УДК 004.622

КЛАСТЕРНЫЙ АНАЛИЗ, МЕТОДЫ И АЛГОРИТМЫ КЛАСТЕРИЗАЦИИ

Тюрин А.Г., к.т.н., E-mail: tyurin@mirea.ru
Зуев И.О., студент,
МГТУ МИРЭА, Москва, Россия

Аннотация. В статье приведен обзор методов кластерного анализа с их особенностями, достоинствами и недостатками.
Ключевые слова. кластерный анализ, кластер, алгоритм, нейронная сеть, поиск, данные, устойчивость, однородность.

CLUSTER ANALYSIS, METHODS AND ALGORITHMS OF THE CLUSTERING

Tyurin A.G., Ph.D., E-mail: tyurin@mirea.ru
Zuyev I.O., student,
MSTU MIREA, Moscow, Russia

Abstract. The review of methods of the cluster analysis is provided in article with their features, merits and demerits.
Keywords. cluster analysis, cluster, algorithm, neural network, search, data, stability, uniformity.

Для «интеллектуальной» группировки результатов при поиске файлов, веб-сайтов, других объектов, используется Кластерный анализ. Он предоставляет пользователю возможность быстрой навигации, выбора заведомо более релевантного подмножества и исключения заведомо менее релевантного — что может повысить юзабилити интерфейса по сравнению с выводом в виде простого сортированного по релевантности списка.

Кластерный анализ — задача разбиения заданной выборки объектов (ситуаций) на подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Задача кластеризации относится к статистической обработке, а также к широкому классу задач обучения без учителя.

Кластерный анализ — это многомерная статистическая процедура, выполняющая сбор данных, содержащих информацию о выборке объектов, и затем упорядочивающая объекты в сравнительно однородные группы (кластеры). Кластер — группа элементов, характеризуемых общим свойством, главная цель кластерного анализа — нахождение групп схожих объектов в выборке. Спектр применений кластерного анализа очень

широк: его используют в археологии, медицине, психологии, химии, биологии, государственном управлении, филологии, антропологии, маркетинге, дистанционном зондировании и других дисциплинах [1-4].

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.
- Определение множества переменных, по которым будут оцениваться объекты в выборке.
- Вычисление значений той или иной меры сходства между объектами.
- Применение метода кластерного анализа для создания групп сходных объектов.
- Проверка достоверности результатов кластерного решения.

Кластерный анализ предъявляет следующие требования к данным: во-первых, показатели не должны коррелировать между собой; во-вторых, показатели должны быть безразмерными; в-третьих, их распределение должно быть близко к нормальному; в-четвёртых, показатели должны отвечать требованию «устойчивости», под которой понимается отсутствие влияния на их значения случайных факторов; в-пятых, выборка должна быть однородна, не содержать «выбросов». Если кластерному анализу предшествует факторный анализ, то выборка не нуждается в «ремонте» — изложенные требования выполняются автоматически самой процедурой факторного моделирования. В противном случае выборку нужно корректировать.

Методы кластерного анализа

Общепринятой классификации методов кластеризации не существует, но можно выделить ряд групп подходов (некоторые методы можно отнести сразу к нескольким группам и потому предлагается рассматривать данную типизацию как некоторое приближение к реальной классификации методов кластеризации).

Вероятностный подход. Предполагается, что каждый рассматриваемый объект относится к одному из классов. Некоторые авторы (например, А. И. Орлов) [5] считают, что данная группа вовсе не относится к кластеризации и противопоставляют

её под названием «дискриминация», то есть выбор отнесения объектов к одной из известных групп (обучающих выборок).

Метод k-средних

Наиболее популярный метод кластеризации. Был изобретён в 1950-х годах математиком Гуго Штейнгаузом и почти одновременно Стюартом Ллойдом [6].

Действие алгоритма таково, что он стремится минимизировать суммарное квадратичное отклонение точек кластеров от центров этих кластеров:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 \qquad (1)$$

где k — число кластеров, S_i — полученные кластеры, $i = 1, 2, \ldots, k$ и μ_i — центры масс векторов $x_j \in S_i$.

