Методы кластеризации

Воронцов Константин Вячеславович
vokov@forecsys.ru
http://www.MachineLearning.ru/wiki?title=User:Vokov

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

Видеолекции: http://shad.yandex.ru/lectures

ШАД Яндекс • 27 апреля 2015

Содержание: методы кластеризации

1. **Статистические методы кластеризации**
   - Постановка задачи кластеризации
   - EM-алгоритм
   - Метод k-средних
2. **Сети Кохонена**
   - Модели конкурентного обучения
   - Карты Кохонена
3. **Иерархическая кластеризация (таксономия)**
   - Агломеративная иерархическая кластеризация
   - Дендрограмма и свойство монотонности
   - Свойства сжатия, растяжения и редуктивности

Постановка задачи кластеризации

Дано:

X — пространство объектов;

$X^\ell = \{x_i\}_{i=1}^\ell$ — обучающая выборка;

$\rho: X \times X \to [0, \infty)$ — функция расстояния между объектами.

Найти:

Y — множество кластеров и

$a: X \to Y$ — алгоритм кластеризации, такие, что:

— каждый кластер состоит из близких объектов;

— объекты разных кластеров существенно различны.

Кластеризация — это *обучение без учителя*.

Некорректность задачи кластеризации

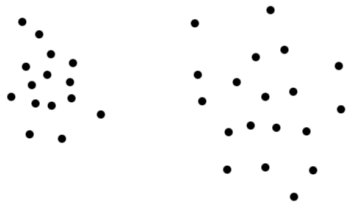
Решение задачи кластеризации принципиально неоднозначно:

- точной постановки задачи кластеризации нет;
- существует много критериев качества кластеризации;
- существует много эвристических методов кластеризации;
- число кластеров $|Y|$, как правило, неизвестно заранее;
- результат кластеризации существенно зависит от метрики ρ, которую эксперт задаёт субъективно.

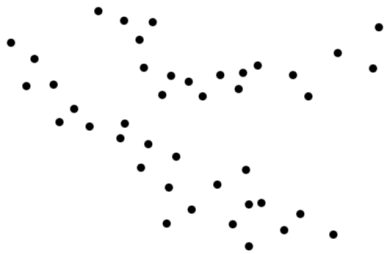
Цели кластеризации

- **Упростить дальнейшую обработку данных**,
  разбить множество X^ℓ на группы схожих объектов
  чтобы работать с каждой группой в отдельности
  (задачи классификации, регрессии, прогнозирования).
- **Сократить объём хранимых данных**,
  оставив по одному представителю от каждого кластера
  (задачи сжатия данных).
- **Выделить нетипичные объекты**,
  которые не подходят ни к одному из кластеров
  (задачи одноклассовой классификации).
- **Построить иерархию множества объектов**
  (задачи таксономии).

