Задача кластерного анализа.

Задача кластерного анализа заключается в том, чтобы на основании данных, содержащихся во множестве X, разбить множество объектов G на m (m – целое) кластеров (подмножеств)

$Q_1, \ldots, Q_m$, так, чтобы каждый объект Q принадлежал одному G_j и только одному подмножеству разбиения. А объекты, принадлежащие одному и тому же кластеру, были сходными, в то время как объекты, принадлежащие разным кластерам, были разнородными.

Решением задачи кластерного анализа являются разбиения, удовлетворяющие некоторому критерию оптимальности. Этот критерий может представлять собой некоторый функционал, выражающий уровни желательности различных разбиений и группировок, который называют целевой функцией. Например, в качестве целевой функции может быть взята внутригрупповая сумма квадратов отклонения:

$$W = \sigma_n = \sum_{j=1}^{n} (x_j - \bar{x})^2 = \sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2,$$

где x_j - представляет собой измерения j-го объекта.

Кластерный анализ. Основные понятия.

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*.

Работа кластерного анализа опирается на два предположения. Первое предположение - рассматриваемые признаки объекта в принципе допускают желательное разбиение совокупности объектов на кластеры.

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.

Кластерный анализ. Нормализация данных

Эта проблема решается при помощи предварительной *стандартизации* переменных. *Стандартизация* или нормирование 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:

$$z = (x - \bar{x}) / \sigma, \quad z = x / \bar{x}, \quad z = x / x_{\max}, \quad z = (x - \bar{x}) / (x_{\max} - x_{\min}),$$

где $\bar{x}$, σ- соответственно среднее и среднеквадратическое отклонение x; $x_{\max}$,-$x_{\min}$наибольшее и наименьшее значение x.

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.

Кластерный анализ. Измерение близости объектов

В кластерном анализе для количественной оценки сходства вводится понятие *метрики.* Сходство или различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними. Если каждый объект описывается k признаками, то он может быть представлен как точка в k-мерном пространстве, и сходство с другими объектами будет определяться как соответствующее расстояние.

Расстоянием (метрикой) между объектами в пространстве параметров называется такая величина d_{ab}, которая удовлетворяет аксиомам:

$$A1.\ d_{ab} > 0,\ d_{ab} = 0,$$
$$A2.\ d_{ab} = d_{ba},$$
$$A3.\ d_{ab} + d_{bc} \geq d_{ac}.$$

Мерой близости (сходства) обычно называется величина μ_{ab}, имеющая предел и возрастающая с возрастанием близости объектов.

$$B1.\ \mu_{ab} \text{ непрерывна},$$
$$B2.\ \mu_{ab} = \mu_{ba},$$
$$B3.\ 0 \leq \mu_{ab} \leq 1.$$

Существует возможность простого перехода от расстояний к мерам близости: $\mu = \dfrac{1}{1+d}$.

Кластерный анализ. Характеристики близости объектов

Объединение или метод древовидной кластеризации используется при формировании кластеров несходства или расстояния между объектами. Эти расстояния могут определяться в одномерном или многомерном пространстве. Например, если вы должны кластеризовать типы еды в кафе, то можете принять во внимание количество содержащихся в ней калорий, цену, субъективную оценку вкуса и т.д. Наиболее прямой путь вычисления расстояний между объектами в многомерном пространстве состоит в вычислении евклидовых расстояний. Если вы имеете двух- или трёхмерное пространство, то эта мера является реальным геометрическим расстоянием между объектами в пространстве. Однако алгоритм объединения не заботится о том, являются ли предоставленные для этого расстояния настоящими или некоторыми другими производными мерами расстояния, что более значимо для исследователя; и задачей исследователей является подобрать правильный метод для специфических применений.

