**КЛАСТЕРНЫЙ АНАЛИЗ**

Кластерный анализ – это анализ, позволяющий получить разбиение большого объема данных на классы или группы согласно некоторому критерию или их совокупности. При этом объединяются в классы объекты близкие между собой в определенном смысле, причем предполагается отсутствие априорной информации о характеристиках объектов внутри классов.

Например, при сравнении стран, регионов, городов имеет смысл разбить всю совокупность изучаемых объектов на классы, близкие, например, по уровню жизни. При производстве товаров массового потребления имеет смысл разбить всех потенциальных покупателей на группы, близкие по потребительской способности и т.д.

В качестве основных задач, решаемых с помощью кластерного анализа, можно выделить:

1. классификацию объектов с учетом большого числа признаков, отражающих их сущность;
2. выявление скрытой структуры изучаемой совокупности объектов;
3. снижение размерности пространства изучаемых объектов.

**МЕРЫ БЛИЗОСТИ ОБЪЕКТОВ**

Пусть дана совокупность из  объектов, каждый из которых характеризуется  признаками. Обозначим:  - вектор значений признаков для i-ого объекта,  - значение j-го признака для i-го объекта. Определим меру близости (или меру однородности) объектов. В качестве меры близости будем использовать расстояние между объектами. Напомним, что расстоянием, или метрикой определенной на множестве  называется числовая функция , удовлетворяющая следующим условиям:

1) ;

2) ;

3) .

В теории кластерного анализа используются следующие метрики для измерения расстояния между отдельными объектами (точками в k-мерном пространстве):

1. Евклидово расстояние:. Используется, если признаки однородны по физическому смыслу и их значения одинаково “весомы”.
2. Квадрат евклидова расстояния:. Применяется для придания большего веса более отдаленным друг от друга объектам.
3. Взвешенное евклидово расстояние:,где – веса, пропорциональные важности признака в задаче классификации.
4. Хеммингово расстояние: . Данную метрику также часто называют расстоянием городских кварталов (city–block). Используется для уменьшения влияния отдельных больших разностей (выбросов).
5. Расстояние Махаланобиса: ,где  – матрица ковариаций признаков (или выборочная матрица ковариаций).
6. Расстояние Чебышева:.
7. Расстояние Минковского:, .
8. Степенное расстояние или обобщенное расстояние Минковского:,,. Применяется в случае, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются.  Параметр p ответственен за постепенное взвешивание разностей по отдельным координатам, параметр r ответственен за прогрессивное взвешивание больших расстояний между объектами.
9. Процент несогласия: , где . Используется для категориальных данных.

В качестве меры тесноты связи между объектами может выступать и величина коэффициента корреляции между объектами, рассматриваемыми как переменные с наблюдаемыми значениями, равными значениям соответствующих признаков:

.

Чем ближе по абсолютной величине значение коэффициента корреляции к единице, тем более близки объекты между собой (обычно в качестве характеристики тесноты связи используют величину ).

Оценка сходства между объектами сильно зависит от абсолютного значения признака и от степени его вариации в совокупности. Если абсолютные значения для различных признаков сильно различаются между собой, следует проводить нормировку исходных данных, преобразуя их, например, к стандартным величинам с выборочным средним равным нулю и выборочной дисперсией равной единице.

Выбор метрики осуществляется исследователем в зависимости от того, какой результат он хочет получить. Этот выбор неформализуем, так как зависит от многих факторов, в частности, от ожидаемого результата, от опыта исследователя, уровня его математической подготовки и т.д.

***,***

**Иерархические агломеративные методы кластерного анализа**

Иерархические (т.е. деревообразующие) агломеративные (т.е. объеъеденяющие) методы являются наиболее простыми и распространенными среди алгоритмов кластерного анализа. На первом шаге каждый из объектов  исходных данных рассматривается как отдельный кластер. По вычисленной матрице расстояний, где  – некоторая метрика, на каждом шаге объединяются наиболее близкие друг к другу кластеры. При этом, каждый раз пересчитывается матрица расстояний (заметим, что при этом размерность матрицы на каждом шаге уменьшается на единицу). Очевидно, что процесс завершится через  шаг, когда в результате все объекты будут объединены в один кластер. Процесс объединения может быть прерван на каком либо шаге, если изначально задать пороговое максимальное значение расстояния для объединения кластеров. Если расстояние между кластерами превосходит эту величину, то объединения кластеров не происходит.