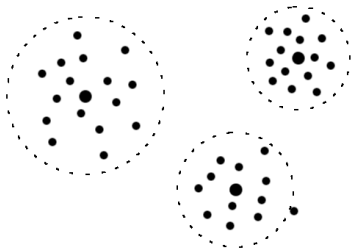
Типы кластерных структур



внутрикластерные расстояния, как правило, меньше межкластерных

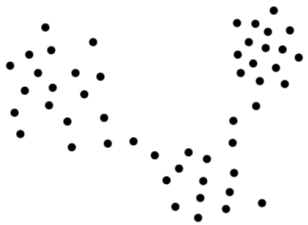


ленточные кластеры

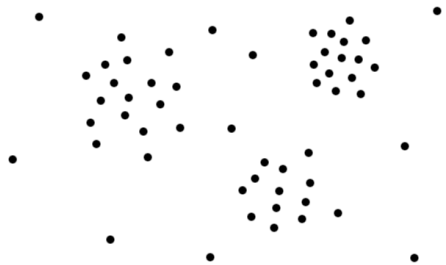


кластеры с центром

Типы кластерных структур



кластеры могут соединяться перемычками



кластеры могут накладываться на разрежен-ный фон из редко расположенных объектов

кластеры могут перекрываться

Типы кластерных структур

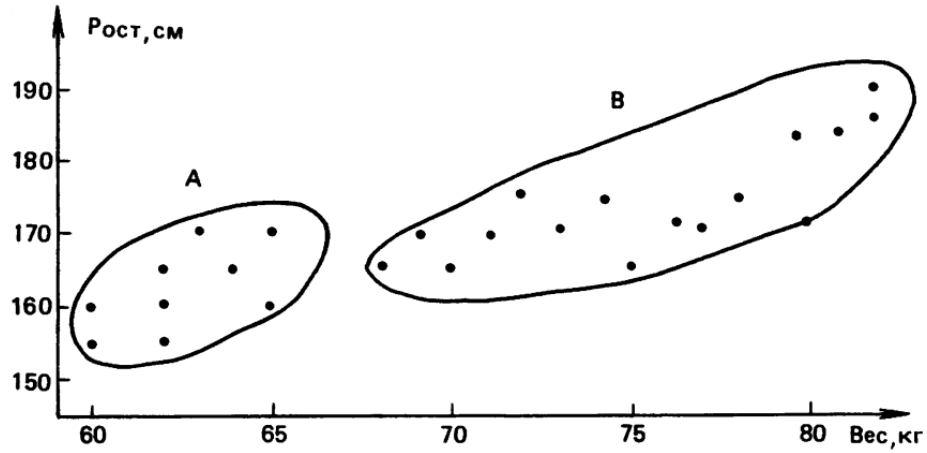
кластеры могут образовываться не по сходству, а по иным типам регулярностей

кластеры могут вообще отсутствовать

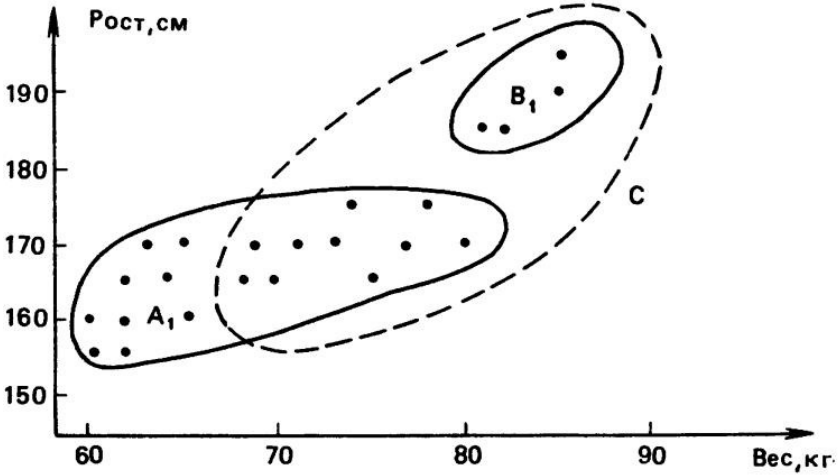
- Каждый метод кластеризации имеет свои ограничения и выделяет кластеры лишь некоторых типов.
- Понятие «тип кластерной структуры» зависит от метода и также не имеет формального определения.

Проблема чувствительности к выбору метрики

Результат зависит от нормировки признаков:



A — студентки,
B — студенты



после перенормировки
(сжали ось «вес» вдвое)

Гипотеза (о вероятностной природе данных)

Выборка X^ℓ случайна, независима, из смеси распределений

$$p(x) = \sum_{y \in Y} w_y p_y(x), \qquad \sum_{y \in Y} w_y = 1,$$

$p_y(x)$ — *плотность*, w_y — *априорная вероятность* кластера y.

