фактических

точки данных. Для обоих наборов точек мы находим расстояние до ближайшего соседа в исходном наборе данных. Пусть ui — расстояния до ближайших соседей искусственно сгенерированных точек, а wi — расстояния до ближайших соседей выборки точек из исходного набора данных. Тогда статистика Хопкинса H определяется уравнением 8.17.

$$Ч = \frac{\sum_{i=1}^{n} wi}{\sum_{i=1}^{n} ui + \sum_{я=1}^{n} ви} \qquad (8,17)$$

Если случайно сгенерированные точки и выборка точек данных имеют примерно одинаковые расстояния до ближайших соседей, то H будет около 0,5. Значения H около 0 и 1 указывают, соответственно, на данные с высокой степенью кластеризации и на данные, которые регулярно распределяются в пространстве данных. В качестве примера можно привести статистику Хопкинса для данных рисунка 8.26, рассчитанную для p = 20 и 100 различных испытаний. Среднее значение H составило 0,56 со стандартным отклонением 0,03. Тот же эксперимент был проведен для хорошо разделенных точек на рисунке 8.30. Среднее значение H составило 0,95 со стандартным отклонением 0,006.

■

8.5.7 Контролируемые меры валидности кластера

Когда у нас есть внешняя информация о данных, она обычно имеет форму меток внешних производных классов для объектов данных. В таких случаях обычной процедурой является измерение степени соответствия между метками кластера и метками классов. Но почему это представляет интерес? Ведь если у нас есть метки классов, то какой смысл проводить кластерный анализ? Мотивами для такого анализа является сравнение методов кластеризации с «настоящими фактами» или оценка степени, в которой процесс ручной классификации может быть автоматически выполнен с помощью кластерного анализа.

Мы рассматриваем два разных подхода. Первый набор методов использует меры классификации, такие как энтропия, чистота и F-мера.
Эти меры оценивают степень, в которой кластер содержит объекты одного класса. Вторая группа методов связана с мерами сходства двоичных данных, такими как мера Жаккара, которую мы видели в главе 2. Эти подходы измеряют степень, в которой два объекта одного класса находятся в одном кластере и наоборот. . Для удобства мы будем называть эти два типа показателей ориентированными на классификацию и ориентированными на сходство соответственно.

Классификационно-ориентированные меры валидности кластера

Существует ряд мер — энтропия, чистота, точность, полнота и
F-мера — обычно используется для оценки эффективности модели классификации. В случае
классификации мы измеряем степень, в которой
прогнозируемые метки классов соответствуют фактическим меткам классов, но для мер
Как только что упоминалось, ничего фундаментального не меняется при использовании меток
кластера вместо прогнозируемых меток классов. Далее мы быстро рассмотрим определения этих
меры, которые обсуждались в главе 4.

Энтропия: степень, в которой каждый кластер состоит из объектов одного класса.

Для каждого кластера сначала рассчитывается распределение данных по классам, т.е.
для кластера j мы вычисляем pij — вероятность того, что член кластера i
принадлежит классу j как pij = mij/mi, где mi — количество объектов в
кластер i и mij — количество объектов класса j в кластере i. С использованием
этого распределения классов, энтропия каждого кластера i рассчитывается с использованием
стандартная формула, ei = где L — количество $-\sum_{j=1}^{L} pij \log_2 pij$,
классы. Полная энтропия набора кластеров рассчитывается как сумма
энтропии каждого кластера, взвешенной по размеру каждого кластера, т. е.
e = $\sum_{i=1}^{K} \frac{mi}{m} ei$, где K — количество кластеров, а m — общее количество
количество точек данных.

Чистота: еще одна мера того, в какой степени кластер содержит объекты
один класс. Используя предыдущую терминологию, чистота кластера i равна
max. $_{j} pij$, общая чистота кластеризации равна чистоте = pi = $\sum_{i=1}^{K} \frac{mi}{m} Pi$.

Точность: доля кластера, состоящая из объектов указанного класса.

Точность кластера i относительно класса j равна Precision(i, j) = pij .

Напомним: степень, в которой кластер содержит все объекты указанного класса.

Отзыв кластера i по отношению к классу j — это отзыв(i, j) = mij/mj ,
где mj — количество объектов в классе j.

F-мера Комбинация точности и полноты, которая измеряет
степень, в которой кластер содержит только объекты определенного класса и все
объекты этого класса. F-мера кластера i относительно класса j равна
F(i, j) = (2 × точность (i, j) × отзыв (i, j))/(точность (i, j) + отзыв (i, j)).

Пример 8.15 (меры контролируемой оценки). Мы представляем пример, иллюстрирующий
эти меры. В частности, мы используем K-средние с косинусом
мера сходства с кластером 3204 газетных статей из Лос-Анджелеса.

Таблица 8.9. Результаты кластеризации K-средних для набора данных документа LA Times.

Кластер Развлечения	Финансовая	Иностранная	Метро	Национальная	Спортивная	Энтропия	Чистота
1	3	5	40	506	96	27	1,2270 0,7474
2	4	7	280	29	39	2	1,1472 0,7756
3						671	0,1813 0,9796
4	1 10	1 162	1 3	7 119	4 73	2	1,7487 0,4390
5	331	22	5	70	13	23	1,3976 0,7134
6 5		358	12	212	48	13	1,5523 0,5525
Всего 354		555	341	943	273	738	1,1450 0,7203

Времена. Эти статьи относятся к шести различным классам: «Развлечения», «Финансы», «Иностранные», «Метро», «Национальные» и «Спорт». В таблице 8.9 показаны результаты K-означает кластеризацию для поиска шести кластеров. Первый столбец обозначает кластер, а следующие шесть столбцов вместе образуют матрицу путаницы; то есть, эти столбцы указывают, как документы каждой категории распределяются между кластеры. Последние два столбца — это энтропия и чистота каждого кластера. соответственно.

В идеале каждый кластер будет содержать документы только одного класса. В действительности, каждый кластер содержит документы из многих классов. Тем не менее, многие кластеры содержат документы преимущественно только одного класса. В частности, кластер 3, содержащий в основном документы из раздела «Спорт», является исключительно хорошо, как с точки зрения чистоты, так и с точки зрения энтропии. Чистота и энтропия другие кластеры не так хороши, но обычно их можно значительно улучшить, если данные разбивается на большее количество кластеров.

