Статистические методы обработки экспериментальных данных

Лекция №7

Санкт-Петербург
2022

Кластерный анализ

Первое применение кластерный анализ нашел в социологии. Название кластерный анализ происходит от английского слова cluster-гроздь, скопление.

Первое описание кластерного анализа было сделано в 1939 году исследователем Трионом (Tryon).

Главное назначение кластерного анализа-разбиение множества исследуемых объектов на однородные в том или ином смысле группы (кластеры).

Другими словами, решается задача классификации множества данных и выявления присущей им структуры. Методы кластерного анализа можно применять в самых различных случаях, даже в тех случаях, когда речь идет о простой группировке, в которой все сводится к образованию групп по заданному критерию сходства.

Кластерный анализ

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 специфику рассматриваемых объектов, и позволяет рассматривать множество исходных данных практически произвольной природы.

Кластерный анализ можно использовать циклически. В этом случае исследование производится до тех пор, пока не будут достигнуты необходимые результаты. При этом каждый цикл может давать информацию, которая способна сильно изменить направленность и подходы дальнейшего применения метода.

Кластерный анализ

Кластерный анализ имеет определенные недостатки и ограничения. В частности, состав и количество кластеров зависит от выбираемых критериев разбиения.

При сведении исходного массива данных к более компактному виду могут возникать определенные искажения, а также могут теряться индивидуальные черты отдельных объектов.

В кластерном анализе предполагается, что:

а) выбранные характеристики в принципе допускают желаемое разбиение на кластеры;

б) единицы измерения (масштаб) выбраны правильно.

Выбор масштаба играет большую роль. Как правило, данные нормализуют вычитанием среднего и делением на стандартное отклонение, так что дисперсия оказывается равной единице.

Кластерный анализ

Задача кластерного анализа заключается в том, чтобы на основании данных, характеризующих исследуемые объекты, разбить множество объектов G на m (m -целое) кластеров (подмножеств G) $G_1, G_2, \ldots, Gm$, таких, что:

$$G_1 \subset G; \quad G_2 \subset G; \ldots G_m \subset G$$

$$G_1 \cup G_2 \cup \ldots \cup G_m = G$$

$$G_i \cap G_j = \varnothing \quad \forall i \neq j, i = 1,2,\ldots,m; j = 1,2,\ldots,m$$

Объекты, принадлежащие одному и тому же кластеру, должны быть однородными в смысле заданного критерия, в то время как объекты, принадлежащие разным кластерам, должны быть разнородными.

Кластерный анализ

Результатом решения задачи кластерного анализа являются разбиения, удовлетворяющие заданной мере внутрикластерного сходства – некоторому критерию оптимальности.

Этот критерий может представлять собой функционал – целевую функцию.

Например, в качестве целевой функции может быть взята внутригрупповая сумма квадратов отклонения:

$$W_k = \sigma_k^2 = \frac{1}{n_k} \sum_{j=1}^{n_k} \left(x_{kj} - \overline{x}_k \right)^2 = \frac{1}{n_k} \left(\sum_{j=1}^{n_k} x_j^2 \right) - \left(\overline{x}_k \right)^2 \quad (7.1)$$

Здесь k – номер (индекс) кластера.

Кластерный анализ

К характеристикам кластера относятся в частности:

центр, радиус; среднеквадратическое отклонение; размер кластера.

Центр кластера – это среднее геометрическое место точек, принадлежащих кластеру, в пространстве данных.
Радиус кластера-максимальное расстояние точек, принадлежащих кластеру, от *центра кластера*.

Кластеры могут быть перекрывающимися. В этом случае невозможно при помощи используемых процедур однозначно отнести объект к одному из двух или более кластеров. Такие объекты называют *спорными*.

Спорный объект-это объект, который по мере сходства может быть отнесен к более, чем одному кластеру.

Кластерный анализ

Размер кластера может быть определен либо по *радиусу кластера*, либо по *среднеквадратичному отклонению* объектов для этого кластера. Объект относится к кластеру, если расстояние от объекта до *центра кластера* меньше или равно *радиусу кластера*. Если это условие выполняется для двух и более кластеров, объект является *спорным*.