Последовательность объединений можно представить в виде дендрограммы или дерева. На рисунке показано, что на первом шаге были объединены объекты, так как расстояние между ними 0,1. На втором шаге к ним был присоединен объект, отстоящий от кластера  на расстояние 0,4 и т.д. Очевидно, на последнем шаге объединяются все объекты в один кластер.

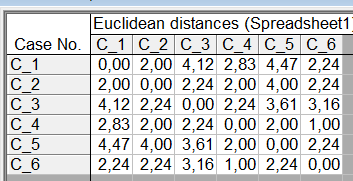
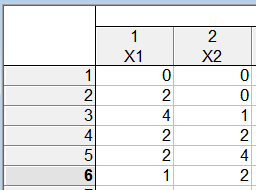


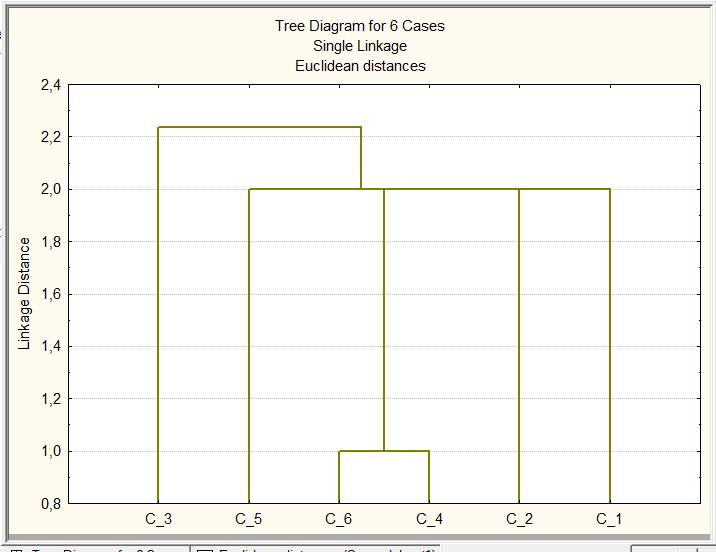
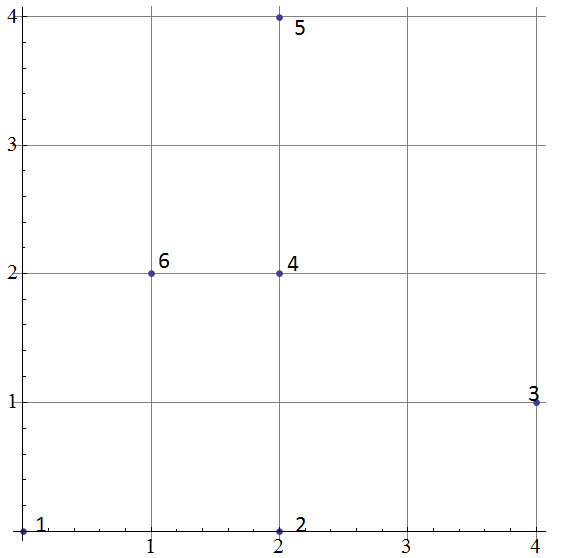
Иерархические методы являются наиболее простыми методами кластерного анализа и используются в первую очередь для анализа структуры множества объектов. В зависимости от того, как определяется расстояние между кластерами (как происходит объединение кластеров), выделяют методы одиночной и полной связи, методы взвешенных и невзвешенных попарных средних, взвешенный и невзвешенный центроидные методы, метод Уорда.

**Метод одиночной связи**

В методе одиночной связи расстояние между кластерами определяется по методу “ближнего соседа”, то есть за расстояние между кластерами принимается расстояние между самыми близкими элементами двух кластеров: .

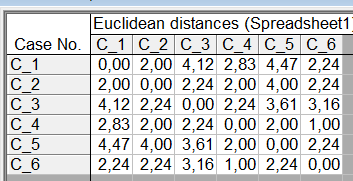
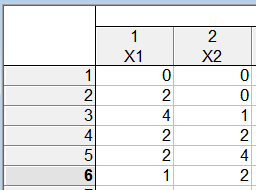
К достоинствам данного метода следует отнести то, что он не чувствителен к монотонному преобразованию матрицы расстояний между исходными объектами. К недостаткам следует отнести то, что зачастую данный метод приводит к появлению длинных цепочек кластеров.

