Методы кластерного анализа. Иерархические методы

В лекции рассматриваются основы кластерного анализа, математические характеристики кластера. Описаны две группы иерархического кластерного анализа: агломеративные и дивизимные методы.

В этой лекции мы опишем понятие "*кластер*" с математической точки зрения, а также рассмотрим методы решения задач кластеризации – методы кластерного анализа.

Термин кластерный *анализ*, впервые введенный Трионом (Tryon) в 1939 году, включает в себя более 100 различных алгоритмов.

В отличие от задач классификации, кластерный *анализ* не требует априорных предположений о наборе данных, не накладывает ограничения на *представление* исследуемых объектов, позволяет анализировать показатели различных типов данных (интервальным данным, частотам, бинарным данным). При этом необходимо помнить, что переменные должны измеряться в сравнимых шкалах.

Кластерный *анализ* позволяет сокращать *размерность* данных, делать ее наглядной.

Кластерный *анализ* может применяться к совокупностям временных рядов, здесь могут выделяться периоды схожести некоторых показателей и определяться группы временных рядов со схожей динамикой.

Кластерный *анализ* параллельно развивался в нескольких направлениях, таких как биология, психология, др., поэтому у большинства методов существует по два и более названий. Это существенно затрудняет работу при использовании кластерного анализа.

Задачи кластерного анализа можно объединить в следующие группы:

1. Разработка типологии или классификации.
2. Исследование полезных концептуальных схем группирования объектов.
3. Представление гипотез на основе исследования данных.

4. Проверка гипотез или исследований для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Как правило, при практическом использовании кластерного анализа одновременно решается несколько из указанных задач.

Рассмотрим пример процедуры кластерного анализа.

Допустим, мы имеем набор данных A, состоящий из 14-ти примеров, у которых имеется по два признака X и Y. Данные по ним приведены в таблице 13.1.

Таблица 13.1. Набор данных A

№ примера	**признак X**	**признак Y**
1	27	19
2	11	46
3	25	15
4	36	27
5	35	25
6	10	43
7	11	44
8	36	24
9	26	14
10	26	14
11	9	45
12	33	23
13	27	16
14	10	47

Данные в табличной форме не носят информативный характер. Представим переменные X и Y в виде диаграммы рассеивания, изображенной на рис. 13.1.

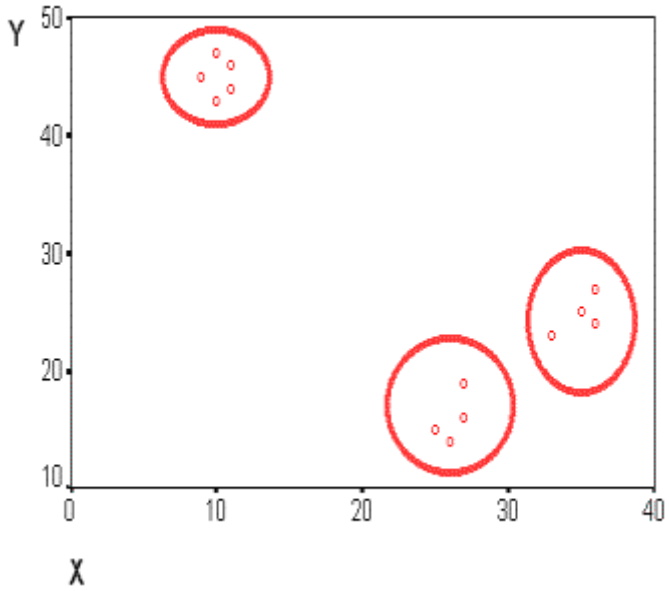

Рис. 13.1. Диаграмма рассеивания переменных X и Y

На рисунке мы видим несколько групп "похожих" примеров. Примеры (объекты), которые по значениям X и Y "похожи" друг на друга, принадлежат к одной группе (кластеру); объекты из разных кластеров не похожи друг на друга.

Критерием для определения схожести и различия кластеров является *расстояние* между точками на диаграмме рассеивания. Это сходство можно "измерить", оно равно расстоянию между точками на графике. Способов определения ***меры расстояния*** между кластерами, называемой еще мерой близости, существует несколько. Наиболее распространенный способ – *вычисление евклидова расстояния* между двумя точками *i* и *j* на плоскости, когда известны их *координаты* X и Y:

$$D_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \tag{13.1}$$

Примечание: чтобы узнать *расстояние* между двумя точками, надо взять разницу их координат по каждой оси, возвести ее в квадрат, сложить полученные значения для всех осей и извлечь квадратный корень из суммы.

