ЛЕКЦИЯ 8

ТЕМА. Кластерный анализ

Вопросы темы. Назначение кластерного анализа. Задача кластерного анализа. Этапы выполнения кластерного анализа. Методы кластерного анализа Иерархические методы кластерного анализа. Выбор метода кластеризации. Принятие решения о количестве кластеров. Интерпретация и профилирование кластеров.

Назначение кластерного анализа

Термин кластерный анализ (впервые ввел Tryon, 1939). В настоящее время известно около 100 различных алгоритмов проведения кластерного анализа. Кластерный анализ не накладывает каких-либо ограничений на представление признаков исследуемых объектов. Желательно только, чтобы переменные измерялись в сравнимых шкалах.

Кластерный анализ позволяет упростить математические модели объектов, входящих в выделенные кластеры за счёт однородности признаков (их меньшей изменчивости).

Кластерный анализ может применяться и к совокупностям временных рядов, здесь могут выделяться периоды схожести некоторых показателей и определяться группы временных рядов со схожей динамикой.

Обычно перед началом классификации данные стандартизируются. Иногда различные независимые переменные измеряются в разных шкалах с различными диапазонами. Если значения одной переменной измеряются в сотнях и изменяются в пределах десяти, в то время как другая переменная в среднем равна нулю и изменяется в пределах единицы, то вклад последней в расстояние между объектами будет пренебрежительно малым. Результатом стандартизации является приведение всех переменных к единой шкале.

Задача кластерного анализа.

Задача кластерного анализа заключается в том, чтобы разбить множество, состоящее из k объектов на m кластеров (m – целое число, меньшее чем k) так, чтобы каждый объект принадлежал только одному кластеру с однородными признаками и чтобы объекты, принадлежащие разным кластерам, были с разнородными признаками. Кластерный анализ

Поставим задачу выделения кластеров по показателям расстояния между признаками в группируемых ОИ с выполнением следующих условий.

$$\sum_{i=1}^{k}\sum_{j=1}^{k}\delta_{ij}d_{ij} \to \min \quad (1)$$

$$\sum_{i=1}^{k}\sum_{i=1}^{k}(1-\delta_{ij})d_{ij} \to \max, \ (2)$$

где k – количество объектов;

d_{ij} - расстояние между i-м и j-м объектами;

δ_{ij} - символ Кронекера, принимающий значение 1, если i-ый и j-ый объекты входят в один и тот же кластер; и значение 0, если не входят.

Признаки представляются либо в натуральных единицах измерения, либо в стандартизированной форме, в которой их средние значения равны нулю, а стандартные отклонения равны единице. В стандартных процедурах для проведения кластерного анализа, как правило задается либо количество кластеров, либо пороговое значение для условия (1). Условие (1) обеспечивает минимум расстояний между признаками объектов, вошедших в один и тот же кластер; а (2) максимум этих расстояний между объектами, вошедшими в разные кластеры.

Этапы выполнения кластерного анализа.

Применения кластерного анализа включает в себя следующие этапы:

1. Стандартизация исходных статистических данных выполняется в случаях, когда учитываемые признаки имеют различные единицы измерения или значительно отличаются по масштабам единиц измерения.
2. Вычисление расстояний между признаками объектов и суммарного расстояния между объектами по всем признакам и составление матрицы расстояний между объектами.
3. Поиск наименьшего расстояния между объектами и объединение двух объектов с наименьшим расстоянием между ними в один кластер.
4. Вычисление расстояний между объектами и формирующимися кластерами и преобразование матрицы расстояний между ними. Переход к пункту 3 и выполнение пунктов 3 и 4 до тех пор, пока не будут сгруппированы все объекты и сформированные кластеры в один общий кластер, после чего переход к пункту 5.
5. Выдача перечней объектов по выделенным кластерам в виде таблицы и соответствующей дендрограммы с указанием расстояний между объектами в выделенных кластерах и сформированными кластерами.