Алгоритм представляет собой версию EM-алгоритма, применяемого также для разделения смеси гауссиан. Он разбивает множество элементов векторного пространства на заранее известное число кластеров k.

Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике.

Алгоритм завершается, когда на какой-то итерации не происходит изменения центра масс кластеров. Это происходит за конечное число итераций, так как количество возможных разбиений конечного множества конечно, а на каждом шаге суммарное квадратичное отклонение V уменьшается, поэтому зацикливание невозможно.

Недостатки:

- Не гарантируется достижение глобального минимума суммарного квадратичного отклонения V, а только одного из локальных минимумов.
- Результат зависит от выбора исходных центров кластеров, их оптимальный выбор неизвестен.
- Число кластеров надо знать заранее.

Метод K - medians

Это вариация k-means метода кластеризации, где для определения центроида кластера вместо среднего вычисляется медиана. Это соответствует минимизации ошибки по всем кластерам в метрике с 1-нормой, вместо метрики с 2-нормой для k-means.

Соответствующая проблема k-median состоит в поиске таких k центров, что сформированные по ним кластеры будут наиболее компактными. Формально, при заданных точках данных x, k центров ci должны быть выбраны так, чтобы минимизировать сумму расстояний от каждой x до ближайшего ci.

K-medians иногда работает лучше чем k-means, где минимизируется сумма квадратов расстояний. Критерий суммы расстояний широко используется для транспортных задач.

EM-алгоритм

Алгоритм, используемый в математической статистике для нахождения оценок максимального правдоподобия параметров вероятностных моделей, в случае, когда модель зависит от некоторых скрытых переменных. Каждая итерация алгоритма состоит из двух шагов. На E-шаге (expectation) вычисляется ожидаемое значение функции правдоподобия, при этом скрытые переменные рассматриваются как наблюдаемые. На M-шаге (maximization) вычисляется оценка максимального правдоподобия, таким образом увеличивается ожидаемое правдоподобие, вычисляемое на E-шаге. Затем это значение используется для E-шага на следующей итерации. Алгоритм выполняется до сходимости.

Пусть $\mathbf{X}$ — некоторые из значений наблюдаемых переменных, а $\mathbf{T}$ — скрытые переменные. Вместе $\mathbf{X}$ и $\mathbf{T}$ образуют полный набор данных. Вообще, $\mathbf{T}$ может быть некоторой подсказкой, которая облегчает решение проблемы в случае, если она известна. Например, если имеется смесь распределений, функция правдоподобия легко выражается через параметры отдельных распределений смеси.

Положим p — плотность вероятности (в непрерывном случае) или функция вероятности (в дискретном случае) полного набора данных с параметрами Θ: $p(\mathbf{X}, \mathbf{T}|\Theta)$. Эту функцию можно понимать как правдоподобие всей модели, если рассматривать её как функцию параметров Θ. Заметим, что условное распределение скрытой компоненты при некотором наблюдении и фиксированном наборе параметров может быть выражено так:

$$p(\mathbf{T}|\mathbf{X},\Theta) = \frac{p(\mathbf{X},\mathbf{T}|\Theta)}{p(\mathbf{X}|\Theta)} = \frac{p(\mathbf{X}|\mathbf{T},\Theta)p(\mathbf{T}|\Theta)}{\int p(\mathbf{X}|\hat{\mathbf{T}},\Theta)p(\hat{\mathbf{T}}|\Theta)d\hat{\mathbf{T}}} \qquad (2)$$

используя расширенную формулу Байеса и формулу полной вероятности. Таким образом, нам необходимо знать только распределение наблюдаемой компоненты при фиксированной скрытой $p(\mathbf{X}|\mathbf{T},\Theta)$ и вероятности скрытых данных $p(\mathbf{T}|\Theta)$.

