

Основные типы кластеризации

- Восходящая/нисходящая кластеризации (hierarcical / partitional)
- Исключающая, перекрывающаяся и нечеткая кластеризации (exclusive / overlapping)
- Полная и частичная кластеризации (complete / partial clustering)

Восходящая/нисходящая кластеризация

- Иерархическая кластеризация (восходящая) - допускаем наличие подкластеров, осуществляется в несколько приемов, в результате образуется иерархическое дерево (дендрограмму).
- Нисходящая (плоская) кластеризация - предполагает разделение на кластеры сразу, причем один объект относится только к одному кластеру.

Исключающая, перекрывающаяся и нечеткая кластеризации (exclusive / overlapping/ fuzzing)

- Исключающая – каждый объект может быть отнесен только к одному кластеру
- Перекрывающаяся - используется, если объект принадлежит к нескольким группам или находится между двумя кластерами.
- Нечеткая или вероятностные кластеризации являются частными случаями перекрывающейся кластеризации. Тогда каждый объект относится к кластеру с определенным весом или вероятностью. Например, вес от 0 до 1, где 0 – абсолютно не принадлежит, 1 – полностью принадлежит.

Полная и частичная кластеризации (complete/ partial)

- Метод полной кластеризации - каждый объект обязательно относится к кластеру
- Частичная кластеризация –некоторые объекты не принадлежат к четко определенным группам, поскольку могут являться выбросами, шумами и т.п.

Определение расстояние между элементами

- Вычисление Евклидова расстояния (если известны координаты точек в пространстве)

$$D(x_i; x_j) = \sqrt{\sum_i^n (x_i - x_j)^2}$$

- Квадрат Евклидова расстояния
- Манхэттенское расстояние (дает те же результаты, что и Евклидово расстояние, но влияние отдельных выбросов уменьшается):

$$D(x_i; x_j) = \sum_i |x_i - x_j|$$

- Расстояние Чебышева
Используется, когда нужно определить два объекта как «различные», если они различаются по какой-либо одной координате

$$D(x_i; x_j) = \max_i |x_i - x_j|$$

Методы объединения кластеров

- **Метод ближнего соседа или одиночная связь.** Расстояние между двумя кластерами определяется расстоянием между двумя наиболее близкими объектами в различных кластерах.
- **Метод наиболее удаленных соседей.** Расстояния между кластерами определяются наибольшим расстоянием между любыми двумя объектами в различных кластерах
- **Метод Варда.** Расстояние как прирост суммы квадратов расстояний объектов до центров кластеров, получаемый в результате их объединения
- **Метод невзвешенного попарного среднего.** В качестве расстояния между двумя кластерами берется среднее расстояние между всеми парами объектов в них.
- **Метод взвешенного попарного среднего.** Метод похож на метод невзвешенного попарного среднего, но в качестве весового коэффициента используется размер кластера
- **Невзвешенный центроидный метод.** В качестве расстояния между двумя кластерами в этом методе берется расстояние между их центрами тяжести.
- **Взвешенный центроидный метод.** Метод похож на предыдущий, но для учета разницы между размерами кластеров (числе объектов