МИНОБРНАУКИ РОССИИ

РГУ нефти и газа (НИУ) имени И.М. Губкина

|  |  |
| --- | --- |
| Факультет | **Автоматики и вычислительной техники** |
| Кафедра | **Информатики** |

**Пояснительная записка к презентации**

|  |  |
| --- | --- |
| по дисциплине | Средства интеллектуального анализа данных и машинное |
| обучение | |

|  |  |
| --- | --- |
| на тему | Кластеризация |

|  |  |  |  |
| --- | --- | --- | --- |
| «К ЗАЩИТЕ» |  | ВЫПОЛНИЛ: |  |
|  |  | Студенты группы | **АА-22-08** |
|  |  |  | (номер группы) |
| к.ф.м.н, доцент кафедры информатики, Вишневская Е.А. |  | Коргин Артём Денисович  Сафуанов Артур Ришатович | |
| (должность, ученая степень; фамилия, и.о.) |  | (фамилия, имя, отчество) | |

|  |  |  |  |
| --- | --- | --- | --- |
|  | Москва, 20 | 25 |  |

Кластеризация - её целью является найти в данных естественное группирование, при котором элементы в одном кластере будут больше похожими друг на друга, чем на элементы из других кластеров.

Данный прием относится к обучению без учителя, позволяющий обнаруживать скрытые структуры в данных, где правильный ответ заранее не известен.

**Типы кластерного анализа:**

1. ***Иерархическая кластеризация:*** Дендрограммы
2. ***Разделительная кластеризация:*** K-средних (k-means)
3. ***Плотностная кластеризация:*** DBSCAN
4. ***Сеточная кластеризация***
5. ***Спектральная кластеризация***

***Иерархическая кластеризация:*** Дендрограммы

Алгоритмы иерархической кластеризации позволяют строить дендрограммы (древовидные диаграммы с визуализацией двоичной иерархической кластеризации), которые могут помочь с интерпретацией

результатов за счет содержательной систематизации. Нам не нужно указывать количество кластеров заранее.

Существуют два основных подхода к иерархической кластеризации **агломеративный** и **дивизивный***.* В дивизивной иерархической кластеризации мы начинаем с одного кластера, который охватывает весь набор данных, и итеративно разделяем его на меньшие кластеры до тех пор, пока каждый кластер не станет содержать один образец. В свою очередь агломеративная иерархическая кластеризация имеет противоположный подход. Мы начинаем с того, что делаем каждый образец индивидуальным кластером и объединяем ближайшие пары кластеров до тех пор, пока не останется один кластер.

***Разделительная кластеризация:*** K-средних (k-means)

Алгоритм **K-Means** относится к категории кластеризации на основе прототипов. Кластеризация на основе прототипов означает, что каждый кластер представляется прототипом, который обычно будет либо центроидом (средним) подобных точек с непрерывными признаками, либо медоидом (наиболее репрезентативной точкой или точкой, сводящей к минимуму расстояние до всех остальных точек, которые принадлежат индивидуальному кластеру) в случае категориальных признаков. Наряду с тем, что алгоритм **K-Means** очень хорош в идентификации кластеров со сферической формой, одним из его недостатков является необходимость заранее указывать количество кластеров *k.*

Алгоритм **K-Means**

1. Случайным образом выбрать *k* центроидов образцов в качестве начальных центров кластеров.
2. Назначить каждый образец ближайшему центроиду µ(j), j ∈ { l, ... , k}.
3. Переместить каждый центроид в центр образцов, которые ему были назначены.
4. Повторять шаги 2 и 3 до тех пор, пока назначения кластеров не перестанут изменяться или не будет достигнуто максимальное количество из итераций.

***Плотностная кластеризация:*** DBSCAN

Давайте посмотрим на еще один подход к проведению кластеризации: основанный на плотности пространственной кластеризации для приложений с шумами. Подход **DBSCAN** не выдвигает допущений о сферических кластерах, как делает алгоритм **K-Means**, и не разделяет набор данных на иерархии, требующие указываемой вручную точки подрезания. Из самого названия следует, что кластеризация на базе плотности назначает метки кластеров, основываясь на плотных областях точек. В **DBSCAN** понятие плотности определяется как количество внутри указанного радиуса ɛ.

Алгоритм **DBSCAN**

После пометки точек как ядерных, граничных или шумовых алгоритм **DBSCAN** сводится к двум простым шагам:

1. Сформировать отдельный кластер для каждой ядерной или связной группы ядерных точек. (Ядерные считаются связными, если точки расположены не дальше чем ɛ.)
2. Назначить каждую граничную точку кластеру, к которому принадлежит соответствующая ей ядерная точка.

