

ИЕРАРХИЧЕСКАЯ И ГРАФОВАЯ КЛАСТЕРИЗАЦИЯ

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ДАННЫХ (DATA MINING)

ЦЕЛЬ ЛЕКЦИИ

- Сформировать у студентов понимание принципов иерархической и графовой кластеризации, изучить основные алгоритмы, методы определения сходства между объектами и способы применения этих методов для анализа структурированных данных и сетей.

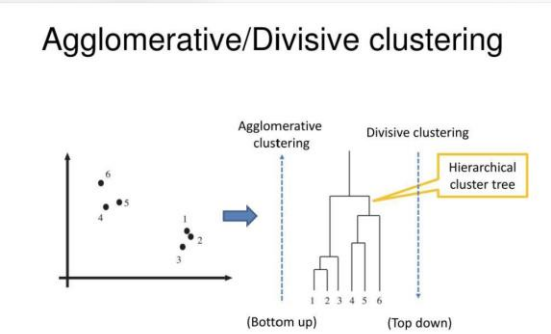
ЗАДАЧИ ЛЕКЦИИ

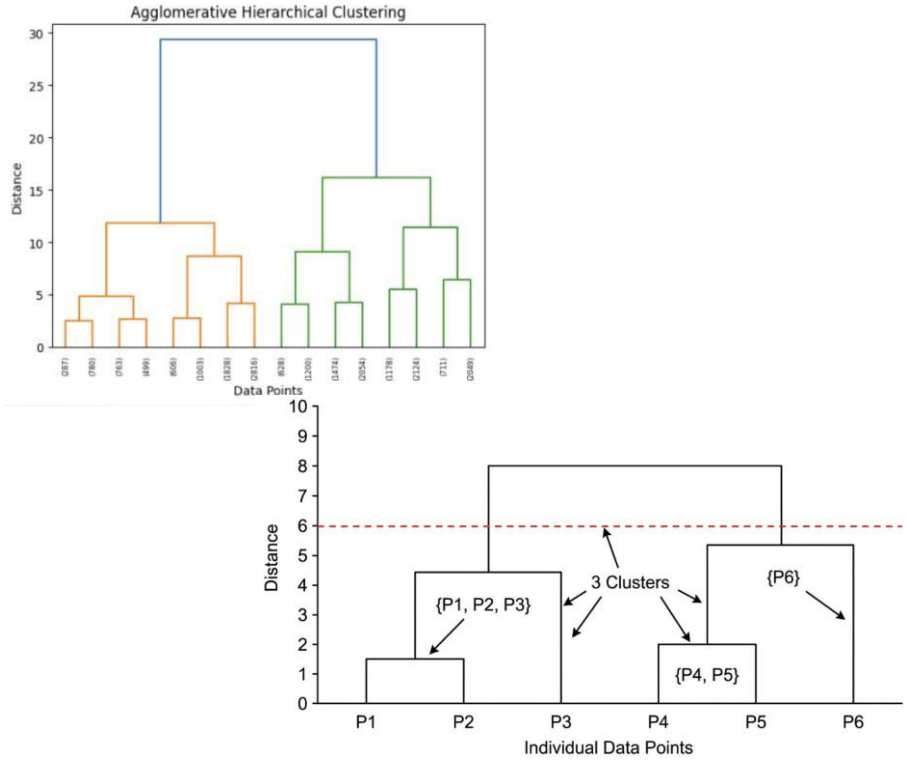
- Рассмотреть понятие и основные виды иерархической кластеризации.
- Изучить алгоритмы агломеративной и дивизивной кластеризации.
- Объяснить методы определения расстояния между кластерами и построения дендрограммы.
- Рассмотреть графовое представление данных и основные алгоритмы графовой кластеризации.
- Сравнить иерархическую и графовую кластеризацию.
- Научиться применять алгоритмы кластеризации на Python.

1. ПОНЯТИЕ ИЕРАРХИЧЕСКОЙ КЛАСТЕРИЗАЦИИ

- Иерархическая кластеризация (Hierarchical Clustering) — это метод интеллектуального анализа данных, который группирует объекты в виде иерархии вложенных кластеров.
- В отличие от алгоритма K-means, где обычно заранее задаётся количество кластеров K, иерархическая кластеризация позволяет изучать структуру групп на разных уровнях детализации.

