Лекция 5. МЕТОДЫ И АЛГОРИТМЫ РЕШЕНИЯ ЗАДАЧ КЛАСТЕРИЗАЦИИ

Кластерный анализ (англ. cluster analysis) — многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.

Кластерный анализ является математическим методом, предназначенным для решения задач классификации; его цель состоит в разделении эмпирической выборки на ряд подмножеств (обычно непересекающихся), которые называются кластерами, а иногда – группами, классами, таксонами. Термин «кластер» (от англ. cluster) означает «гроздь, пучок, скопление, группа элементов, характеризуемых каким-либо общим свойством»; а термин «таксон» (от англ. taxon - систематизированная группа любой категории; термин биологического происхождения. Подразумевается, что элементы, объединенные в один кластер, являются в том или ином смысле более похожими друг на друга, более близкими по каким-либо характеристикам по сравнению с остальными. При этом кластерный анализ опирается, как правило, только на данные о самой эмпирической выборке и не использует никаких дополнительных априорных предположений: например, о характере распределения вероятностей в генеральной совокупности. Более того, после построения классификации ее результаты считаются окончательными и не пересматриваются (для данной эмпирической выборки данных и примененного конкретного метода кластерного анализа, хотя при получении дополнительных данных или при выборе другого метода классификация, естественно, может быть построена заново)

Кластерный анализ выполняет следующие основные задачи:

- Разработка типологии или классификации.
- Исследование полезных концептуальных схем группирования объектов.
- Порождение гипотез на основе исследования данных.
- Проверка гипотез или исследования для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Независимо от предмета изучения применение кластерного анализа предполагает следующие этапы:

- Отбор выборки для кластеризации. Подразумевается, что имеет смысл кластеризовать только количественные данные.
- Определение множества переменных, по которым будут оцениваться объекты в выборке, то есть признакового пространства.
- Вычисление значений той или иной меры сходства (или различия) между объектами.
- Применение метода кластерного анализа для создания групп сходных объектов.
- Проверка достоверности результатов кластерного решения.

Можно встретить описание двух фундаментальных требований, предъявляемых к данным — однородность и полнота. Однородность требует, чтобы все кластеризуемые сущности были одной природы, описывались сходным набором характеристик. Если кластерному анализу предшествует факторный анализ, то выборка не нуждается в «ремонте» — изложенные требования выполняются автоматически самой процедурой факторного моделирования (есть ещё одно достоинство — z-стандартизация без негативных последствий для выборки; если её проводить непосредственно для кластерного анализа, она может повлечь за собой уменьшение чёткости разделения групп). В противном случае выборку нужно корректировать.

Кластеризация отличается от классификации тем, что для проведения анализа требуется иметь выделенную целевую переменную, с этой точки зрения относится к классу unsupervised learning. Эта задача решается на начальных этапах исследования, когда о данных мало что известно. Ее решение помогает лучше понять данные, и с этой точки зрения задача кластеризации является описательной задачей (descriptive).

Для этапа кластеризации характерно отсутствие каких-либо различий как между переменными, так и между записями. Напротив, ищутся группы наиболее близких, похожих записей. Методы автоматического разбиения на кластеры редко используются сами по себе, просто для получения групп схожих объектов. Анализ только начинается с разбиения на кластеры. После определения кластеров используются другие методы Data Mining, для того чтобы попытаться

установить, а что означает такое разбиение на кластеры, чем оно вызвано.

Большое достоинство кластерного анализа в том, что он позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ, в отличие от большинства математико-статистических методов, не накладывает никаких ограничений на вид рассматриваемых объектов и позволяет рассматривать множество исходных данных практически произвольной природы. Это имеет большое значение, например, для прогнозирования конъюнктуры при наличии разнородных показателей, затрудняющих применение традиционных эконометрических подходов.

Кластерный анализ позволяет рассматривать достаточно большой объем информации и резко сокращать, сжимать большие массивы информации, делать их компактными и наглядными.