Точность, полноту и F-меру можно рассчитать для каждого кластера. K приведем конкретный пример, мы рассматриваем кластер 1 и класс Metro таблицы 8.9. Точность 506/677 = 0,75, полнота 506/943 = 0,26, следовательно, Значение F составляет 0,39. Напротив, значение F для кластера 3 и спорта составляет 0,94. ■

Ориентированные на сходство меры валидности кластера

Все меры, которые мы обсуждаем в этом разделе, основаны на предпосылке что любые два объекта, находящиеся в одном кластере, должны принадлежать к одному и тому же классу. и наоборот. Мы можем рассматривать этот подход к валидности кластера как включающий сравнение двух матриц: (1) идеальная матрица сходства кластеров обсуждавшийся ранее, который имеет 1 в ij-й записи, если два объекта, i и j, находятся в одном кластере и 0, в противном случае, и (2) сходство идеального класса матрица, определенная относительно меток классов, которая имеет 1 в i-й записи, если

два объекта, i и j, принадлежат к одному и тому же классу, в противном случае — 0. Как и прежде, мы можем принять корреляцию этих двух матриц в качестве меры валидности кластера. Эта мера известна как статистика Г в литературе по проверке кластеризации.

Пример 8.16 (Корреляция между матрицами кластеров и классов). Чтобы продемонстрировать эту идею более конкретно, мы приведем пример, включающий пять точек данных: p1, p2, p3, p4, p5, два кластера, C1 = {p1, p2, p3} и C2 = {p4, p5}, и два класса. , L1 = {p1, p2} и L2 = {p3, p4, p5}. Идеальные матрицы сходства кластеров и классов приведены в таблицах 8.10 и 8.11. Корреляция между элементами этих двух матриц составляет 0,359.

Таблица 8.10. Матрица сходства идеального кластера.

Точка	p1	p2	p3	p4	p5
p1	1	1	1	0	0
p2	1	1	1	0	0
p3	1	1	1	0	0
p4	0	0	0	1	1
p5	0	0	0	1	1

Таблица 8.11. Матрица сходства идеальных классов.

Точка	p1	p2	p3	p4	p5
p1	1	1	0	0	0
p2	1	1	0	0	0
p3	0	0	1	1	1
p4	0	0	1	1	1
p5	0	0	1	1	1

■

В более общем смысле мы можем использовать любые меры двоичного сходства, которые мы видели в разделе 2.4.5. (Например, мы можем преобразовать эти две матрицы в двоичные векторы, добавив строки.) Мы повторяем определения четырех величин, используемых для определения этих мер сходства, но изменяем наш описательный текст, чтобы он соответствовал текущему контексту. В частности, нам нужно вычислить следующие четыре величины для всех пар различных объектов. (Таких пар m(m 1)/2, если m — количество объектов.)

f00 = количество пар объектов разного класса и разного кластера f01 = количество пар объектов разного класса и одного кластера f10 = количество пар объектов одного класса и разного кластера f11 = количество пар объектов одного класса и одного кластера

В частности, простой коэффициент соответствия, который в этом контексте известен как статистика Рэнда, и коэффициент Жаккара являются двумя наиболее часто используемыми показателями валидности кластера.

$$\text{Статистика Рэнда} = \frac{\phi 00 + \phi 11}{f00 + f01 + f10 + f11} \tag{8.18}$$

$$\text{Коэффициент Жаккара} = \frac{f11}{f01 + f10 + f11} \tag{8,19}$$

Пример 8.17 (меры Ранда и Жаккара). На основании этих формул мы можем легко вычислить статистику Рэнда и коэффициент Жаккара для пример на основе таблиц 8.10 и 8.11. Учитывая, что f00 = 4, f01 = 2, f10 = 2, и f11 = 2, статистика Рэнда = (2 + 4)/10 = 0,6 и коэффициент Жаккара = 2/(2+2+2) = 0,33. ■

Отметим также, что четыре величины f00, f01, f10 и f11 определяют кон- таблицу тингентности, как показано в Таблице 8.12.

Таблица 8.12. Двусторонняя таблица сопряженности для определения принадлежности пар объектов к одному классу. и тот же кластер.

	Тот же кластер	Другой кластер
Тот же класс	ф11 ф10	
Другой класс	ф01 ф00	

Ранее, в контексте анализа ассоциаций (см. раздел 6.7.1), мы представил обширное обсуждение мер ассоциации, которые могут быть использованы для этого типа таблицы непредвиденных обстоятельств. (Сравните Таблицу 8.12 с Таблицей 6.7.) меры также могут быть применены к валидности кластера.

Валидность кластера для иерархических кластеров

До сих пор в этом разделе мы обсуждали контролируемые меры валидности кластеров только для секционированных кластеров. Контролируемая оценка иерархической кластеризация сложнее по ряду причин, включая тот факт, что ранее существовавшая иерархическая структура часто не существует. Здесь мы дадим пример подхода к оценке иерархической кластеризации с точки зрения (плоский) набор меток классов, которые с большей вероятностью будут доступны, чем уже существующий иерархическая структура.

Основная идея этого подхода состоит в том, чтобы оценить, содержит ли иерархическая кластеризация для каждого класса хотя бы один кластер, который является относительно чистым и включает большинство объектов этого класса. Чтобы оценить иерархическую кластеризацию с точки зрения этой цели, мы вычисляем для каждого класса F-меру для каждый кластер в иерархии кластеров. Для каждого класса мы берем максимальную F-меру, достигнутую для любого кластера. Наконец, мы вычисляем общую F-меру для иерархическая кластеризация путем вычисления средневзвешенного значения всех показателей каждого класса F-меры, где веса основаны на размерах классов. Более формально,

эта иерархическая F-мера определяется следующим образом:

$$F = \sum_{j} \frac{m_j}{m} \max_i F(i, j)$$

где максимум берется по всем кластерам i на всех уровнях, mj — количество объектов в классе j, а m — общее количество объектов.

8.5.8 Оценка значимости показателей валидности кластера

Меры валидности кластера призваны помочь нам оценить качество полученных кластеров. Действительно, они обычно дают нам одно число как меру этой доброты. Однако тогда мы сталкиваемся с проблемой интерпретации значения этого числа, задачей, которая может оказаться еще более сложной.

Во многих случаях минимальные и максимальные значения показателей кластерной оценки могут служить некоторым ориентиром. Например, по определению чистота 0 — это плохо, а чистота 1 — хорошо, по крайней мере, если мы доверяем меткам наших классов и хотим, чтобы структура нашего кластера отражала структуру классов. Точно так же энтропия, равная 0, хороша, как и SSE, равная 0.

Однако иногда минимальное или максимальное значение может отсутствовать, или масштаб данных может повлиять на интерпретацию. Кроме того, даже если существуют минимальные и максимальные значения с очевидной интерпретацией, промежуточные значения все равно необходимо интерпретировать. В некоторых случаях мы можем использовать абсолютный стандарт. Если, например, мы кластеризуемся ради полезности, мы, возможно, готовы допустить только определенный уровень ошибок при аппроксимации наших точек центроидом кластера.

Но если это не так, то мы должны сделать что-то другое. Распространенный подход заключается в интерпретации значения нашей меры достоверности в статистических терминах. В частности, мы пытаемся оценить, насколько вероятно, что наблюдаемое нами значение может быть получено случайно. Ценность хороша, если она необычна; т. е. если маловероятно, что это будет результатом случайной случайности. Мотивацией для этого подхода является то, что нас интересуют только кластеры, которые отражают неслучайную структуру данных, и такие структуры должны генерировать необычно высокие (низкие) значения нашей меры достоверности кластера, по крайней мере, если меры достоверности предназначены для отражения наличие сильной кластерной структуры.