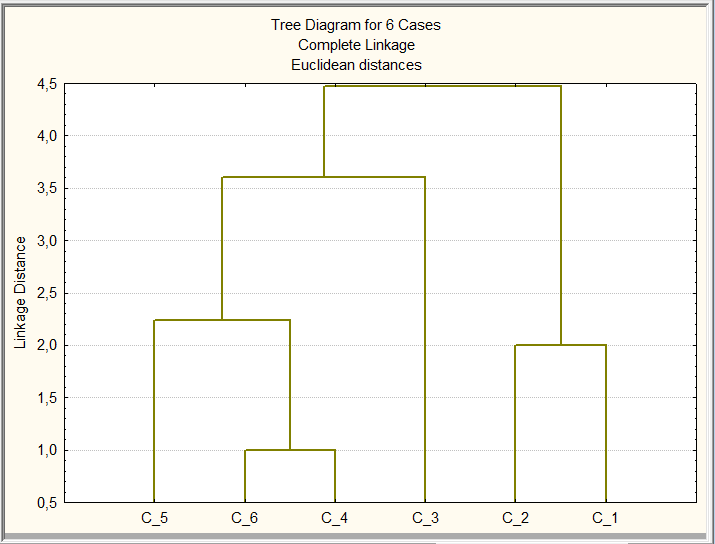
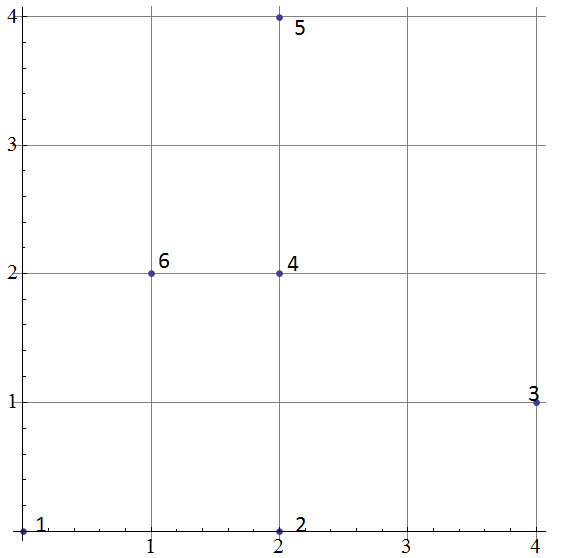




**Метод полной связи**

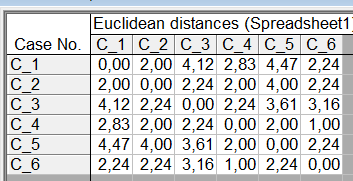
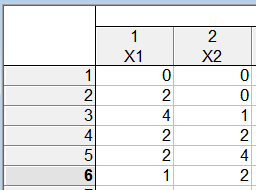
В методе полной связи расстояние между кластерами вычисляется по принципу «дальнего соседа», то есть за расстояние между кластерами принимается расстояние между самыми удаленными элементами двух кластеров: .

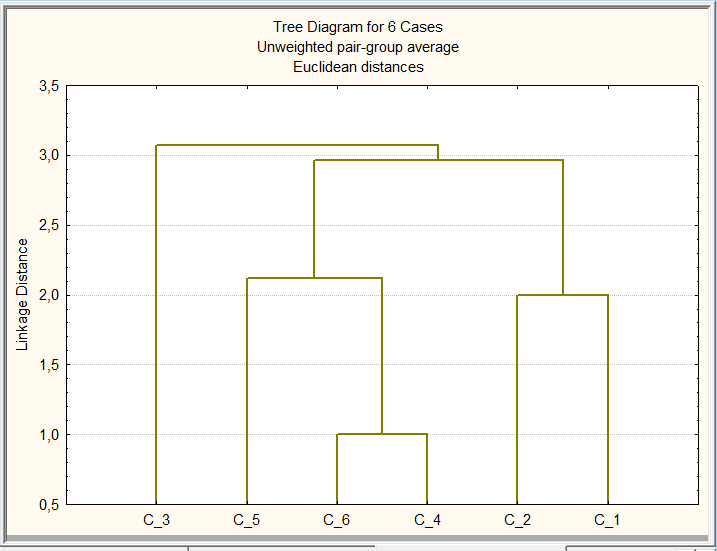
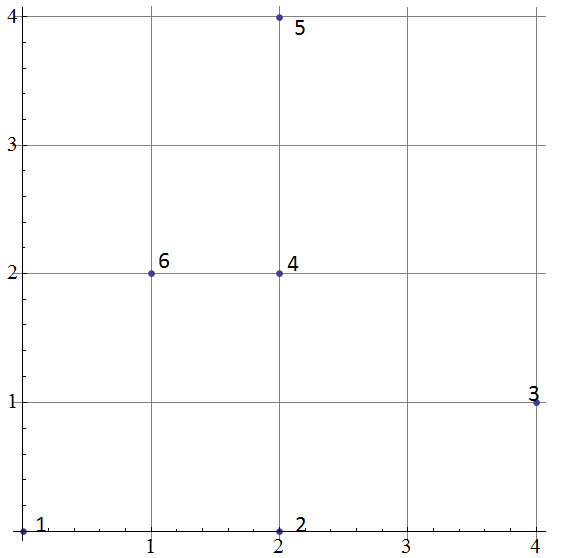




