Кластеризация

Понятие кластеризации

- Кластеризация (или кластерный анализ) — это задача разбиения множества объектов на группы, называемые кластерами.

1. Внутри каждой группы должны оказаться «похожие» объекты, а объекты разных групп должны быть как можно более отличны.
2. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы

Этапы кластеризации

A. Отбор выборки объектов для кластеризации.
B. Определение множества переменных, по которым будут оцениваться объекты в выборке. При необходимости – нормализация значений переменных.
C. Вычисление значений меры сходства между объектами.
D. Применение метода кластерного анализа для создания групп сходных объектов (кластеров).
E. Представление результатов анализа

Меры расстояний

- составить вектор характеристик для каждого объекта
- можно провести нормализацию, чтобы все компоненты давали одинаковый вклад при расчете «расстояния».
- для каждой пары объектов измеряется «расстояние» между ними — степень похожести.

Примеры формул для вычислений

- Евклидово расстояние

$$\rho(x, x') = \sqrt{\sum_{i}^{n} (x_i - x_i')^2}$$

- Квадрат евклидова расстояния

$$\rho(x, x') = \sum_{i}^{n} (x_i - x_i')^2$$

Примеры формул для вычислений (2)

- Расстояние городских кварталов (манхэттенское расстояние) – среднее разностей по координатам. В большинстве случаев эта мера расстояния приводит к таким же результатам, как и для обычного расстояния Евклида. Однако для этой меры влияние отдельных больших разностей (выбросов) уменьшается .

$$\rho(x,x') = \sum_{i}^{n} | x_i - x_i' |$$

Примеры формул для вычислений (3)

- Расстояние Чебышева. Это расстояние может оказаться полезным, когда нужно определить два объекта как «различные», если они различаются по какой-либо одной координате.

$$\rho(x, x') = \max(|x_i - x_i'|)$$

Примеры формул для вычислений (4)

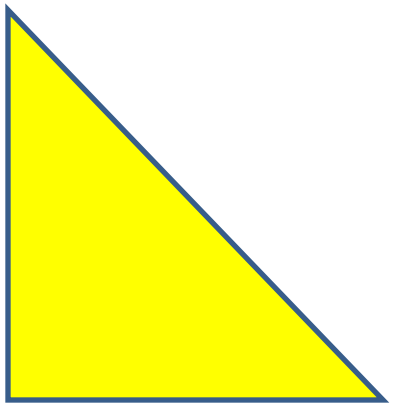
- Степенное расстояние. Применяется в случае, когда необходимо увеличить или уменьшить вес, относящийся к размерности, для которой соответствующие объекты сильно отличаются

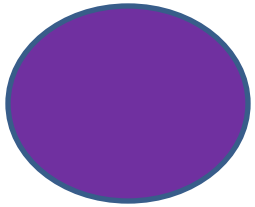
$$\rho(x, x') = \sqrt[r]{\sum_{i}^{n} (x_i - x_i')^p}$$

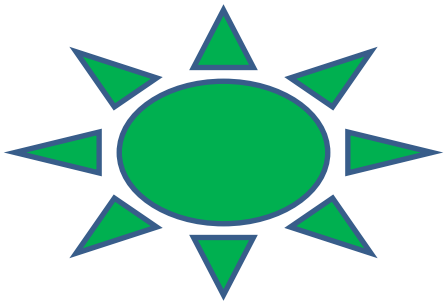
Примеры формул для вычислений (5)

- где r и p – параметры, определяемые пользователем. Параметр p ответственен за постепенное взвешивание разностей по отдельным координатам, параметр r ответственен за прогрессивное взвешивание больших расстояний между объектами.

Практическое задание







Практическое задание (2)

- Сформулировать 5-10 характеристических свойств для картинок.
- Определить их значения для каждого изображения.
- Посчитать расстояние между картинками, используя разные меры.

Алгоритмы кластеризации

- **Алгоритмы иерархической кластеризации**

восходящие и нисходящие алгоритмы.

Нисходящие алгоритмы работают по принципу «сверху-вниз»: в начале все объекты помещаются в один кластер, который затем разбивается на все более мелкие кластеры.

Более распространены восходящие алгоритмы, которые в начале работы помещают каждый объект в отдельный кластер, а затем объединяют кластеры во все более крупные, пока все объекты выборки не будут содержаться в одном кластере. Таким образом строится система вложенных разбиений.

Алгоритмы кластеризации (2)

- **Алгоритмы квадратичной ошибки**

Задачу кластеризации можно рассматривать как построение оптимального разбиения объектов на группы. При этом оптимальность может быть определена как требование минимизации среднеквадратической ошибки разбиения

$$e^2(X,L) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \| x_i^{(j)} - c_j \|^2$$

Алгоритмы кластеризации (3)

- **Нечеткие алгоритмы**

Наиболее популярным алгоритмом нечеткой кластеризации является алгоритм c-средних (c-means). Он представляет собой модификацию метода k-средних.

Алгоритмы кластеризации (4)

- **Алгоритмы, основанные на теории графов**

Суть таких алгоритмов заключается в том, что выборка объектов представляется в виде графа $G=(V, E)$, вершинам которого соответствуют объекты, а ребра имеют вес, равный «расстоянию» между объектами.

Алгоритмы кластеризации (5)

- **Алгоритм выделения связных компонент**

В алгоритме выделения связных компонент задается входной параметр *R* и в графе удаляются все ребра, для которых «расстояния» больше *R*. Соединенными остаются только наиболее близкие пары объектов. Смысл алгоритма заключается в том, чтобы подобрать такое значение *R*, лежащее в диапазон всех «расстояний», при котором граф «развалится» на несколько связных компонент. Полученные компоненты и есть кластеры.

Алгоритмы кластеризации (6)

- **Алгоритм минимального покрывающего дерева**

Алгоритм минимального покрывающего дерева сначала строит на графе минимальное покрывающее дерево, а затем последовательно удаляет ребра с наибольшим весом.

Алгоритмы кластеризации (7)

- **Послойная кластеризация**

Алгоритм послойной кластеризации основан на выделении связных компонент графа на некотором уровне расстояний между объектами (вершинами). Уровень расстояния задается порогом расстояния *c*.