Большое значение в кластерном анализе имеет **выбор масштаба**. Пусть, например, значения переменной x превышают 100, а переменной y - в интервале от 0 до 1.

Тогда, при расчете расстояния между точками переменная x, будет практически полностью доминировать над переменной y. В результате практически невозможно корректно рассчитать расстояния между точками.

Диаметр кластера – максимальное расстояние между элементами кластера.

Кластерный анализ

Эта проблема решается при помощи предварительной стандартизации (нормировки) переменных. Существуют различные способы нормировки данных:

$$z=\frac{(x-\bar{x})}{\sigma};\ z=\frac{x}{\bar{x}};\ z=\frac{x}{x_{max}};\ z=\frac{(x-\bar{x})}{x_{max}-x_{min}}$$

Наряду со нормировкой переменных, существует вариант придания каждой из них определенного **коэффициента важности**, или **веса**, отражающего значимость соответствующей переменной. В качестве весов могут выступать экспертные оценки. Полученные произведения нормированных переменных на соответствующие веса позволяют получать расстояния между точками в многомерном пространстве с учетом важности переменных.

Кластерный анализ

При различных методах классификации и в кластерном анализе в том числе неизбежно возникает проблема **измерения степени близости (сходства) объектов**. Сходство или 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, и сходство с другими объектами будет определяться как соответствующее расстояние.

Для количественной оценки сходства (расстояния) введем понятие **метрики**.

Кластерный анализ

Расстоянием (метрикой) между объектами a и b в пространстве параметров называется такая величина d_{ab}, которая удовлетворяет аксиомам:

1. $d_{ab}>0$, если $a\neq b$, 2. $d_{ab}=0$, если $a=b$
3. $d_{ab}= d_{ba}$; 4. $d_{ab}+d_{bc}\geq d_{ac}$.

Мерой близости (сходства) называется величина μ_{ab}, имеющая предел и возрастающая с возрастанием близости объектов и удовлетворяющая условиям:

μ_{ab} непрерывна; $\mu_{ab}= \mu_{ba}$; $0\leq \mu_{ab}\leq 1$.

Существует возможность простого перехода от расстояния к мерам близости:

$$\mu = \frac{1}{1+d}$$

Кластерный анализ

К наиболее часто используемым способам определения расстояния между объектами относятся следующие:

Показатели	Формулы
Для количественных шкал	
Евклидово расстояние	$d_{Eij}=\left(\sum_{k=1}^{m}\left(x_k^{(i)}-x_k^{(j)}\right)^2\right)^{\frac{1}{2}}$
Квадрат евклидового расстояния	$d_{Eij}^2=\sum_{k=1}^{m}\left(x_k^{(i)}-x_k^{(j)}\right)^2$
Обобщенное степенное расстояние Минковского	$d_{Pij}=\left(\sum_{k=1}^{m}\left(x_k^{(i)}-x_k^{(j)}\right)^P\right)^{\frac{1}{P}}$
Расстояние Чебышева	$d_{ij}=\max_{1\leq k\leq m}\left\|x_k^{(i)}-x_k^{(j)}\right\|$
Расстояние городских кварталов (Манхэттенское расстояние)	$d_H\left(x^{(i)},x^{(j)}\right)=\sum_{k=1}^{m}\left\|x_k^{(i)}-x_k^{(j)}\right\|$

Кластерный анализ

Расстояния между кластерами

1. Расстояние «**Ближайшего соседа**»:

$$\rho_{\min}\left(K_i, K_j\right) = \min_{x_i \in K_i, x_j \in K_j} \rho\left(x_i, x_j\right)$$

2. Расстояние «**Дальнего соседа**»:

$$\rho_{\max}\left(K_i, K_j\right) = \max_{x_i \in K_i, x_j \in K_j} \rho\left(x_i, x_j\right)$$

3. **Невзвешенный центроидный метод.**
Расстояние кластерами определяется как расстояние между их центрами тяжести.