Гипотеза (о пространстве объектов и форме кластеров)

$X = \mathbb{R}^n$, $x_i \equiv \big(f_1(x_i), \ldots, f_n(x_i)\big)$; кластеры n-мерные гауссовские:

$$p_y(x) = (2\pi)^{-\frac{n}{2}} (\sigma_{y1} \cdots \sigma_{yn})^{-1} \exp\left(-\tfrac{1}{2}\rho_y^2(x, \mu_y)\right),$$

$\mu_y = (\mu_{y1}, \ldots, \mu_{yn})$ — центр кластера y;
$\Sigma_y = \operatorname{diag}(\sigma_{y1}^2, \ldots, \sigma_{yn}^2)$ — диагональная матрица ковариаций;

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

EM-алгоритм (повторение)

1: начальное приближение w_y, μ_y, Σ_y для всех $y \in Y$;
2: **повторять**
3: E-шаг (expectation):

$$g_{iy} := P(y|x_i) \equiv \frac{w_y p_y(x_i)}{\sum_{z \in Y} w_z p_z(x_i)}, \quad y \in Y, \quad i = 1, \ldots, \ell;$$

4: M-шаг (maximization):

$$w_y := \frac{1}{\ell} \sum_{i=1}^{\ell} g_{iy}, \quad y \in Y;$$

$$\mu_{yj} := \frac{1}{\ell w_y} \sum_{i=1}^{\ell} g_{iy} f_j(x_i), \quad y \in Y, \quad j = 1, \ldots, n;$$

$$\sigma^2_{yj} := \frac{1}{\ell w_y} \sum_{i=1}^{\ell} g_{iy} (f_j(x_i) - \mu_{yj})^2, \quad y \in Y, \quad j = 1, \ldots, n;$$

5: $y_i := \arg\max_{y \in Y} g_{iy}, \quad i = 1, \ldots, \ell;$
6: **пока** y_i не перестанут изменяться;

Метод k-средних (k-means)

$X = \mathbb{R}^n$. Упрощённый аналог EM-алгоритма:

1: начальное приближение центров μ_y, $y \in Y$;
2: **повторять**
3: аналог E-шага:
отнести каждый x_i к ближайшему центру:
$y_i := \arg\min_{y \in Y} \rho(x_i, \mu_y)$, $i = 1, \ldots, \ell$;
4: аналог M-шага:
вычислить новые положения центров:
$$\mu_{yj} := \frac{\sum_{i=1}^{\ell} [y_i = y] f_j(x_i)}{\sum_{i=1}^{\ell} [y_i = y]}, \quad y \in Y, \quad j = 1, \ldots, n;$$
5: **пока** y_i не перестанут изменяться;

Модификации и обобщения

Варианты k-means:

- вариант Болла-Холла (на предыдущем слайде);
- вариант МакКина: при каждом переходе объекта из кластера в кластер их центры пересчитываются;

Основные отличия EM и k-means:

- EM: мягкая кластеризация: $g_{iy} = \mathsf{P}\{y_i = y\}$;
  k-m: жёсткая кластеризация: $g_{iy} = [y_i = y]$;
- EM: форма кластеров эллиптическая, настраиваемая;
  k-m: форма кластеров жёстко определяется метрикой ρ;

Гибридные варианты по пути упрощения EM:

- EM с жёсткой кластеризацией на E-шаге;
- EM без настройки дисперсий (сферические гауссианы);

Недостатки k-means

- Чувствительность к выбору начального приближения.
- Необходимость задавать k;

Способы устранения этих недостатков:

- Несколько случайных кластеризаций;
  выбор лучшей по функционалу качества.
- Постепенное наращивание числа кластеров k
  (аналогично EM-алгоритму)

Оптимизационная задача кластеризации

Дано:
$X = \mathbb{R}^n$, $Y = \{1, \ldots, M\}$ — множество кластеров;
$X^\ell = \{x_i\}_{i=1}^\ell$ — обучающая выборка объектов;
$\rho\colon X \times X \to [0, \infty)$ — функция расстояния между объектами.

Алгоритм кластеризации $a\colon X \to Y$
относит объект $x \in X$ к ближайшему кластеру
(правило жёсткой конкуренции WTA — Winner Takes All):

$$a(x) = \arg\min_{m \in Y} \rho(x, w_m),$$

где $w_m \in \mathbb{R}^n$, $m = 1, \ldots, M$ — центры кластеров.