Кластерный анализ. Характеристики близости объектов

Таблица 1. способы определения близости между объектами

Показатели	Формулы
Линейное расстояние	$d_{lij} = \sum_{l=1}^{m} \lvert x_i^l - x_j^l \rvert$
Евклидово расстояние	$d_{Eij} = \left(\sum_{l=1}^{m} (x_i^l - x_j^l)^2 \right)^{\frac{1}{2}}$
Квадрат евклидова расстояния	$d_{Eij}^2 = \sum_{l=1}^{m} (x_i^l - x_j^l)^2$
Обобщенное степенное расстояние Минковского	$d_{Pij} = \left(\sum_{l=1}^{m} (x_i^l - x_j^l)^P \right)^{\frac{1}{P}}$
Расстояние Чебышева	$d_{ij} = \max_{1 \le i,j \le l} \lvert x_i - x_j \rvert$
Расстояние городских кварталов (Манхэттенское расстояние)	$d_H(x_i, x_j) = \sum_{l=1}^{k} \lvert x_i^l - x_j^l \rvert$

Кластерный анализ. Характеристики близости объектов

Евклидово расстояние является самой популярной метрикой в кластерном анализе. Оно попросту является геометрическим расстоянием в многомерном пространстве. Геометрически оно лучше всего объединяет объекты в шарообразных скоплениях.

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

Обобщенное степенное расстояние представляет только математический интерес как универсальная метрика.

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

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.

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

Кластерный анализ. 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.

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.

Кластерный анализ. Иерархическая кластеризация

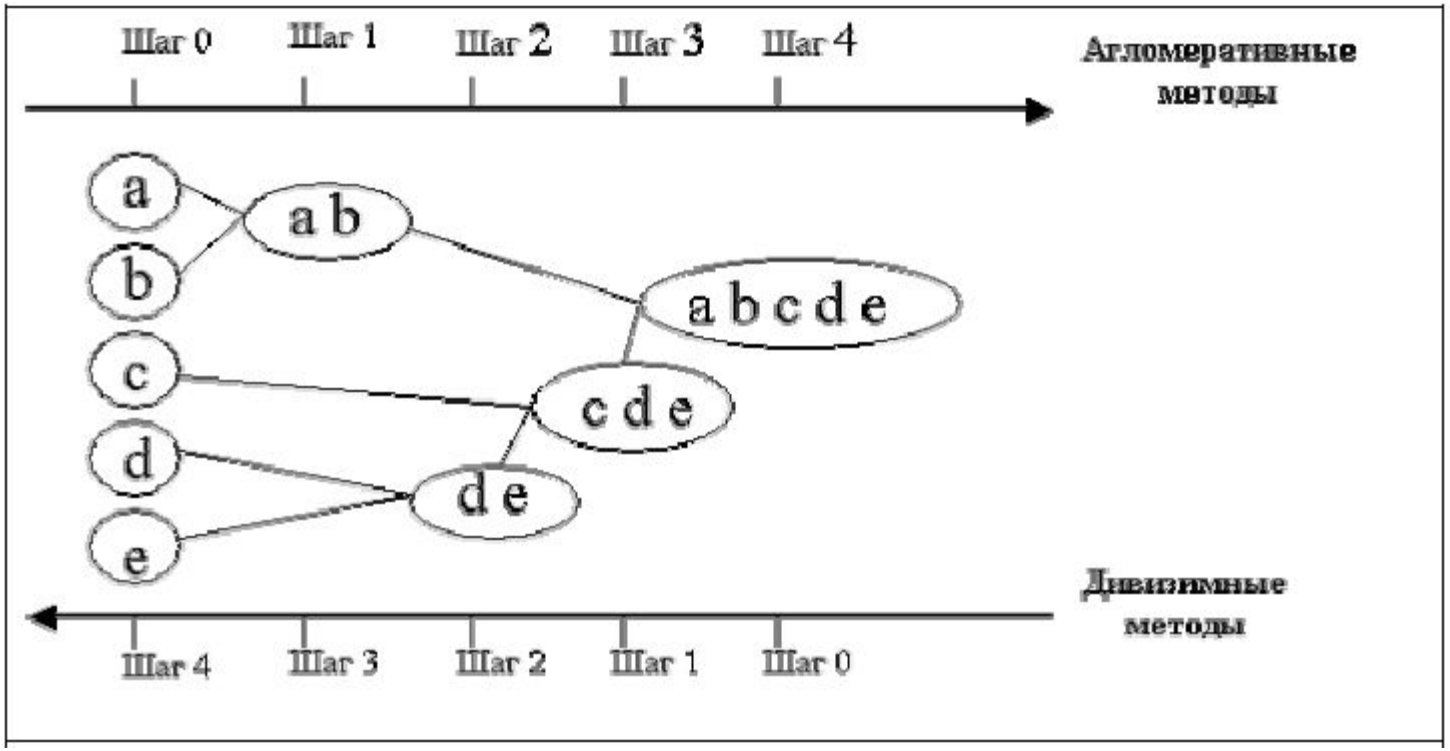

Рис.1. Дендрограмма агломеративных и дивизимных методов

Кластерный анализ. Иерархическая кластеризация

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*, 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.

