# Обзор существующих методов иерархической кластеризации

Основная цель иерархической кластеризации состоит в группировании объектов в соответствии с их сходством на основе заданных мер или функций расстояния.

Существует два основных подхода к иерархической кластеризации: агломеративный и дивизивный. В агломеративном подходе каждый объект начинает в качестве отдельного кластера, затем на каждом шаге алгоритм объединяет два наиболее близких кластера, продолжая этот процесс до тех пор, пока все объекты не будут объединены в один кластер (Рисунок 1). Дивизивный подход, напротив, начинается с одного крупного кластера, который затем рекурсивно разделяется на более мелкие кластеры до достижения заданного критерия останова (Рисунок 2).

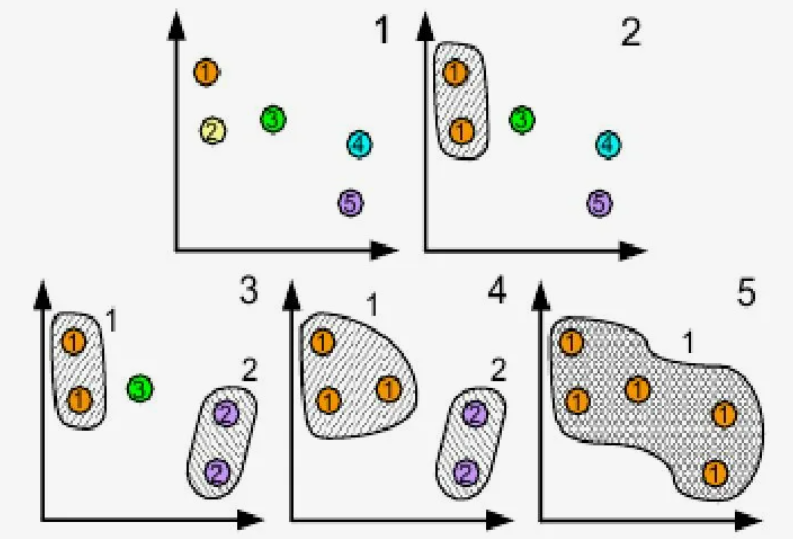


Рисунок 1 - Общая схема агломеративной иерархической кластеризации

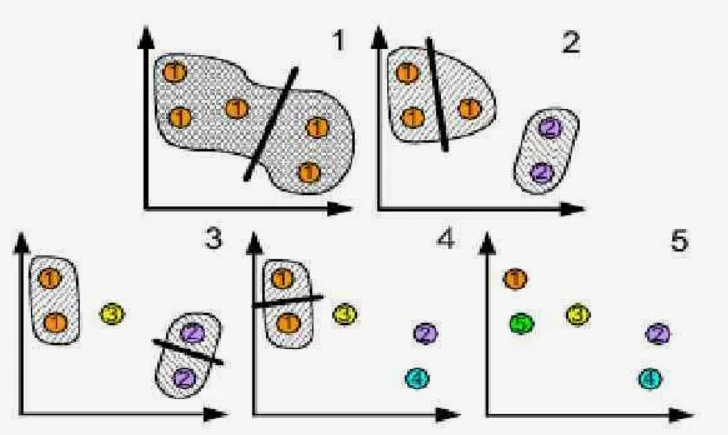


Рисунок 2 - Общая схема дивизивной иерархической кластеризации

В процессе иерархической кластеризации образуется дерево или дендрограмма, которая представляет собой иерархическую структуру кластеров (Рисунок 3). Дендрограмма позволяет визуально оценить структуру данных и определить оптимальное число кластеров.