Например, при сегментации клиентов можно сначала выделить две большие группы, затем разделить каждую из них на более мелкие сегменты.



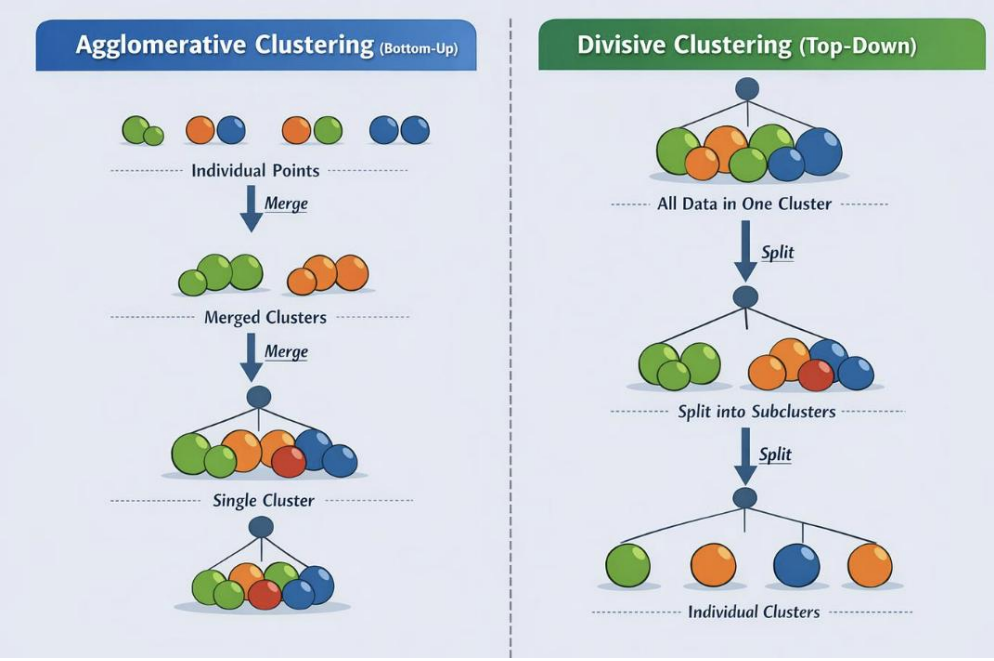


1.1. ОСНОВНЫЕ ПОНЯТИЯ

- Объект — отдельная запись или наблюдение в наборе данных.
- Кластер — группа похожих объектов.
- Расстояние между объектами — мера различия между двумя наблюдениями.
- Расстояние между кластерами — мера различия между двумя группами объектов.
- Дендрограмма (Dendrogram) — дерево, показывающее, как кластеры объединяются или разделяются на разных уровнях.

Иерархическая кластеризация особенно полезна, когда необходимо не только получить группы объектов, но и понять отношения между этими группами.