Кластерный анализ. Иерархическая кластеризация

Существует много способов построения дендограмм. В дендограмме объекты могут располагаться вертикально или горизонтально.

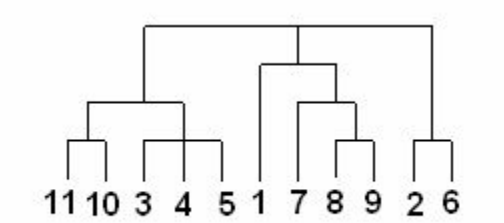

Рис. 2. Пример вертикальной дендрограммы

Числа 11, 10, 3 и т.д. соответствуют номерам объектов или наблюдений исходной выборки. 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.

Пусть

K_i - *i*-я группа (класс, кластер), состоящая из *n* объектов;

$\bar{x}_i$ - среднее арифметическое векторных наблюдений K_i группы, т.е. «центр тяжести» *i*-й группы;

$\rho(K_i, K_j) = \rho_{ij}$ – расстояние между группами K_i и K_j.

Кластерный анализ. Иерархическая кластеризация

Обобщенная алгомеративная процедура. На первом шаге каждый объект считается отдельным кластером. На следующем шаге объединяются два ближайших объекта, которые образуют новый класс, определяются расстояния от этого класса до всех остальных объектов, и размерность матрицы расстояний D сокращается на единицу. На p-ом шаге повторяется та же процедура на матрице $D(n- p)(n- p)$, пока все объекты необъединятся в один класс.

Если сразу несколько объектов (классов) имеют минимальное расстояние, то возможны две стратегии: выбрать одну случайную пару или объединить сразу все пары. Первый способ является классическим и реализован во всех процедурах (иногда его называют восходящей иерархической классификацией). Второй способ называют методом ближайших соседей (не путать с алг. *“Ближайшего соседа”*) и используют реже.

Кластерный анализ. Расстояния между кластерами

На первом шаге, когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой.

Далее необходимо правило объединения или связи для двух кластеров. Здесь имеются различные возможности: например, можно связать два кластера вместе, когда любые два объекта в двух кластерах ближе друг к другу, чем соответствующее расстояние связи. Т.е. используется правило ближайшего соседа для определения расстояния между кластерами; этот метод называется методом одиночной связи. Это правило строит волокнистые кластеры, т.е. кластеры сцепленные вместе только отдельными элементами, случайно оказавшимися ближе остальных друг к другу. Как альтернативу можно использовать соседей в кластерах, которые находятся дальше всех остальных пар объектов друг от друга. Этот метод называется метод полной связи. Существует также множество других методов объединения кластеров.

1. *Расстояние "Ближайшего соседа" (Одиночная связь).* Первый шаг совпадает с первым шагом алгоритма *Обобщенная алгомеративная процедура*. Расстояние равно расстоянию между ближайшими объектами классов.

$$\rho_{\min}(K_i, K_j) = \min_{x_i \in K_i, x_j \in K_j} \rho(x_i, x_j)$$

2. *Расстояние "Дальнего соседа" (Полная связь).* Расстояние равно расстоянию между самыми дальними объектами классов.

$$\rho_{\max}(K_i, K_j) = \max_{x_i \in K_i, x_j \in K_j} \rho(x_i, x_j)$$

Кластерный анализ. Расстояния между кластерами

3. *Невзвешенное попарное среднее.* В этом методе расстояние между двумя различными кластерами вычисляется как среднее расстояние между всеми парами объектов в них. Метод эффективен, когда объекты в действительности формируют различные рощи, однако он работает одинаково хорошо и в случаях протяженных (цепочного типа) кластеров.