Минимизация среднего внутрикластерного расстояния:

$$Q(w; X^\ell) = \frac{1}{2} \sum_{i=1}^{\ell} \rho^2(x_i, w_{a(x_i)}) \to \min_w, \quad w = (w_1, \ldots, w_M);$$

Метод стохастического градиента

Минимизация среднего внутрикластерного расстояния:

$$Q(w; X^\ell) = \frac{1}{2}\sum_{i=1}^{\ell} \rho^2(x_i, w_{a(x_i)}) \to \min_w, \quad w = (w_1, \ldots, w_M);$$

Пусть метрика евклидова, $\rho^2(x_i, w_m) = \|w_m - x_i\|^2$.

$$\frac{\partial Q(w; X^\ell)}{\partial w_m} = \sum_{i=1}^{\ell} (w_m - x_i)\big[a(x_i) = m\big].$$

Градиентный шаг в методе SG: для случайного $x_i \in X^\ell$

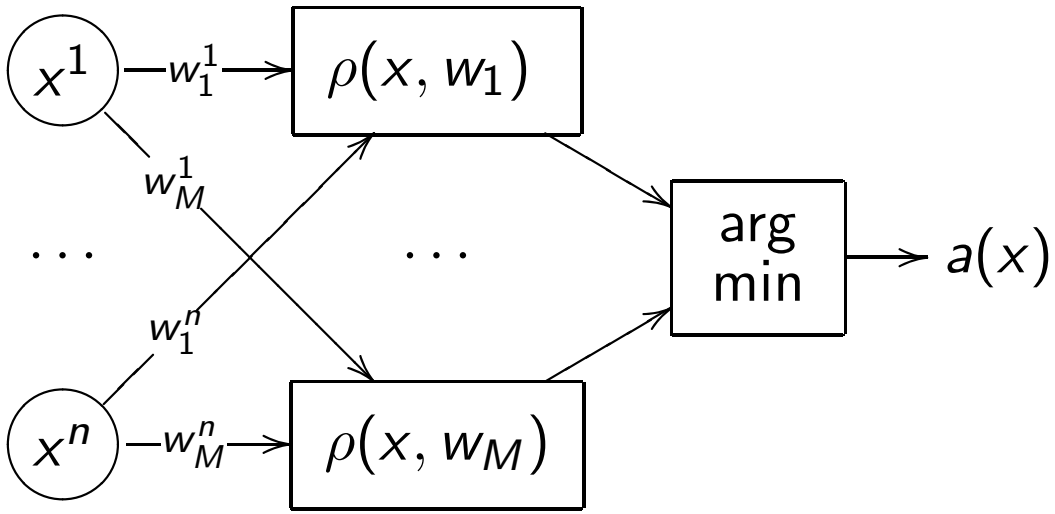
$$w_m := w_m + \eta(x_i - w_m)\big[a(x_i) = m\big]$$

(если x_i относится к кластеру m, то w_m сдвигается в сторону x_i).

Сеть Кохонена (сеть с конкурентным обучением)

Структура алгоритма — двухслойная нейронная сеть:

$$a(x) = \arg\min_{m \in Y} \rho(x, w_m) :$$



Градиентное правило обучения напоминает персептрон:

$$\text{если} \quad a(x_i) = m, \quad \text{то} \quad w_m := w_m + \eta(x_i - w_m).$$

Алгоритм SG (Stochastic Gradient)

Вход: выборка X^ℓ; темп обучения η; параметр λ;
Выход: центры кластеров $w_1, \ldots, w_M \in \mathbb{R}^n$;

1: инициализировать центры w_m, $m = 1, \ldots, M$;
2: инициализировать текущую оценку функционала:
$Q := \sum_{i=1}^{\ell} \rho^2(x_i, w_{a(x_i)})$;
3: **повторять**
4: выбрать объект x_i из X^ℓ (например, случайно);
5: вычислить кластеризацию: $m := \arg\min_{m \in Y} \rho(x_i, w_m)$;
6: градиентный шаг: $w_m := w_m + \eta(x_i - w_m)$;
7: оценить значение функционала:
$Q := (1 - \lambda)Q + \lambda\rho^2(x_i, w_m)$;
8: **пока** значение Q и/или веса w не стабилизируются;

Жёсткая и мягкая конкуренция

Правило жёсткой конкуренции WTA (winner takes all):

$$w_m := w_m + \eta(x_i - w_m)\big[a(x_i) = m\big], \quad m = 1, \ldots, M,$$

Недостатки правила WTM:

- медленная скорость сходимости;
- некоторые w_m могут никогда не выбираться.