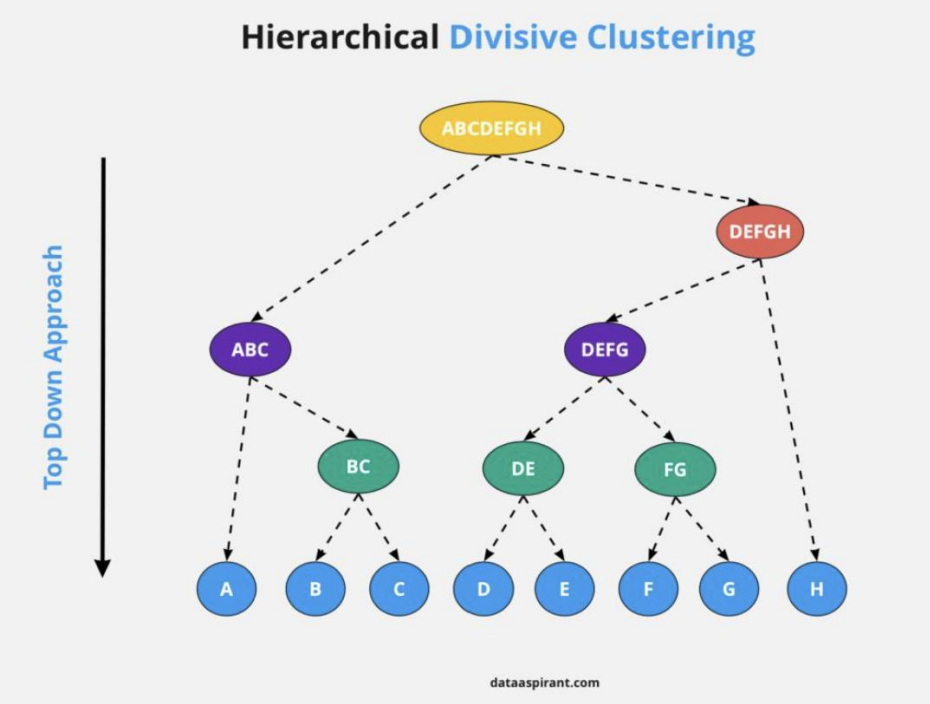
2. ВИДЫ ИЕРАРХИЧЕСКОЙ КЛАСТЕРИЗАЦИИ



2.1. Агломеративная кластеризация (Agglomerative)

Подход «снизу вверх». Вначале каждый объект образует отдельный кластер. Затем на каждом шаге объединяются наиболее близкие кластеры. Процесс продолжается до достижения заданного количества кластеров или объединения всех объектов.

2.2. ДИВИЗИВНАЯ КЛАСТЕРИЗАЦИЯ (DIVISIVE)



- Подход «сверху вниз». Сначала все объекты относятся к одному кластеру. Затем кластер последовательно разделяется на более мелкие группы, пока не будет достигнут заданный уровень детализации.

На практике агломеративный подход используется чаще, поскольку его проще реализовать и он представлен в распространённых библиотеках машинного обучения.

2.3. АЛГОРИТМ АГЛОМЕРАТИВНОЙ КЛАСТЕРИЗАЦИИ

Основные этапы:

- Каждый объект объявляется отдельным кластером.
- Рассчитываются расстояния между объектами или кластерами.
- Находятся два наиболее близких кластера.
- Эти кластеры объединяются.
- Пересчитываются расстояния между полученными группами.
- Процесс повторяется до достижения условия остановки.

Например, если имеется пять объектов A, B, C, D, E, алгоритм сначала может объединить A и B, затем D и E, после чего объединять полученные группы в соответствии с выбранной метрикой расстояния.

3. МЕТОДЫ ОПРЕДЕЛЕНИЯ РАССТОЯНИЯ МЕЖДУ КЛАСТЕРАМИ

Результат иерархической кластеризации зависит от способа вычисления расстояния между группами объектов. Этот способ называется правилом связи (linkage criterion).

Рассмотрим четыре распространённых метода.

3.1. МЕТОД БЛИЖАЙШЕГО СОСЕДА — SINGLE LINKAGE

- Расстояние между двумя кластерами определяется как минимальное расстояние между любой парой объектов из этих кластеров.

$$D(A, B) = \min_{x \in A,\, y \in B} d(x, y)$$

Особенность: позволяет обнаруживать группы сложной формы, но может приводить к эффекту цепочки, когда кластеры объединяются через последовательность близких объектов.

3.2. МЕТОД ДАЛЬНЕГО СОСЕДА — COMPLETE LINKAGE

- Расстояние определяется как максимальное расстояние между объектами двух кластеров.

$$D(A, B) = \max_{x \in A,\ y \in B} d(x, y)$$

Особенность: чаще формирует компактные кластеры, но может разделять группы, внутри которых есть удалённые друг от друга объекты.

3.3. МЕТОД СРЕДНЕГО РАССТОЯНИЯ — AVERAGE LINKAGE

- Вычисляется среднее расстояние между всеми парами объектов двух кластеров.

$$D(A,B) = \frac{1}{|A||B|} \sum_{x \in A} \sum_{y \in B} d(x,y)$$

Особенность: представляет компромисс между Single Linkage и Complete Linkage.