**Метод невзвешенных попарных средних (метод средней связи)**

В методе невзвешенных попарных средних за расстояние между кластерами берется среднее расстояние между всеми парами объектов двух кластеров: , где  - количество объектов в кластерах  соответственно.





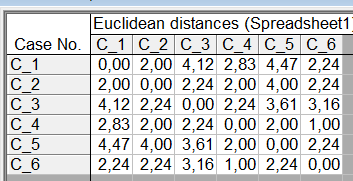
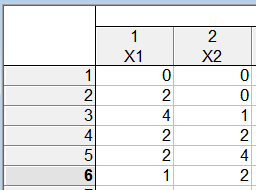
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | X1 | X2 | X3 | X5 |
| X4,X6 | 2,53 | 2,12 | 2,70 | 2,12 |

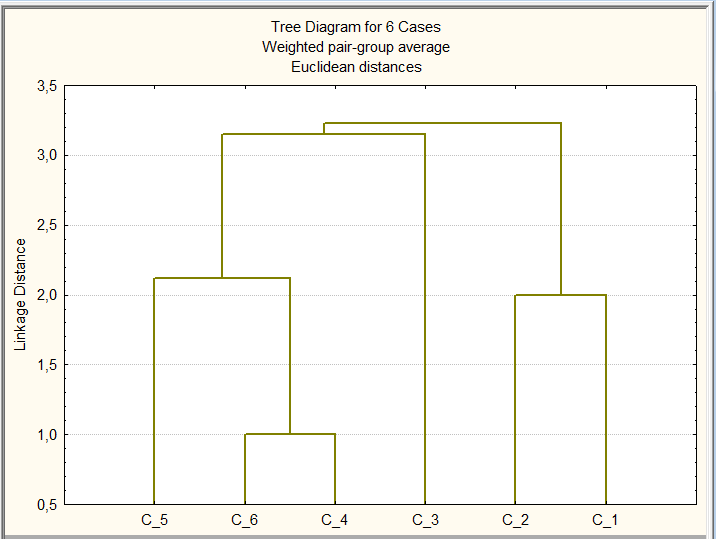
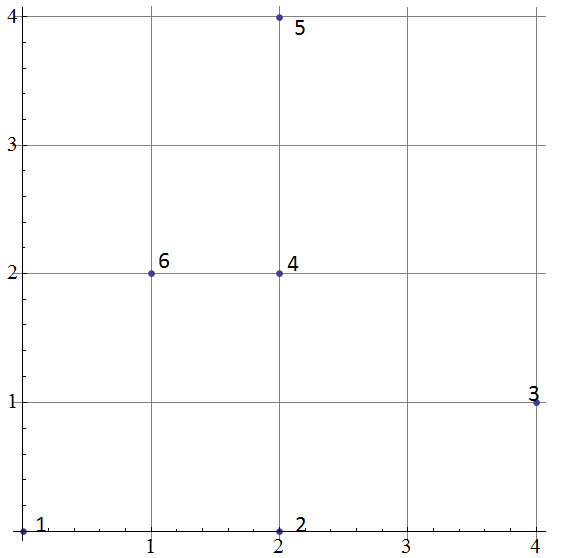
|  |  |  |  |
| --- | --- | --- | --- |
|  | X3 | X5 | X4,X6 |
| X1,X2 | 4,30 | 4,24 | 3,06 |

|  |  |  |
| --- | --- | --- |
|  | X3 | X1,X2 |
| X4,X5,X6 | 3,0 | 2,96 |

**Метод взвешенных попарных средних**

В методе взвешенных попарных средних расстояние между кластерами определяется следующим образом: пусть кластер  (u,v)  образован путем объединения кластеров u и v,тогда расстояние от кластера w до кластера (u,v) равно: .