Когда осей больше, чем две, *расстояние* рассчитывается таким образом: сумма квадратов разницы координат состоит из стольких слагаемых, сколько осей (измерений) присутствует в нашем пространстве. Например, если нам

нужно найти *расстояние* между двумя точками в пространстве трех измерений (такая ситуация представлена на рис. 13.2), формула (13.1) приобретает вид:

$$D = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}, \qquad (13.2)$$

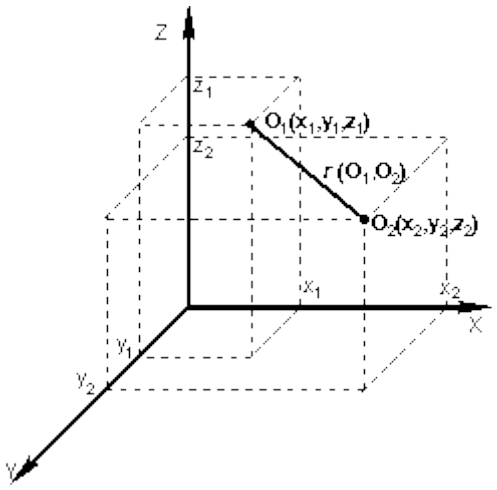

Рис. 13.2. Расстояние между двумя точками в пространстве трех измерений

Кластер имеет следующие ***математические характеристики***: *центр*, *радиус*, *среднеквадратическое отклонение*, *размер кластера*.

Центр кластера – это среднее геометрическое *место* точек в пространстве переменных.

Радиус кластера – максимальное *расстояние* точек от *центра кластера*.

Кластеры могут быть перекрывающимися. Такая ситуация возникает, когда обнаруживается перекрытие кластеров. В этом случае невозможно при помощи математических процедур однозначно отнести *объект* к одному из двух кластеров. Такие объекты называют *спорными*.

Спорный объект – это *объект*, который по мере сходства может быть отнесен к нескольким кластерам.

Размер кластера может быть определен либо по *радиусу кластера*, либо по *среднеквадратичному отклонению* объектов для этого кластера. *Объект* относится к кластеру, если *расстояние* от объекта до *центра кластера* меньше *радиуса кластера*. Если это условие выполняется для двух и более кластеров, *объект* является *спорным*.

Неоднозначность данной задачи может быть устранена экспертом или аналитиком.

Работа кластерного анализа опирается на два предположения. Первое предположение – рассматриваемые признаки объекта в принципе допускают желательное *разбиение* пула (совокупности) объектов на кластеры. В начале лекции мы уже упоминали о сравнимости шкал, это и есть второе предположение – правильность выбора масштаба или единиц измерения признаков.

Выбор масштаба в кластерном анализе имеет большое *значение*. Рассмотрим пример. Представим себе, что данные признака *x* в наборе данных A на два порядка больше данных признака *y*: значения переменной *x* находятся в диапазоне от 100 до 700, а значения переменной *y* – в диапазоне от 0 до 1.

Тогда, при расчете величины расстояния между точками, отражающими положение объектов в пространстве их свойств, *переменная*, имеющая большие значения, т.е. *переменная x*, будет практически полностью доминировать над переменной с малыми значениями, т.е. переменной *y*. Таким образом, из-за неоднородности единиц измерения признаков становится невозможно корректно рассчитать расстояния между точками.

Эта проблема решается при помощи предварительной *стандартизации* переменных. ***Стандартизация*** (*standardization*) или нормирование (*normalization*), приводит значения всех преобразованных переменных к единому диапазону значений путем выражения через *отношение* этих значений к некой величине, отражающей определенные свойства конкретного признака. Существуют различные способы нормирования исходных данных.

Два наиболее распространенных способа:

- деление исходных данных на *среднеквадратичное отклонение* соответствующих переменных;
- вычисление Z-вклада или стандартизованного вклада.

Наряду со *стандартизацией* переменных, существует вариант придания каждой из них определенного коэффициента важности, или веса, который бы

отражал *значимость* соответствующей переменной. В качестве весов могут выступать *экспертные оценки*, полученные в ходе опроса экспертов – *специалистов предметной области*. Полученные произведения нормированных переменных на соответствующие веса позволяют получать расстояния между точками в многомерном пространстве с учетом неодинакового веса переменных.

В ходе экспериментов возможно сравнение результатов, полученных с учетом экспертных оценок и без них, и выбор лучшего из них.

Методы кластерного анализа

Методы кластерного анализа можно разделить на две группы:

- иерархические;
- неиерархические.

Каждая из групп включает множество подходов и алгоритмов.

Используя различные методы кластерного анализа, *аналитик* может получить различные решения для одних и тех же данных. Это считается нормальным явлением.

Рассмотрим иерархические и неиерархические методы подробно.

Иерархические методы кластерного анализа

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие.

Иерархические агломеративные методы (Agglomerative Nesting, AGNES)

Эта группа методов характеризуется последовательным объединением исходных элементов и соответствующим уменьшением числа кластеров.

В начале работы алгоритма все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

Иерархические дивизимные (делимые) методы

(DIvisive ANAlysis, DIANA)

Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Принцип работы описанных выше групп методов в виде *дендрограммы* показан на рис. 13.3.