3.4. МЕТОД УОРДА — WARD LINKAGE

- Метод Уорда объединяет кластеры так, чтобы увеличение внутрикластерной суммы квадратов отклонений было минимальным.

$$\Delta(A,B) = \frac{|A||B|}{|A|+|B|}\|\mu_A - \mu_B\|^2$$

Здесь:

- |A| и |B| — размеры кластеров;
- μ_A и μ_B — центры кластеров;
- $\Delta(A,B)$ — увеличение внутрикластерной суммы квадратов при объединении.

СРАВНЕНИЕ МЕТОДОВ СВЯЗИ

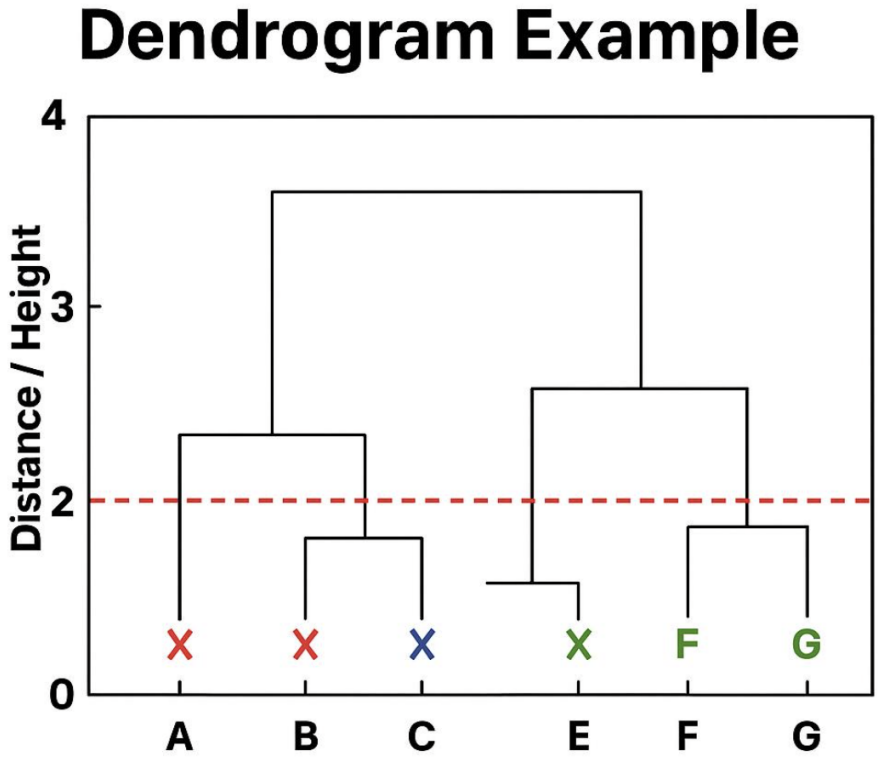
Метод	Принцип	Особенность
Single Linkage	Минимальное расстояние	Может создавать цепочечные кластеры
Complete Linkage	Максимальное расстояние	Предпочитает компактные группы
Average Linkage	Среднее расстояние	Компромиссный подход
Ward Linkage	Минимальный прирост внутрикластерной дисперсии	Подходит для компактных кластеров

4. ДЕНДРОГРАММА И ВЫБОР ЧИСЛА КЛАСТЕРОВ

- Дендрограмма — графическое представление процесса объединения кластеров. По ней можно определить, на каком уровне объединялись отдельные объекты и группы.

На дендрограмме:

- по горизонтальной оси располагаются объекты или конечные группы;
- по вертикальной оси отображается расстояние, на котором происходило объединение;
- горизонтальные соединения показывают, какие группы были объединены.



КАК ВЫБРАТЬ КОЛИЧЕСТВО КЛАСТЕРОВ?

Один из распространённых способов — провести горизонтальную линию на выбранной высоте дендрограммы и посчитать число ветвей, которые она пересекает.