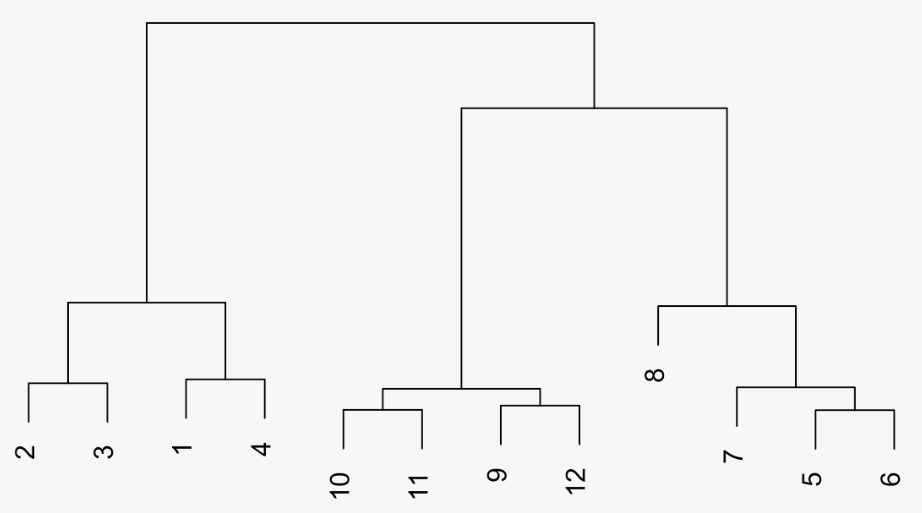


Рисунок 3 - Пример дендрограммы

## Классические методы иерархической кластеризации

Кластерные расстояния, также известные как меры близости или расстояния между кластерами, используются для измерения степени сходства или различия между кластерами в результате кластеризации данных. Эти меры играют важную роль в оценке качества кластеризации, а также в решении задач, таких как объединение или разделение кластеров.

Некоторые из классических методов кластерных расстояний (Рисунок 4):

* метод односвязной (single-linkage) кластеризации: этот метод использует минимальное расстояние между парами точек из двух кластеров. Другими словами, расстояние между двумя кластерами определяется как минимальное расстояние между всеми парами точек, одна из которых принадлежит первому кластеру, а другая – второму (4):

(4)

где W, S – кластеры,

d(w, s) – расстояние между объектами w и s, принадлежащие кластерам W и S соответственно.

* метод полносвязной (complete-linkage) кластеризации: этот метод использует максимальное расстояние между парами точек из двух кластеров, то есть расстояние между двумя кластерами определяется как максимальное расстояние между всеми парами точек, одна из которых принадлежит первому кластеру, а другая – второму (5):

(5)

где W, S – кластеры,

d(w, s) – расстояние между объектами w и s, принадлежащие кластерам W и S соответственно.

* метод среднего расстояния (average linkage): этот метод использует среднее расстояние между всеми парами точек из двух кластеров. Он вычисляет среднее арифметическое расстояний между всеми парами точек, одна из которых принадлежит первому кластеру, а другая – второму (6):

(6)

где W, S – кластеры,

|W|, |S| - мощности кластеров W и S,

d(w, s) – расстояние между объектами w и s, принадлежащие кластерам W и S соответственно.

* метод ускоренного увеличения (Ward's method): этот метод оптимизирует объединение кластеров так, чтобы минимизировать увеличение общей суммы квадратов расстояний между каждым объектом и центроидом его кластера (7):

(7)

где W, S – кластеры,

|W|, |S| - мощности кластеров W и S,

d(w, s) – расстояние между объектами w и s, принадлежащие кластерам W и S соответственно.

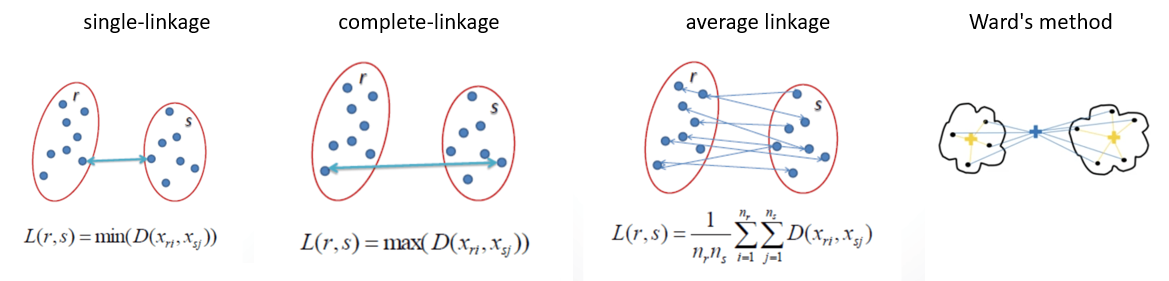


Рисунок 4 - Иллюстрация связей

## Новые методы иерархической кластеризации

Наряду с классическими методами иерархической кластеризации, особый интерес представляют новые методы, которые внедряют усовершенствованные алгоритмы для объединения данных в группы. Эти методы обещают повысить точность и эффективность кластеризации по сравнению с традиционными методами.