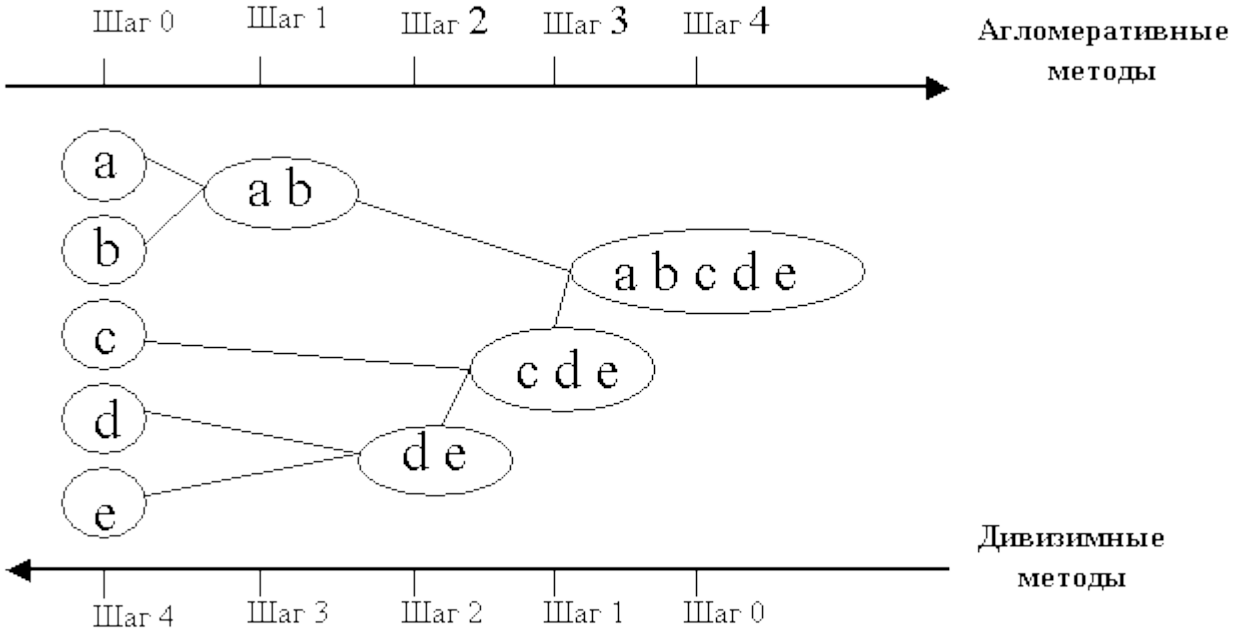

Рис. 13.3. Дендрограмма агломеративных и дивизимных методов

Программная реализация алгоритмов кластерного анализа широко представлена в различных инструментах *Data Mining*, которые позволяют решать задачи достаточно большой размерности. Например, агломеративные методы реализованы в пакете SPSS, дивизимные методы – в пакете Statgraf.

Иерархические методы кластеризации различаются правилами построения кластеров. В качестве правил выступают критерии, которые используются при решении вопроса о "схожести" объектов при их объединении в группу (агломеративные методы) либо разделения на группы (дивизимные методы).

Иерархические методы кластерного анализа используются при небольших объемах наборов данных.

Преимуществом иерархических методов кластеризации является их наглядность.

Иерархические алгоритмы связаны с построением *дендрограмм* (от греческого dendron – "дерево"), которые являются результатом *иерархического кластерного анализа*. ***Дендрограмма*** описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.

Дендрограмма (*dendrogram*) – древовидная диаграмма, содержащая n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров.

Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры.

Дендрограмма представляет собой вложенную группировку объектов, которая изменяется на различных уровнях иерархии.

Существует много способов построения *дендрограмм*. В *дендрограмме* объекты могут располагаться вертикально или горизонтально. Пример вертикальной *дендрограммы* приведен на рис. 13.4.

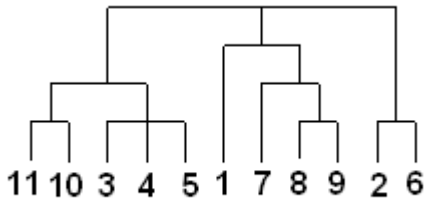

Рис. 13.4. Пример дендрограммы

Числа 11, 10, 3 и т.д. соответствуют номерам объектов или наблюдений исходной выборки. Мы видим, что на первом шаге каждое наблюдение представляет один кластер (вертикальная линия), на втором шаге наблюдаем объединение таких наблюдений: 11 и 10; 3, 4 и 5; 8 и 9; 2 и 6. На втором шаге продолжается объединение в кластеры: наблюдения 11, 10, 3, 4, 5 и 7, 8, 9.

Данный процесс продолжается до тех пор, пока все наблюдения не объединятся в один кластер.