Расстояние между объектами по **Евклиду** вычисляется по формуле:

$$d_{ij} = \sqrt{\sum_{g=1}^{l} (x_{ig} - x_{jg})^2} . \quad)$$

Для придания больших весов более отдаленным друг от друга объектам можно воспользоваться **квадратом евклидова расстояния** путем возведения в квадрат стандартного евклидова расстояния.

$$d_{ij} = \sum_{g=1}^{l} (x_{ig} - x_{jg})^2 .$$

Манхэттенское расстояние (расстояние городских кварталов), также называемое "хемминговым". Это расстояние рассчитывается как среднее разностей по координатам. В большинстве случаев эта мера расстояния приводит к результатам, подобным при использовании евклидова расстояния. Для расстояния хемминга влияние имеющихся «выбросов» (больших отклонений) меньше, чем при использовании евклидова расстояния, поскольку в этом случае координаты не возводятся в квадрат.

$$d_{ij} = \sum_{g=1}^{l} |x_{ig} - x_{jg}|; \; i = \overline{1,k} \; ; \; j = \overline{1,k} .$$

Расстояние Чебышева. Это расстояние следует использовать, когда необходимо определить два объекта как "различные", если они сильно отличаются отличаются по какому-то одному измерению.

$$d_{ij} = \max \left| x_{ig} - x_{jg} \right|; \; i = \overline{1,k} ; \; j = \overline{1,k}; \; g = \overline{1,l}.$$

Степенное расстояние (обобщенное степенное расстояние Минковского) используется в том случае, если необходимо прогрессивно увеличить или уменьшить вес, относящийся к размерности, для которой соотвествующие объекты сильно отличаются.

$$d_{ij} = \left(\sum_{g=1}^{l} |x_{ig} - x_{jg}|^{p} \right)^{1/r}; \; i = \overline{1,k} \; ; \; j = \overline{1,k}.$$

В формуле параметры r и p, определяются в зависимости от задачи. Параметр p ответственен за постепенное взвешивание разностей по отдельным измерениям, параметр r влияет на прогрессивное взвешивание больших расстояний между объектами.

В формулах приняты следующие обозначения:

d_{ij} – расстояние между i-ым и j-ым объектами;

k – количество объектов;

l – количество признаков;

x_{ig} – значение i-го признака g-го объекта;

x_{jg} – значение j-го признака g-го объекта.

Расстояние от формирующегося кластера с вошедшими в него объектами до других объектов может вычисляться по следующим правилам.

Принцип ближайшего соседа.

$d_{qg} = d_{ig}$, при $d_{ig} \le d_{jg}$;

$d_{qg} = d_{jg}$,при $d_{ig} > d_{jg}$.

Принцип наиболее удаленного соседа.

$d_{qg} = d_{ig}$, при $d_{jg} \le d_{ig}$;

$d_{qg} = d_{jg}$,при $d_{jg} > d_{ig}$.

3.Принцип среднего расстояния.

$$d_{qg} = \frac{1}{2}(d_{ig} + d_{jg}).$$

Принцип медианы.

$$d_{qg} = \frac{1}{2}\sqrt{2(d_{ig}^2 + d_{jg}^2) - d_{ij}^2}.$$

В формулах приняты следующие обозначения:

d_{qg} - расстояние между q-ым кластером, к которому подсоединен еще один объект, и g-ым объектом или кластером;

d_{ig} - расстояние между i-ым и g-ым объектами или кластерами;

d_{jg} - расстояние между j-ым и g-ым объектами или кластерами;

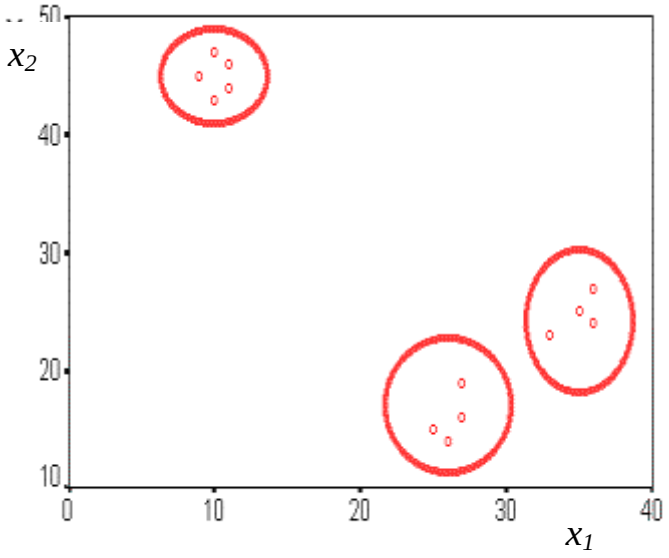
d_{ij} - расстояние между i-ым и j-ым объектами или кластерами.