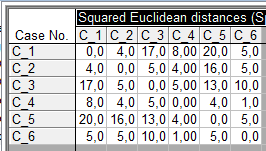
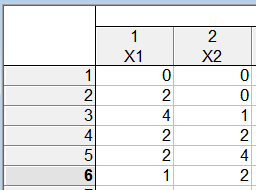
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | X1 | X2 | X3 | X5 |
| X4,X6 | 2,53 | 2,12 | 2,70 | 2,12 |

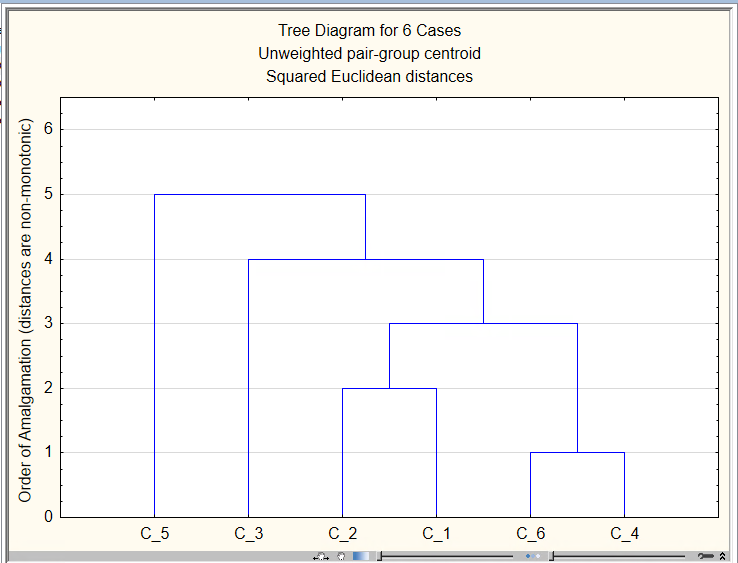
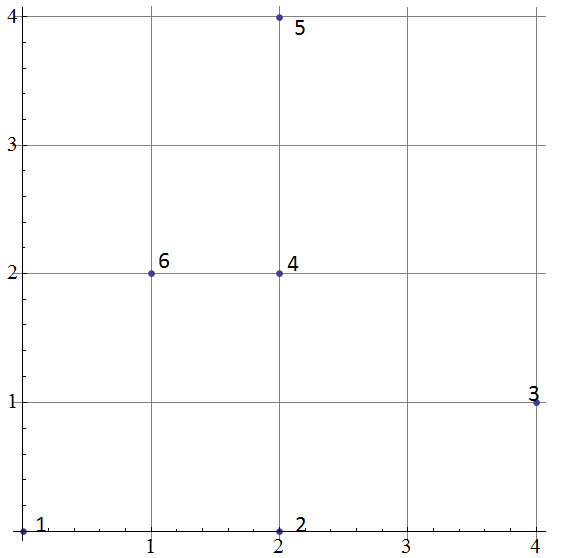
|  |  |  |  |
| --- | --- | --- | --- |
|  | X3 | X5 | X4,X6 |
| X1,X2 | 4,30 | 4,24 | 3,28 |

|  |  |  |
| --- | --- | --- |
|  | X3 | X1,X2 |
| X4,X5,X6 | 3,15 | 3,26 |

**Невзвешенный центроидный метод**

В невзвешенном центроидном методе за расстояние между кластерами берется расстояние между центрами тяжести кластеров: , где  вектор выборочных средних для кластера .





|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | X1 | X2 | X3 | X5 |
| X4,X6: (1,5;2) | 6,25 | 4,25 | 7,25 | 4,25 |