Меры сходства

Для вычисления расстояния между объектами используются различные меры сходства (меры подобия), называемые также метриками или функциями расстояний. В начале лекции мы рассмотрели *евклидово расстояние*, это наиболее популярная *мера* сходства.

Квадрат *евклидова расстояния*.

Для придания больших весов более отдаленным друг от друга объектам можем воспользоваться квадратом *евклидова расстояния* путем возведения в квадрат стандартного *евклидова расстояния*.

Манхэттенское расстояние (*расстояние* городских кварталов), также называемое "хэмминговым" или "сити-блок" расстоянием.

Это *расстояние* рассчитывается как среднее разностей по координатам. В большинстве случаев эта *мера* расстояния приводит к результатам, подобным расчетам расстояния евклида. Однако, для этой меры влияние отдельных выбросов меньше, чем при использовании *евклидова расстояния*, поскольку здесь *координаты* не возводятся в квадрат.

Расстояние Чебышева. Это *расстояние* стоит использовать, когда необходимо определить два объекта как "различные", если они отличаются по какому-то одному измерению.

Процент несогласия. Это *расстояние* вычисляется, если данные являются категориальными.

Методы объединения или связи

Когда каждый *объект* представляет собой отдельный *кластер*, расстояния между этими объектами определяются выбранной мерой. Возникает следующий вопрос – как определить расстояния между кластерами? Существуют различные правила, называемые методами объединения или связи для двух кластеров.

Метод ближнего соседа или **одиночная связь**. Здесь *расстояние* между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. Этот метод позволяет выделять кластеры сколь угодно сложной формы при условии, что различные части таких кластеров соединены цепочками близких друг к другу элементов. В результате работы этого метода кластеры представляются длинными "цепочками" или "волокнистыми" кластерами, "сцепленными вместе" только отдельными элементами, которые случайно оказались ближе остальных друг к другу.

Метод наиболее удаленных соседей или полная *связь*. Здесь расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями"). Метод хорошо использовать, когда объекты действительно происходят из различных "рощ". Если же кластеры имеют в некотором роде удлиненную форму или их естественный тип является "цепочечным", то этот метод не следует использовать.

Метод Варда (Ward's *method*). В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до *центров кластеров*, получаемый в результате их объединения (Ward, 1963). В отличие от других методов кластерного анализа для оценки расстояний между кластерами, здесь используются методы *дисперсионного анализа*. На каждом шаге алгоритма объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов. Этот метод направлен на *объединение* близко расположенных кластеров и "стремится" создавать кластеры малого размера.

Метод невзвешенного попарного среднего (метод невзвешенного попарного арифметического среднего – unweighted pair-group *method* using *arithmetic* averages, UPGMA (Sneath, Sokal, 1973)).

В качестве расстояния между двумя кластерами берется среднее *расстояние* между всеми парами объектов в них. Этот метод следует

использовать, если объекты действительно происходят из различных "рощ", в случаях присутствия кластеров "цепочного" типа, при предположении неравных размеров кластеров.

Метод взвешенного попарного среднего (метод взвешенного попарного арифметического среднего – weighted pair-group *method* using *arithmetic* averages, WPGM A (Sneath, Sokal, 1973)). Этот метод похож на метод невзвешенного попарного среднего, разница состоит лишь в том, что здесь в качестве весового коэффициента используется ***размер кластера*** (число объектов, содержащихся в кластере).

Этот метод рекомендуется использовать именно при наличии предположения о кластерах разных размеров.

Невзвешенный центроидный метод (метод невзвешенного попарного центроидного усреднения – unweighted pair-group *method* using the *centroid average* (Sneath *and* Sokal, 1973)).

В качестве расстояния между двумя кластерами в этом методе берется *расстояние* между их центрами тяжести.

Взвешенный центроидный метод (метод взвешенного попарного центроидного усреднения – weighted pair-group *method* using the *centroid average*, WPGMC (Sneath, Sokal 1973)). Этот метод похож на предыдущий, разница состоит в том, что для учета разницы между *размерами кластеров* (числе объектов в них), используются веса. Этот метод предпочтительно использовать в случаях, если имеются предположения относительно существенных отличий в *размерах кластеров*.

Методы кластерного анализа. Итеративные методы

При большом количестве наблюдений иерархические методы кластерного анализа не пригодны. В таких случаях используют неиерархические методы, основанные на разделении, которые представляют собой ***итеративные методы*** дробления исходной совокупности. В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено *правило остановки*.

Такая неиерархическая *кластеризация* состоит в разделении набора данных на определенное количество отдельных кластеров. Существует два подхода. Первый заключается в определении границ кластеров как наиболее плотных участков в многомерном пространстве исходных данных, т.е. *определение* кластера там, где имеется большое "сгущение точек". Второй подход заключается в минимизации меры различия объектов

Алгоритм k-средних (k-means)

Наиболее распространен среди неиерархических методов *алгоритм k-средних*, также называемый **быстрым кластерным анализом**. Полное описание алгоритма можно найти в работе Хартигана и Вонга (Hartigan *and* Wong, 1978). В отличие от иерархических методов, которые не требуют предварительных предположений относительно числа кластеров, для возможности использования этого метода необходимо иметь гипотезу о наиболее вероятном количестве кластеров.

Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает *алгоритм k-средних*, – наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга.

Описание алгоритма

1. Первоначальное распределение объектов по кластерам.

Выбирается число *k*, и на первом шаге эти точки считаются "центрами" кластеров. Каждому кластеру соответствует один центр.

Выбор начальных центроидов может осуществляться следующим образом:

- выбор *k*-наблюдений для максимизации начального расстояния;

- случайный выбор *k*-наблюдений;
- выбор первых *k*-наблюдений.

В результате каждый объект назначен определенному кластеру.

2. Итеративный процесс.

Вычисляются *центры кластеров*, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются.

Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий:

- кластерные центры стабилизировались, т.е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации;
- число итераций равно максимальному числу итераций.

На рис. 14.1 приведен пример работы *алгоритма k-средних* для *k*, равного двум.

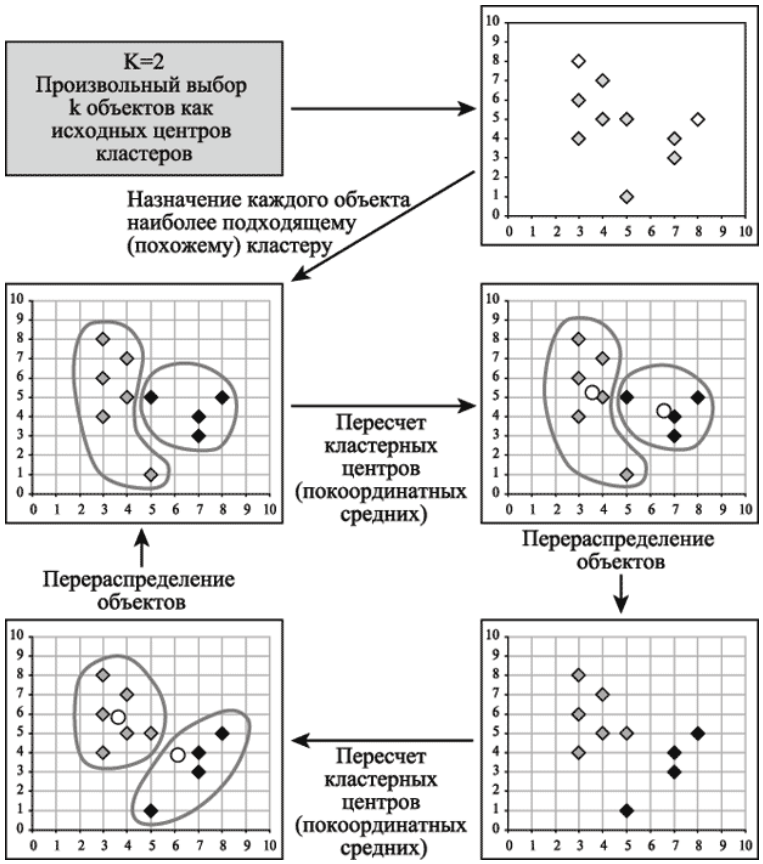

Рис. 14.1. Пример работы алгоритма *k*-средних (*k*=2)

Выбор числа кластеров является сложным вопросом. Если нет предположений относительно этого числа, рекомендуют создать 2 кластера, затем 3, 4, 5 и т.д., сравнивая полученные результаты.

Проверка качества кластеризации

После получений результатов кластерного анализа методом *k*-средних следует проверить правильность кластеризации (т.е. оценить, насколько кластеры отличаются друг от друга). Для этого рассчитываются средние значения для каждого кластера. При хорошей кластеризации должны быть получены сильно отличающиеся средние для всех измерений или хотя бы большей их части.

Достоинства *алгоритма k-средних*:

- простота использования;
- быстрота использования;
- понятность и прозрачность алгоритма.

Недостатки *алгоритма k-средних*:

- алгоритм слишком чувствителен к выбросам, которые могут искажать среднее. Возможным решением этой проблемы является использование модификации алгоритма – алгоритм *k*-медианы;
- алгоритм может медленно работать на больших базах данных. Возможным решением данной проблемы является использование выборки данных.

Алгоритм PAM (partitioning around Medoids)

PAM является модификацией алгоритма *k*-средних, алгоритмом *k*-медианы (*k*-medoids).

Алгоритм менее чувствителен к шумам и выбросам данных, чем алгоритм *k*-means, поскольку медиана меньше подвержена влияниям выбросов.

PAM эффективен для небольших баз данных, но его не следует использовать для больших наборов данных.

