Кластерный анализ

- Проблемы кластерного анализа
- Сущность кластерного анализа
- Разновидности кластерного анализа

Прежде чем перейти к обсуждению основных методологических этапов проведения кластерного анализа, необходимо сделать несколько предостережений общего характера.

1. *Многие методы кластерного анализа – довольно простые процедуры, которые, как правило, не имеют достаточного статистического обоснования.* Другими словами, большинство методов кластерного анализа являются эвристическими (подкрепленными лишь опытом разработчиков). Они - не более чем правдоподобные алгоритмы, используемые для создания кластеров объектов. В этом резкое отличие, например, от методов факторного анализа, который хорошо обоснован статистически. Хотя многие кластерные методы обладают важными, подробно исследованными математическими свойствами (см. Jardin and Sibson, 1971), все же важно сознавать их простоту. В этом случае маловероятно, что пользователь допустит ошибку при трактовке результата кластерного анализа.

2. *Методы кластерного анализа разрабатывались для многих научных дисциплин, а потому несут на себе отпечатки специфики этих дисциплин.* Это важно отметить, потому что каждая дисциплина предъявляет свои требования к отбору данных, к форме их представления, к предполагаемой структуре классификации. Что может быть полезным в психологии, может оказаться ненужным для биологов, а так как кластерные методы порой не более чем правила для создания групп, то пользователь должен знать те особенности, которые часто сопровождают обсуждение и описание методов кластеризации.

3. *Разные кластерные методы могут порождать и порождают различные решения для одних и тех же данных.* Это обычное явление в большинстве прикладных исследований. Одной из причин неодинаковых решений является то, что кластерные методы получены из разных источников, которые предопределяли использование различных правил формирования групп. Данная ситуация вносит в работу с кластерным анализом путаницу не только для начинающих, но и для опытных пользователей. Кроме того, желательно иметь специальную методику, позволяющую проверить, насколько “естественны” группы, выделенные методом кластеризации в наборе данных. Было разработано несколько процедур, способных помочь в решении этой задачи.

4. *Цель кластерного анализа заключается в поиске существующих структур. В то же время его действие состоит в привнесении структуры в анализируемые данные,* т.е. методы кластеризации необходимы для обнаружения структуры в данных, которую нелегко найти при визуальном обследовании или с помощью экспертов. Эта ситуация отличается от ситуации дискриминантного анализа, который более точно определяется как процедура идентификации. Последний приписывает объекты к уже существующим группам, а не создает новые группы. Хотя цель кластеризации и заключается в нахождении структуры, на деле кластерный метод привносит структуру в данные и эта структура может не совпадать с искомой, “реальной”. Кластерный метод всегда размещает объекты по группам, которые могут радикально различаться по составу, если применяются различные методы кластеризации. Ключом к использованию кластерного анализа является умение отличать “реальные” группировки от навязанных методом кластеризации данных.

Кластер–анализ – это способ группировки многомерных объектов, основанный на представлении результатов отдельных наблюдений точками подходящего геометрического пространства с последующим выделением групп, как “сгустков” этих точек.

Другими словами, кластерный анализ – анализ неклассифицированных объектов, проводимый с целью выделения структур, классов образов, множеств подобных объектов,

таких что объекты внутри групп были бы похожи в некотором смысле друг на друга, а объекты из разных групп - непохожи.

Термин кластер (cluster) с английского языка переводится как “сгусток”, “гроздь”, “скопление” и т.д. Этот термин весьма удачно вписался в научную терминологию, поскольку его первый слог соответствует традиционному термину “класс”, а второй как производная от первого.

При решении задачи кластерного анализа молчаливо принимается, что, во-первых, выбранное множество объектов в принципе допускают желательное разбиение на кластеры, во-вторых, единицы измерения (масштаб) выбраны правильно. Первая проблема называется проблемой выбора свойств или характеристик объектов. Вообще предполагается, что проблема выбора характеристик решена до начала процесса кластеризации. Однако следует предупредить, что этим вносится некоторый произвол, что в отдельных случаях требует дополнительного рассмотрения.

Другой вопрос, который всегда сопутствует измерению, - выбор масштаба также играет большую роль. Как правило, данные нормализуют вычитанием среднего и делением на стандартное отклонение; так что дисперсия оказывается равной единице. В случае же, когда исходят из непосредственных (обычных) единиц измерения, возникает проблема интерпретации. Однако наиболее серьезная проблема возникает в связи с тем, что разбиение на кластеры зависит от выбора масштаба. Было бы желательно иметь такой метод кластеризации, который был бы инвариантен к изменению масштабов измерения.

Определение кластерного анализа в различных литературных источниках дается по-разному. Одной из причин этого является то обстоятельство, что кластерный анализ используется и совершенствуется в столь различных областях, как социология, психология, экология, геология, медицина и т.д. В кластерном анализе не существует однозначного количественного критерия, поскольку в различных прикладных задачах различными могут быть и цели анализа. Иногда необходимо выделить группы с высокой плотностью распределения и малой дисперсией, а иногда необходимо обнаружить связанные точные структуры.

В литературе описаны сотни методов кластеризации. Кластерный анализ тесно связан с исследованием процесса порождения данных и важную роль играет спецификация создания того или иного метода.

Рассмотрим три разновидности кластерного анализа.

Первая группа методов основана на отыскании моды распределения. Идея этих методов состоит в том, что кластерам соответствуют максимум плотности распределения данных.

Поэтому здесь необходимо оценивать плотность распределения и отыскать все максимумы, соответствующие модам. Каждый максимум плотности распределения соответствует некоторому кластеру. Каждый объект относится к одному из кластеров. Естественно, что некоторые кластеры могут быть объединены. Результаты кластеризации при этом тесно связаны со способом оценивания плотности распределения. Число мод может изменяться от одной (при большом коэффициенте сглаживания) до значения, равного числу объектов (при малом коэффициенте сглаживания). Для выбора коэффициента сглаживания можно воспользоваться априорными (экспертными) данными о числе кластеров.

Если число кластеров известно, можно воспользоваться второй разновидностью анализа - методом, в котором в качестве критерия используется отношение внутрикластерной дисперсии к межкластерной дисперсии. Могут использоваться и другие критерии. Как явный представитель такого подхода во многих источниках описывается итеративный самоорганизующийся метод анализа данных - ИСОМАД. В его основу положена идея о том, что объекты принадлежат кластеру с наиболее близким средним значением. Результат кластеризации может в значительной степени зависеть от выбора начальных средних значений для кластеров. Ниже приводится принцип работы такого алгоритма.

И, наконец, можно отметить иерархические схемы кластеризации. При их использовании обычно вначале каждый объект рассматривается как отдельный кластер. Затем два ближайших кластера объединяются и т.д. Этот шаг повторяется до получения единственного кластера. Результат работы такого подхода зависит от выбора расстояния или меры близости между кластерами. Существует много способов задания расстояний, которые отдельно рассматриваются ниже.