Задача кластеризации состоит в разделении исследуемого множества объектов на группы "похожих" объектов, называемых кластерами. Слово кластер английского происхождения (cluster), переводится как сгусток, пучок, группа. Родственные понятия, используемые в литературе, — класс, таксон, сгущение. Часто решение задачи разбиения множества элементов на кластеры называют кластерным анализом.

Решением задачи классификации является отнесение каждого из объектов данных к одному (или нескольким) из заранее определенных классов и построение, в конечном счете, одним из методов классификации модели данных, определяющей разбиение множества объектов данных на классы. В задаче кластеризации отнесение каждого из объектов данных осуществляется к одному (или нескольким) из заранее неопределенных классов. Разбиение объектов данных по кластерам осуществляется при одновременном их формировании. Определение кластеров и разбиение по ним объектов данных выражается в итоговой модели данных, которая является решением задачи кластеризации.

Отметим ряд особенностей, присущих задаче кластеризации. Во-первых, решение сильно зависит от природы объектов данных (и их атрибутов). Так, с одной стороны, это могут быть однозначно определенные, четко количественно очерченные объекты, а с другой — объекты, имеющие вероятностное или нечеткое описание.

Во-вторых, решение значительно зависит также и от представления классов (кластеров) и предполагаемых отношений объектов данных и классов (кластеров). Так, необходимо учитывать такие свойства, как возможность/невозможность принадлежности объектов нескольким классам (кластерам). Необходимо определение самого понятия принадлежности классу (кластеру): однозначная (принадлежит/не принадлежит), вероятностная (вероятность принадлежности), нечеткая (степень принадлежности).

Ввиду особого положения задачи кластеризации в списке задач интеллектуального анализа данных было разработано множество способов ее решения. Один из них — построение набора характеристических функций классов, которые показывают, относится ли объект данных к данному классу или нет. Характеристическая функция класса может быть двух типов:

1. дискретная функция, принимающая одно из двух определенных значений, смысл которых в принадлежности/непринадлежности объекта данных заданному классу;
2. функция, принимающая вещественные значения, например, из интервала 0...1. Чем ближе значение функции к единице, тем больше объект данных принадлежит заданному классу.

Общий подход к решению задачи кластеризации стал возможен после развития Л. Заде теории нечетких множеств. В рамках данного подхода удается формализовать качественные понятия, неопределенность, присущую реальным данным и процессам. Успех этого подхода объясняется еще и тем, что в процессе анализа данных участвует человек, оценки и суждения которого расплывчаты и субъективны. Уместно привести высказывание Л. Заде, основоположника теории нечетких множеств: "...нужна новая точка зрения, новый комплекс понятий и методов, в которых нечеткость принимается как универсальная реальность человеческого существования".

Применяя теорию нечетких множеств для решения задачи кластеризации, возможны различные варианты введения нечеткости в методы, решающие данную задачу. Нечеткость может учитываться как в представлении данных, так и при описании их взаимосвязи. Кроме того, данные могут как обладать, так и не обладать количественной природой. Тем не менее, во многих практических задачах данные, которые необходимо исследовать, являются результатом накопленного опыта в той или иной сфере человеческой

деятельности и часто имеют количественное представление. Учет нечеткости самих исследуемых данных, в общем случае, — серьезная проблема. Поэтому как в существующих алгоритмах, так и в подходе, предлагаемом в данном издании, не делается никаких допущений о нечеткости самих исходных данных. Считается, что данные являются четкими и выражены количественно.

Описывать нечеткие взаимосвязи данных можно разными способами. Одним из таких способов, нашедших широкое распространение в используемых в настоящее время алгоритмах нечеткой кластеризации данных, является описание взаимосвязи данных через их отношение к некоторым эталонным образцам — центрам кластеров. В данных алгоритмах нечеткость проявляется в описании кластеров как нечетких множеств, имеющих ядро в центре кластера. С другой стороны, взаимосвязь данных в условиях неопределенности можно учитывать при помощи аппарата нечетких отношений между отдельными образцами данных, не прибегая при этом к понятию центра кластера. Такой подход не нашел еще широкого распространения на практике, хотя, очевидно, является более универсальным. В данном издании делается попытка восполнить этот пробел и показать те преимущества, которые дает использование указанного понятия для задачи кластеризации. Итак, перейдем к постановке задачи кластеризации.

