Кластерный анализ

Автор: Куликова Маргарита Александровна
Группа: ИВМО-04-22

Оглавление

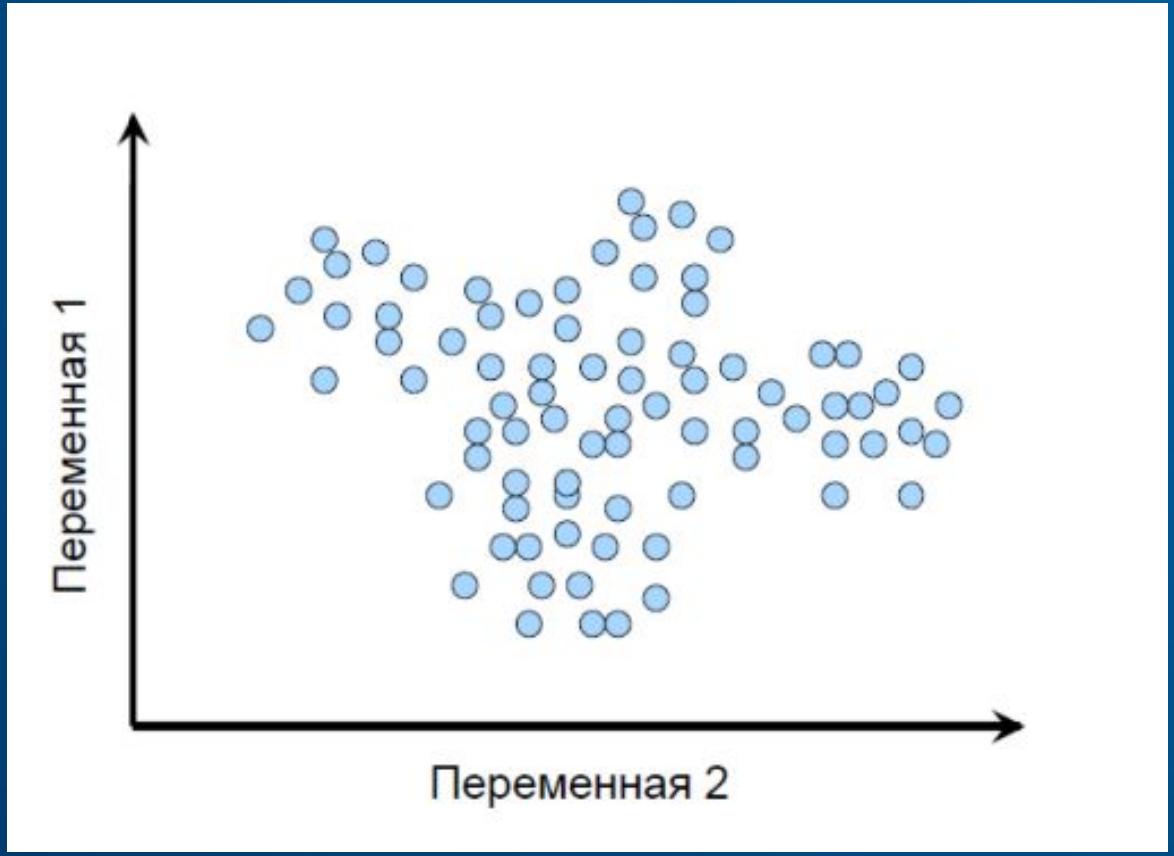
1. Кластерный анализ: понятие и применение

Кластерный анализ – группа методов, используемых для классификации объектов или событий в относительно гомогенные (однородные) группы, которые называют кластерами (clusters).

Кластерный анализ применяется для разбиения исходных данных на поддающиеся интерпретации группы, таким образом, чтобы элементы, входящие в одну группу были максимально «схожи», а элементы из разных групп были максимально «отличными» друг от друга.

Кластерный анализ в теории

Кластерный анализ на практике



2. История возникновения метода

- Первые работы, описывающие методы кластерного анализа относятся к концу 30-х годов.
- Считается, что термин «кластерный анализ» первым в употребление ввёл американский психолог из университета Беркли Роберт Трайон (Robert C. Tryon) в 1939.
- Однако активный интерес к данной теме пришёлся на период 60-80 гг.
- Импульсом для разработки многих кластерных методов послужила книга «Начала численной таксономии», опубликованная в 1963 г. Двумя биологами — Робертом Сокэлом и Петером Снитом
- (Sneath, Sokal).

4. Методы кластерного анализа и его специфика

Кластерный анализ делится на несколько этапов.