4. **Ввешенный центроидный метод.**
Метод идентичен предыдущему, за исключением того, что при вычислениях учитывается число объектов в них.

Кластерный анализ

Расстояния между кластерами

5. **Метод Уорда (Ward).** В этом методе в качестве целевой функции применяют внутригрупповую сумму квадратов отклонений каждого элемента кластера от средней точки этого кластера. Мерой расстояния между двумя кластерами считается увеличение целевой функции, т. е. внутригрупповой суммы квадратов, при объединении этих кластеров.

Кластерный анализ

Оценка качества многомерной классификации

Для оценки полученных результатов кластеризации используются так называемые **функционалы качества**. Наилучшим считается такое разбиением, при котором достигается минимальное или максимальное значение выбранного функционала качества.

В качестве таких функционалов могут быть использованы:

1. Сумма квадратов расстояний до центров кластеров

$$F_1 = \sum_{k=1}^{K} \sum_{i=1}^{N_k} d^2(X_i^{(k)}, \bar{X}^{(k)}) \Rightarrow \min$$

2. Сумма внутрикластерных расстояний между объектам

$$F_2 = \sum_{k=1}^{K} \sum_{X_i, X_j \in S_k} d^2\left(X_i, X_j\right) \Rightarrow \min$$

Кластерный анализ

Оценка качества многомерной классификации

3. Сумма внутрикластерных дисперсий

$$F_3 = \sum_{k=1}^{K} \sum_{j=1}^{N_k} \sigma_{kj}^2 \Rightarrow \min$$

Здесь σ_{kj}^2 - дисперсия j-й переменной в k-м кластере.

Оптимальным следует считать разбиение, при котором сумма внутрикластерных (внутригрупповых) дисперсий будет минимальной.

Судить о качестве разбиения позволяют и **некоторые простейшие приемы**. Например, можно сравнивать средние значения признаков в отдельных кластерах (группах) со средними значениями в целом по всей совокупности объектов. Если групповые средние существенно отличаются от общего среднего значения, то это может являться признаком хорошего разбиения.

Методы кластерного анализа

Методы кластерного анализа можно разделить на две группы:

иерархические;

неиерархические.

Каждая из групп включает множество подходов и алгоритмов. Используя различные методы кластерного анализа, аналитик может получить различные решения для одних и тех же данных.

Иерархические методы.

Суть иерархической кластеризации состоит в последовательном объединении меньших кластеров в большие или разделении больших кластеров на меньшие.