Рассмотрим два новых метода (метод 1 и метод 2), которые представляют собой новаторские подходы к решению задачи кластеризации данных. Они основаны на уникальных принципах вычисления расстояний между кластерами и имеют потенциал для улучшения качества кластеризации.

**Метод 1** основан на идее определения расстояния между кластерами как полусуммы минимального и максимального расстояний между объектами из двух кластеров (8, 9, 10):

(8)

(9)

(10)

где A, B – кластеры,

D(A, B) - расстояние между кластерами A и B.

**Метод 2** представляет расстояние между кластерами как медианное расстояние между всеми парами объектов из двух кластеров. Формула для вычисления расстояния между двумя кластерами выглядит следующим образом (11):

(11)

где D(A, B)-расстояние между кластерами A и B.

Оба метода представляют собой новые подходы к определению расстояния между кластерами и могут быть эффективными для различных типов данных и структур кластеров. Их эффективность и сравнение с традиционными методами оцениваются с использованием различных метрик качества, что будет рассмотрено подробнее далее.

**2. Агломеративная иерархическая кластеризация**

**2.1. Алгоритм расчета иерархической кластеризации**

Ограничимся только агломеративным типом иерархической кластеризации, т.е. дерево строится от листьев к дереву.

1. На начальном этапе каждый объект принимается за отдельный кластер;
2. Расстояние между каждым кластером определяется некоторой метрикой. Чаще всего используются Евклидово, Манхэттенское и Чебышевское расстояние. Эти и другие метрики мы обсудим в следующем параграфе.
3. Расстояния сравниваются, и два наиболее похожих кластера во всем наборе данных сначала группируются, становясь единым кластером.
4. Расстояния между всеми кластерами пересчитываются на основе выбранного метода связи – одиночная связь, полная связь, средняя связь и другие, все они будут обсуждаться в следующих параграфах.
5. Следующий набор наиболее похожих кластеров группируется, и процесс повторяется до тех пор, пока не появится один кластер, включающий все остальные.

**2.2. Методы вычисления расстояний между синглтонами**

Одна из наиболее важных задач при кластеризации данных — решить, какую метрику использовать для расчета расстояния между каждым объектом данных. Существует множество различных метрик, применимых к разным типам данных, вот некоторые из них:

* *Евклидово расстояние* – расстояние между двумя точками в -мерном пространстве, вычисляемое по формуле:

|  |  |  |
| --- | --- | --- |
|  |  | (1.1) |

* *Манхэттенское расстояние* – также известное как расстояние городских кварталов, определяется как сумма абсолютных различий между соответствующими координатами точек:

|  |  |  |
| --- | --- | --- |
|  |  | (1.2) |

* *Расстояние Чебышева* – также известное как максимум-минимум расстояние или -метрикой, определяется как максимальная разница между координатами точек в каждом измерении:

|  |  |  |
| --- | --- | --- |
|  |  | (1.3) |

* *Косинусное расстояние* – мера сходства между двумя векторами в многомерном пространстве, определяемая как косинус угла между ними:

|  |  |  |
| --- | --- | --- |
|  |  | (1.4) |

* *Корреляционное расстояние* – может быть определено как евклидово расстояние между двумя векторами, полученными из оригинальных наборов данных, путем вычисления корреляции между их признаками и замены значений корреляции на расстояния.

|  |  |  |
| --- | --- | --- |
|  |  | (1.4) |

* *Расстояние Минковского* – это обобщение евклидова, манхэттенского и чебышевского расстояний расстояния для более высоких размерностей и вычисляется по формуле:

|  |  |  |
| --- | --- | --- |
|  |  | (1.5) |

где – это порядок расстояния. Если , то это манхэттенское расстояние, если , то это евклидово расстояние, а если , то это расстояние Чебышева.

