Оценка качества четкой кластеризации

Елена Сивоголовко

Санкт-Петербургский государственный университет
математико-механический факультет

План доклада

- **Кластеризация: основные понятия**
- Оценка качества кластеризации
  - Внешние метрики
  - Внутренние метрики
  - Относительные метрики
- Сравнение метрик оценки качества
  - Тестовые множества
  - Алгоритмы кластеризации
  - Индексы
- Эксперименты
- Заключение

Кластер

Кластер (cluster) — объединение нескольких однородных элементов, которое может рассматриваться как самостоятельная единица, обладающая определёнными свойствами.

- Информатика
- Астрономия
- Химия
- Экономика
- Лингвистика
- Музыка

Кластер: информатика

- единица хранения данных на диске;
- группа компьютеров, использующихся как единый ресурс;
- специализированный объект базы данных для физически совместного хранения одной или нескольких таблиц

Кластер: data mining

Кластер — объединение нескольких однородных элементов, которое может рассматриваться как самостоятельная единица.
Кластеризация – задача разбиения множества на однородные группы, так, чтобы элементы в одной группе были максимально схожи друг с другом, а элементы из разных групп значительно отличались.
Data Mining
Machine learning
Unsupervised learning

Кластеризация: общая схема

1. Выделение значимых характеристик
2. Определение метрики схожести
3. Разбиение на группы
4. Оценка качества результатов
5. Представление и интерпретация результатов

План доклада

- Кластеризация: основные понятия
- **Оценка качества кластеризации**
  - Внешние метрики
  - Внутренние метрики
  - Относительные метрики
- Сравнение метрик оценки качества
  - Тестовые множества
  - Алгоритмы кластеризации
  - Индексы
- Эксперименты
- Заключение

Качество кластеризации

- **Качество ПО стандарт ISO 9000**:
  "The totality of features and characteristics of a product or service that bear on its ability to satisfy stated or implied needs"
- **Качество кластеризации (Cluster validity)**:
  "The adequacy of a clustering structure refers to the sense in which the clustering structure provides true information about the data, or the ability of recovered structure to reflect the intrinsic character of the data"

Оценка качества кластеризации

Методы (индексы) оценки качества (cluster validity methods) — инструментарий для количественной оценки результатов кластеризации.

Методы оценки качества

- Для четкой кластеризации: кластеры не пересекаются.
- для нечеткой кластеризации: допускается пересечение кластеров.

- Для иерархических структур.
- Для не иерархических структур.
- Для отдельных кластеров

- Внешние (external)
- Внутренние (internal)
- Относительные (relative)

Rand, Jaccard, FM

Рассмотрим пары (x_i, x_j) из элемнтов X. Подсчитаем количество пар, в которых :

1. элементы принадлежат одному кластеру и одному классу: SS.
2. элементы принадлежат одному кластеру, но разным классам: SD
3. элементы принадлежат разным кластерам, но одному классу: DS
4. элементы принадлежат разным классам и разным кластерам: DD

Rand, Jaccard, FM

$$Rand = \frac{SS + DD}{SS + DS + SD + DD} \tag{1}$$

$$Jaccard = \frac{SS}{SS + SD + DS} \tag{2}$$

$$FM = \sqrt{\frac{SS}{SS + SD} * \frac{SS}{SS + DS}} \tag{3}$$

RS (R Squared) Индекс

Пусть

1. SS_w сумма квадратов расстояний внутри кластера
2. SS_b сумма квадратов расстояний между кластерами
3. SS_t сумма квадратов расстояний по всему множеству, причем $SS_t = SS_w + SS_b$

Формула индекса RS:

$$RS = \frac{SS_b}{SS_t} = \frac{SS_t - SSw}{SS_t} \tag{16}$$