Если линия пересекает три независимые ветви, то при таком срезе получаются три кластера.

Однако выбор высоты не всегда очевиден. Обычно учитывают:

1. большие скачки расстояния между последовательными объединениями;
2. предметную область и назначение кластеризации;
3. интерпретируемость полученных групп;
4. дополнительные показатели качества кластеризации.

ПРИМЕР

Предположим, что при объединении кластеров расстояния составили:

0,8; 1,1; 1,4; 4,7; 5,2.

Большой скачок от 1,4 до 4,7 может указывать на то, что до этого объединения существовали достаточно различающиеся группы. Поэтому срез перед большим скачком может быть разумным вариантом для выбора количества кластеров.

Это ориентир, а не универсальное правило: полученное решение необходимо проверить с учётом структуры данных.

5. ПРЕИМУЩЕСТВА И ОГРАНИЧЕНИЯ ИЕРАРХИЧЕСКОЙ КЛАСТЕРИЗАЦИИ

Преимущества:

- не обязательно заранее задавать количество кластеров;
- можно анализировать структуру данных на разных уровнях;
- дендрограмма помогает визуально интерпретировать результаты;
- доступны разные методы определения расстояния между группами.

Ограничения:

- вычислительные и временные затраты могут быть высокими на больших наборах данных;
- результат зависит от метрики расстояния и правила связи;
- ранние объединения в классическом агломеративном алгоритме обычно не пересматриваются;
- шум и выбросы могут влиять на структуру кластеров;
- выбор окончательного количества кластеров требует обоснования.

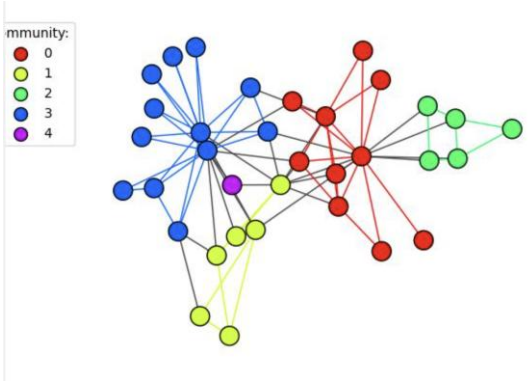
6. ПОНЯТИЕ ГРАФОВОЙ КЛАСТЕРИЗАЦИИ

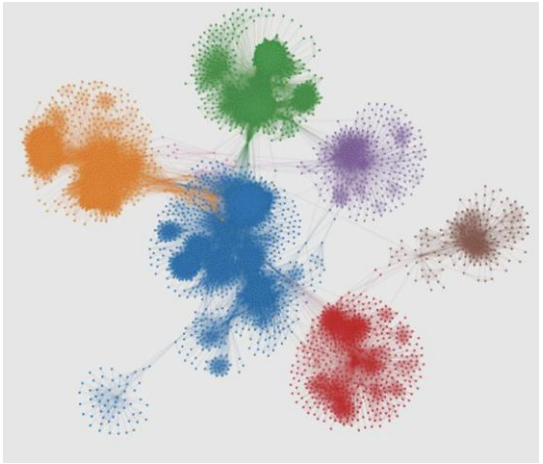
- Графовая кластеризация (Graph Clustering) — это группировка вершин графа в кластеры с учётом структуры связей между ними.
- Граф записывается как:

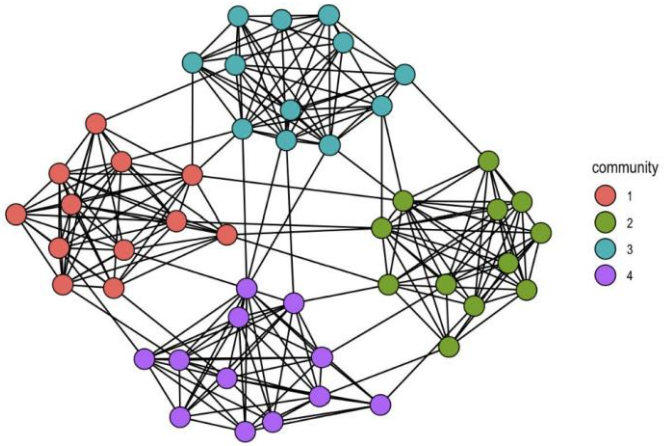
$$G = (V, E)$$

где:

- V — множество вершин;
- E — множество рёбер, соединяющих вершины.