***Сеточная кластеризация:*** CLIQUE

Сеточная кластеризация (Grid-based Clustering) — это метод кластеризации, который делит пространство данных на ячейки (сетку) и группирует их, основываясь на плотности точек в этих ячейках. Этот подход особенно эффективен для работы с большими объемами данных и многомерными пространствами, так как оперирует не отдельными точками, а ячейками сетки, что снижает вычислительную сложность.

Алгоритм **CLIQUE**

Шаги алгоритма сеточной кластеризации (на примере простой реализации):

1. Создание сетки - разбиение пространства на ячейки.

2. Подсчет точек в каждой ячейке - для каждой точки определить, в какую ячейку она попадает.

3. Определение плотных ячеек - задать минимальный порог плотности и отметить ячейки как плотные, если они превышают порог.

4. Объединение соседних плотных ячеек - проверить соседние ячейки (по горизонтали, вертикали, диагонали) и объединить их в кластеры.

***Спектральная кластеризация***

**Спектральная кластеризация** — это метод, основанный на спектральном анализе графов, который позволяет находить кластеры произвольной формы. В отличие от **K-means** или **DBSCAN**, он работает с **матрицей схожести** данных, преобразуя задачу кластеризации в задачу разделения графа на подграфы.

Алгоритм **SpectralClustering**

1. Построение матрицы схожести между точками, используя различные метрики схожести

2. Построение матрицы Лапласа - используется для нормализации графа

3. Вычисление собственных векторов - находятся первые k собственных векторов матрицы Лапласа (где k — число кластеров). Эти векторы задают новое пространство, где кластеры лучше разделены.

4. Кластеризация в новом пространстве - собственные векторы объединяются в матрицу, и к ним применяется K-means для финальной кластеризации.

***Основные этапы кластерного анализа:***

1. Подготовка данных
2. Выбор меры близости:
   * Выбор меры близости (или метрики расстояния) в методах кластеризации определяет, как вычисляется "похожесть" или "различие" между объектами данных. Это критически важный шаг, так как от него зависит форма и качество кластеров. Разные метрики подходят для разных типов данных и задач.
   * **K-Means** по умолчанию использует евклидово расстояние. В реализации scikit-learn метрика жестко задана и не может быть изменена (алгоритм основан на минимизации внутрикластерной дисперсии).
   * **DBSCAN** позволяет выбирать метрику через параметр metric.
3. Выбор алгоритма
4. Определение числа кластеров
5. Определение числа кластеров:
   * Внутренние метрики: **Индекс силуэта**
   * Внешние метрики: **Adjusted Rand Index**

***Внутренние метрики: Индекс силуэта***

Чтобы вычислить коэффициент силуэта одиночного образца в наборе данных, мы можем выполнить описанные ниже три шага.

Вычислить связность кластера *a(i)* как среднее расстояние между образцом *x(i)* и всеми остальными точками в том же самом кластере.

Вычислить отделение кластера *b(i)* от следующего ближайшего кластера как среднее расстояние между образцом x*(i)* и всеми образцами в ближайшем кластере.

Вычислить силуэт s*(i)* как разность между связностью и отделением кластера, деленную на большее среди них значение.

Коэффициент силуэта ограничен диапазоном от -1 до 1. Основываясь на предыдущем уравнении, мы можем отметить, что коэффициент силуэта равен 0 в случае равенства отделения и связности кластера (b*(i)* = *a(i)).* Кроме того, мы приближаемся к идеальному коэффициенту силуэта, если *b(i)* >> *a(i),* т.к. b*(i)* количественно определяет, насколько образец не похож на образцы из других кластеров, и *a(i)* сообщает о том, в какой степени он похож на остальные образцы в своем кластере.

***Внешние метрики: Adjusted Rand Index***

**Adjusted Rand Index (ARI)** — это метрика для оценки качества кластеризации, которая сравнивает предсказанные кластеры с истинными (эталонными) метками. Она корректирует Rand Index (**Rand Index (RI)** измеряет процент пар точек, которые согласуются в истинной и предсказанной кластеризации) для учета случайного совпадения кластеризаций, что делает ее более интерпретируемой.

* + ARI = 1: Идеальное совпадение кластеризаций.
  + ARI = 0: Случайное совпадение.
  + ARI < 0: Хуже случайного (редко встречается на практике).

***Проблемы и ограничения***

1. **Субъективность выбора числа кластеров**

Выбор оптимального числа кластеров — одна из ключевых проблем в кластеризации. Несмотря на существование методов, таких как локтевой метод, силуэтный анализ или индекс Калински-Харабаса, ни один из них не даёт однозначного ответа.