Пример 8.18 (Значение SSE). Чтобы показать, как это работает, мы представляем пример, основанный на K-средних и SSE. Предположим, нам нужна мера того, насколько хорошо разделены кластеры на рис. 8.30 по отношению к случайным данным. Мы генерируем множество случайных наборов по 100 точек, имеющих тот же диапазон, что и

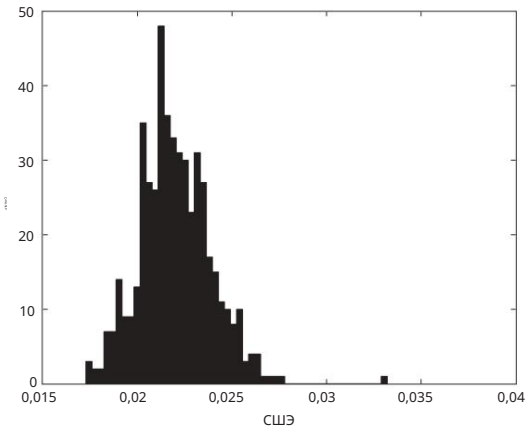


Рисунок 8.34. Гистограмма SSE для 500 наборов случайных данных.

точки в трех кластерах, найдите три кластера в каждом наборе данных с помощью K-средних и накопите распределение значений SSE для этих кластеризаций. K используя это распределение значений SSE, мы можем затем оценить вероятность значения SSE для исходных кластеров. На рисунке 8.34 представлена гистограмма SSE из 500 случайных прогонов. Самый низкий SSE, показанный на рисунке 8.34, равен 0,0173. Для трех кластеров на рисунке 8.30 SSE составляет 0,0050. Поэтому мы могли бы консервативно утверждают, что существует менее 1% вероятности того, что такая кластеризация как показано на рис. 8.30, могло произойти случайно. ■

В заключение мы подчеркиваем, что кластерная оценка — это нечто большее, чем контролируемая оценка. или без присмотра — чем получение численной меры валидности кластера. Если это значение не имеет естественной интерпретации, основанной на определении меры, нам необходимо каким-то образом интерпретировать это значение. Если наша оценка кластера мера определяется так, что более низкие значения указывают на более сильные кластеры, тогда мы можем использовать статистику, чтобы оценить, является ли полученное нами значение необычным низкий, при условии, что у нас есть распределение показателя оценки. Мы представили пример того, как найти такое распределение, но здесь значительно

подробнее по этой теме, и мы отсылаем читателя к библиографическим примечаниям для получения более подробной информации. указатели.

Наконец, даже когда мера оценки используется как относительная мера, т.е. для сравнения двух кластеризаций нам все равно необходимо оценить значимость в разница между оценочными показателями двух кластеров. Хотя одно значение почти всегда будет лучше другого, может быть сложно определить, существенна ли разница. Обратите внимание, что существует два аспекта эта значимость: является ли разница статистически значимой (повторяемой)

и имеет ли значение разница с точки зрения применения. Многие не считают разницу в 0,1% существенной, даже если она стабильно воспроизводима.

8.6 Библиографические примечания

На обсуждение в этой главе наибольшее влияние оказали книги по кластерному анализу, написанные Джайном и Дюбсом [396], Андербергом [374], а также Кауфманом и Руссиу [400]. Дополнительные книги по кластеризации, которые также могут представлять интерес, включают книги Альдендерфера и Блэшфилда [373], Everitt et al. [388], Хартиган [394], Миркин [405], Мурта [407], Ромесбург [409] и Спат [413].
Более статистически ориентированный подход к кластеризации представлен в книге Дуды и др. по распознаванию образов. [385], книга Митчелла по машинному обучению [406] и книга Хасти и др. по статистическому обучению. [395]. Общий обзор кластеризации дан Jain et al. [397], а обзор методов интеллектуального анализа пространственных данных предоставлен Han et al. [393]. Беркин [379] представляет обзор методов кластеризации для интеллектуального анализа данных. Хорошим источником ссылок на кластеризацию за пределами области интеллектуального анализа данных является статья Араби и Хьюберта [376]. В статье Кляйнберга [401] обсуждаются некоторые компромиссы, на которые идут алгоритмы кластеризации, и доказывается, что алгоритм кластеризации не может одновременно обладать тремя простыми свойствами.

Алгоритм K-средних имеет долгую историю, но до сих пор является предметом текущих исследований. Оригинальный алгоритм K-средних был предложен Маккуином [403]. Алгоритм ISODATA Болла и Холла [377] был ранней, но сложной версией K-средних, в которой использовались различные методы предварительной и постобработки для улучшения базового алгоритма. Алгоритм K-средних и многие его варианты подробно описаны в книгах Андерберга [374] и Джайна и Дюбса [396]. Алгоритм деления K-средних пополам, обсуждаемый в этой главе, был описан в статье Steinbach et al. [414], а реализация этого и других подходов к кластеризации свободно доступна для академического использования в пакете CLUTO (CLUstering TOolkit), созданном Карипсом [382]. Боли [380] создал алгоритм кластеризации с разделительным разделением (PDDP), основанный на поиске первого главного направления (компонента) данных, а Саварези и Боли [411] исследовали его связь с делением K-средних пополам. Недавние вариации K-средних представляют собой новую дополнительную версию K-средних (Dhillon et al. [383]), X-средние (Пеллег и Мур [408]) и K-гармонические средние (Чжан и др. [416]). Хамерли и Элкан [392] обсуждают некоторые алгоритмы кластеризации, которые дают лучшие результаты, чем K-средние. Хотя некоторые из ранее упомянутых подходов каким-то образом решают проблему инициализации K-средних,

другие подходы к улучшению инициализации K-средних можно также найти в работе Брэдли и Файяда [381]. Диллон и Модха [384] представляют обобщение K-средних, называемое сферическими K-средними, которое работает с широко используемыми функциями сходства. Общая основа кластеризации K-средних, в которой используются функции несходства на основе расхождений Брегмана, была построена Banerjee et al. [378].

Методы иерархической кластеризации также имеют долгую историю. Большая часть первоначальной деятельности относилась к области таксономии и описана в книгах Джардина и Сибсона [398] и Снита и Сокала [412]. Общие обсуждения иерархической кластеризации также доступны в большинстве книг по кластеризации, упомянутых выше. Агломеративная иерархическая кластеризация находится в центре внимания большинства работ в области иерархической кластеризации, но определенное внимание также уделяется разделительным подходам. Например, Зан [415] описывает метод разделительной иерархии, который использует минимальное остовное дерево графа. Хотя и разделительный, и агломеративный подходы обычно придерживаются мнения, что решения о слиянии (разделении) являются окончательными, Фишер [389] и Карипис и др. провели некоторую работу. [399] для преодоления этих ограничений.