EM-алгоритм итеративно улучшает начальную оценку Θ_0, вычисляя новые значения оценок $\Theta_1, \Theta_2,$ и так далее. На каждом шаге переход к Θ_{n+1} от Θ_n выполняется следующим образом:

$$\Theta_{n+1} = \arg\max_{\Theta} Q(\Theta) \qquad (3)$$

где $Q(\Theta)$ — матожидание логарифма правдоподобия. Другими словами, мы не можем сразу вычислить точное правдоподобие, но по известным данным (X) мы можем найти апостериорную оценку вероятностей для различных значений скрытых переменных T. Для каждого набора значений T и параметров Θ мы можем вычислить матожидание функции правдоподобия по данному набору X. Оно зависит от предыдущего значения Θ, потому что это значение влияет на вероятности скрытых переменных T.

$Q(\Theta)$ вычисляется следующим образом:

$$Q(\Theta) = E_{\mathbf{T}}\Big[\log p\,(\mathbf{X}, \mathbf{T} \mid \Theta)\,\Big|\mathbf{X}\Big] \qquad (4)$$

то есть это условное матожидание $\log p\,(\mathbf{X}, \mathbf{T} \mid \Theta)$ при условии Θ.

Другими словами, Θ_{n+1} — это значение, максимизирующее (M) условное матожидание (E) логарифма правдоподобия при данных значениях наблюдаемых переменных и предыдущем значении параметров. В непрерывном случае значение $Q(\Theta)$ вычисляется так:

$$Q(\Theta) = E_{\mathbf{T}}\Big[\log p\,(\mathbf{X}, \mathbf{T} \mid \Theta)\,\Big|\mathbf{X}\Big] = \int_{-\infty}^{\infty} p\,(\mathbf{T} \mid \mathbf{X}, \Theta_n) \log p\,(\mathbf{X}, \mathbf{T} \mid \Theta)\, d\mathbf{T}$$

(5)

Алгоритмы семейства FOREL

Алгоритм кластеризации, основанный на идее объединения в один кластер объектов в областях их наибольшего сгущения [7].

Цель кластеризации: Разбить выборку на такое (заранее неизвестное число) таксонов, чтобы сумма расстояний от объектов кластеров до центров кластеров была минимальной по всем кластерам. То есть наша задача — выделить группы максимально близких друг к другу объектов, которые в силу гипотезы схожести и будут образовывать наши кластеры.

Минимизируемый алгоритмом функционал качества:

$$F = \sum_{j=1}^{k} \sum_{x \in K_j} \rho(x, W_j),$$

(6)

где первое суммирование ведется по всем кластерам выборки, второе суммирование — по всем объектам x, принадлежащим текущему кластеру Kj, а Wi — центр текущего кластера, p(x,y) — расстояние между объектами.

На каждом шаге мы случайным образом выбираем объект из выборки, раздуваем вокруг него сферу радиуса R, внутри этой сферы выбираем центр тяжести и делаем его центром новой сферы. Т.о. мы на каждом шаге двигаем сферу в сторону локального сгущения объектов выборки, то есть стараемся захватить как можно больше объектов выборки сферой фиксированного радиуса. После того как центр сферы стабилизируется, все объекты внутри сферы с этим центром мы помечаем как кластеризованные и выкидываем их из выборки. Этот процесс мы повторяем до тех пор, пока вся выборка не будет кластеризована.

Преимущества:

- Точность минимизации функционала качества (при удачном подборе параметра R)
- Наглядность визуализации кластеризации
- Сходимость алгоритма
- Возможность операций над центрами кластеров — они известны в процессе работы алгоритма
- Возможность подсчета промежуточных функционалов качества, например, длины цепочки локальных сгущений
- Возможность проверки гипотез схожести и компактности в процессе работы алгоритма

Недостатки:

- Относительно низкая производительность (решается введение функции пересчета поиска центра при добавлении 1 объекта внутрь сферы)
- Плохая применимость алгоритма при плохой разделимости выборки на кластеры
- Неустойчивость алгоритма (зависимость от выбора начального объекта)
- Произвольное по количеству разбиение на кластеры
- Необходимость априорных знаний о ширине (диаметре) кластеров

Подходы на основе систем искусственного интеллекта: весьма условная группа, так как методов очень много и методически они весьма различны.