Решением задачи кластерного анализа является разбиение, удовлетворяющее заданным условиям. Рассмотрим пример процедуры кластерного анализа.

Пример

Допустим, мы имеем совокупность, состоящую из 13-ти объектов, кото-рые характеризуются двумя признаками: x_1 и x_2. Данные по ним приведены в таблице.

Таблица 5.1		
№объ-екта	Признак x1	Признак x2
1	27	19
2	11	46
3	25	15
4	36	27
5	35	25
6	10	43
7	11	44
8	36	24
9	26	14
10	26	14
11	9	45
12	33	23
13	10	47



Представим переменные x_1 и x_2 в виде диаграммы рассеивания, приведённой на рис.. В таблице объекты вошедшие в одни и те же кластеры выделены одинаковыми цветами. На рисунке мы видим три кластера. Объекты, которые по значениям x_1 и x_2 "похожи" друг на друга, принадлежат к одному кластеру; объекты из разных кластеров не похожи друг на друга.

Кластер имеет следующие **математические характеристики**: центр, радиус, среднеквадратическое отклонение, размер кластера.

Центр кластера - это среднее геометрическое место точек в пространстве переменных. Среднее геометрическое n значений определяется по формуле: $g = \sqrt[n]{x_1, x_2, ...x_n}$.

Радиус кластера - максимальное расстояние точек от центра кластера.

Кластеры могут быть перекрывающимися. Такая ситуация возникает, когда обнаруживается перекрытие кластеров. В этом случае невозможно при помощи математических процедур однозначно отнести объект к одному из двух кластеров. Такие объекты называют спорными.

Спорный объект - это объект, который по мере сходства может быть отнесен к нескольким кластерам.

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 радиуса кластера. Если это условие выполняется для двух и более кластеров, объект является спорным.

Методы кластерного анализа

На практике наиболее часто используются следующие методы кластерного анализа:

- иерархический;
- k-средних;
- двувходовое объединение.

Иерархические методы кластерного анализа

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

Иерархические агломеративные методы (Agglomerative Nesting, AGNES)

Эта группа методов характеризуется последовательным объединением исходных объектов и соответствующим уменьшением числа кластеров. В начале работы алгоритма 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 (DIvisive ANAlysis, 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.

Принцип работы описанных выше групп методов в виде дендрограммы показан на рис 2.

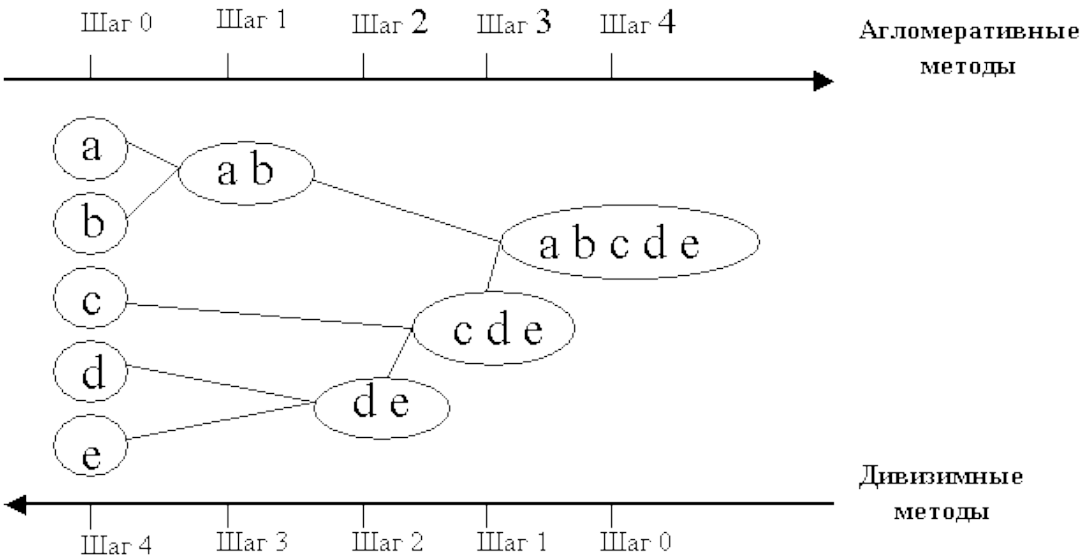


Рис.2

Иерархические методы кластерного анализа используются при сравнительно небольших объемах наборов данных. Преимуществом иерархических 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.