Эстер и др. предложил DBSCAN [387], который позже был обобщен на алгоритм GDBSCAN Сандером и др. [410] для обработки более общих типов данных и мер расстояния, таких как многоугольники, близость которых измеряется степенью пересечения. Инкрементная версия DBSCAN была разработана Kriegel et al. [386]. Одним из интересных разработок DBSCAN является ОПТИКА (упорядочение точек для идентификации структуры кластеризации) (Анкерст и др. [375]), которая позволяет визуализировать структуру кластера, а также может использоваться для иерархической кластеризации.

Авторитетное обсуждение валидности кластеров, которое сильно повлияло на обсуждение в этой главе, представлено в главе 4 книги Джайна и Дьюбса по кластеризации [396]. Более поздние обзоры валидности кластеров принадлежат Halkidi et al. [390, 391] и Миллиган [404]. Коэффициенты силуэта описаны в книге Кауфмана и Руссиу по кластеризации [400]. Источником мер сцепления и разделения в таблице 8.6 является статья Чжао и Кариписа [417], которая также содержит обсуждение энтропии, чистоты и иерархической F-меры. Первоисточником иерархической F-меры является статья Ларсена и Аоне [402].

Библиография [373]

М. С. Альдендерфер и Р. К. Блашфилд. Кластерный анализ. Sage Publications, Лос-Анджелес Анхелес, 1985 год.

[374] Г-н Андерберг. Кластерный анализ для приложений. Академическое издательство, Нью-Йорк, Декабрь 1973 года.

[375] М. Анкерст, М.М. Брюниг, Х.-П. Кригель и Дж. Сандер. ОПТИКА: точки упорядочения для определения структуры кластеризации. В Proc. 1999 г. ACM-SIGMOD Intl. Конф. по управлению данными, страницы 49–60, Филадельфия, Пенсильвания, июнь 1999 г. ACM Press.

[376] П. Араби, Л. Юбер и Г. Д. Соете. Обзор комбинаторного анализа данных. В книге П. Араби, Л. Хьюберта и Г. Д. Соете, редакторов, «Кластеризация и классификация», страницы 188–217. World Scientific, Сингапур, январь 1996 г.

[377] Г. Болл и Д. Холл. Метод кластеризации для суммирования многомерных данных. Наука о поведении, 12: 153–155, март 1967 г.

[378] А. Банерджи, С. Меругу, И. С. Диллон и Дж. Гош. Кластеризация с помощью дивергенций Брегмана. В Proc. Международного конкурса SIAM 2004 г. Конф. по интеллектуальному анализу данных, страницы 234–245, Лейк-Буэна-Виста, Флорида, апрель 2004 г.

[379] П. Берхин. Обзор методов кластерного анализа данных. Технический отчет, Accrue Программное обеспечение, Сан-Хосе, Калифорния, 2002 г.

[380] Д. Боли. Основное направление разделительного разделения. Интеллектуальный анализ данных и знания Дискавери, 2 (4): 325–344, 1998.

[381] П.С. Брэдли и У.М. Файяд. Уточнение начальных точек для кластеризации K-средних. В Proc. 15-го Международного конкурса. Конф. по машинному обучению, страницы 91–99, Мэдисон, Висконсин, июль 1998 г. Morgan Kaufmann Publishers Inc.

[382] CLUTO 2.1.1: Программное обеспечение для кластеризации многомерных наборов данных. /www.cs.umn.edu/karypis, ноябрь 2003 г.

[383] И.С. Диллон, Ю. Гуан и Дж. Коган. Итеративная кластеризация многомерных текстовых данных, дополненная локальным поиском. В Proc. международного стандарта IEEE 2002 г. Конф. по интеллектуальному анализу данных, стр. 131–138. Компьютерное общество IEEE, 2002.

[384] И.С. Диллон и Д.С. Модха. Декомпозиция понятий для больших разреженных текстовых данных Использование кластеризации. Машинное обучение, 42(1/2):143–175, 2001.

[385] Р.О. Дуда, П.Е. Харт и Д.Г. Сторк. Классификация шаблонов. John Wiley & Sons, Inc., Нью-Йорк, второе издание, 2001 г.

[386] М. Эстер, Х.-П. Кригель, Дж. Сандер, М. Виммер и К. Сюй. Инкрементальная кластеризация для майнинга в среде хранения данных. В Proc. 24-й конференции VLDB, стр. 323–333, Нью-Йорк, август 1998 г. Морган Кауфманн.

[387] М. Эстер, Х.-П. Кригель, Дж. Сандер и К. Сюй. Алгоритм на основе плотности для обнаружения кластеров в больших пространственных базах данных с шумом. В Proc. 2-го Международного Конф. по обнаружению знаний и интеллектуальному анализу данных, страницы 226–231, Портленд, Орегон, август 1996 г. AAAI Press.

[388] Б.С. Эверитт, С. Ландау и М. Лиз. Кластерный анализ. Arnold Publishers, Лондон, четвертое издание, май 2001 г.

[389] Д. Фишер. Итеративная оптимизация и упрощение иерархических кластеризаций. Журнал исследований искусственного интеллекта, 4:147–179, 1996.

[390] М. Халкиди, Ю. Батистакис и М. Вазиргианнис. Методы кластерной валидности: часть I. SIGMOD Record (Специальная группа ACM по управлению данными), 31 (2): 40–45, июнь 2002 г.

[391] М. Халкиди, Ю. Батистакис и М. Вазиргианнис. Методы проверки валидности кластеризации: часть II. SIGMOD Record (Специальная группа ACM по управлению данными), 31 (3): 19–27, сентябрь 2002 г.

[392] Г. Хамерли и К. Элкан. Альтернативы алгоритму k-средних, которые находят лучшие кластеризации. В Proc. 11-го Международного конкурса. Конф. по управлению информацией и знаниями, страницы 600–607, Маклин, Вирджиния, 2002. ACM Press.

[393] Дж. Хан, М. Камбер и А. Тунг. Методы пространственной кластеризации в интеллектуальном анализе данных: обзор. В книге Х. Дж. Миллера и Дж. Хана, редакторов журнала Geographic Data Mining and Knowledge Discovery, страницы 188–217. Тейлор и Фрэнсис, Лондон, декабрь 2001 г.

[394] Дж. Хартиган. Алгоритмы кластеризации. Уайли, Нью-Йорк, 1975 год.

[395] Т. Хасти, Р. Тибширани и Дж. Х. Фридман. Элементы статистического обучения: интеллектуальный анализ данных, логический вывод, прогнозирование. Спрингер, Нью-Йорк, 2001 г.

[396] А.К. Джейн и Р.С. Дубес. Алгоритмы кластеризации данных. Серия расширенных справочных материалов Прентис Холла. Прентис Холл, март 1988 г. Книга доступна в Интернете по адресу http://www.cse.msu.edu/ jain/Clustering Jain Dubes.pdf.