Метод нечеткой кластеризации С-средних

позволяет разбить имеющееся множество векторов (точек) мощностью p на заданное число нечетких множеств. Особенностью метода является использование

нечеткой матрицы принадлежности U с элементами uij, определяющими принадлежность i-го элемента исходного множества векторов - j-му кластеру. Кластеры описываются своими центрами cj - векторами того же пространства, которому принадлежит исходное множество векторов.

В ходе решения задачи нечеткой кластеризации C-means решается задача минимизации следующей целевой функции $E=\Sigma\Sigma uijm\cdot||xi-cj||^2$ при ограничениях $\Sigma juij=1$, i=1..p

Нейронная сеть Кохонена

Класс нейронных сетей, основным элементом которых является слой Кохонена. Слой Кохонена состоит из адаптивных линейных сумматоров («линейных формальных нейронов»). Как правило, выходные сигналы слоя Кохонена обрабатываются по правилу «победитель забирает всё»: наибольший сигнал превращается в единичный, остальные обращаются в ноль.

По способам настройки входных весов сумматоров и по решаемым задачам различают много разновидностей сетей Кохонена. Наиболее известные из них:

- Сети векторного квантования сигналов, тесно связанные с простейшим базовым алгоритмом кластерного анализа (метод динамических ядер или K-средних)
- Самоорганизующиеся карты Кохонена (Self-Organising Maps, SOM)
- Сети векторного квантования, обучаемые с учителем (Learning Vector Quantization)

Слой Кохонена состоит из некоторого количества n параллельно действующих линейных элементов. Все они имеют одинаковое число входов m и получают на свои входы один и тот же вектор входных сигналов $x = (x_1, ...x_m)$. На выходе jго линейного элемента получаем сигнал

$$y_j = w_{j0} + \sum_{i=1}^{m} w_{ji} x_i, \qquad (7)$$

где w_{ji} — весовой коэффициент iго входа jго нейрона, w_{j0} — пороговый коэффициент.

После прохождения слоя линейных элементов сигналы посылаются на обработку по правилу «победитель забирает всё»: среди выходных сигналов y_jищется максимальный; его номер

$$j_{\max} = \arg\max_j \{y_j\} \qquad (8)$$

Окончательно, на выходе сигнал с номером $j_{\max}$ равен единице, остальные — нулю. Если максимум одновременно достигается для нескольких $j_{\max}$, то либо принимают все соответствующие сигналы равными единице, либо только первый в списке (по соглашению).

Большое распространение получили слои Кохонена, построенные следующим образом: каждому (jму) нейрону сопоставляется точка $W_j = (w_{j1}, ...w_{jm})$ в m-мерном пространстве (пространстве сигналов). Для входного вектора $x = (x_1, ...x_m)$ вычисляются его евклидовы расстояния $\rho_j(x)$ до точек W_j и «ближайший получает всё» — тот нейрон, для которого это расстояние минимально, выдаёт единицу, остальные — нули. Следует заметить, что для сравнения расстояний достаточно вычислять линейную функцию сигнала:

$$\rho_j(x)^2 = \|x - W_j\|^2 = \|W_j\|^2 - 2\sum_{i=1}^{m} w_{ji}x_i + \|x\|^2 \qquad (9)$$

(здесь $\|y\|$ - Евклидова длина вектора: $\|y\|^2 = \sum_i y_i^2$). Последнее слагаемое $\|x\|^2$ одинаково для всех нейронов, поэтому для нахождения ближайшей точки оно не нужно. Задача сводится к поиску номера наибольшего из значений линейных функций:

$$j_{\max} = \arg\max_j \left\{ \sum_{i=1}^{m} w_{ji}x_i - \frac{1}{2}\|W_j\|^2 \right\}. \qquad (10)$$

Таким образом, координаты точки $W_j = (w_{j1}, ...w_{jm})$ совпадают с весами линейного нейрона слоя Кохонена (при этом значение порогового коэффициента $w_{j0} = -\|W_j\|^2/2$).