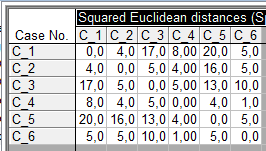
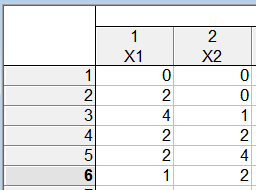
|  |  |  |  |
| --- | --- | --- | --- |
|  | X4,X6 | X3 | X5 |
| X1,X2:(1;0) | 4,25 | 10 | 17 |

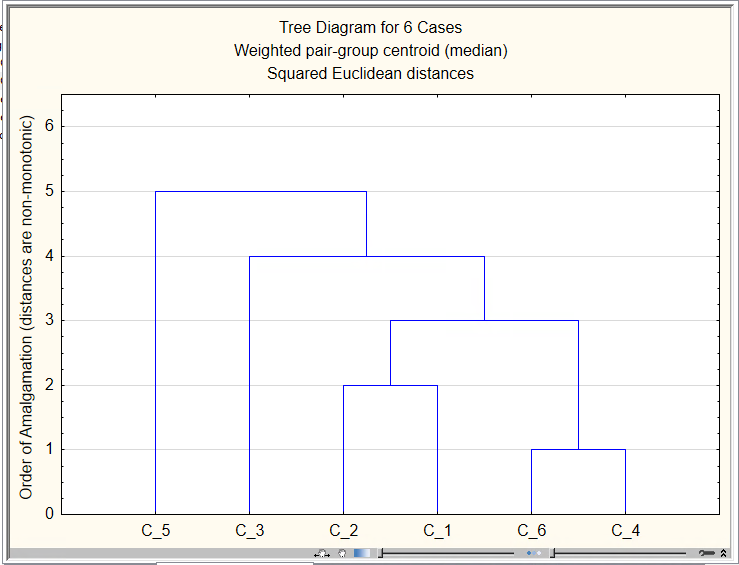
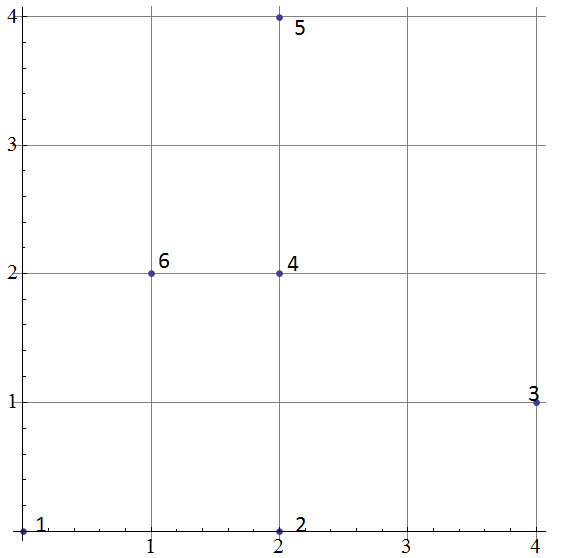
|  |  |  |
| --- | --- | --- |
|  | X3 | X5 |
| X1,X2,X4,X6: (1,25;1) | 7,5625 | 9,5625 |

Центроидный метод в отличии от предыдущих методов не обладает свойством монотонности: в результате объединения минимальное расстояние между кластерами может уменьшаться. Поэтому на диаграмме по оси OY указан порядковый номер объединения, а не расстояние между кластерами.

**Взвешенный центроидный метод (медианный метод)**

В взвешенном центроидном методе при объединении кластеров центр нового кластера определяется просто усреднением центров объединяемых кластеров. Расстояние между кластерами можно определить следующим образом: пусть кластер  (u,v)  образован путем объединения кластеров u и v,тогда расстояние от кластера w до кластера (u,v) равно: .





|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | X1 | X2 | X3 | X5 |
| X4,X6: (1,5;2) | 6,25 | 4,25 | 7,25 | 4,25 |

|  |  |  |  |
| --- | --- | --- | --- |
|  | X4,X6 | X3 | X5 |
| X1,X2:(1;0) | 4,25 | 10 | 17 |

|  |  |  |
| --- | --- | --- |
|  | X3 | X5 |
| X1,X2,X4,X6: (1,25;1) | 7,5625 | 9,5625 |

**Метод Уорда**

В методе Уорда объединение кластеров происходит таким образом, что сумма квадратов расстояний объектов до центров кластеров получает наименьшее приращение. То есть, для каждого кластера , состоящего из объектов , в соответствии с выбранной метрикой, вычисляется величина, где  - центр кластера . Данная величина, по сути, характеризует внутрикластерную дисперсию для кластера . Для совокупности кластеров , соответственно, вычисляется суммарная сумма квадратов отклонений . На первом шаге все величины , и на каждом шаге объединяются те кластеры, которые приводят к наименьшему изменению величины  после объединения (для этого потребуется рассмотреть все возможные объединения кластеров). Метод Уорда на каждом шаге приводит к формированию кластеров с наименьшей внутрикластерной дисперсией.