Предварительное сокращение размерности

Рассмотрим пример. Есть база данных клиентов фирмы, которых следует разбить на однородные группы. Каждый клиент описывается при помощи 25 переменных. Использование такого большого числа переменных приводит к выделению кластеров нечеткой структуры. В результате аналитику достаточно сложно интерпретировать полученные кластеры.

Более понятные и прозрачные результаты кластеризации могут быть получены, если вместо множества исходных переменных использовать некие обобщенные переменные или критерии, содержащие в сжатом виде информацию о связях между переменными. Т.е. возникает задача понижения размерности данных. Она может решаться при помощи различных методов; один из наиболее распространенных – факторный анализ. Остановимся на нем более подробно.

Факторный анализ

Факторный анализ – это метод, применяемый для изучения взаимосвязей между значениями переменных.

Вообще, факторный анализ преследует две цели:

– сокращение числа переменных;

– классификацию переменных – определение структуры взаимосвязей между переменными.

Соответственно, факторный анализ может использоваться для решения задач сокращения размерности данных или для решения задач классификации.

Критерии или главные факторы, выделенные в результате факторного анализа, содержат в сжатом виде информацию о существующих связях между переменными. Эта информация позволяет получить лучшие результаты кластеризации и лучше объяснить семантику кластеров. Самим факторам может быть сообщен определенный смысл.

При помощи факторного анализа большое число переменных сводится к меньшему числу независимых влияющих величин, которые называются факторами.

Фактор в "сжатом" виде содержит информацию о нескольких переменных. В один фактор объединяются переменные, которые сильно коррелируют между собой. В результате факторного анализа отыскиваются такие комплексные факторы, которые как можно более полно объясняют связи между рассматриваемыми переменными.

На первом шаге факторного анализа осуществляется стандартизация значений переменных, необходимость которой была рассмотрена в предыдущей лекции.

Факторный анализ опирается на гипотезу о том, что анализируемые переменные являются косвенными проявлениями сравнительно небольшого числа неких скрытых факторов.

Факторный анализ – это совокупность методов, ориентированных на выявление и анализ скрытых зависимостей между наблюдаемыми переменными. Скрытые зависимости также называют латентными.

Один из методов факторного анализа – метод главных компонент - основан на предположении о независимости факторов друг от друга.

Заключение. Таким образом, *алгоритм k-средних* делит совокупность исходных данных на заданное количество кластеров. Для возможности визуализации полученных результатов следует воспользоваться одним из графиков, например, диаграммой рассеивания. Однако традиционная *визуализация* возможна для ограниченного количества измерений, ибо, как известно, человек может воспринимать только трехмерное *пространство*. Поэтому, если мы анализируем более трех переменных, следует использовать специальные многомерные методы представления информации, о них будет рассказано в одной из последующих лекций курса.

Итеративные методы кластеризации различаются выбором следующих параметров:

- начальной точки;
- правилом формирования новых кластеров;
- правилом остановки.

Выбор метода кластеризации зависит от количества данных и от того, есть ли необходимость работать одновременно с несколькими типами данных.

В пакете SPSS, например, при необходимости работы как с количественными (например, доход), так и с категориальными (например, семейное положение) переменными, а также если объем данных достаточно велик, используется метод Двухэтапного кластерного анализа, который представляет собой масштабируемую процедуру кластерного анализа, позволяющую работать с данными различных типов.

Для этого на первом этапе работы записи предварительно кластеризуются в большое количество суб-кластеров. На втором этапе полученные суб-кластеры группируются в необходимое количество. Если это количество неизвестно, процедура сама автоматически определяет его. При помощи этой процедуры банковский работник может, например, выделять группы людей, одновременно используя такие показатели как возраст, пол и уровень дохода. Полученные результаты позволяют определить клиентов, входящих в группы риска невозврата кредита.

Процесс кластерного анализа. Рекомендуемые этапы

В общем случае все этапы кластерного анализа взаимосвязаны, и решения, принятые на одном из них, определяют действия на последующих этапах.

Аналитику следует решить, использовать ли все наблюдения либо же исключить некоторые данные или выборки из набора данных.

Выбор метрики и метода стандартизации исходных данных.

Определение количества кластеров (для итеративного кластерного анализа).

Определение метода кластеризации (правила объединения или связи).

По мнению многих специалистов, выбор метода кластеризации является решающим при определении формы и специфики кластеров.

Анализ результатов кластеризации. Этот этап подразумевает решение таких вопросов: не является ли полученное разбиение на кластеры случайным;

является ли разбиение надежным и стабильным на подвыборках данных; существует ли взаимосвязь между результатами кластеризации и переменными, которые не участвовали в процессе кластеризации; можно ли интерпретировать полученные результаты кластеризации.