[397] А.К. Джайн, М.Н. Мерти и П.Дж. Флинн. Кластеризация данных: обзор. ACM Computing Surveys, 31(3):264–323, сентябрь 1999 г.

[398] Н. Джардин и Р. Сибсон. Математическая таксономия. Уайли, Нью-Йорк, 1971 год.

[399] Карипис Г., Э.-Х. Хан и В. Кумар. Многоуровневое уточнение иерархической кластеризации. Технический отчет TR 99-020, Университет Миннесоты, Миннеаполис, Миннесота, 1999 г.

[400] Л. Кауфман и П.Дж. Руссиу. Поиск групп в данных: введение в кластерный анализ. Ряд Уайли по вероятности и статистике. Джон Уайли и сыновья, Нью-Йорк, ноябрь 1990 г.

[401] Дж. М. Кляйнберг. Теорема невозможности кластеризации. В Proc. 16-й ежегодной конференции. по нейронным системам обработки информации, 9–14 декабря 2002 г.

[402] Б. Ларсен и К. Аоне. Быстрый и эффективный анализ текста с использованием кластеризации документов за линейное время. В Proc. 5-го Международного Конф. по обнаружению знаний и интеллектуальному анализу данных, страницы 16–22, Сан-Диего, Калифорния, 1999. ACM Press.

[403] Дж. Маккуин. Некоторые методы классификации и анализа многомерных наблюдений. В Proc. 5-го симпозиума Беркли. по математической статистике и вероятности, страницы 281–297. Калифорнийский университет Press, 1967.

[404] Г.В. Миллиган. Валидация кластеризации: результаты и последствия для прикладного анализа. В П. Араби, Л. Хьюберте и Г. Д. Соете, редакторах, «Кластеризация и классификация», страницы 345–375. World Scientific, Сингапур, январь 1996 г.

[405] Б. Миркин. Математическая классификация и кластеризация, том 11 книги «Невыпуклая оптимизация и ее приложения». Kluwer Academic Publishers, август 1996 г.

[406] Т. Митчелл. Машинное обучение. МакГроу-Хилл, Бостон, Массачусетс, 1997 г.

[407] Ф. Муртаг. Алгоритмы многомерной кластеризации. Physica-Verlag, Гейдельберг и Вена, 1985 год.

[408] Д. Пеллег и А. В. Мур. X-средние: расширение K-средних за счет эффективной оценки количества кластеров. В Proc. 17-го Международного турнира. Конф. по машинному обучению, страницы 727–734. Морган Кауфманн, Сан-Франциско, Калифорния, 2000 г.

[409] К. Ромесбург. Кластерный анализ для исследователей. Обучение на протяжении всей жизни, Белмонт, Калифорния, 1984.

[410] Дж. Сандер, М. Эстер, Х.-П. Кригель и Х. Сюй. Кластеризация на основе плотности в пространственных базах данных: алгоритм GDBSCAN и его приложения. Интеллектуальный анализ данных и обнаружение передовых знаний, 2 (2): 169–194, 1998.

[411] С.М. Саварези и Д. Боли. Сравнительный анализ биссектрисы K-средних и алгоритмов кластеризации PDDP. Интеллектуальный анализ данных, 8(4):345–362, 2004.

[412] PHA Sneath и RR Sokal. Числовая таксономия. Фриман, Сан-Франциско, 1971 год.

[413] Х. Спат. Алгоритмы кластерного анализа для обработки данных и классификации объектов, том 4 «Компьютеры и их применение». Издательство Эллис Хорвуд, Чич-Эстер, 1980. ISBN 0-85312-141-9.

[414] М. Штайнбах, Г. Карипис и В. Кумар. Сравнение методов кластеризации документов. В Proc. семинара KDD по интеллектуальному анализу текста, учеб. 6-го Международного Конф. по обнаружению знаний и интеллектуальному анализу данных, Бостон, Массачусетс, август 2000 г.

[415] К. Т. Зан. Теоретико-графовые методы обнаружения и описания гештальт-кластеров. Транзакции IEEE на компьютерах, C-20(1):68–86, январь 1971 г.

[416] Б. Чжан, М. Сюй и У. Даял. K-гармонические средние — алгоритм кластеризации данных. Технический отчет HPL-1999-124, Hewlett Packard Laboratories, 29 октября 1999 г.

[417] Ю. Чжао и Г. Карипис. Эмпирические и теоретические сравнения выбранных целевых функций для кластеризации документов. Машинное обучение, 55(3):311–331, 2004.

8.7 Упражнения

1. Рассмотрим набор данных, состоящий из 220 векторов данных, где каждый вектор имеет 32 компонента, и каждый компонент представляет собой 4-байтовое значение. Предположим, что для сжатия используется векторное квантование и используется 216 векторов-прототипов. Сколько байт памяти занимает этот набор данных до и после сжатия и какова степень сжатия?

2. Найдите все хорошо разделенные кластеры в множестве точек, показанном на рисунке 8.35.

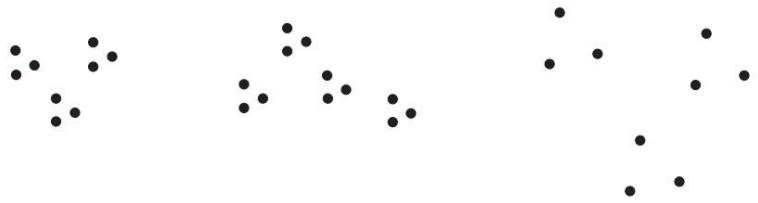


Рисунок 8.35. Баллы за упражнение 2.

3. Многие алгоритмы секционной кластеризации, которые автоматически определяют количество кластеров, утверждают, что это преимущество. Назовите две ситуации, в которых это не так.

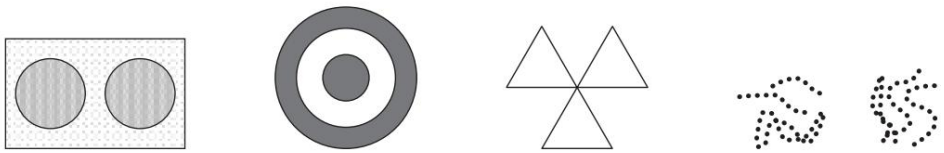
4. Учитывая K кластеров одинакового размера, вероятность того, что случайно выбранный начальный центроид будет происходить из любого данного кластера, равна 1/K, но вероятность того, что каждый кластер будет иметь ровно один начальный центроид, намного ниже. (Должно быть ясно, что наличие одного начального центроида в каждом кластере является хорошей стартовой ситуацией для K-средних.) В общем, если имеется K кластеров и каждый кластер имеет n точек, то вероятность p выбора в выборке размера K один начальный центроид из каждого кластера определяется уравнением 8.20. (Это предполагает выборку с заменой.)
   Из этой формулы мы можем вычислить, например, что вероятность наличия одного начального центроида из каждого из четырех кластеров равна 4!/44 = 0,0938.

$$p = \frac{\text{количество способов выбрать один центроид из каждого кластера}}{\text{количество способов выбрать K центроидов}} = \frac{K!^K}{(K^K)} = \frac{K!}{K^K} \quad (8.20)$$

(а) Постройте график вероятности получения одной точки из каждого кластера выборки. размера K для значений K от 2 до 100.

