Иерархическая кластеризация

ВЫПОЛНИЛА:
ЗУБРИЛИНА ОЛЬГА

Задача кластеризации — задача разбиения заданной выборки *объектов* на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Постановка задачи кластеризации

Дано:

$\boldsymbol{X}$ – пространство объектов;

$\boldsymbol{X^l = \{x_i\}_{i=1}^l}$ – обучающая выборка;

$\boldsymbol{p}: \boldsymbol{X \times X \to [0, \infty)}$ – функция расстояния между объектами.

Найти:

$\boldsymbol{Y}$ – множество кластеров

$\boldsymbol{a}: \boldsymbol{X \to Y}$ – алгоритм кластеризации такой, что:

- каждый кластер состоит из близких объектов
- объекты разных кластеров существенно различны.

Типы входных данных

- Признаковое описание объектов $x^{(1)}, x^{(2)}, \dots, x^{(N)} \in R^d$
- Матрица расстояний между объектами

Количество кластеров может быть известно заранее, а может определяться в процессе работы алгоритма.

Если число кластеров известно заранее, то алгоритм находит отображение

$$C: \{1, \dots, N\} \to \{1, \dots, K\}$$

где N – число имеющихся объектов, К – число кластеров

Виды кластеризации

- Иерархическая – последовательное построение кластеров из уже найденных кластеров.
  - Агломеративная
  - Разделительная
- Неиерархическая – оптимизируем некую целевую функцию.
  - Алгоритмы теории графов
  - EM алгоритм
  - K – means
  - Нечеткие алгоритмы

Иерархическая кластеризация

Алгоритмы иерархической кластеризации, или таксономии, находят иерархическое представление, такое, что кластеры на каждом уровне получаются объединением кластеров на более низком уровне (кластеры на более низком уровне получаются дроблением кластеров на более высоком уровне).

Иерархические структуры удобно представлять в виде **дендрограмм** (корневых деревьев). Верхний уровень классификации содержит один кластер, включающий все объекты. На нижнем уровне – N кластеров, содержащих один объект.

Агломеративные методы

Строят дерево в направлении от листьев к корню.

Агломеративные методы

Строят дерево в направлении от листьев к корню.

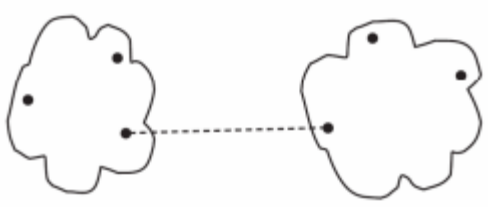
$\delta_{ii'}$ - мера различия («расстояние») между i – м и i' -м объектами (i=1,…, N)
δ_{RS} - мера различия между кластерами R и S, каким-либо образом определяется исходя из попарных различий между элементами из R и из S.

Агломеративные методы

Способы определения δ_{RS} :

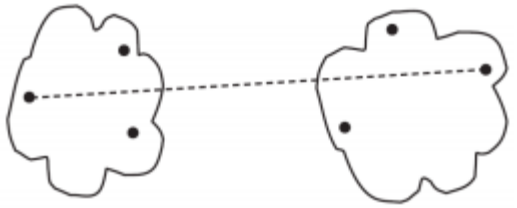
По минимальному различию (по принципу «ближнего соседа»)

$$\delta_{RS} = \min_{i \in R,\, i' \in S} \delta_{ii'}$$



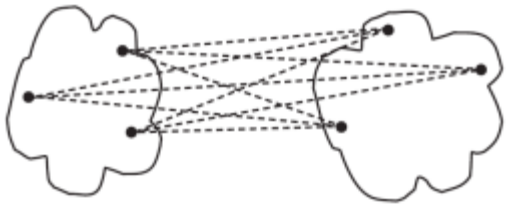
По максимальному различию (по принципу «дальнего соседа»)

$$\delta_{RS} = \max_{i \in R,\, i' \in S} \delta_{ii'}$$



По усредненному различию (по принципу «средней связи»)

$$\delta_{RS} = \frac{1}{|R| * |S|} \sum_{i \in R} \sum_{i' \in S} \delta_{ii'}$$



Агломеративные методы начинают работу с N кластеров, каждый из которых содержит 1 объект.

На некотором этапе имеется некоторый набор кластеров.