**Формула Ланса-Уильямса**

В общем случае процесс кластеризации – весьма трудоемкий процесс. Число необходимых операций можно уменьшить, если на каждом шаге использовать формулы, позволяющие подсчитывать расстояние от образовавшегося кластера  до любого другого кластера s, используя известные расстояния с предыдущих шагов. В частности, это можно сделать, с помощью формулы, предложенной Лансом и Уильямсом в 1967 году:



где  - числовые параметры.

Расстояние между кластерами для рассмотренных выше методов кластеризации удовлетворяет формуле Ланса-Уильямса со следующими коэффициентами:

* **Метод одиночной связи** (англ. *single linkage*)



* **Метод полной связи** (англ. *complete linkage*)



* **Невзвешенный метод средней связи** (англ. *UPGMA (Unweighted Pair Group Method with Arithmetic mean)*)



* **Взвешенный метод средней связи** (англ. *WPGMA (Weighted Pair Group Method with Arithmetic mean)*)



* **Невзвешенный центроидный метод** (англ. *UPGMC (Unweighted Pair Group Method with Centroid average)*)



* **Взвешенный центроидный метод** (англ. *WPGMC (Weighted Pair Group Method with Centroid average)*)



* **Метод Уорда** (англ. *Ward’s method*)



Формула Ланса-Уильямса для центроидного метода справедлива, если в качестве меры используется квадрат евклидова расстояния! Пакет Statistica считает расстояние между кластерами по формуле Ланса-Уильямса, поэтому, если выбрать центроидный метод, и, в качестве метрики выбрать евклидово расстояние, результат может получиться парадоксальным!!!

**Итеративные методы кластеризации**

В итеративных методах кластерного анализа на каждой итерации каждый из исходных объектов относится к одному из кластеров в соответствии с заданными условиями кластеризации, при этом одновременно происходит перерасчет параметров кластеров. Затем процесс кластеризации повторяется до тех пор, пока не получатся устойчивые образования кластеров, не меняющиеся от итерации к итерации. В качестве условий кластеризации используются, например, число кластеров, которые необходимо получить, предельный радиус кластеров и т.д. Данные условия выбираются обычно по результатам иерархических методов.

**Метод *k* – средних**

Пусть необходимо совокупность объектов  разбить на *k* кластеров.

1) На первом шаге из *n* векторов случайным образом (или не случайным) выбираем *k* из них, считая, что каждый из них образует центр кластера.

2) На следующем шаге каждый вектор из оставшегося набора данных относится к тому из кластеров, расстояние до центра которого от данного вектора минимально.

3) Пересчитываем центры кластеров.

4) Повторяем пункты 2 и 3 до тех пор пока не получится устойчивое разбиение, то есть составы кластеров на i-ой и i+1-ой итерации не будут различаться.

Представленный алгоритм называется алгоритмом Болла-Холла. Существует разновидность алгоритма, при которой центр кластера пересчитывается при каждом добавлении в кластер нового вектора (вариант Маккуина).

Можно показать, что при большом  и малом  алгоритм быстро сходится к устойчивому решению, т.е. к решению, в котором кластеры, полученные после применения алгоритма на i-ой итерации, совпадают по количеству и составу с кластерами, найденными на i+1-ой итерации. Тем не менее, алгоритм применяют несколько раз, используя различные начальные векторы в качестве центров кластеров и, если результаты различаются, выбирают оптимальное разбиение, используя один из критериев качества разбиения. Также при определении порядка выбора объектов на каждой итерации, рекомендуется этот порядок определять случайным образом.