Понятие меры близости. Большинство алгоритмов кластерного анализа полностью исходит из матрицы расстояний (или близостей) либо требует вычисления отдельных ее элементов, поэтому если данные представлены в форме X, то первым этапом решения задачи поиска кластеров будет выбор способа вычисления расстояний, или близости, между объектами или признаками.

Относительно проще решается вопрос об определении близости между признаками. Как правило, кластерный анализ признаков преследует те же цели, что и факторный анализ – выделение групп связанных между собой признаков, отражающих определенную сторону изучаемых объектов. Мерами близости в этом случае служат различные статистические коэффициенты связи

В кластерном анализе для количественной оценки близости вводится понятие метрики. Сходство и различие между классифицируемыми объектами устанавливается в зависимости от метрического расстояния между ними. Если каждый объект описывается k признаками, то он может быть представлен как точка в k-мерном пространстве. Сходство с другими объектами будет

определяться как соответствующее расстояние. В кластерном анализе используют различные меры расстояния между объектами. Существует множество метрик, вот лишь основные из них:

1. Евклидово расстояние. Наиболее распространенная функция расстояния. Представляет собой геометрическим расстоянием в многомерном пространстве:

$$\rho(x, x') = \sqrt{\sum_{i}^{n} (x_i - x_i')^2}$$

2. Квадрат евклидова расстояния. Применяется для придания большего веса более отдаленным друг от друга объектам. Это расстояние вычисляется следующим образом:

$$\rho(x, x') = \sum_{i}^{n} (x_i - x_i')^2$$

3. Расстояние городских кварталов (манхэттенское расстояние). Это расстояние является средним разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако для этой меры влияние отдельных больших разностей (выбросов) уменьшается (т.к. они не возводятся в квадрат). Формула для расчета манхэттенского расстояния:

$$\rho(x, x') = \sum_{i}^{n} x_i - x_i' \vee$$

4. Расстояние Чебышева. Это расстояние может оказаться полезным, когда нужно определить два объекта как «различные», если они различаются по какой-либо одной координате. Расстояние Чебышева вычисляется по формуле:

$$\rho(x, x') = max\left(\left|x_i - x_i'\right|\right)$$

5. Степенное расстояние. Применяется в случае, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются. Степенное расстояние вычисляется по следующей формуле:

$$\rho(x, x') = r\sqrt{\sum_{i}^{n} (x_i - x_i')^p}$$

где r и p – параметры, определяемые пользователем. Параметр p ответственен за постепенное взвешивание разностей по отдельным

координатам, параметр r ответственен за прогрессивное взвешивание больших расстояний между объектами. Если оба параметра – r и p — равны двум, то это расстояние совпадает с расстоянием Евклида.

Выбор метрики полностью лежит на исследователе, поскольку результаты кластеризации могут существенно отличаться при использовании разных мер.

Классификация алгоритмов кластеризации. Можно выделить две основные классификации алгоритмов кластеризации

1. Иерархические и плоские. Иерархические алгоритмы (также называемые алгоритмами таксономии) строят не одно разбиение выборки на непересекающиеся кластеры, а систему вложенных разбиений. Т.е. на выходе мы получаем дерево кластеров, корнем которого является вся выборка, а листьями — наиболее мелкие кластера. Плоские алгоритмы строят одно разбиение объектов на кластеры.

2. Четкие и нечеткие. Четкие (или непересекающиеся) алгоритмы каждому объекту выборки ставят в соответствие номер кластера, т.е. каждый объект принадлежит только одному кластеру. Нечеткие (или пересекающиеся) алгоритмы каждому объекту ставят в соответствие набор вещественных значений, показывающих степень отношения объекта к кластерам. Т.е. каждый объект относится к каждому кластеру с некоторой вероятностью.