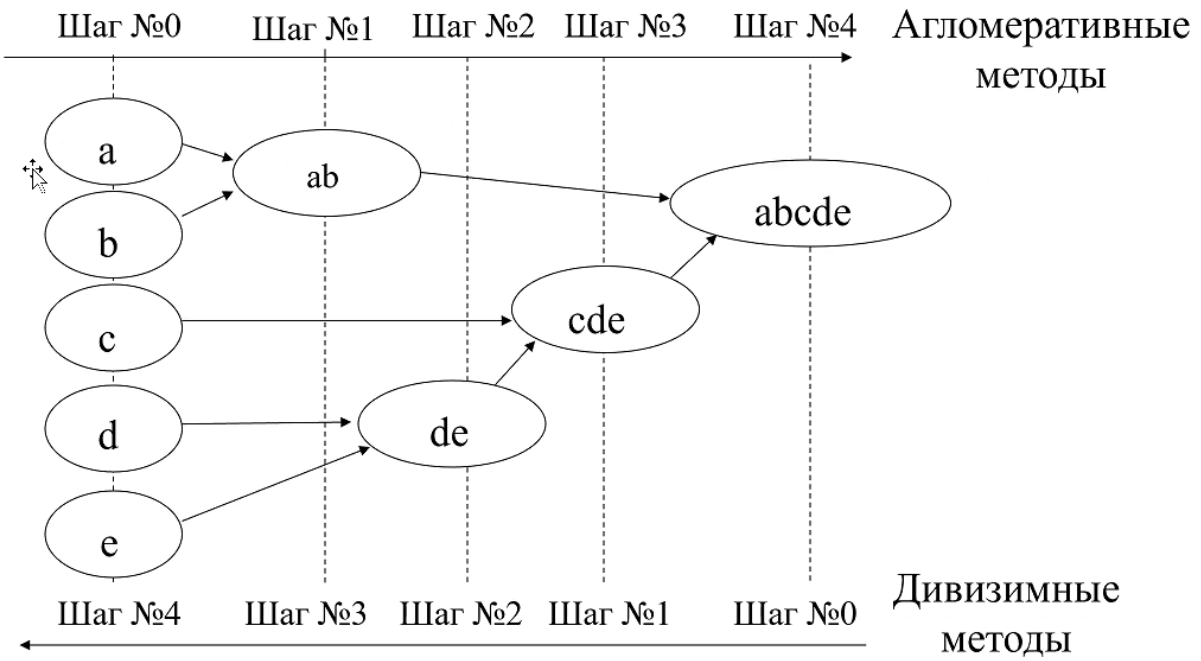
Методы кластерного анализа

Иерархические агломеративные методы *(AgglomerativeNesting,AGNES)*.
В начале работы метода все объекты являются отдельными кластерами. На первом шаге наиболее похожие объекты объединяются в кластер. На последующих шагах объединение продолжается до тех пор, пока все объекты не будут составлять один кластер.

Иерархические дивизимные (делимые) методы *(DIvisiveAN Alysis, DIANA)*.
Эти методы являются логической противоположностью агломеративным методам. В начале работы алгоритма все объекты принадлежат одному кластеру, который на последующих шагах делится на меньшие кластеры, в результате образуется последовательность расщепляющих групп.

Методы кластерного анализа

Принцип работы описанных выше групп методов в виде *дендрограммы* показан на рисунке.



Методы кластерного анализа

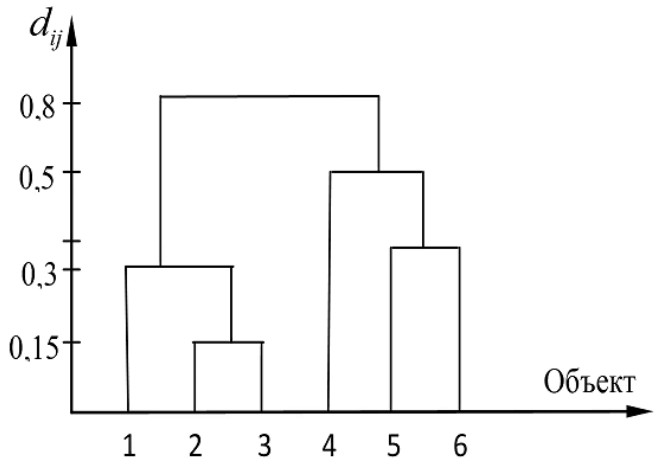
Иерархические методы кластерного анализа используются, как правило, при небольших объемах наборов данных. Преимуществом этих методов кластеризации является их наглядность.

Иерархические алгоритмы связаны с построением **дендрограмм** (от греческого dendron - «дерево»), которые являются результатом иерархического кластерного анализа. Дендрограмма описывает близость отдельных точек и кластеров друг к другу, представляет в графическом виде последовательность объединения (разделения) кластеров.

Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры.

Методы кластерного анализа

Пример вертикальной дендрограммы:



Дендрограмма иерархического кластерного анализа

Методы иерархического кластерного анализа отличаются алгоритмами классификации, из которых наиболее распространенными являются: метод одиночной связи, метод полных связей, метод средней связи, метод Уорда

Методы кластерного анализа

Методы эталонного типа

Существует многочисленная группа так называемых итеративных методов кластерного анализа (метод *k*-средних.).

Процесс классификации начинается с задания некоторых начальных условий (количество образуемых кластеров, порог завершения процесса классификации,..).