Дендрограмма (dendrogram) - древовидная диаграмма, содержащая n уровней, каждый из которых соответствует одному из шагов процесса последовательного укрупнения кластеров. Дендрограмму также называют древовидной схемой, деревом объединения кластеров, деревом иерархической структуры.

Существует много способов построения дендрограмм. В дендрограмме объекты могут располагаться вертикально или горизонтально. Пример вертикальной дендрограммы приведен на Рис. 3.

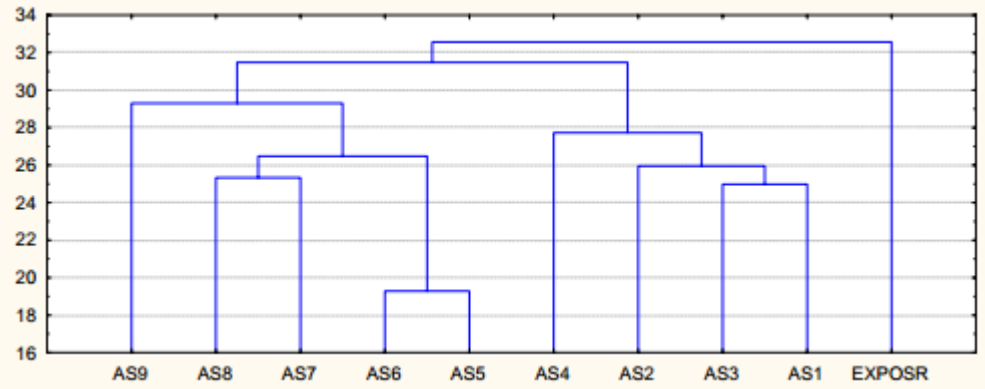


Рис. 3

Если рис. 3 повернуть по часовой стрелке на угол 90 градсов,то получим горизонтальную дендрограмму.

Иерархические методы кластеризации различаются правилами построения кластеров. В качестве правил выступают критерии, которые используются при решении вопроса о "схожести" объектов при их объединении в группу (агломеративные методы) либо разделения на группы (дивизимные методы). Из них наиболее распространены методы: одиночной связи, полных связей, средней связи, Уорда.

Метод ближнего соседа или одиночная связь. Здесь расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами (ближайшими соседями) в различных кластерах. Этот метод позволяет выделять кластеры сколь угодно сложной формы при условии, что различные части таких кластеров соединены цепочками близких друг к другу элементов. В этом методе кластеры представляются длинными "цепочками" или "волокнистыми" кластерами, "сцепленными вместе" только отдельными элементами, которые случайно оказались ближе остальных друг к другу.

Метод наиболее удаленных соседей или полная связь. Расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями"). Метод хорошо использовать, когда объекты являются представителями различных "рощ". Если же кластеры имеют в некотором роде удлиненную форму или их естественный тип является "цепочечным", то этот метод не следует использовать.

Метод средней связи. Для решения вопроса о включении нового объекта в уже существующий кластер вычисляется среднее значение меры сходства, которое затем сравнивается с заданным пороговым уровнем.

Метод Уорда (Ward's method). В качестве расстояния между кластерами берется прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения (Ward, 1963). 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, т.е. внутригрупповой суммы квадратов. Этот метод направлен на объединение близко расположенных кластеров и "стремится" создавать кластеры малого размера.

Метод невзвешенного попарного среднего. В качестве расстояния между двумя кластерами берется среднее расстояние между всеми парами объектов в них. Этот метод следует использовать, если объекты являются представителями различных "рощ", а в случаях присутствия кластеров "цепочного" типа, при предположении неравных размеров кластеров.

