Обучение без учителя

Интеллектуальный анализ данных, 20!7

Малютин Евгений Алексеевич

Планчик

Сегондя в программе:

- Как выжить в суровом мире без Y
- Графовые методы выживания
- Статистические методы выживания
- Иерархические методы выживания
- Качественная оценка уцелевших

Задача обучения без учителя

Отсутствует целевая переменная

Вопросы:

- Существует информативный способ визуализации данных?
- Можем ли мы выделить подгруппы среди переменных?
- Существуют ли незаметные зависимости или паттерны поведения? Построить иерархию?
- Хранить меньше данных?
- Устранить шумы?

Где ипользуют:

- Анализ изображений
- Визуализация
- Биоинформатика
- Маркетинг
- Составные части сложных алгоритмов
- Везде

Задача обучения без учителя

Постановка задачи кластеризации:

Дано:

1. Пространство объектов X
2. Обучающая выборка X^l
3. Функция расстояния между объектами $\rho: X \times X \to [0, \infty)$

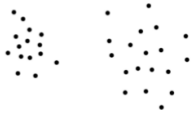
Найти:

- Множество кластеров Y: каждый кластер состоит из близких объектов, а объекты разных кластеров существенно различны
- Алгоритм кластеризации: $\alpha : X \to Y$

Проблемы?

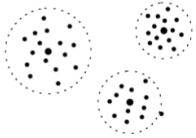
- Существует множество критериев качества
- Число кластеров обычно заранее неизвестно
- Результат существенно зависит от функции расстояния, которую задает эксперт
- Точной постановки задачи нет

Задача обучения без учителя:



внутрикластерные расстояния, как правило, меньше межкластерных

ленточные кластеры



кластеры с центром

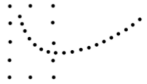
Задача обучения без учителя:

кластеры могут соединяться перемычками

кластеры могут накладываться на разреженный фон из редко расположенных объектов

кластеры могут перекрываться

Задача обучения без учителя:



кластеры могут образовываться не по сходству, а по иным типам регулярностей

кластеры могут вообще отсутствовать

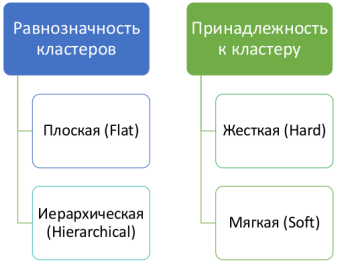
При этом

- Каждый метод имеет свои ограничения и выделяет лишь некоторые типы кластеров
- Понятие "тип кластерной структуры"так же зависит от метода и может вообще отсутствовать

Задача обучения без учителя:

Требования к алгоритмам

- масштабируемость
- способность обрабатывать различные типы атрибутов
- работать с шумными данными
- инкрементные обновления
- кластеры произвольной формы и ограничения
- интерпретируемость и удобство использования



Задача обучения без учителя

Типология по алгоритмам:

- **Hierarchical clustering:** BIRCH, CURE, SLINK, Single-link Algorithms Based on Minimum Spanning Trees, CLINK, DIANA, DISMEA
- **Centroid-based clustering:** K-means, variations of the k-means
- **Distribution-based clustering:** The EM Algorithm
- **Density-based clustering:** DBSCAN, OPTICS, DENCLUE, BRIDGE, DBCLASD
- **Graph-based Clustering Algorithms:** Chameleon, CACTUS, A Dynamic System–based Approach, ROCK
- **Grid-based Clustering Algorithms:** STING, OptiGrid, GRIDCLUS, GDILC, WaveCluster
- **Subspace Clustering:** CLIQUE, PROCLUS, ORCLUS, ENCLUS, FINDIT, MAFIA, DOC, CLTree, PART, SUBCAD

Графовые методы:

Выделение связных компонент

Представим обучающую выборку в виде графа: вершины – обучающие объекты x_i, ребра задают расстояния между соответствующими объектами $\rho_{ij} = \rho(x_i, x_j)$. Пусть задан параметр K.

- Удалить из графа все ребра, веса которых больше R.
- Посчитать число компонент связности K графа.
- Если K меньше искомого, то уменьшить R и повторить.