Так же выделяются методы по способу обработки данных:

1. Иерархические методы:

   Агломеративные методы AGNES (Agglomerative Nesting):

   - CURE;
   - ROCK;
   - CHAMELEON и т.д.

   Дивизимные методы DIANA (Divisive Analysis):

   - BIRCH;
   - MST и т.д.

2. Неиерархические методы.

   Итеративные

   - K-средних (k-means)
   - PAM (k-means + k-medoids)
   - CLOPE
   - LargeItem и т.д.

3. Методы по способу анализа данных:

   - Четкие;

- Нечеткие.

4. Методы по количеству применений алгоритмов кластеризации:
   - С одноэтапной кластеризацией;
   - С многоэтапной кластеризацией.
5. Методы по возможности расширения объема обрабатываемых данных:
   - Масштабируемые;
   - Немасштабируемые.
6. Методы по времени выполнения кластеризации [1]:
   - Потоковые (on-line);
   - Не потоковые (off-line).

Алгоритм к-средних. Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции.

Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга.

В качестве меры близости используется Евклидово расстояние:

$$\rho(x,y)=\|x-y\|=\sqrt{\sum_{p=1}^{n}(x_p-y_p)^2}, \text{где } x,y\in R^n$$

Описание алгоритма

1. Первоначальное распределение объектов по кластерам.

Выбирается число k, и на первом шаге эти точки считаются "центрами" кластеров. Каждому кластеру соответствует один центр. Выбор начальных центроидов может осуществляться следующим образом:

- выбор k-наблюдений для максимизации начального расстояния;
- случайный выбор k-наблюдений;
- выбор первых k-наблюдений.

В результате каждый объект назначен определенному кластеру.

2. Итеративный процесс.

Вычисляются центры кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются.

Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий:

- кластерные центры стабилизировались, т.е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации;
- число итераций равно максимальному числу итераций.

На рис. 5.1 приведен пример работы алгоритма k-средних для k, равного двум.

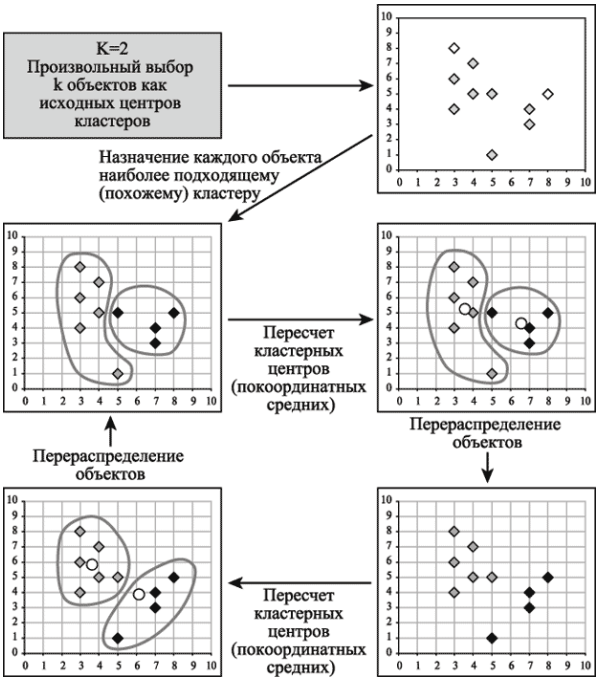

Рис. 5.1. Пример работы алгоритма k-средних (k=2)

Выбор числа кластеров является сложным вопросом. Если нет предположений относительно этого числа, рекомендуют создать 2 кластера, затем 3, 4, 5 и т.д., сравнивая полученные результаты.

Наиболее популярным алгоритмом нечеткой кластеризации является алгоритм c-средних (c-means). Он представляет собой модификацию метода k-средних. Шаги работы алгоритма:

1. Выбрать начальное нечеткое разбиение *n* объектов на *k* кластеров путем выбора матрицы принадлежности *U* размера *n x k*.