Метод взвешенного попарного среднего. Этот метод похож на метод невзвешенного попарного среднего, разница состоит лишь в том, что здесь в качестве весового коэффициента используется размер кластера (число объектов, содержащихся в кластере). Этот метод рекомендуется использовать при наличии предположения о кластерах разных размеров.

Невзвешенный центроидный метод. В качестве расстояния между двумя кластерами в этом методе берется расстояние между их центрами тяжести.

Взвешенный центроидный метод. Этот метод похож на предыдущий, разница состоит в том, что для учета разницы между размерами кластеров (числе объектов в них), используются веса. Этот метод предпочтительно использовать в случаях, если имеются предположения относительно существенных отличий в размерах кластеров.

Этапы выполнения иерархического кластерного анализа включают следующие процедуры:

1. Стандартизации исходных статистических данных выполняется в случаях, когда учитываемые признаки имеют различные единицы измерения или значительно отличаются по масштабам единиц измерения.
2. Вычисления расстояний между признаками объектов и суммарного расстояния между объектами по всем признакам и составление матрицы расстояний между объектами.
3. Поиска наименьшего расстояния между объектами и объединение двух объектов с наименьшим расстоянием между ними в один кластер.
4. Вычисления расстояний между объектами и формирующимися кластерами и преобразование матрицы расстояний между ними. Переход к пункту 3 и выполнение пунктов 3 и 4 до тех пор, пока не будут сгруппированы все объекты и сформированные кластеры в один общий кластер, после чего переход к пункту 5.
5. Выдачи перечней объектов по выделенным кластерам в виде таблицы и соответствующей дендрограммы с указанием расстояний между объектами в выделенных кластерах и сформированными кластерами.

Метод k-средних

При большом количестве объектов иерархические методы кластерного анализа не эффективны, или даже совсем не пригодны. В таких случаях используют неиерархические методы, основанные на разделении, которые представляют собой итеративные методы дробления исходной совокупности. В процессе деления новые кластеры формируются до тех пор, пока не будет выполнено правило остановки.

Такая неиерархическая кластеризация состоит в разделении набора данных на определенное количество отдельных кластеров. Существует два подхода. Первый заключается в определении границ кластеров как наиболее плотных участков в многомерном пространстве исходных данных, т.е. определение кластера там, где имеется большое "сгущение точек". Второй подход заключается в минимизации меры различия объектов.

Наиболее распространен среди неиерархических методов **алгоритм k-средних**, также называемый быстрым кластерным анализом. В отличие от иерархических методов, которые не требуют предварительных предположений относительно числа кластеров, для возможности использования этого метода необходимо выдвинуть гипотезу о наиболее вероятном количестве кластеров.

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: для заданного фиксированного количества кластеров - k средние значения признаков максимально возможно отличаются друг от друга.

Первоначальное распределение объектов по кластерам.

По выбранному количеству кластеров - k на первом шаге в качестве их центров выбирается k объектов. Каждому кластеру соответствует один центр. Выбор начальных центров может осуществляться следующими методами:

- выбор первых k-объектов;
- случайный выбор k-объектов;
- выбор первых двух объектов в качестве центров двух первых кластеров по максимальному расстоянию между ними (по k-средним); в качестве центра третьего кластера выбирается объект суммарное расстояние которого от объектов, ранее выбранных в качестве центров кластеров, максимально. И таким

образом определяются объекты в качестве центров всех остальных кластеров с общим количеством – k кластеров. В результате каждому из k кластеров назначется объект в качестве первоначального центра.

Затем все объекты распределяются по k кластерам по наименьшему расстоянию к объекта м, выбранным в качестве центров кластеров.

Итеративный процесс перераспределения объектов в кластерах

Вычисляются центры сформированных кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты перераспределяют-ся по наименьшему расстоянию их покоординатного среднего и поокорди-натных средних каждого кластера.

Итеративный процесс вычисления центров и перераспределения объектов в кластерах продолжается до тех пор, пока не выполнено одно из условий:

- кластерные центры стабилизировались, т.е. все объекты входят в кластеры, в которые они входили до текущей итерации;
- число итераций равно максимальному заданному числу итераций.