Особенности

- задаётся неудобный параметр R
- высокая чувствительность к шуму

Алгоритм КНП – Кратчайший незамкнутый путь

Исходное состояние графа: известна матрица расстояний, но ребра отсутствуют.

- Найти пару вершин с наименьшим расстоянием $\rho(x_i, x_j)$ и соединить их ребром.
- Пока остаются изолированные вершины:
  1. Найти вершину, ближайшую к некоторой неизолированной
  2. Соединить их ребром.
- Удалить $K - 1$ самых длинных ребер.

Особенности

- задаётся число кластеров K
- высокая чувствительность к шуму

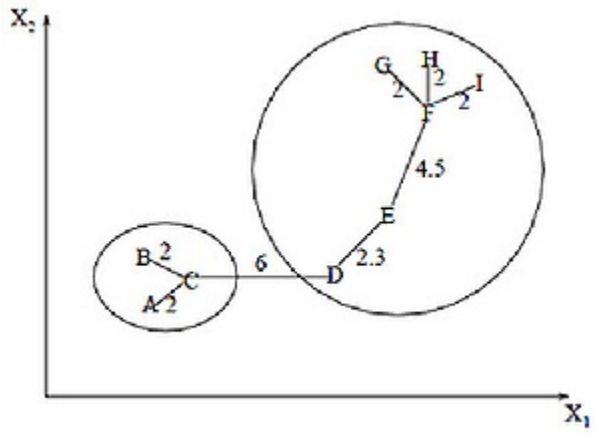


Рис.: Графовый метод кластеризации (угадайте какой?)

Статистические методы

EM:

- Мягкая $Py_i = y$
- Форма кластеров настраивается

K-Means:

- Жесткая $y_i = y$
- Форма кластеров задана метрикой
- Чувствительность к начальному приближению
- Необходимость задания числа кластеров

Обозначения:

- Пусть пространство $X = R^n$. Тогда обучающая выборка состоит из: $x_i = (f_1(x_i), ...f_n(x_i))$; "задано"множество кластеров Y
- Обозначим центры кластеров как $\mu_y = (\mu_{y1}, ..., \mu_{yn})$
- Пусть ω_y - априорная вероятность кластера y, $\sum_{y \in Y} \omega_y = 1$
- Будем считать что X^l независима, случайна, и пришла из смеси распределений:

$$p(x) = \sum y \in Y \omega_y p_y(x)$$

- Полагая кластера n-мерными и гауссовскими имеем:

$$p(x) = (2\pi)^{\frac{n}{2}}(\sigma_{y1}, ..., \sigma_{yn}) exp(-\frac{1}{2}\rho_y^2(x, \mu_y));$$

Σ – диагональная матрица ковариаций

$$\rho_y^2(x, x') = \sum_{i=1}^{n} \sigma_{yi}^{-2} |f_j(x) - f_j(x')|^2$$

EM-алгоритм:

1. Пусть дано начальное приближение $\omega_y,\ \mu_y, \Sigma_y\ \forall\ y \in Y$.
2. E-шаг: $\forall\ y, i = 1..l$:
$$g_i y \leftarrow P(y|x_i) = \frac{\omega_i p_y(x_i)}{\sum\limits_{z \in Y} w_z p_z(x)}$$
3. M-шаг $\forall\ y, i = 1..l$:
$$w_y \leftarrow \frac{1}{l}\sum_{i=1}^{l} g_{iy} \qquad \mu_{yi} \leftarrow \frac{1}{l * w_y}\sum_{i=1}^{l} g_{iy} f_j(x_i)$$
$$\sigma^2_{yj} =\leftarrow \frac{1}{l * w_y}\sum_{i=1}^{l} g_{iy}(f_j(x_i) - \mu_{yj})^2$$
4. $y_i = arg\ \max_{y \in Y} g_{iy}; i = 1..l$
5. Повторять пока не фиксируются y_i

Вариант Болла-Холла:

Упрощение EM-алгоритми, $X = R^n$

1. Начальное приближение центров кластеров μ_y
2. Повторяем:
3. Вычислить кластера $y_i = arg \min_{y \in Y} \rho(x_i, y_)$
4. Пересчитать центры кластеров: $\mu_{yj} = \frac{\sum_{i=1}^{l} [y=y_i] f_j(x_i)}{\sum_{i=1}^{l} [y=y_i]}$
5. Пока не устаканится

Иерархическая кластеризация:

Обобщенный алгоритм Ланса-Уильямса

1. Сформировать одноэлементные кластеры:

$$C_1 = \{\{x_1\}, \{x_2\}, ..., \{x_n\}\}; \quad R(\{x_i\}) = \rho(x_i, x_j)$$

2. $\forall t = 2, ..., l$:
   1. найти в C_{t-1} два ближайшиъ кластера (U, V)
   2. найти в C_{t-1} два ближайших кластера (U, V)
   3. Сформировать новый кластер: $W = (U \cup V)$

Иерархическая кластеризация:

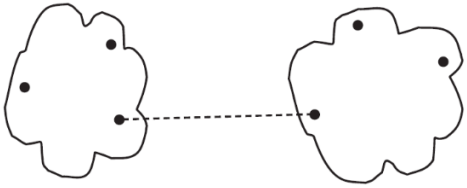
Поиск расстояния:

Известны расстояния $R(U,S), R(U,V), R(V,S)$
Сформировали $W = (U \cup V)$
Как определить $R(W,S)$?
Обобщенная формула расстояния:

$$R(U \cup V, S) = \alpha_u \cdot R(U,S) + \alpha_v \cdot R(V,S) + \beta \cdot R(U,V) + \gamma\,[R(U,S) - R(V,S)[$$

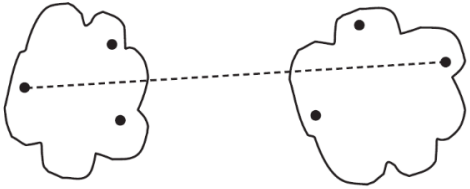
Расстояние ближнего соседа

$$\alpha_U = \alpha_V = \frac{1}{2}$$
$$\beta = 0 \quad \gamma = -\frac{1}{2}$$



Расстояние дальнего соседа

$$\alpha_U = \alpha_V = \frac{1}{2}$$
$$\beta = 0 \quad \gamma = \frac{1}{2}$$

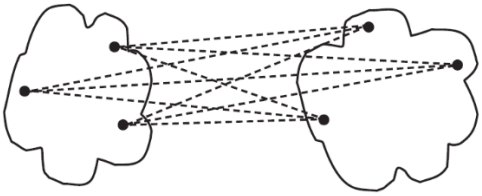


Групповое среднее расстояние

$$\alpha_U = \frac{|U|}{|W|}$$

$$\alpha_V = \frac{|V|}{|W|}$$

$$\beta = \gamma = 0$$



5. Расстояние Уорда:

$$R^y(W,S) = \frac{|S||W|}{|S|+|W|}\,\rho^2\Big(\sum_{w\in W}\frac{w}{|W|},\sum_{s\in S}\frac{s}{|S|}\Big);$$

$$\alpha_U = \frac{|S|+|U|}{|S|+|W|},\ \alpha_V = \frac{|S|+|V|}{|S|+|W|},\ \beta = \frac{-|S|}{|S|+|W|},\ \gamma = 0.$$

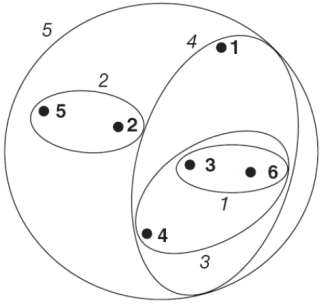
Иерархическая K - кластеризация

Советы

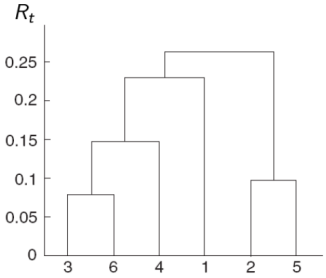
- пользуйтесь Уордом
- пользуйтесь модификациями
- подумайте когда "срезать"

5. Расстояние Уорда:

Диаграмма вложения



Дендрограмма



End

THE END