Если заданы точки $W_j = (w_{j1}, ...w_{jm})$, то m-мерное пространство разбивается на соответствующие многогранники Вороного-Дирихле V_j: многогранник V_j состоит из точек, которые ближе к W_j, чем к другим W_k ($k \neq j$).

Генетический алгоритм

Это эвристический алгоритм поиска, используемый для решения задач оптимизации и моделирования путём случайного подбора, комбинирования и вариации искомых параметров с использованием механизмов, аналогичных естественному отбору в природе. Является разновидностью эволюционных вычислений, с помощью которых решаются оптимизационные задачи с использованием методов естественной

эволюции, таких как наследование, мутации, отбор и кроссинговер. Отличительной особенностью генетического алгоритма является акцент на использование оператора «скрещивания», который производит операцию рекомбинации решений-кандидатов, роль которой аналогична роли скрещивания в живой природе.

Задача формализуется таким образом, чтобы её решение могло быть закодировано в виде вектора («генотипа») генов, где каждый ген может быть битом, числом или неким другим объектом. В классических реализациях ГА предполагается, что генотип имеет фиксированную длину. Однако существуют вариации ГА, свободные от этого ограничения.

Некоторым, обычно случайным, образом создаётся множество генотипов начальной популяции. Они оцениваются с использованием «функции приспособленности», в результате чего с каждым генотипом ассоциируется определённое значение («приспособленность»), которое определяет насколько хорошо фенотип, им описываемый, решает поставленную задачу.

При выборе «функции приспособленности» (или fitness function в англоязычной литературе) важно следить, чтобы её «рельеф» был «гладким».

Из полученного множества решений («поколения») с учётом значения «приспособленности» выбираются решения (обычно лучшие особи имеют большую вероятность быть выбранными), к которым применяются «генетические операторы» (в большинстве случаев «скрещивание» — crossover и «мутация» — mutation), результатом чего является получение новых решений. Для них также вычисляется значение приспособленности, и затем производится отбор («селекция») лучших решений в следующее поколение.

Этот набор действий повторяется итеративно, так моделируется «эволюционный процесс», продолжающийся несколько жизненных циклов (поколений), пока не будет выполнен критерий остановки алгоритма. Таким критерием может быть:

- нахождение глобального, либо субоптимального решения;
- исчерпание числа поколений, отпущенных на эволюцию;
- исчерпание времени, отпущенного на эволюцию.

Генетические алгоритмы служат, главным образом, для поиска решений в многомерных пространствах поиска.

Таким образом, можно выделить следующие этапы генетического алгоритма:

- Задать целевую функцию (приспособленности) для особей популяции
- Создать начальную популяцию
- (Начало цикла)

- Размножение (скрещивание)
- Мутирование
- Вычислить значение целевой функции для всех особей
- Формирование нового поколения (селекция)
- Если выполняются условия остановки, то (конец цикла), иначе (начало цикла).

Иерархическая кластеризация

Совокупность алгоритмов упорядочивания данных, визуализация которых обеспечивается с помощью графов.

Алгоритмы упорядочивания данных указанного типа исходят из того, что некое множество объектов характеризуется определённой степенью связности. Предполагается наличие вложенных групп (кластеров различного порядка). Алгоритмы, в свою очередь, подразделяются на агломеративные (объединительные) и дивизивные (разделяющие). По количеству признаков иногда выделяют монотетические и политетические методы классификации. Как и большинство визуальных способов представления зависимостей графы быстро теряют наглядность при увеличении числа объектов. Существует ряд специализированных программ для построения графов.

Под дендрограммой обычно понимается дерево, то есть граф без циклов, построенный по матрице мер близости. Дендрограмма позволяет изобразить взаимные связи между объектами из заданного множества. Для создания дендрограммы требуется матрица сходства (или различия), которая определяет уровень сходства между парами объектов. Чаще используются агломеративные методы.