* *Расстояние Хэмминга* – это метрика расстояния между двумя строками равной длины, определяющаяся как количество позиций, в которых символы обеих строк различаются:

|  |  |  |
| --- | --- | --- |
|  |  | (1.6) |

* *Расстояние Гаверса* – это мера расстояния между двумя объектами в многомерном прострастве, которая учитывает как числовые, так и порядковые, категориальные признаки:

|  |  |  |
| --- | --- | --- |
|  |  | (1.7) |

функция вычисляется для каждого типа данных отдельно

* *Расстояние Махаланобиса* – мера расстояния между двумя точками в многомерном пространстве, которая учитывает корреляции и ковариации признаков:

|  |  |  |
| --- | --- | --- |
|  |  | (1.8) |

* *Расстояние Жаккара* – метрика сходства между множествами, варьируемая от 0 до 1 и определяемая как:

|  |  |  |
| --- | --- | --- |
|  |  | (1.9) |

* *Расстояние Левенштейна* – известное как редакционное расстояние или расстояние редактирования, используется для измерения разницы между двумя строками. Оно определяется как минимальное количество операций вставки, удаления и замены символов, необходимых для преобразования одной строки в другую.

Выбор конкретной метрики зависит от области и задачи применения. Например, евклидово расстояние обычно используется в задачах классификации и кластеризации объектов в евклидовом пространстве. Расстояние Манхэттенское более подходит для задач, связанных с геометрическими приложениями, где расстояние между точками должно основываться на их геометрическом расположении. Расстояние Чебышева - подходит для задач, в которых важно знать максимальную разницу между переменными. Например, при анализе временных рядов или задачах, связанных с компьютерным зрением и распознаванием образов, где важно знать, какое значение имеет максимальное отклонение между точками. Расстояния Жаккара используется в задачах, связанных с анализом текстовых данных, а расстояние Левенштейна - в задачах редактирования строк.

Ограничимся тремя наиболее популярными метриками, применимые к геометрическим приложениям, а именно: Евклидовым расстоянием, Манхэттенским расстоянием и расстоянием Чебышева.

**2.3 Методы вычисления расстояний между кластерами**

После того, как была проведена кластеризация синглтонов, необходимо объединить полученные кластеры в более крупные группы. Для этого применяются различные методы (рис. 6) вычисления расстояний между кластерами, но некоторые из них наиболее распространены:

* *Одиночная связь (Single Linkage)* – минимальное расстояние, достигаемое между парой объектов, где два объекта из пары принадлежат отдельным кластерам:

|  |  |  |
| --- | --- | --- |
|  |  | (2.1) |

* *Полная связь (Complete Linkage)* – максимальное расстояние, достигаемое между парой объектов, где два объекта из пары принадлежат отдельным кластерам:

|  |  |  |
| --- | --- | --- |
|  |  | (2.2) |

* *Групповая средняя связь (Group Average Linkage, UPGMA)* – среднее расстояние между элементами двух кластеров:

|  |  |  |
| --- | --- | --- |
|  |  | (2.3) |

* *Взвешенная средняя связь (Weighted Average Linkage, WPGMA)* – среднее арифметическое расстояний от кластеров объединенной пары до кластера, до которого необходимо рассчитать расстояние:

|  |  |  |
| --- | --- | --- |
|  |  | (2.4) |

* *Уордовская связь (Minimum Variance, Ward)* – прирост суммы квадратов расстояний объектов до центра кластера, получаемого в результате их объединения:

|  |  |  |
| --- | --- | --- |
|  |  | (2.5) |

* *Центройдная связь (Centroid, UPGMC)* – определяет расстояние между кластерами на основе расстояния между их центрами тяжести или же центроидами. Центроид кластера представляет собой среднее арифметическое всех точек в кластере.

|  |  |  |
| --- | --- | --- |
|  |  | (2.6) |

* *Медианная связь (Median, WPGMC)* – аналогичен предыдущему, за исключением того, что при вычислениях используются веса для учета разницы между размерами кластеров.