(б) Для кластеров K, K = 10, 100 и 1000, найдите вероятность того, что выборка размером 2K содержит хотя бы одну точку из каждого кластера. Вы можете использовать либо математические методы, либо статистическое моделирование для определения отвечать.

5. Определите кластеры на рисунке 8.36, используя определения на основе центра, смежности и плотности. Также укажите количество кластеров для каждого случая и дайте краткое изложение ваших рассуждений. Обратите внимание, что темнота или количество точек указывает на плотность. Если это помогает, предположим, что «центральное» означает K-среднее, «непрерывное» означает «одиночное соединение», а «основанное на плотности» означает DBSCAN.

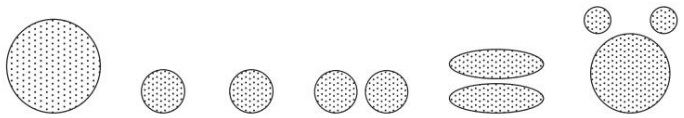


(а) (б) (с) (г)

Рисунок 8.36. Кластеры для упражнения 5.

6. Для следующих наборов двумерных точек (1) дайте схему того, как они будут разделены на кластеры с помощью K-средних для заданного количества кластеров и (2) приблизительно указать, где будут находиться результирующие центроиды. Предполагать что мы используем целевую функцию квадрата ошибки. Если ты думаешь, что там существует более одного возможного решения, тогда укажите, каждое ли решение является глобальным или локальным минимумом. Обратите внимание, что метка каждой диаграммы на рис. 8.37 соответствует соответствующей части этого вопроса, например, на рисунке 8.37(а) показано: с частью (а).

(а) K = 2. Предполагая, что точки распределены по окружности равномерно, сколько возможных способов (теоретически) существует для разделения точек на два кластера? Что вы можете сказать о позициях двух центроиды? (Опять же, вам не нужно указывать точное местоположение центроида, просто качественное описание.)



(а) (б) (в) (г) (е)

Рисунок 8.37. Схемы к упражнению 6.

(б) K = 3. Расстояние между краями кругов немного больше чем радиусы окружностей.

(в) K = 3. Расстояние между краями окружностей много меньше радиусы кругов.

(d) K = 2. (e) K = 3. Подсказка: используйте симметрию ситуации и помните, что мы ищем приблизительный набросок того, каким будет результат.

7. Предположим, что для набора данных

- имеется m точек и K кластеров,
- половина точек и кластеров находятся в «более плотных» регионах,
- половина точек и кластеров находятся в «менее плотных» регионах, и
- эти две области хорошо отделены друг от друга.

Для данного набора данных, что из следующего должно произойти, чтобы минимизировать квадратичную ошибку при поиске K-кластеров:

(а) Центроиды должны быть равномерно распределены между более плотными и менее плотными регионы.

(б) Больше центроидов следует разместить в менее плотной области. (с) Больше центроидов должно быть выделено в более плотную область.

Примечание. Не отвлекайтесь на особые случаи и не учитывайте другие факторы, кроме плотности. Однако, если вы считаете, что истинный ответ отличается от приведенного выше, обоснуйте свой ответ.

8. Рассмотрим среднее значение кластера объектов из набора данных двоичных транзакций. Каковы минимальные и максимальные значения компонентов среднего? Что означают интерпретации компонентов кластера? Какие компоненты наиболее точно характеризуют объекты кластера?

9. Приведите пример набора данных, состоящего из трех естественных кластеров, для которого (почти всегда) K-средние, скорее всего, найдут правильные кластеры, а разделение K-средних пополам — нет.

10. Будет ли косинусная мера подходящей мерой сходства для использования с кластеризацией K-средних для данных временных рядов? Почему или почему нет? Если нет, то какая мера сходства была бы более подходящей?

11. Суммарный SSE – это сумма SSE по каждому отдельному атрибуту. Что это значит, если SSE для одной переменной низкий для всех кластеров? Мало для одного кластера? Высокий для всех кластеров? Высокое значение только для одного кластера? Как вы могли бы использовать информацию SSE для каждой переменной, чтобы улучшить кластеризацию?

12. Алгоритм лидера (Хартиган [394]) представляет каждый кластер с использованием точки, известной как лидер, и присваивает каждую точку кластеру, соответствующему ближайшему лидеру, если только это расстояние не превышает заданного пользователем порога. В этом случае точка становится лидером нового кластера.

    а) Каковы преимущества и недостатки алгоритма лидера?
    по сравнению с K-средствами?

    (б) Предложите способы улучшения алгоритма лидера.

13. Диаграмма Вороного для набора K точек плоскости представляет собой разбиение всех точек плоскости на K областей так, что каждой точке (плоскости) соответствует ближайшая точка среди K заданных точек. (См. рис. 8.38.) Какова связь между диаграммами Вороного и кластерами K-средних? Что диаграммы Вороного говорят нам о возможных формах кластеров K-средних?

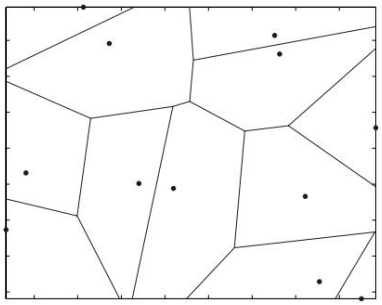


Рисунок 8.38. Диаграмма Вороного к упражнению 13.

14. Вам предоставляется набор данных из 100 записей и предлагается кластеризовать данные.
    Вы используете K-средние для кластеризации данных, но для всех значений K, 1 K 100, алгоритм K-средних возвращает только один непустой кластер. Затем вы применяете инкрементальную версию K-средних, но получаете точно такой же результат. Как это возможно? Как один канал или DBSCAN будут обрабатывать такие данные?

15. Традиционные процедуры агломеративной иерархической кластеризации объединяют два кластера на каждом этапе. Представляется ли вероятным, что такой подход точно отражает

(вложенная) кластерная структура набора точек данных? Если нет, объясните, как можно выполнить постобработку данных, чтобы получить более точное представление о структуре кластера.

16. Используйте матрицу сходства в таблице 8.13 для выполнения однозвенной иерархической кластеризации. Покажите свои результаты, нарисовав дендрограмму. Дендрограмма должна четко показывать порядок объединения точек.