Далее необходимо выбрать метод построения дендрограммы, который определяет способ пересчёта матрицы сходства (различия) после объединения (или разделения) очередных двух объектов в кластер.

1. В работах по кластерному анализу описан довольно внушительный ряд способов построения (англ. sorting strategies) дендрограмм:

2. Метод одиночной связи (англ. single linkage). Также известен, как «метод ближайшего соседа».

3. Метод полной связи (англ. complete linkage). Также известен, как «метод дальнего соседа».

4. Метод средней связи (англ. pair-group method using arithmetic averages).

- Невзвешенный (англ. unweighted).
- Взвешенный (англ. weighted).

5. Центроидный метод (англ. pair-group method using the centroid average).

- Невзвешенный.
- Взвешенный (медианный).

6. Метод Уорда (англ. Ward's method).

Для первых трёх методов существует общая формула, предложенная А. Н. Колмогоровым для мер сходства:

$$K_\eta([i,j],k) = \left[\frac{(n_i K(i,k)^\eta + (n_j K(j,k)^\eta)}{n_i + n_j}\right]^{\frac{1}{\eta}}, -\infty \leqslant \eta \leqslant +\infty \qquad (11)$$

где $[i,j]$ — группа из двух объектов (кластеров) i и j; k — объект (кластер), с которым ищется сходство указанной группы; n_i — число элементов в кластере i; n_j — число элементов в кластере j. Для расстояний имеется аналогичная формула Ланса — Вильямса[8].

Центроидный метод использует для пересчёта матрицы расстояний[9]. В качестве расстояния между двумя кластерами в этом методе берётся расстояние между их центрами тяжести.

В методе Уорда в качестве расстояния между кластерами берётся прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения [10]. В отличие от других методов кластерного анализа, для оценки расстояний между кластерами здесь используются методы дисперсионного анализа. На каждом шаге алгоритма объединяются такие два кластера, которые приводят к минимальному увеличению целевой функции, то есть внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров.

Выводы

В результате дано определение кластерному анализу, рассмотрены методы и алгоритмы кластеризации данных, такие как: метод k-средних, метод K - medians, EM-алгоритм, алгоритмы семейства FOREL, метод нечеткой кластеризации C-средних, нейронная сеть Кохонена, генетический алгоритм кластеризации и иерархическая кластеризация.

Список литературы

1. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. — М.: Финансы и статистика, 1989. — 607 с.

2. Мандель И. Д. Кластерный анализ. — М.: Финансы и статистика, 1988. — 176 с.

3. Хайдуков Д. С. Применение кластерного анализа в государственном управлении// Философия математики: актуальные проблемы. — М.: МАКС Пресс, 2009. — 287 с.

4. Савиных В.П. Цветков В.Я. Геоинформационный анализ данных дистанционного зондирования. - М.: Картоцентр-Геодезиздат, 2001. - 224с.

5. Орлов А. И. Теория принятия решений //М.: Экзамен. – 2006. –473с.

6. Классификация и кластер. Под ред. Дж. Вэн Райзина. М.: Мир, 1980. 390 с.

7. Поляков А.А., Цветков В.Я. Прикладная информатика: Учебно-методическое пособие: В 2-х частях: Часть.1 / Под общ.ред. А.Н. Тихонова- М.: МАКС Пресс. 2008 - 788 с

8. Tryon R.C. Cluster analysis. — London: Ann Arbor Edwards Bros, 1939. — 139 p.

9. Вятченин Д. А. Нечёткие методы автоматической классификации. — Минск: Технопринт, 2004. — 219 с.

10. Олдендерфер М. С., Блэшфилд Р. К. Кластерный анализ / Факторный, дискриминантный и кластерный анализ: пер. с англ.; Под. ред. И. С. Енюкова. — М.: Финансы и статистика, 1989—215 с.