**Кластерный анализ** – это совокупность методов, позволяющих классифицировать многомерные наблюдения. Термин кластерный анализ, впервые введенный Трионом (Tryon) в 1939 году, включает в себя более 100 различных алгоритмов.

В отличие от задач классификации, кластерный анализ не требует априорных предположений о наборе данных, не накладывает ограничения на представление исследуемых объектов, позволяет анализировать показатели различных типов данных (интервальным данным, частотам, бинарным данным). При этом необходимо помнить, что переменные должны измеряться в сравнимых шкалах.

Кластерный анализ позволяет сокращать размерность данных, делать ее наглядной.

**задачи кластерного анализа**[**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=2)

Задачи кластерного анализа можно объединить в следующие группы:

1. Разработка типологии или классификации.

2. Исследование полезных концептуальных схем группирования объектов.

3. Представление гипотез на основе исследования данных.

4. Проверка гипотез или исследований для определения, действительно ли типы (группы), выделенные тем или иным способом, присутствуют в имеющихся данных.

Как правило, при практическом использовании кластерного анализа одновременно решается несколько из указанных задач.

**схожесть кластеров**[**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=3)

Критерием для определения схожести и различия кластеров является расстояние между точками на диаграмме рассеивания. Это сходство можно "измерить", оно равно расстоянию между точками на графике. Способов определения меры расстояния между кластерами, называемой еще мерой близости, существует несколько. Наиболее распространенный способ - вычисление евклидова расстояния между двумя точками i и j на плоскости, когда известны их координаты X и Y:

D_{ij} = \sqrt{ (x_i - x_j)^2+ ( y_i - y_j )^2}

*Примечание:* чтобы узнать расстояние между двумя точками, надо взять разницу их координат по каждой оси, возвести ее в квадрат, сложить полученные значения для всех осей и извлечь квадратный корень из суммы.

**Математические характеристики кластера**[**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=4)

Кластер имеет следующие математические характеристики:

**Центр кластера** - это среднее геометрическое место точек в пространстве переменных.

x_{kj} = \frac {\sum_{j=1}^n w_jx_{ij}}{I_k}

**Дисперсия кластера** - это мера рассеяния точек в пространстве относительно центра кластера:

D_k = \frac{\sum_{i=1}^{I_1} \sum_{j=1}^n w_j(x_{ij}-\overline x_{kj})^2}{I_k -1}

**Среднеквадратичное отклонение (СКО) объектов** относительно центра кластера:

S_k = \sqrt{D_k}

**Радиус кластера** - максимальное расстояние точек от центра кластера:

R_k = max \sqrt{\sum_{j=1}^n w_j(x_{ij}-\overline x_{kj})^2}

**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.

Неоднозначность данной задачи может быть устранена экспертом или аналитиком.

Работа кластерного анализа опирается на два предположения: **Первое предположение** - рассматриваемые признаки объекта в принципе допускают желательное разбиение совокупности объектов на кластеры. **Второе предположение** - правильность выбора масштаба или единиц измерения признаков.

**Методы кластерного анализа**[**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=5)

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. Это считается нормальным явлением. Рассмотрим иерархические и неиерархические методы подробно.

**Иерархические методы кластерного анализа**[**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=6)

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 элементов 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.

Программная реализация алгоритмов кластерного анализа широко представлена в различных инструментах Data Mining, которые позволяют решать задачи достаточно большой размерности. Например, агломеративные методы реализованы в пакете SPSS, дивизимные методы - в пакете Statgraf.

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).

**Меры сходства кластеров**[**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=7)

Для вычисления расстояния между объектами используются различные меры сходства (меры подобия), называемые также метриками или функциями расстояний. Евклидово расстояние, это наиболее популярная мера сходства.

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

**Манхэттенское расстояние** (расстояние городских кварталов), также называемое "хэмминговым" или "сити-блок" расстоянием.

Это расстояние рассчитывается как среднее разностей по координатам. ***В большинстве случаев эта мера расстояния приводит к результатам, подобным расчетам расстояния евклида.*** Однако, для этой меры влияние отдельных выбросов меньше, чем при использовании евклидова расстояния, поскольку здесь координаты не возводятся в квадрат. **Расстояние Чебышева.** Это расстояние стоит использовать, когда необходимо определить два объекта как "различные", если они отличаются по какому-то одному измерению.