Таблица 8.13. Матрица подобия для упражнения 16.

		p2 p3	p4 p5 p1	p1 1,00 0,10	
0,41	0,55 0,35	p2 0,10 1,00	0,64 0,47	0,98 p3 0,41	
0,64	1,00 0,44	0,85 p4 0,55	0,47 0,44	1,00 0,76 p5	
0,3 5	0,98 0,85	0,76 1,00			

17. Иерархическая кластеризация иногда используется для создания K кластеров, K > 1, путем взятия кластеров на K-м уровне дендрограммы. (Корень находится на уровне 1.) Глядя на кластеры, созданные таким образом, мы можем оценить поведение иерархической кластеризации на разных типах данных и кластеров, а также сравнить иерархические подходы с K-средними.

Ниже приведен набор одномерных точек: {6, 12, 18, 24, 30, 42, 48}.

(а) Для каждого из следующих наборов начальных центроидов создайте два кластера, назначив каждую точку ближайшему центроиду, а затем рассчитайте общий квадрат ошибки для каждого набора из двух кластеров. Покажите и кластеры, и общую квадратичную ошибку для каждого набора центроидов.

я. {18, 45} ii.

{15, 40}

(б) представляют ли оба набора центроидов устойчивые решения; т. е., если бы алгоритм K-средних был запущен на этом наборе точек с использованием данных центроидов в качестве начальных центроидов, произошли ли бы какие-либо изменения в сгенерированных кластерах?

(с) Какие два кластера создаются по одной ссылке? (г) Какой метод, K-

средние или одиночная связь, по-видимому, обеспечивает «наиболее естественную» кластеризацию в этой ситуации? (Для K-средних возьмите кластеризацию с наименьшей квадратичной ошибкой.) (е) Какому

определению(ям) кластеризации соответствует эта естественная кластеризация? (Хорошо разделенные, основанные на центре, непрерывные или плотные.) (f)

Какая известная характеристика алгоритма K-средних объясняет предыдущее поведение?

18. Предположим, мы находим K кластеров, используя метод Уорда, разделяя K-средние пополам и обычные K-средние. Какое из этих решений представляет собой локальный или глобальный минимум? Объяснять.

19. Алгоритмы иерархической кластеризации требуют времени O(m2 log(m)) и, следовательно, их непрактично использовать непосредственно на больших наборах данных. Одним из возможных способов сокращения требуемого времени является выборка набора данных. Например, если требуется K кластеров и из m точек выбрано $\overline{\text{m}}$ точек, то алгоритм иерархической кластеризации создаст иерархическую кластеризацию примерно за O($\overline{\text{m}}$) время. K-кластеры можно извлечь из этой иерархической кластеризации, взяв кластеры на K-м уровне дендрограммы. Остальные точки затем можно отнести к кластеру за линейное время, используя различные стратегии. В качестве конкретного примера можно вычислить центроиды K-кластеров, а затем каждую из оставшихся точек m - $\overline{\text{m}}$ можно присвоить кластеру, связанному с ближайшим центроидом.

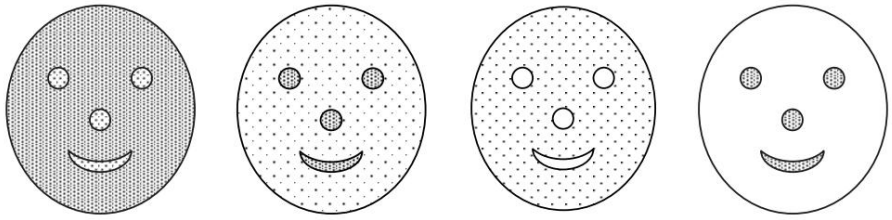
Для каждого из следующих типов данных или кластеров кратко обсудите, вызовет ли (1) выборка проблемы для этого подхода и (2) каковы эти проблемы. Предположим, что метод выборки случайным образом выбирает точки из общего набора из m точек и что любые неупомянутые характеристики данных или кластеров являются максимально оптимальными. Другими словами, сосредоточьтесь только на проблемах, вызванных упомянутой конкретной характеристикой. Наконец, предположим, что K намного меньше m.

(а) Данные с кластерами очень разного размера. (б) Многомерные данные. (c) Данные с выбросами, т.е. нетипичными точками. (d) Данные с крайне нерегулярными областями. (д) Данные с шаровыми скоплениями. (f) Данные с очень разной плотностью. (g) Данные с небольшим процентом шумовых точек. (h) Неевклидовы данные. (i) Евклидовы данные. (j) Данные со многими и смешанными типами атрибутов.

20. Рассмотрим следующие четыре грани, показанные на рис. 8.39. Опять же, темнота или количество точек представляют плотность. Линии используются только для выделения регионов и не обозначают точки.

(а) Не могли бы вы использовать одну ссылку для каждого рисунка, чтобы найти представленные модели? носом, глазами и ртом? Объяснять.

(б) Можете ли вы использовать K-средние для каждого рисунка, чтобы найти представленные закономерности? носом, глазами и ртом? Объяснять.



(а) (б) (с) (г)

Рисунок 8.39. Рисунок к упражнению 20.

(c) Какие ограничения имеет кластеризация при обнаружении всех сформированных паттернов по точкам на рисунке 8.39(c)?

21. Вычислите энтропию и чистоту матрицы путаницы в таблице 8.14.

Таблица 8.14. Матрица ошибок для упражнения 21.

Кластер Развлечения Финансовые Иностранные Метро Национальные Спорт Итого							
№1		1	0 4 693	11		676	
№2	1 27	89	333	827	253	33	1562 г.
№3	326	465	8	105	16	29	949
Общий	354	555	341	943	273	738	3204

22. Вам даны два набора по 100 очков, попадающих в пределах единичного квадрата. Один комплект точек расположены так, что точки расположены равномерно. Другой набор точки создаются из равномерного распределения по единичной площади.

а) Есть ли разница между двумя наборами точек?

(б) Если да, то какой набор точек обычно будет иметь меньшее SSE для K = 10? кластеры?

(c) Как будет вести себя DBSCAN в едином наборе данных? случайный набор данных?

23. Используя данные упражнения 24, вычислите коэффициент силуэта для каждой точки, каждый из двух кластеров и общая кластеризация.

24. Учитывая набор меток кластера и матрицу сходства, показанные в таблицах 8.15 и 8.16 соответственно, вычислите корреляцию между матрицей сходства и идеальная матрица сходства, т. е. матрица, ij-й элемент которой равен 1, если два объекта принадлежат одному и тому же кластеру, и 0 в противном случае.

Таблица 8.15. Таблица меток кластеров для упражнения 24.

Метка кластера точек	
П1	1
П2	1
П3	2
П4	2

Таблица 8.16. Матрица подобия для упражнения 24.