Например, в социальной сети люди образуют вершины, а связи между ними — рёбра. Графовая кластеризация позволяет обнаружить сообщества, в которых участники связаны между собой сильнее, чем с участниками других сообществ.

6.1. ОСНОВНАЯ ИДЕЯ

В графовой кластеризации учитываются не только свойства отдельных объектов, но и отношения между ними.

Для многих задач желательно, чтобы:

- внутри кластера было много связей;
- между разными кластерами было сравнительно мало связей.

Это принцип обнаружения сообществ (Community Detection).

Графы могут быть:

- неориентированными — ребро не имеет направления;
- ориентированными — ребро имеет направление;
- взвешенными — ребру соответствует вес, например сила сходства;
- невзвешенными — все рёбра рассматриваются без различия по весу.

Выбор модели графа зависит от решаемой задачи.

7. ОСНОВНЫЕ МЕТОДЫ ГРАФОВОЙ КЛАСТЕРИЗАЦИИ

7.1. Кластеризация по связным компонентам

- Связная компонента — это максимальная группа вершин, в которой между любыми двумя вершинами существует путь.
- Если граф состоит из нескольких несвязанных частей, каждую такую часть можно рассматривать как отдельную группу.

Преимущества: простой и быстрый метод.

Ограничения: если граф имеет одну большую связную компоненту, метод не разделит её на более мелкие сообщества, даже если внутри существуют явно выраженные группы.

7.2. АЛГОРИТМ ЛУВЕНА (LOUVAIN)

Алгоритм Louvain применяется для обнаружения сообществ в больших сетях. Он стремится увеличить показатель модулярности (modularity) — меру того, насколько связи внутри сообществ отличаются от ожидаемых при определённой случайной модели сети.

Основные этапы алгоритма:

- Каждая вершина первоначально образует отдельное сообщество.
- Вершины последовательно перемещаются в соседние сообщества, если это увеличивает модульность.
- Каждое обнаруженное сообщество можно представить одной новой вершиной.
- Формируется укрупнённый граф.
- Процесс повторяется до тех пор, пока дальнейшее улучшение не становится существенным.

Лувена часто используется для анализа социальных сетей, графов взаимодействий и сетей взаимосвязанных объектов.

7.3. СПЕКТРАЛЬНАЯ КЛАСТЕРИЗАЦИЯ (SPECTRAL CLUSTERING)

Спектральная кластеризация использует матричное представление графа и собственные векторы матриц, связанных с его структурой.

- Пусть W — матрица весов рёбер, где w_{ij} отражает силу связи между вершинами i и j.
- Матрица степеней D — диагональная матрица, на диагонали которой находятся суммы весов связей каждой вершины:

$$D_{ii} = \sum_{j} w_{ij}$$

- Одна из основных матриц, применяемых в спектральной кластеризации, — графовый лапласиан:

$$L = D - W$$

Общая схема метода:

1. Построить граф сходства объектов и матрицу весов \(W\).
2. Вычислить матрицу степеней \(D\) и лапласиан \(L\).
3. Найти собственные векторы выбранной матрицы или её нормализованного варианта.
4. Представить вершины в новом пространстве признаков.
5. Применить алгоритм, например K-means, к полученному представлению.

7.4. СРАВНЕНИЕ МЕТОДОВ ГРАФОВОЙ КЛАСТЕРИЗАЦИИ

Метод	Основной принцип	Подходящие задачи
Связные компоненты	Поиск вершин, соединённых путями	Разделение несвязанных частей графа
Louvain	Максимизация модульности	Обнаружение сообществ в больших сетях
Spectral Clustering	Спектральный анализ матриц графа	Выделение групп по структуре связей и сходству

8. МОДУЛЬНОСТЬ ГРАФА И ОЦЕНКА КАЧЕСТВА СООБЩЕСТВ

Для оценки качества графовой кластеризации часто используют модульность (modularity).