|  |  |  |
| --- | --- | --- |
|  |  | (2.7) |

* *Минимальная сумма медойдов (Minimum Sum Medoid linkage)* – данный метод использует понятие медойда, т.е. объекта, принадлежащего кластеру, различие которого с другими минимально. Здесь считается сумма расстояний между медойдом и всеми объектами, объединяющихся на данном этапе кластеров:

|  |  |  |
| --- | --- | --- |
|  |  | (2.8) |

* *Медойдная связь Medoid linkage* – данный метод считает расстояние между кластерами как расстояние между медойдами объединяющихся кластеров:

|  |  |  |
| --- | --- | --- |
|  |  | (2.9) |

|  |
| --- |
|  |
| Рисунок 6. Визуализация различных видов иерархической кластеризации |

Каждый из этих методов имеет свои преимущества и недостатки, и выбор метода зависит от конкретной задачи и данных, с которыми вы работаете.

**2.4. Оптимизация вычислений – рекуррентная формула Ланса-Уильмса**

В 1966-1967 годах ученые Г. Ланс и В. Уильямс представили обобщенную рекуррентную формулу, которая может быть использована для вычисления межкластерного расстояния. Данная формула дает обновленное расстояние между кластером и объединенным кластером . Обозначим формально расстояние между объединенным кластером и кластером за , рекуррентная формула рассчитывается следующим образом:

|  |  |  |
| --- | --- | --- |
|  |  | (3.1) |

Параметры – отвечают за метод объединения кластеров. Было показано, что по меньшей мере восемь хорошо известных алгоритмов иерархической кластеризации удовлетворяют рекуррентному соотношению [Кормак, 1971]. Значения параметров показаны в Таблице 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Таблица 1  Иерархические алгоритмы | | | | |
|  |  |  |  |  |
| Метод |  |  |  |  |
| Single Linkage |  |  |  |  |
| Complete Linkage |  |  |  |  |
| Group Average |  |  |  |  |
| Weighted Average |  |  |  |  |
| Centroid |  |  |  |  |
| Median |  |  |  |  |
| Ward |  |  |  |  |
| Flexible (Lance and Williams) |  |  |  |  |

В Таблице 1 последним пунктом приведен новый метод Flexible , который является методом иерархической кластеризации и был предложен Г. Лансом и В. Уильямсом. Идея заключается в том, чтобы вычислить расстояние между двумя кластерами на основе расстояний между их элементами. Данный метод является более универсальным методом, чем Single Linkage и Complete Linkage. Он может быть настроен в зависимости от данных и целей анализа. Этот метод позволяет управлять тем, какие расстояния будут использоваться при объединении кластеров, что может помочь уменьшить искажения и улучшить качество кластеризации.

Важность рекуррентного соотношениях для настоящей работы заключается в том, что формула была использована для оптимизации времени и памяти вычисления матриц близости.

**2.5. Свойство монотонности методов иерархической кластеризации**

Пусть – это расстояние между кластерами, возникающее при их объединении, на уровне иерархии . Если представить, что мы не будем накладывать на никаких дополнительных условий и оно может принимать любое значение, то в общем случае дендрограмма будет иметь огромное множество самопересечений. Такой график не будет иметь никакой полезной информации.

Г. Миллиган показал, что если на каждом уровне иерархии расстояния уменьшаются или остаются неизменными, т.е. если , то дендрограмму можно построить таким образом, что она не будет иметь самопересечения. Величина обладающая таким свойством называется монотонной.

|  |
| --- |
|  |
| Рисунок 7. Визуализация непересекающейся и  самопересекащейся дендрограмм |

Кроме того, свойство монотонности позволяет использовать меры расстояния, которые могут не удовлетворять неравенству треугольника, такие как расстояние Канторовича или расстояние Левенштейна. Эти меры могут быть полезны для работы с категориальными данными или при сравнении строк символов. Однако они не могут быть использованы в алгоритмах кластеризации, которые не удовлетворяют свойству монотонности.

Также в 1979 году Г. Миллиган показал, что если параметры рекуррентной формулы Ланса-Уильямса (3.1) удовлетворяют условиям, указанным в формуле (4.1), то кластеризация будет монотонной:

|  |  |  |
| --- | --- | --- |
|  | 1. ; 2. ; | (4.1) |

Из перечисленных в Таблице 1 методов, монотонными являются все, кроме Centroid и Median. По этой причине в данной работе мы отбросим их и будем рассматривать только монотонные методы.