Точка P1 P2 P3 P4				
П1	1	0,8 0,65 0,55		
П2 0,8		1	0,7	0,6
П3 0,65 0,7 П4 0,55			1	0,9
0,6 0,9				1

25. Вычислить иерархическую F-меру для восьми объектов {p1, p2, p3, p4, p5, p6, p7, p8} и иерархическая кластеризация, показанная на рисунке 8.40. Класс A содержит точки p1, p2 и p3, а p4, p5, p6, p7 и p8 принадлежат классу B.

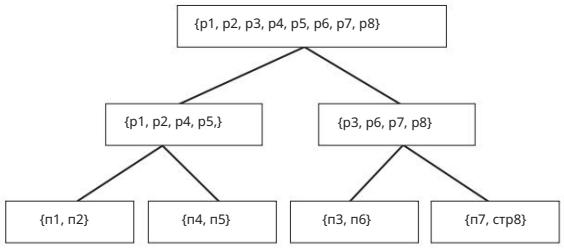


Рисунок 8.40. Иерархическая кластеризация для упражнения 25.

26. Вычислите коэффициент кофенетической корреляции для иерархических кластеризаций. в упражнении 16. (Вам нужно будет преобразовать сходства в различия.)

27. Докажите уравнение 8.14.

28. Докажите уравнение 8.16.

29. Докажите, $\sum_{i=1}^{K} \sum_{x \in C_i} (x - m_i)(m - m_i) = 0$. Этот факт был использован при доказательстве что TSS = SSE + SSB в разделе 8.5.2.

30. Кластеры документов можно обобщить, найдя верхние термины (слова) по документов в кластере, например, взяв наиболее частые k терминов, где k — константа, скажем, 10, или если взять все члены, которые встречаются чаще, чем определенный порог. Предположим, что K-средние используются для поиска кластеров обоих документы и слова для набора данных документа.

    (a) Как может набор кластеров терминов, определяемый главными терминами в документе, кластер отличается от кластеров слов, найденных путем кластеризации терминов с K-значит?

    (b) Как можно использовать термин «кластеризация» для определения групп документов?

31. Мы можем представить набор данных как набор узлов объектов и набор узлы атрибутов, где существует связь между каждым объектом и каждым атрибутом,

и где вес этой ссылки равен значению объекта для этого атрибута. Для разреженных данных, если значение равно 0, ссылка опускается. Двудольная кластеризация пытается разделить этот граф на непересекающиеся кластеры, где каждый кластер состоит из набора узлов объектов и набора узлов атрибутов. Цель состоит в том, чтобы максимизировать вес связей между объектными и атрибутными узлами кластера, минимизируя при этом вес связей между объектными и атрибутными связями в разных кластерах. Этот тип кластеризации также известен как совместная кластеризация, поскольку объекты и атрибуты кластеризуются одновременно.

(а) Чем двусторонняя кластеризация (совместная кластеризация) отличается от кластеризации наборы объектов и атрибутов отдельно?

(б) Существуют ли случаи, когда эти подходы дают одни и те же кластеры? (с) Каковы сильные и слабые стороны совместной кластеризации по сравнению с обычной кластеризацией?

32. На рисунке 8.41 сопоставьте матрицы сходства, отсортированные по меткам кластеров, с наборами точек. Различия в штриховке и форме маркеров различают кластеры, и каждый набор точек содержит 100 точек и три кластера. В наборе точек, отмеченных цифрой 2, есть три очень плотных кластера одинакового размера.

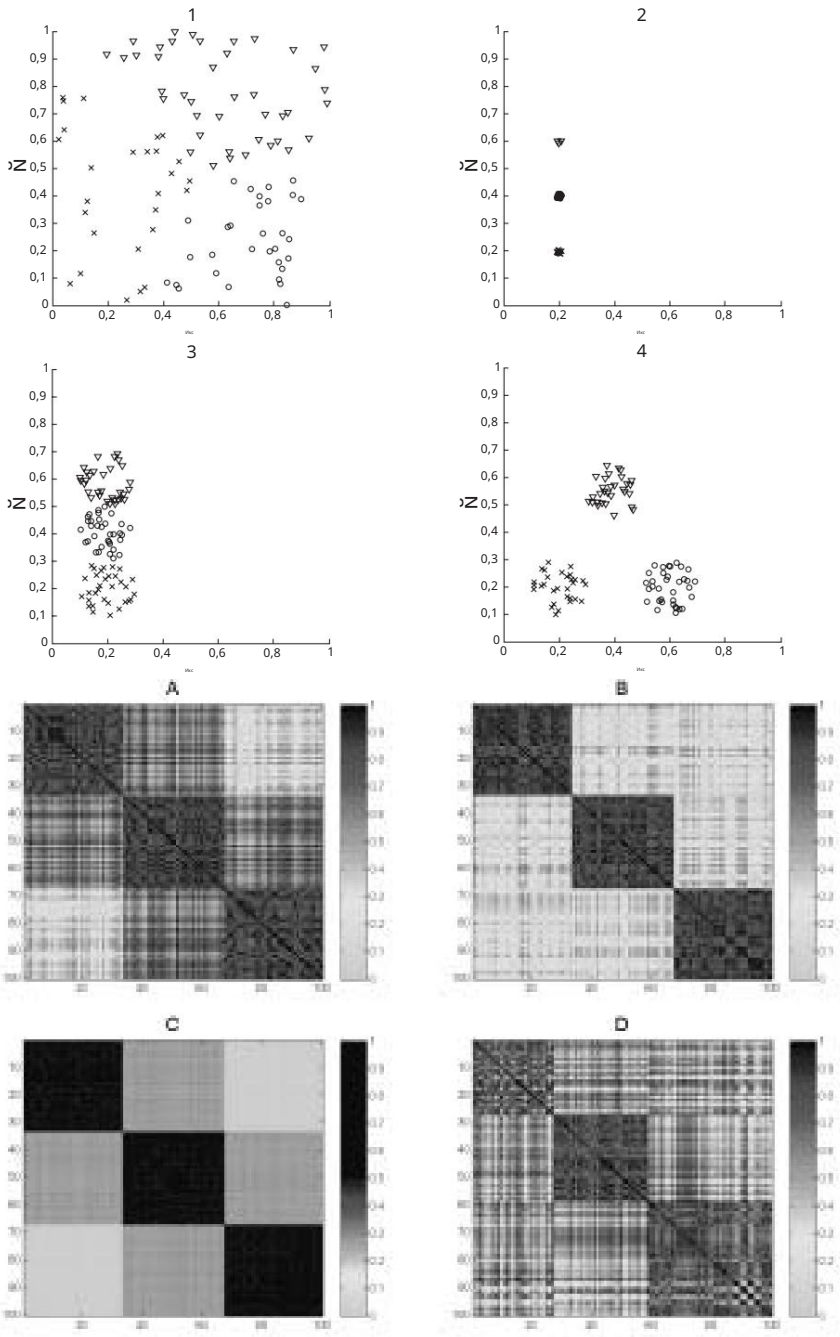


Рисунок 8.41. Точки и матрицы подобия для упражнения 32.