Модульность показывает, насколько плотность связей внутри выделенных сообществ превышает ожидаемую плотность связей в случайной сети с сопоставимой структурой степеней вершин.

Для неориентированного взвешенного или невзвешенного графа стандартная формула имеет вид:

$$Q = \frac{1}{2m} \sum_{i,j} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta(c_i, c_j)$$

где:

- A_{ij} — вес ребра между вершинами i и j;
- k_i, k_j — степени вершин;
- m — суммарный вес рёбер, причём для невзвешенного графа это количество рёбер;
- c_i, c_j — сообщества, к которым относятся вершины;
- $\delta(c_i, c_j) = 1$, если вершины принадлежат одному сообществу, и 0 в противном случае.

ПРИМЕР

- Предположим, социальная сеть содержит 100 пользователей. Между пользователями внутри двух групп существует много связей, а между группами — мало.
- Алгоритм графовой кластеризации может выделить два сообщества. Если связи внутри сообществ значительно плотнее, чем ожидается в случайной модели, такое разбиение может иметь высокую модульность.

СРАВНЕНИЕ ИЕРАРХИЧЕСКОЙ И ГРАФОВОЙ КЛАСТЕРИЗАЦИИ

Критерий	Иерархическая кластеризация	Графовая кластеризация
Основные данные	Объекты и их признаки	Вершины и связи между ними
Основной принцип	Последовательное объединение или разделение групп	Выделение групп по структуре связей
Представление результата	Дерево кластеров, дендрограмма	Сообщества вершин графа
Основные алгоритмы	Agglomerative, Divisive	Louvain, Spectral Clustering, связные компоненты
Необходимость задавать число кластеров	Не всегда; зависит от алгоритма и способа среза	Зависит от выбранного алгоритма
Основные ограничения	Вычислительные затраты, чувствительность к метрике и правилу связи	Зависимость от структуры графа, способа построения рёбер и критерия качества
Примеры применения	Сегментация объектов, группировка документов	Социальные сети, анализ взаимодействий, сети рекомендаций

КОНТРОЛЬНЫЕ ВОПРОСЫ

- Что такое иерархическая кластеризация?
- Чем отличаются агломеративный и дивизивный подходы?
- Для чего используется дендрограмма?
- Чем отличаются Single Linkage, Complete Linkage, Average Linkage и Ward Linkage?
- Как можно выбрать количество кластеров по дендрограмме?
- Что такое графовая кластеризация?
- Что представляют собой вершины и рёбра графа?
- Что такое обнаружение сообществ?
- Как работает алгоритм Louvain?
- Какую роль играет модульность в графовой кластеризации?
- Что такое спектральная кластеризация и как используется графовый лапласиан?
- В каких случаях целесообразно использовать иерархическую кластеризацию, а в каких — графовую?

ЗАКЛЮЧЕНИЕ

- Иерархическая и графовая кластеризация являются важными методами интеллектуального анализа данных, но решают задачи группировки с разных позиций.
- Иерархическая кластеризация позволяет изучать отношения между объектами на разных уровнях детализации и представлять результат в виде дендрограммы. Графовая кластеризация ориентирована на выявление сообществ по структуре связей между объектами и широко применяется для анализа социальных и информационных сетей.
- Выбор метода определяется структурой данных, размером набора, особенностями прикладной задачи и требованиями к интерпретации результата.

РЕКОМЕНДУЕМАЯ ЛИТЕРАТУРА И РЕСУРСЫ

- Müllner, D. Modern hierarchical, agglomerative clustering algorithms. arXiv, 2011.
- Newman, M. E. J. Networks: An Introduction. Oxford University Press, 2010.
- von Luxburg, U. A Tutorial on Spectral Clustering. Statistics and Computing, 2007.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. Fast unfolding of communities in large networks. Journal of Statistical Mechanics, 2008.
- Документация Scikit-learn: Clustering.
- Документация SciPy: Hierarchical clustering.
- Документация NetworkX: Community detection.