2. Используя матрицу U, найти значение критерия нечеткой ошибки:

$$E^2(X,U)=\sum_{i=1}^{N}\sum_{k=1}^{K}U_{ik}\left\|x_i^{(k)}-c_k\right\|^2$$

где c_k — «центр масс» нечеткого кластера *k*:

$$c_k=\sum_{i=1}^{N}U_k x_i$$

3. Перегруппировать объекты с целью уменьшения этого значения критерия нечеткой ошибки.

4. Возвращаться в п. 2 до тех пор, пока изменения матрицы *U* не станут незначительными.

Этот алгоритм может не подойти, если заранее неизвестно число кластеров, либо необходимо однозначно отнести каждый объект к одному кластеру.

Оценки качества моделей. Основная задача кластеризации формулируется так: разделить объекты на группы таким образом, чтобы объекты одной группы имели большое сходство, а сходство между объектами разных групп было малым. Согласно приведенного определения формулируются основные критерии качества кластеризации: компактность, отделимость и, редко используемый, концентрация.

Компактность означает, что элементы одного кластера должны быть как можно ближе друг к другу (обладать высокой степенью сходства). Это свойство можно выразить через расстояния между элементами в кластере, плотностью внутри кластера или же объемом, занимаемым кластером в пространстве.

Свойство отделимости значит, что элементы разных кластеров должны быть как можно дальше друг от друга (обладать низкой степенью сходства). Расстояние между кластерами обычно измеряется одним из трех способов:

- расстояние между ближайшими элементами кластеров;
- расстояние между наиболее удаленными элементами кластеров;
- расстояние между кластерными центрами.

Концентрация означает, что элементы кластера должны быть сконцентрированы вокруг центра кластера. Этот пункт используется гораздо реже, потому что далеко не во всех алгоритмах кластеризации используется понятие центра кластера.

Для самих показателей качества кластеризации обычно вводят следующую классификацию: внешние, внутренние и относительные. К внутренним показателям относятся те, которые учитывают априорную информацию о структуре кластеров в рассматриваемом множестве данных. К внешним относят показатели, которые не имеют априори знаний о структуре классов и при оценке опираются только на ту информацию, которую можно получить из самого разбиения. Относительные показатели оценивают качество, сравнивая несколько кластерных структур между собой, не имея априорной информации.

Качество кластеризации — степень приближения результата кластеризации к идеальному решению. Для большинства задач идеальное решение неизвестно. Оценка качества кластеризации может быть произведена двумя способами:

- Формальный способ. Формальный способ основан на определении формальных критериев. Наилучшим считается решение, для которого значение формального критерия максимально.
- Экспертный способ. Решение оценивается специалстами заданной предметной области.

Основные этапы оценки качества кластеризации:

1. Алгоритм кластеризации, построение модели данных.
2. Вычисление критерия качества кластеризации. Критерии вычисляются на основе получившейся в ходе работы алгоритма кластеризации матрицы принадлежности и/или множества кластерных центров.
3. Определение параметров настройки алгоритма.

Критерии качества:

1. Показатели четкости: коэффициент разбиения, модифицированный коэффициент разбиения, индекс четкости.
2. Энтропийные критерии: энтропия разбиения, модифицированная энтропия.
3. Показатель компактности и изолированности
4. Индекс эффективности

Можно выделить несколько подходов к валидации кластеров:

- внешняя валидация, которая заключается в сравнении итогов кластерного анализа с заранее известным результатом (т.е. метки кластеров известны априори);
- относительная валидация, которая оценивает структуру кластеров, изменяя различные параметры одного и того же алгоритма (например, число групп kk);
- внутренняя валидация, которая использует внутреннюю информацию процесса объединения в кластеры (если внешняя информация отсутствует);
- оценка стабильности объединения в кластеры (или специальная версия внутренней валидации), использующая методы ресэмплинга.

Одна из проблем машинного обучения без учителя состоит в том, что методы кластеризации будут формировать группы, даже если анализируемый набор данных представляет собой полностью случайную структуру. Поэтому первой задачей валидации, которую рекомендуется выполнить перед началом кластерного анализа, является оценка общей предрасположенности имеющихся данных к объединению в кластеры (clustering tendency).