* Локтевой метод анализирует изменение суммы внутрикластерных расстояний при увеличении числа кластеров. «Локоть» на графике — точка, после которой добавление новых кластеров не даёт значимого улучшения. Однако определение положения этого «локтя» часто субъективно: визуально он может выглядеть как плавный изгиб, что приводит к разным интерпретациям.
* Силуэтный коэффициент оценивает компактность и разделимость кластеров. Максимальное значение силуэта считается оптимальным, но на практике максимум может соответствовать слишком детализированному или, наоборот, обобщённому разбиению.
* Экспертная оценка играет ключевую роль. Например, в маркетинге выбор числа кластеров может зависеть от бизнес-задач: выделение 3–5 сегментов часто удобнее для интерпретации, даже если математически оптимальным является 7 кластеров.

Таким образом, число кластеров остаётся компромиссом между математической строгостью, визуальной логикой и предметной целесообразностью.

1. **Чувствительность к начальным условиям (на примере K-means)**

K-means — алгоритм, который сильно зависит от начального положения центроидов. Это связано с его итеративной природой: центры кластеров обновляются на основе средних значений точек, что может приводить к попаданию в локальные минимумы.

* Пример: Если начальные центроиды случайно попадут в плотную область одного кластера, алгоритм может объединить два естественных кластера в один или, наоборот, разделить один кластер на части.
* K-means++ — улучшенная инициализация, которая уменьшает эту проблему, выбирая начальные центры так, чтобы они находились далеко друг от друга. Однако даже это не гарантирует полной устойчивости.
* Решение: Многократный запуск алгоритма с разными начальными условиями и выбор лучшего результата по критерию инерции (сумме квадратов расстояний).

Эта чувствительность делает K-means менее надёжным для данных с неочевидной структурой или большим количеством шума.

1. **Проблема «шума» в данных**

Шумовые точки (выбросы) могут серьёзно исказить результаты кластеризации, особенно в алгоритмах, которые предполагают, что все точки принадлежат какому-либо кластеру.

* В K-means выбросы «притягивают» центроиды, смещая их положение. Например, один удалённый выброс может создать отдельный кластер, даже если он не имеет смысла.
* DBSCAN и другие алгоритмы, основанные на плотности, более устойчивы. Они помечают выбросы как шум (метка -1), не включая их в кластеры. Однако настройка параметров eps и min\_samples критична: слишком маленький eps приведёт к тому, что разреженные области будут считаться шумом, а слишком большой — к объединению разных кластеров.
* Стратегии борьбы с шумом:
  + Предварительная обработка данных: удаление выбросов, нормализация.
  + Использование алгоритмов, устойчивых к шуму (DBSCAN, HDBSCAN, OPTICS).
  + Анализ кластеров на устойчивость: если удаление части данных меняет результат, кластеризация ненадёжна.

Шум особенно опасен в задачах, где кластеры используются для принятия решений (например, сегментация клиентов), так как ложные кластеры могут привести к ошибочным выводам.

1. **Интерпретируемость результатов**

**Проблема:** Кластеры могут быть статистически корректными, но бессмысленными в реальном контексте.

**Причины:**

* Использование нерелевантных признаков (например, группировка по случайным или шумовым данным).
* Отсутствие связи с предметной областью (алгоритм выявляет математические, а не содержательные паттерны).
* Переобучение на шум или артефакты данных.

**Решение:**

1. Отбор признаков — фокусироваться на показателях, значимых для задачи.
2. Экспертная валидация — проверять, соответствуют ли кластеры реальным категориям.
3. Анализ характеристик — сравнивать средние значения, распределения признаков внутри кластеров.
4. Объяснимые алгоритмы — использовать методы, которые формируют прозрачные правила (например, иерархическая кластеризация).

***Заключение***

Кластерный анализ остается одним из фундаментальных инструментов в исследовании данных. Его ключевые этапы включают:

1. Выбор метода (K-means, DBSCAN, иерархическая кластеризация) в зависимости от структуры данных и целей.
2. Предобработку данных (нормализация, борьба с шумом), критичную для качества результатов.
3. Оценку кластеров с использованием метрик (Adjusted Rand Index, силуэтный коэффициент) и экспертной интерпретации.

Кластеризация находит применение в самых разных областях: от сегментации клиентов в маркетинге до анализа геномных данных в биологии. Её ценность заключается в способности выявлять скрытые паттерны и структурировать информацию без явных меток.