**2.6. Ультраметрическая матрица**

На каждом уровне иерархии мы будем получать значения близости , которые будут возникать при возникновении новых подмножеств, путем объединения старых. Все эти значения мы положим матрицу размера . В общем случае, в полученной матрице будет различных ненулевых значений мер близости , которые определяют уровни, на которых формируется новых подмножеств в иерархии.

Остановимся подробней на составлении ультраметрической матрицы (рис. 8). Для начала определим все необходимые структуры. Представим, что у нас имеется:

* Представим, что у нас имеется упорядоченный набор , где каждый элемент представляет собой минимальные значения в матрице близости найденное на некотором этапе иерархии, причем первый элемент набора соответствует первому этапу иерархии, второй элемент – второму этапу и т.д.
* Бинарное дерево , в котором на узлах указаны индексы исходного набора данных, объединенных в кластеры на некотором этапе. В корневом узле дерева находятся индексы всех элементов исходного набора, так как он соответствует последнему этапу кластеризации, когда все исходные объекты образуют один кластер, а в листьях (терминальных узлах) находятся по одному индексу объектов из исходного набора.
* Пустая матрица размера размера , где – количество объектов исходного набора данных.

Теперь опишем сам алгоритм. Вначале берем внутренние узлы, образующие корневой узел. Различные комбинации элементов из каждого множества, представленных в узлах, будут использоваться в качестве индексов для заполнения соответствующих ячеек ультраметрической матрицы, значением равным последнему элементу упорядоченного набора .

|  |
| --- |
|  |
| Рисунок 8. Визуализация алгоритма построения ультраметрической матрицы |

Далее продолжаем алгоритм, применяя его к каждой вершине, пока размер множества в вершине не станет равным 1, что означает достижение листовых узлов дерева и завершение процесса составления ультраметрической матрицы.

Матрица называется ультраметрической, если для каждой тройки индексов и выполняется (неравенство треугольника). Другими словами, расстояние между кластерами на каждом следующем уровне иерархии не уменьшается по сравнению с предыдущим уровнем. Другими словами, если метод иерархической кластеризации является монотонным, то для нее можно построить ультраметрическую матрицу.

Ультраметрическая матрица может быть использованы для определения иерархии кластеров, а также для анализа сходства и различия между группами объектов. На дендрограмме с ультраметрическими расстояниями между кластерами, мы будем видеть, что расстояния между кластерами на каждом уровне иерархии не уменьшается по сравнению с предыдущим уровнем.

Важно, чтобы матрица расстояний сохраняла свои свойства, в частности ультраметрические свойства, чтобы результаты иерархической кластеризации были корректными. Если матрица расстояний не является ультраметрической, то полученное дерево кластеров может не соответствовать иерархическим отношениям между объектами, что может привести к ошибочным выводам при анализе данных.

**2.7. Метрики близости начальной и ультраметрической матриц**

Для обеспечения правильной работы алгоритмов кластеризации, важно, чтобы ультраметрическая матрица расстояний была как можно ближе к исходной матрице расстояний, которая была получена из данных. Чем ближе ультраметрическая матрица расстояний к исходной матрице расстояний, тем более точным и корректным будет результат иерархической кластеризации. Для этих целей мы введем метрики, для определения степени близости матрицы близости и ультраметрической матрицы:

1. *Максимальное отклонение* – максимальное значение отношения абсолютного значения разницы первоначальной матрицы расстояний и ультраметрической матрицы к абсолютному значению первоначальной матрицы расстояний:

|  |  |  |
| --- | --- | --- |
|  |  | (6.1) |

1. Сумма отклонений – сумма абсолютных значений разницы первоначальной матрицы расстояний и ультраметрической матрицы деленное на количество ребер между всеми синглтонами:

|  |  |  |
| --- | --- | --- |
|  |  | (6.2) |

Помимо вышеописанных введенных нами метрик, существуют также другие, представленные в литературе. Например, метрика VAF (Variance Accounted For) описывающаяся следующим образом:

|  |  |  |
| --- | --- | --- |
|  |  | (6.3) |

В данной работе мы будем использовать эти метрики для оценки близости ультраметрической матрицы и начальной матрицы расстояний.