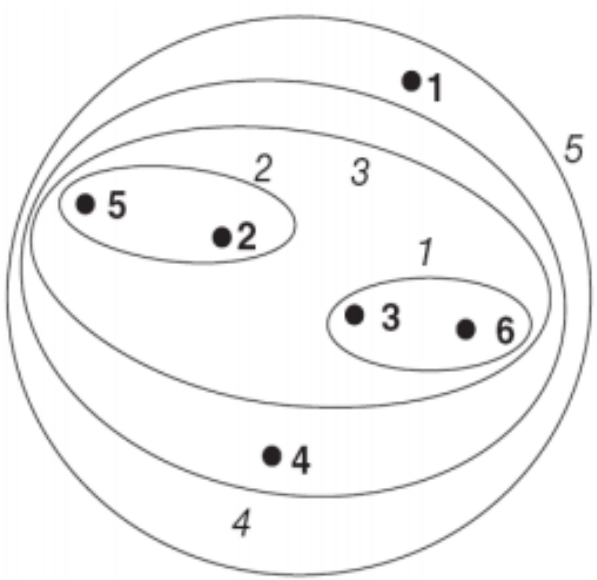
Этот набор содержит пару кластеров R и S, для которой δ_{RS} минимально.

Объекты из этих групп объединяются в один кластер и т.д.

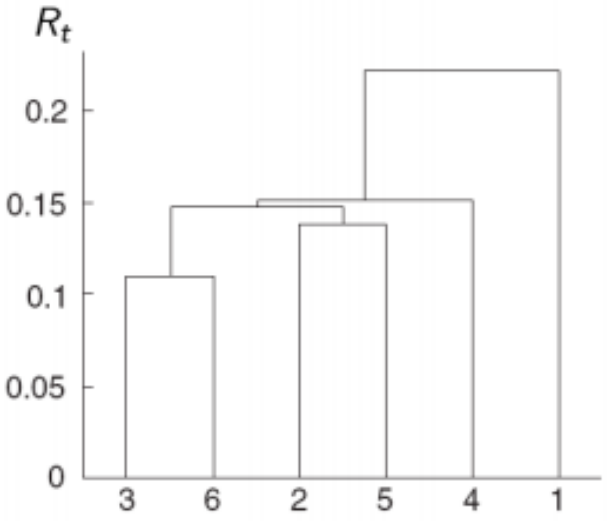
Использование различных функций δ_{RS} приводит к различным уточнениям данного метода