4. *Взвешенное попарное среднее.* Метод идентичен методу невзвешенного попарного среднего, за исключением того, что при вычислениях размер соответствующих кластеров (т.е. число объектов, содержащихся в них) используется в качестве весового коэффициента. Поэтому предлагаемый метод должен быть использован, когда предполагаются неравные размеры кластеров.

5. *Невзвешенный центроидный метод*. В этом методе расстояние между двумя кластерами определяется как расстояние между их центрами тяжести.

6. *Взвешенный центроидный метод (медиана)*. Метод идентичен предыдущему, за исключением того, что при вычислениях используются веса для учёта разницы между размерами кластеров (т.е. числами объектов в них). Поэтому, если имеются (или подозреваются) значительные отличия в размерах кластеров, этот метод оказывается предпочтительнее предыдущего.

7. *Метод Варда.* В этом методе в качестве целевой функции применяют внутригрупповую сумму квадратов отклонений, которая есть ни что иное, как сумма квадратов расстояний между каждой точкой (объектом) и средней по кластеру, содержащему этот объект. На каждом шаге объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, т.е. внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров.

Кластерный анализ. Эталонные методы

Наряду с иерархическими методами классификации, существует многочисленная группа итеративных методов кластерного анализа (метод *k*-средних.). Сущность их заключается в том, что процесс классификации начинается с задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации и т.д.). В отличие от иерархических процедур метод *k*-средних не требует вычисления и хранения матрицы расстояний или сходств между объектами. Алгоритм этого метода предполагает использование только исходных значений переменных. Для начала процедуры классификации должны быть заданы *k* выбранных объектов, которые будут служить эталонами, т.е. центрами кластеров. Считается, что алгоритмы эталонного типа удобные и быстродействующие. В этом случае важную роль играет выбор начальных условий, которые влияют на длительность процесса классификации и на его результаты. Метод *k*-средних удобен для обработки больших статистических совокупностей.

Математическое описание алгоритма метода *k* - средних.

Пусть имеется *n* наблюдений, каждое из которых характеризуется *m* признаками $X_1, \ldots, X_m$. Эти наблюдения необходимо разбить на *k* кластеров. Для начала из *n* точек исследуемой совокупности отбираются случайным образом или задаются исследователем исходя из каких-либо априорных соображений *k* точек (объектов). Эти точки принимаются за эталоны. Каждому эталону присваивается порядковый номер, который одновременно является и номером кластера.

Кластерный анализ. Эталонные методы

На первом шаге из оставшихся (*n*-*k*) объектов извлекается точка с координатами и проверяется, к какому из эталонов (центров) она находится ближе всего. Для этого используется одна из метрик, например, евклидово расстояние. Проверяемый объект присоединяется к тому центру (эталону), которому соответствует минимальное из расстояний. Эталон заменяется новым, пересчитанным с учетом присоединенной точки, и вес его (количество объектов, входящих в данный кластер) увеличивается на единицу. Если встречаются два или более минимальных расстояния, то *i*-ый объект присоединяют к центру с наименьшим порядковым номером. На следующем шаге выбираем точку и для нее повторяются все процедуры. Таким образом, через (*n*-*k*) шагов все точки (объекты) совокупности окажутся отнесенными к одному из *k* кластеров, но на этом процесс разбиения не заканчивается. Для того чтобы добиться устойчивости разбиения по тому же правилу, все точки опять подсоединяются к полученным кластером, при этом веса продолжают накапливаться. Новое разбиение сравнивается с предыдущим. Если они совпадают, то работа алгоритма завершается. В противном случае цикл повторяется. Окончательное разбиение имеет центры тяжести, которые не совпадают с эталонами, их можно обозначить . При этом каждая точка ( ) будет относиться к такому кластеру (классу) *l* , для которого расстояние минимально. Возможны две модификации метода *k*-средних. Первая предполагает пересчет центра тяжести кластера после каждого изменения его состава, а вторая – лишь после того, как будет завершен просмотр всех данных. В обоих случаях итеративный алгоритм этого метода минимизирует дисперсию внутри каждого кластера, хотя в явном виде такой критерий оптимизации не используется.