Проверка результатов кластеризации. Результаты кластеризации также должны быть проверены формальными и неформальными методами. Формальные методы зависят от того метода, который использовался для кластеризации. Неформальные включают следующие процедуры проверки качества кластеризации:

- анализ результатов кластеризации, полученных на определенных выборках набора данных;
- *кросс-проверка*;
- проведение кластеризации при изменении порядка наблюдений в наборе данных;
- проведение кластеризации при удалении некоторых наблюдений;
- проведение кластеризации на небольших выборках.

Один из вариантов проверки качества кластеризации – использование нескольких методов и сравнение полученных результатов. Отсутствие подобия не будет означать некорректность результатов, но присутствие похожих групп считается признаком качественной кластеризации.

Сложности и проблемы, которые могут возникнуть при применении кластерного анализа

Как и любые другие методы, методы кластерного анализа имеют определенные слабые стороны, т.е. некоторые сложности, проблемы и ограничения.

При проведении кластерного анализа следует учитывать, что результаты кластеризации зависят от критериев разбиения совокупности исходных данных. При понижении размерности данных могут возникнуть определенные искажения, за счет обобщений могут потеряться некоторые индивидуальные характеристики объектов.

Существует ряд сложностей, которые следует продумать перед проведением кластеризации.

- Сложность выбора характеристик, на основе которых проводится кластеризация. Необдуманный выбор приводит к неадекватному разбиению на кластеры и, как следствие, – к неверному решению задачи.
- Сложность выбора метода кластеризации. Этот выбор требует неплохого знания методов и предпосылок их использования. Чтобы проверить эффективность конкретного метода в определенной предметной области, целесообразно применить следующую процедуру: рассматривают несколько априори различных между собой групп и перемешивают их представителей между собой случайным образом. Далее проводится кластеризация для восстановления исходного разбиения на кластеры. Доля совпадений объектов в выявленных и исходных группах является показателем эффективности работы метода.
- Проблема выбора числа кластеров. Если нет никаких сведений относительно возможного числа кластеров, необходимо провести ряд экспериментов и, в результате перебора различного числа кластеров, выбрать оптимальное их число.
- Проблема интерпретации результатов кластеризации. Форма кластеров в большинстве случаев определяется выбором метода объединения. Однако следует учитывать, что конкретные методы стремятся создавать кластеры определенных форм, даже если в исследуемом наборе данных кластеров на самом деле нет.

Сравнительный анализ иерархических и неиерархических методов кластеризации

Перед проведением кластеризации у аналитика может возникнуть вопрос, какой группе методов кластерного анализа отдать предпочтение. Выбирая между иерархическими и неиерархическими методами, необходимо учитывать следующие их особенности.

Неиерархические методы выявляют более высокую устойчивость по отношению к шумам и выбросам, некорректному выбору метрики, включению незначимых переменных в набор, участвующий в кластеризации. Ценой, которую приходится платить за эти достоинства метода, является слово "априори". Аналитик должен заранее определить количество кластеров, количество итераций или правило остановки, а также некоторые другие параметры кластеризации. Это особенно сложно начинающим специалистам.

Если нет предположений относительно числа кластеров, рекомендуют использовать иерархические алгоритмы. Однако если объем выборки не позволяет это сделать, возможный путь - проведение ряда экспериментов с различным количеством кластеров, например, начать разбиение совокупности данных с двух групп и, постепенно увеличивая их количество, сравнивать результаты. За счет такого "варьирования" результатов достигается достаточно большая гибкость кластеризации.

Иерархические методы, в отличие от неиерархических, отказываются от определения числа кластеров, а строят полное дерево вложенных кластеров.

Сложности иерархических методов кластеризации: ограничение объема набора данных; выбор меры близости; негибкость полученных классификаций.

Преимущество этой группы методов в сравнении с неиерархическими методами – их наглядность и возможность получить детальное представление о структуре данных.

При использовании иерархических методов существует возможность достаточно легко идентифицировать выбросы в наборе данных и, в результате, повысить качество данных. Эта процедура лежит в основе двухшагового алгоритма кластеризации. Такой набор данных в дальнейшем может быть использован для проведения неиерархической кластеризации.

Существует еще один аспект, о котором уже упоминалось в этой лекции. Это вопрос кластеризации всей совокупности данных или же ее выборки. Названный аспект существенен для обеих рассматриваемых групп методов, однако он более критичен для иерархических методов. Иерархические методы

не могут работать с большими наборами данных, а использование некоторой выборки, т.е. части данных, могло бы позволить применять эти методы.

Результаты кластеризации могут не иметь достаточного статистического обоснования. С другой стороны, при решении задач кластеризации допустима нестатистическая интерпретация полученных результатов, а также достаточно большое разнообразие вариантов понятия кластера. Такая нестатистическая интерпретация дает возможность аналитику получить удовлетворяющие его результаты кластеризации, что при использовании других методов часто бывает затруднительным.

Новые алгоритмы и некоторые модификации алгоритмов кластерного анализа

Методы, которые мы рассмотрели в этой и предыдущей лекциях, являются "классикой" кластерного анализа. До последнего времени основным критерием, по которому оценивался *алгоритм* кластеризации, было качество кластеризации: полагалось, чтобы весь набор данных умещался в оперативной памяти.