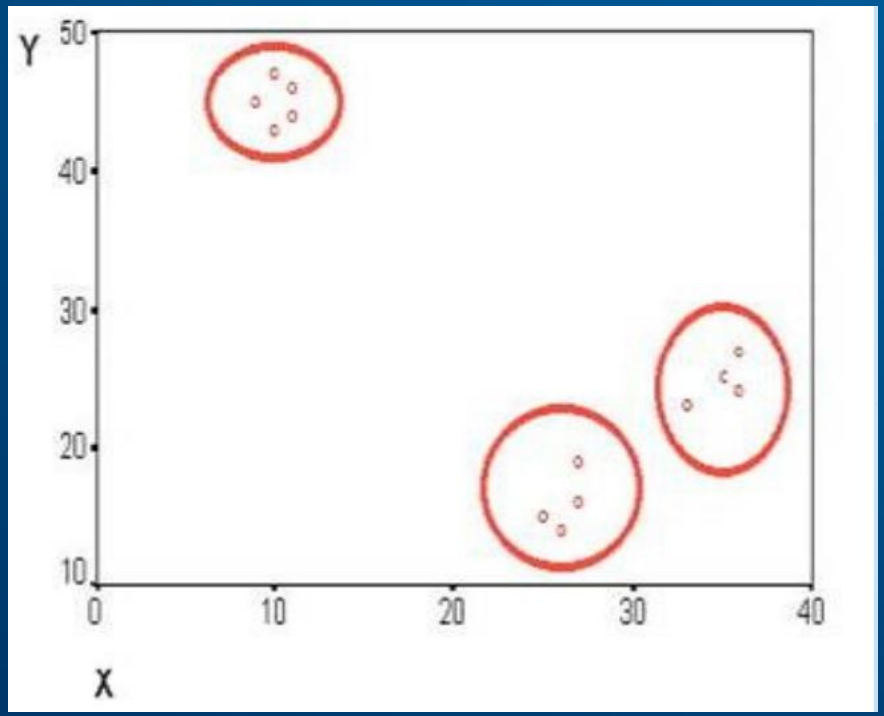
1. Спецификация проблемы, т. е. выбор переменных, на основе которых будет производиться кластеризация.
2. Выбор меры расстояния между объектами.
3. Преобразование переменных.
4. Выбор метода кластеризации.
5. Задание количества кластеров.
6. Интерпретация полученных результатов.
7. Оценка эффективности кластерного анализа.

Методы кластерного анализа

1. АГГЛОМЕРАТИВНЫЕ
2. ДИВИЗИВНЫЕ

5.Меры расстояния

Для того чтобы определить близость, или схожесть, различных объектов, необходимо ввести количественную величину, характеризующую эту близость (схожесть). Естественным представляется ввести некоторую меру расстояния между объектами, аналогичную обычному физическому пространству.



В кластерном анализе используют следующие меры для измерения расстояний.

1. Евклидово расстояние *(Euclidean distances)*. Вычисляется по формуле (по исходным, а не по стандартизованным данным):

$$\text{расстояние}(x,y) = [\Sigma_i (x_i - y_i)^2]^{1/2}$$

2. Квадрат евклидова расстояния *(Squared Euclidean distances)*.

$$\text{расстояние}(x,y) = \Sigma_i (x_i - y_i)^2$$

3. Расстояние городских кварталов *(City-block (Manhattan) distances)*.

$$\text{расстояние}(x,y) = \Sigma_i |x_i - y_i|$$

4. Расстояние Чебышева *(Chebychev ditances metric)*.

$$\text{расстояние}(x,y) = \text{Максимум}|x_i - y_i|$$

5. Степенное расстояние.

$$\text{расстояние}(x,y) = (\Sigma_i |x_i - y_i|^p)^{1/r}$$

где *r* и *p* - параметры, определяемые пользователем. Если оба они равны 2, то это расстояние совпадает с расстоянием Евклида.

6. Процент несогласия *(Percent disagreement)*.

$$\text{расстояние}(x,y) = (\text{Количество } x_i \neq y_i)/ i$$

6. Алгоритмы объединения в кластеры

Существует ряд методов для объединения в кластеры.

1. Метод ближайшего соседа (Euclidean distances) одиночная связь, Single linkage).
2. Метод наиболее удаленного соседа (полная связь, Complete linkage).
3. Невзвешенное попарное среднее (Unweighted pair-group average).
4. Взвешенное попарное среднее (Weighted pair-group average).
5. Невзвешенный центроидный метод (Unweighted pair-group centroid).
6. Взвешенный центроидный метод (Euclidean distances) медиана).
7. Метод Варда (Ward's method).

Спасибо за внимание!