**Метод поиска сгущений**

Данный алгоритм ориентирован на построение кластеров, радиус которых не превышает заданный. Радиус кластера – это расстояние от центра до наиболее удаленного объекта кластера. На первом шаге случайным образом выбирают один вектор, который будет играть роль центра первого кластера. После этого определяются вектора , расстояние от которых до центра первого кластера не превосходит R (т.е. попавшие внутрь сферы радиуса R с центром в точке, являющейся центром первого кластера) и помещаются в данный кластер. Затем пересчитывается центр получившегося кластера и процедура повторяется для всех объектов без исключения. Условием окончания итераций является равенство по составу кластеров, найденных на *m* и (*m+1*) шагах.

Попавшие внутрь первой сферы элементы  заключаем в один кластер и исключаем их из дальнейшего исследования. Для оставшихся точек алгоритм повторяется.

Можно показать /2/, что алгоритм сходится при любом выборе начального приближения и любом объеме исходных данных. Однако для получения устойчивого разбиения (т.е. разбиения, в котором кластеры, найденные после первого применения алгоритма, совпадают по количеству и составу с кластерами, найденными при повторном применении метода) рекомендуется повторить алгоритм несколько раз при различном выборе центров кластеров. Для поиска оптимального разбиения можно также варьировать значения радиуса R

**Функционалы качества разбиения**

Заметим, что задача кластеризации не имеет единственного решения. Как следствие, перебрать все допустимые разбиения данных на классы достаточно сложно и не всегда представляется возможным. Для того чтобы оценить качество различных способов кластеризации вводят понятие функционала качества разбиения, который принимает минимальное значение на наилучшем (с точки зрения исследователя) разбиении.

Пусть  – некоторая совокупность наблюдений, которая разбивается на классы , причем *q* заранее известно. Тогда основные функционалы качества разбиения при известном числе кластеров имеют вид:

1. Cумма квадратов расстояний до центров кластеров:

,

где  – центр кластера .

1. Сумма средних квадратов расстояний до центров кластеров:

,

где – число элементов в кластере .

1. Сумма внутриклассовых квадратов расстояний между элементами



1. Сумма средних внутриклассовых квадратов расстояний между элементами



где – число элементов в кластере .

1. Обобщенная внутриклассовая дисперсия

,

где  – число элементов в,  – выборочная ковариационная матрица для .

Если число кластеров в  заранее неизвестно, то используют следующие функционалы качества разбиения при произвольно выбираемом целом *m*:

1. , 
2. ,

где: – средняя мера внутриклассового рассеяния,  – мера концентрации точек множества ,  – число элементов в кластере, содержащем точку .

Заметим, что при произвольном значении параметра *m*, функционал  достигает минимума, равного , если исходное разбиение на кластеры  является разбиением на одноточечные кластеры:, так как . В то же время,  достигает максимума, равного 1, если  – один кластер, содержащий все исходные данные, так как . В частных случаях можно показать, что , где *q* – число различных кластеров в , , где  – число элементов в кластере , , .

Функционал  является убывающей функцией числа классов , а функционал  – возрастающей функцией числа классов . Такое поведение , гарантирует существование экстремумов , при подходящем выборе коэффициентов . Для нормированных данных можно рекомендовать значения . Для произвольных данных можно взять, например, .

Можно для оценивания качества разбиения использовать статистические методы. Например, проверять значимо ли различаются значения признаков для различных классов, сравнивая характеристики межгруппового и внутригруппового разбросов (такой подход используется в дисперсионном анализе). Пусть  - выбранный признак. Обозначим: - среднее значение признака по всем объектам (наблюдениям),  - среднее значение признака  для кластера ,  - значение признака  для i-го элемента кластера , ,  - число элементов в кластере . Рассмотрим следующие суммы:

 - полная сумма квадратов отклонений;

 сумма квадратов отклонений групповых средних от общей средней (межгрупповой разброс);

 сумма квадратов отклонений наблюдений в группах от групповых средних (внутригрупповой разброс).

Соответственно, величина  характеризует средний межгрупповой разброс, а величина  средний внутригрупповой разброс. Если средние для кластеров значимо различаются, то средний межгрупповой разброс должен значимо отличаться от среднего внутригруппового разброса. Для проверки значимости различия используем критерий Фишера. В предположении нормальности исходных данных статистика критерия: , при истинности  приближенно имеет распределение Фишера с  степенями свободы. Если наблюдаемое значение статистики превышает критическое для заданного уровня значимости, то  отвергается и различие признаков для классов признается значимым. Чем больше признаков, значения которых значимо различаются для различных классов, тем лучше кластеризация.