Однако сейчас, в связи с появлением сверхбольших баз данных, появились новые требования, которым должен удовлетворять *алгоритм* кластеризации. Основное из них, как уже упоминалось в предыдущих лекциях, – это *масштабируемость* алгоритма.

Отметим также другие свойства, которым должен удовлетворять *алгоритм* кластеризации: независимость результатов от порядка входных данных; независимость параметров алгоритма от входных данных.

В последнее время ведутся активные разработки новых алгоритмов кластеризации, способных обрабатывать сверхбольшие *базы данных*. В них основное внимание уделяется масштабируемости. К таким алгоритмам относятся обобщенное *представление* кластеров (summarized *cluster* representation), а также *выборка* и использование структур данных, поддерживаемых нижележащими *СУБД* [33].

Разработаны алгоритмы, в которых методы иерархической кластеризации интегрированы с другими методами. К таким алгоритмам относятся: *BIRCH, CURE, CHAMELEON, ROCK*.

Алгоритм BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

Алгоритм предложен Тьян Зангом и его коллегами [55].

Благодаря обобщенным представлениям кластеров, скорость кластеризации увеличивается, алгоритм при этом обладает большим масштабированием.

В этом алгоритме реализован двухэтапный процесс кластеризации.

В ходе первого этапа формируется предварительный набор кластеров. На втором этапе к выявленным кластерам применяются другие алгоритмы кластеризации - пригодные для работы в оперативной памяти.

В [33] приведена следующая аналогия, описывающая этот алгоритм. Если каждый элемент данных представить себе как бусину, лежащую на поверхности стола, то кластеры бусин можно "заменить" теннисными шариками и перейти к более детальному изучению кластеров теннисных шариков. Число бусин может оказаться достаточно велико, однако диаметр теннисных шариков можно подобрать таким образом, чтобы на втором этапе можно было, применив традиционные алгоритмы кластеризации, определить действительную сложную форму кластеров.

Алгоритм WaveCluster

WaveCluster представляет собой алгоритм кластеризации на основе волновых преобразований [56]. В начале работы алгоритма данные обобщаются путем наложения на пространство данных многомерной решетки. На дальнейших шагах алгоритма анализируются не отдельные точки, а обобщенные характеристики точек, попавших в одну ячейку решетки. В результате такого обобщения необходимая информация умещается в оперативной памяти. На последующих шагах для определения кластеров алгоритм применяет волновое преобразование к обобщенным данным.

Главные особенности WaveCluster:

1. сложность реализации;
2. алгоритм может обнаруживать кластеры произвольных форм;
3. алгоритм не чувствителен к шумам;
4. алгоритм применим только к данным низкой размерности.

Алгоритм CLARA (Clustering LARge Applications)

Алгоритм CLARA был разработан Kaufmann и Rousseeuw в 1990 году для кластеризации данных в больших базах данных. Данный алгоритм строится в статистических аналитических пакетах, например, таких как S+.

Изложим кратко суть алгоритма. Алгоритм CLARA извлекает множество образцов из базы данных. Кластеризация применяется к каждому из образцов, на выходе алгоритма предлагается лучшая кластеризация.

Для больших баз данных этот алгоритм эффективнее, чем алгоритм *PAM*. Эффективность алгоритма зависит от выбранного в качестве образца набора данных. Хорошая кластеризация на выбранном наборе может не дать хорошую кластеризацию на всем множестве данных.

Алгоритмы Clarans, CURE, DBScan

Алгоритм Clarans (Clustering Large Applications based upon RANdomized Search) [14] формулирует задачу кластеризации как случайный поиск в графе. В результате работы этого алгоритма совокупность узлов графа представляет собой разбиение множества данных на число кластеров, определенное пользователем. "Качество" полученных кластеров определяется при помощи критериальной функции. Алгоритм Clarans сортирует все возможные разбиения множества данных в поисках приемлемого решения. Поиск решения останавливается в том узле, где достигается минимум среди предопределенного числа локальных минимумов.

Среди новых масштабируемых алгоритмов также можно отметить алгоритм ***CURE*** [57] - алгоритм иерархической кластеризации, и алгоритм DBScan [58], где понятие кластера формулируется с использованием концепции плотности (density).

Основным недостатком алгоритмов *BIRCH*, Clarans, *CURE*, DBScan является то обстоятельство, что они требуют задания некоторых порогов плотности точек, а это не всегда приемлемо. Эти ограничения обусловлены тем, что описанные алгоритмы ориентированы на сверхбольшие базы данных и не могут пользоваться большими вычислительными ресурсами [59].

Над масштабируемыми методами сейчас активно работают многие исследователи, основная задача которых - преодолеть недостатки алгоритмов, существующих на сегодняшний день.