Визуализация кластерной структуры Расстояние ближнего соседа

Диаграмма вложения

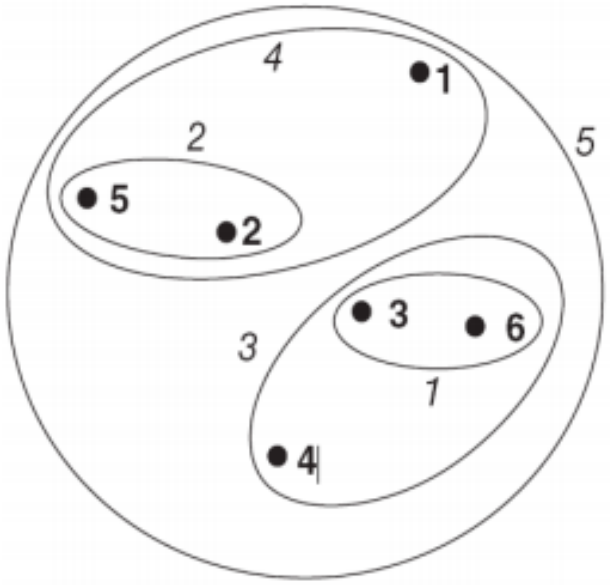


Дендрограмма

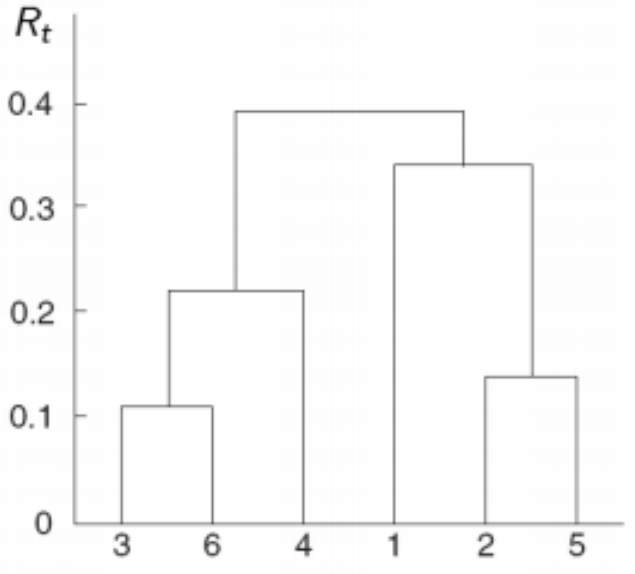


Визуализация кластерной структуры Расстояние дальнего соседа

Диаграмма вложения

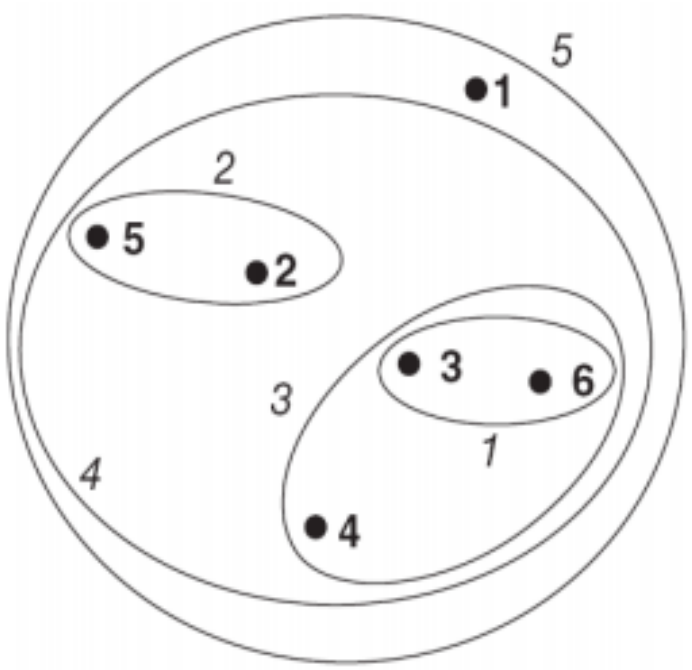


Дендрограмма

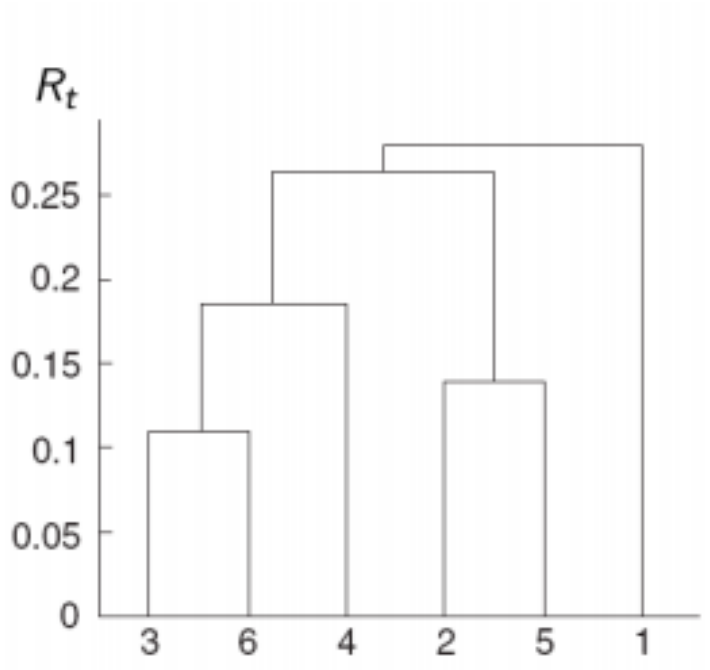


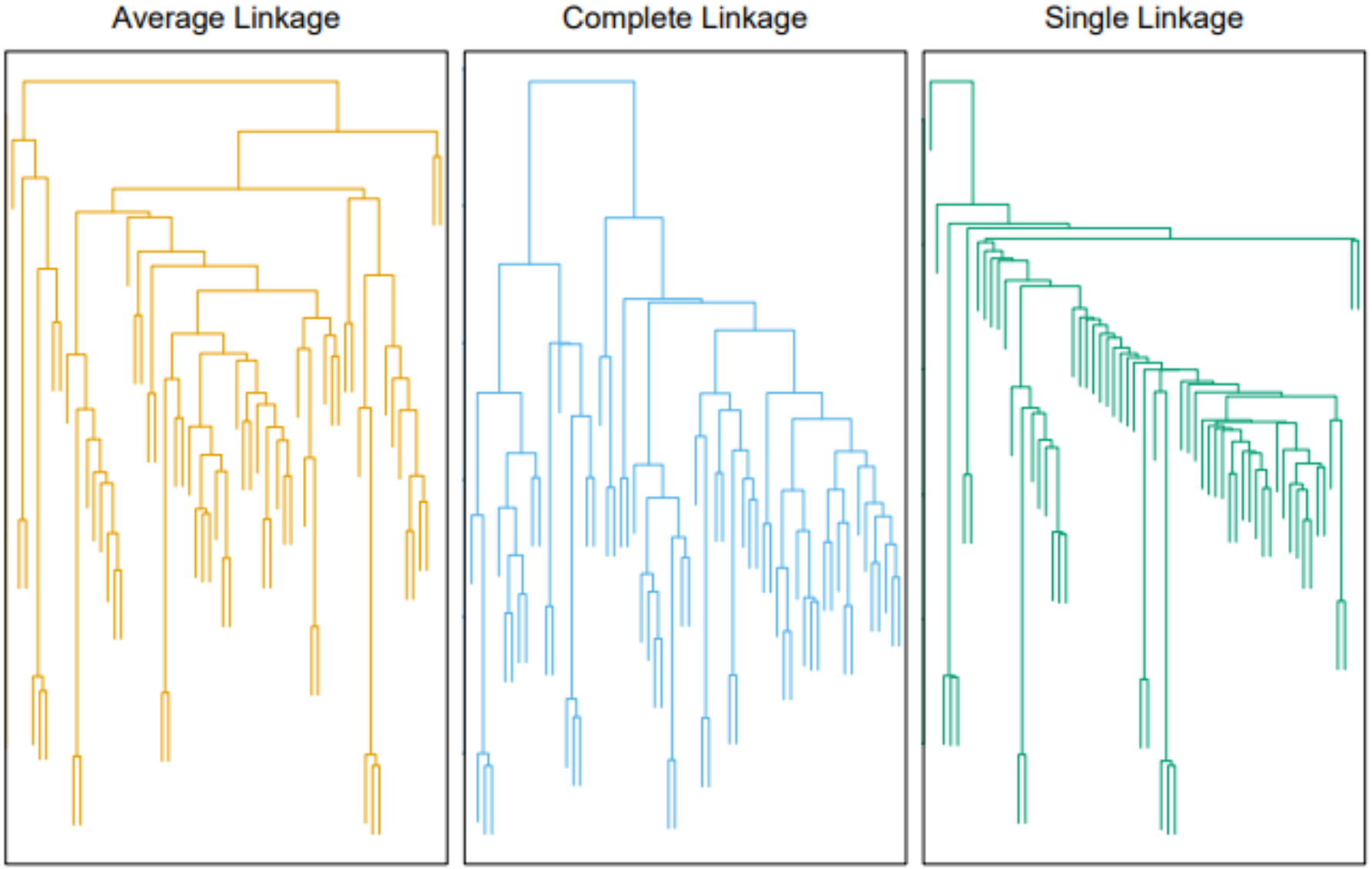
Визуализация кластерной структуры Групповое среднее расстояние

Диаграмма вложения



Дендрограмма




Average Linkage
Complete Linkage
Single Linkage

Разделяющие методы

Строят дерево в направлении от корня к листьям.

Разделяющие методы

Строят дерево в направлении от корня к листьям.

Первый подход:

Ко множеству объектов применить алгоритм (k - means, K-medoids), разбивающий множество на два кластера. Затем разбиваем каждый из полученных кластеров и т.д.

Другой подход (Macnaughton Smith at al. (1965)):

Найдём в R объект, для которого среднее расстояние до всех остальных объектов максимально:

$$i^* = \underset{i \in \mathrm{R}}{\operatorname{argmax}} \frac{1}{|R| - 1} \sum_{i' \in \mathrm{R}, i' \neq i} \delta_{ii'}$$

Удалим i^* из R и поместим в S.

Затем выполняем следующие итерации:

Среди оставшихся в R объектов найдём

$$i^* = \underset{i \in \mathrm{R}}{\operatorname{argmax}} \left(\frac{1}{|R| - 1} \sum_{i' \in \mathrm{R}, i' \neq i} \delta_{ii'} - \frac{1}{|S|} \sum_{i'' \in \mathrm{S}} \delta_{ii''} \right)$$

Поместим i^* в S и удалим из R

Будем продолжать итерацию до тех пор, пока

$$\max \left(\frac{1}{|R| - 1} \sum_{i' \in \mathrm{R}, i' \neq i} \delta_{ii'} - \frac{1}{|S|} \sum_{i'' \in \mathrm{S}} \delta_{ii''} \right) \geq 0$$

Получаем разбиение исходного кластера на два подкластера R и S.

Спасибо за внимание!