На рис. 4 приведен пример алгоритма k-средних для k, равного двум.

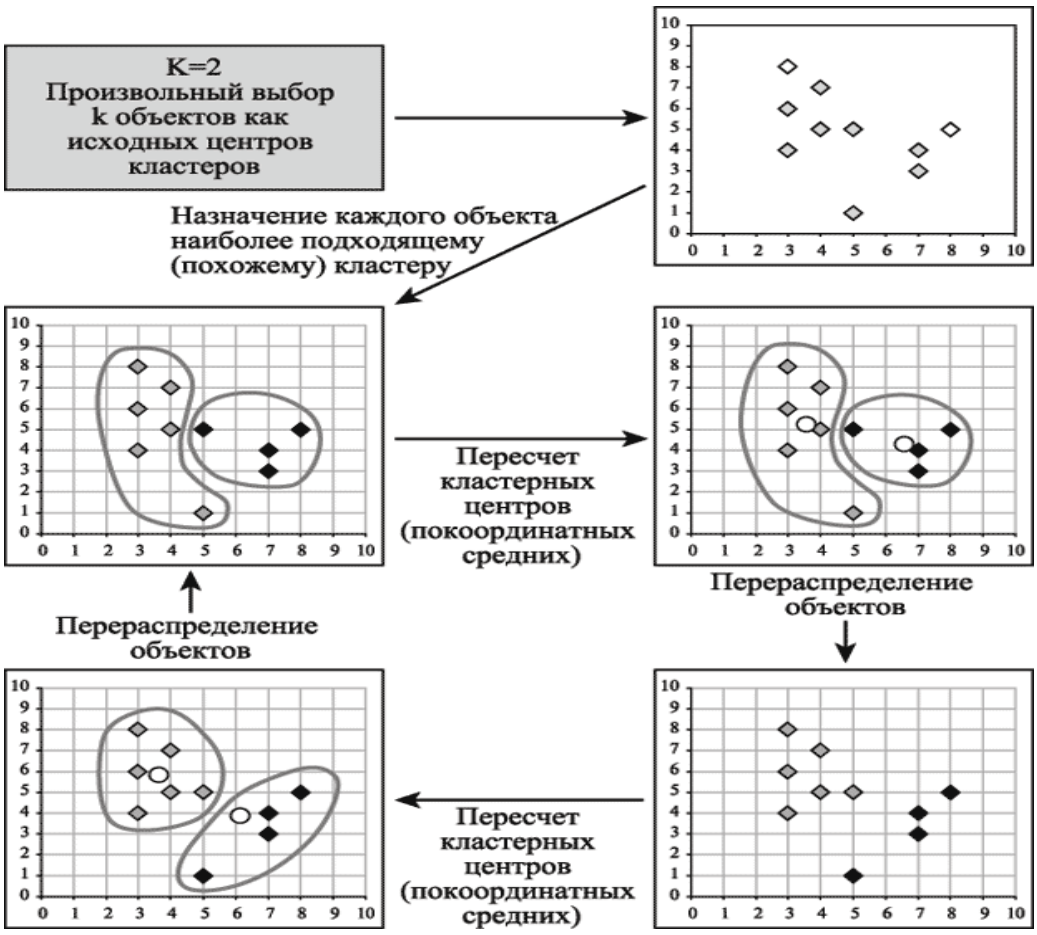


Рис. 4

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

Проверка качества кластеризации

После получений результатов кластерного анализа методом k-средних проверяется качество кластеризации (т.е. оценивается, насколько кластеры отличаются друг от друга). Для этого рассчитываются средние значения для каждого кластера и определяется разница средняя разница между ними. При хорошей кластеризации должны быть получены сильно отличающиеся средние для всех объектов, или хотя бы большей их части.

Достоинства алгоритма k-средних:

- простота использования;
- быстрота использования;
- понятность и прозрачность алгоритма.

Недостатки алгоритма k-средних:

- алгоритм слишком чувствителен к выбросам, которые могут искажать среднее. Возможным решением этой проблемы является использование модификации алгоритма - алгоритм k-медианы;
- алгоритм может медленно работать при большом количестве объектов и большом количестве признаков объектов.