Метод *k* –средних не требует вычисления и хранения матрицы расстояний между объектами. Предполагается использование только исходных значений переменных. Для начала процедуры классификации должны быть заданы *k* выбранных объектов, которые будут служить центрами кластеров – эталонами.

Считается, что алгоритмы эталонного типа удобны для обработки больших статистических совокупностей. и быстро действующие.

АААААААААААААААААА ЗАБЫЛ 23 СЛАЙД ААААААААААА

23

Метод *k* –средних

Краткое описание метода.

Пусть имеется *n* наблюдений, каждое из которых характеризуется *m* признаками $X_1, X_2, \ldots, X_m$. Эти наблюдения необходимо разбить на *k* кластеров.

Из *n* точек исследуемой совокупности отбираются случайным образом или задаются, исходя из каких-либо априорных соображений, *k* точек (объектов). Эти точки принимаются за эталоны –центры кластеров. Каждому эталону присваивается порядковый номер, который одновременно является и номером кластера.

На первом шаге из оставшихся (*n*-*k*) объектов извлекается точка X_i с координатами $(x_{i1}, x_{i2}, \ldots, x_{im})$ и проверяется, к какому из эталонов (центров) она находится ближе всего.

24

Метод *k* –средних

Для этого используется одна из метрик, например, евклидово расстояние.

Проверяемый объект присоединяется к тому центру (эталону), которому соответствует минимальное из расстояний.

Эталон заменяется новым (корректируется), пересчитанным с учетом присоединенной точки (вычисляется новое значение среднего арифметического всех включенных в кластер элементов), вес кластера (количество объектов, входящих в данный кластер) увеличивается на единицу.

Если встречаются два и более минимальных расстояния, то *i*-й объект присоединяют к центру (кластеру) с наименьшим порядковым номером.

Метод *k* –средних

На следующем шаге выбирают точку X_{i+1} и для нее повторяются все процедуры.

Таким образом, через (n-k) шагов все точки (объекты) совокупности окажутся отнесенными к одному из k кластеров. Цикл процедуры завершается, но на этом метод работу не заканчивает.

Для того чтобы добиться устойчивости все кластеры считаются пустыми с центрами (эталонами), полученными в конце предыдущего цикла. Все точки $X_1, X_2, \ldots, X_n$ снова последовательно подсоединяются к этим кластерам по рассмотренным правилам. Цикл повторяется. По его завершению новое разбиение сравнивается с полученным в предыдущем цикле. Если они совпадают, работа алгоритма завершается. В противном случае цикл снова повторяется.

Метод *k* –средних

Окончательное разбиение имеет центры тяжести, которые, как правило, не совпадают с первоначальными эталонами. Каждая точка X_i (i=1,2,…,n) будет относиться к тому кластеру, расстояние до центра которого от этой точки минимально.

Возможны две разновидности метода k-средних.

Первая предполагает пересчет центра кластера после каждого изменения его состава, как рассмотрено выше, а **вторая** –лишь после завершения цикла.

В обоих случаях итеративный алгоритм этого метода минимизирует дисперсию внутри каждого кластера, хотя в явном виде такой критерий оптимизации не используется.

Метод *k* –средних

Перед началом работы метода целесообразно нормировать характеристики объектов: $\hat{X} = \frac{x - \bar{x}_в}{S_x}$; $\hat{Y} = \frac{y - \bar{y}_в}{S_y}$

Задание количества кластеров является сложным вопросом. Если нет разумных соображений на этот счет, рекомендуется первоначально создать 2 кластера, затем 3, 4, 5 и т.д., сравнивая полученные результаты.

Существуют модификации метода *k*-средних – ***k*-medians и *k*-medoids**. Обе модификации для определения расстояния используют «манхэттенское расстояние».

В методе ***k*-medians** при определении центра кластера вместо среднего используется медиана.

В методе **k-medoids** центром кластера считается объект, для которого среднее расстояние до других объектов кластера минимально.