**Процент несогласия.** Это расстояние вычисляется, если данные являются категориальными.

**Методы объединения или связи**[**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=8)

Когда каждый объект представляет собой отдельный кластер, расстояния между этими объектами определяются выбранной мерой. Возникает следующий вопрос - как определить расстояния между кластерами? Существуют различные правила, называемые методами объединения или связи для двух кластеров.

**Метод ближнего соседа или одиночная связь.**[**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=9)

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 результате работы этого метода 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.**[**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=10)

Здесь расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах (т.е. "наиболее удаленными соседями"). Метод хорошо использовать, когда объекты действительно происходят из различных "рощ". Если же кластеры имеют в некотором роде удлиненную форму или их естественный тип является "цепочечным", то этот метод не следует использовать.

**Метод Варда (Ward's method).**[**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=11)

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**[**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=12)

(метод невзвешенного попарного арифметического среднего - unweighted pair-group method using arithmetic averages, UPGMA (Sneath, Sokal, 1973)). В качестве расстояния между двумя кластерами берется среднее расстояние между всеми парами объектов в них. Этот метод следует использовать, если объекты действительно происходят из различных "рощ", в случаях присутствия кластеров "цепочного" типа, при предположении неравных размеров кластеров.

**Метод взвешенного попарного среднего**[**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=13)

(метод взвешенного попарного арифметического среднего - weighted pair-group method using arithmetic averages, WPGM A (Sneath, Sokal, 1973)). 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 именно при наличии предположения о кластерах разных размеров. Невзвешенный центроидный метод (метод невзвешенного попарного центроидного усреднения - unweighted pair-group method using the centroid average (Sneath and Sokal, 1973)).

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

**Взвешенный центроидный метод** [**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=14)

метод взвешенного попарного центроидного усреднения -weighted pair-group method using the centroid average, WPGMC (Sneath, Sokal 1973)). 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.

**Иерархический кластерный анализ в SPSS**[**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=15)

Рассмотрим процедуру иерархического кластерного анализа в пакете SPSS (SPSS). Процедура иерархического кластерного анализа в SPSS предусматривает группировку как объектов (строк матрицы данных), так и переменных (столбцов). Можно считать, что в последнем случае роль объектов играют переменные, а роль переменных - столбцы. В этом методе реализуется иерархический агломеративный алгоритм, смысл которого заключается в следующем. Перед началом кластеризации все объекты считаются отдельными кластерами, в ходе алгоритма они объединяются. Вначале выбирается пара ближайших кластеров, которые объединяются в один кластер. В результате количество кластеров становится равным N-1. Процедура повторяется, пока все классы не объединятся. На любом этапе объединение можно прервать, получив нужное число кластеров. Таким образом, результат работы алгоритма агрегирования зависит от способов вычисления расстояния между объектами и определения близости между кластерами. Для определения расстояния между парой кластеров могут быть сформулированы различные подходы. С учетом этого в SPSS предусмотрены следующие методы:

• Среднее расстояние между кластерами (Between-groups linkage), устанавливается по умолчанию.

• Среднее расстояние между всеми объектами пары кластеров с учетом расстояний внутри кластеров (Within-groups linkage).

• Расстояние между ближайшими соседями - ближайшими объектами кластеров (Nearest neighbor).

• Расстояние между самыми далекими соседями (Furthest neighbor).

• Расстояние между центрами кластеров (Centroid clustering) или центроидный метод. Недостатком этого метода является то, что центр объединенного кластера вычисляется как среднее центров объединяемых кластеров, без учета их объема.

• Метод медиан - тот же центроидный метод, но центр объединенного кластера вычисляется как среднее всех объектов (Median clustering).

• Метод Варда.

**Определение количества кластеров**[**Править**](http://ru.science.wikia.com/wiki/%D0%9A%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%BD%D1%8B%D0%B9_%D0%B0%D0%BD%D0%B0%D0%BB%D0%B8%D0%B7?action=edit&section=16)

Существует проблема определения числа кластеров. Иногда можно априорно определить это число. Однако в большинстве случаев число кластеров определяется в процессе агломерации/разделения множества объектов. Процессу группировки объектов в иерархическом кластерном анализе соответствует постепенное возрастание коэффициента, называемого критерием Е. Скачкообразное увеличение значения критерия Е можно определить как характеристику числа кластеров, которые действительно существуют в исследуемом наборе данных. Таким образом, этот способ сводится к определению скачкообразного увеличения некоторого коэффициента, который характеризует переход от сильно связанного к слабо связанному состоянию объектов.