Правило мягкой конкуренции WTM (winner takes most):

$$w_m := w_m + \eta(x_i - w_m)\, K\big(\rho(x_i, w_m)\big), \quad m = 1, \ldots, M,$$

где ядро $K(\rho)$ — неотрицательная невозрастающая функция.

Теперь центры *всех* кластеров смещаются в сторону x_i,
но чем дальше от x_i, тем меньше величина смещения.

Карта Кохонена (Self Organizing Map, SOM)

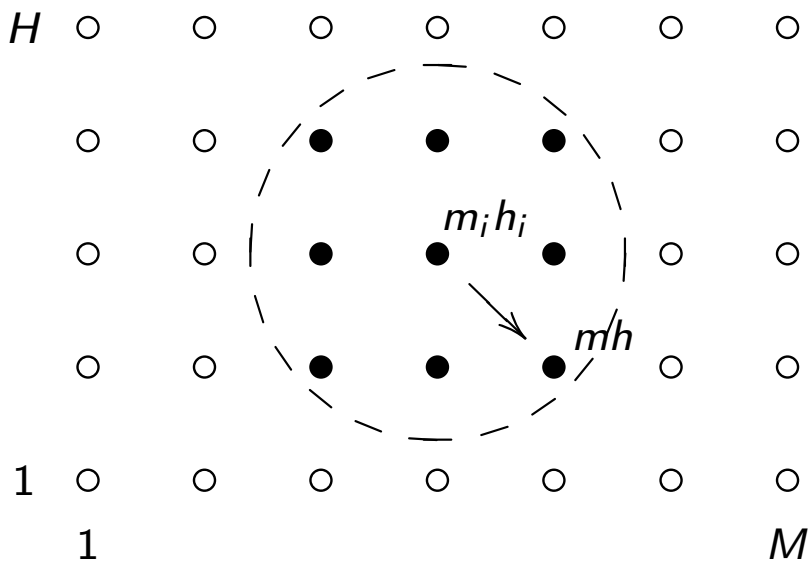
$Y = \{1, \ldots, M\} \times \{1, \ldots, H\}$ — прямоугольная сетка кластеров.

Каждому узлу (m, h) приписан нейрон Кохонена $w_{mh} \in \mathbb{R}^n$.

Наряду с метрикой $\rho(x_i, x)$ на X вводится метрика на сетке Y:

$$r\big((m_i, h_i), (m, h)\big) = \sqrt{(m - m_i)^2 + (h - h_i)^2}.$$

Окрестность(m_i, h_i):



Обучение карты Кохонена

Вход: X^ℓ — обучающая выборка; η — темп обучения;
Выход: $w_{mh} \in \mathbb{R}^n$ — векторы весов, $m = 1..M$, $h = 1..H$;

1: $w_{mh} := \text{random}\left(-\frac{1}{2MH}, \frac{1}{2MH}\right)$ — инициализация весов;
2: **повторять**
3: выбрать объект x_i из X^ℓ случайным образом;
4: WTA: вычислить координаты кластера:
$(m_i, h_i) := a(x_i) \equiv \arg\min_{(m,h)\in Y} \rho(x_i, w_{mh})$;
5: **для всех** $(m, h) \in$ Окрестность(m_i, h_i)
6: WTM: сделать шаг градиентного спуска:
$w_{mh} := w_{mh} + \eta(x_i - w_{mh})\, K\big(r((m_i, h_i), (m, h))\big)$;
7: **пока** кластеризация не стабилизируется;

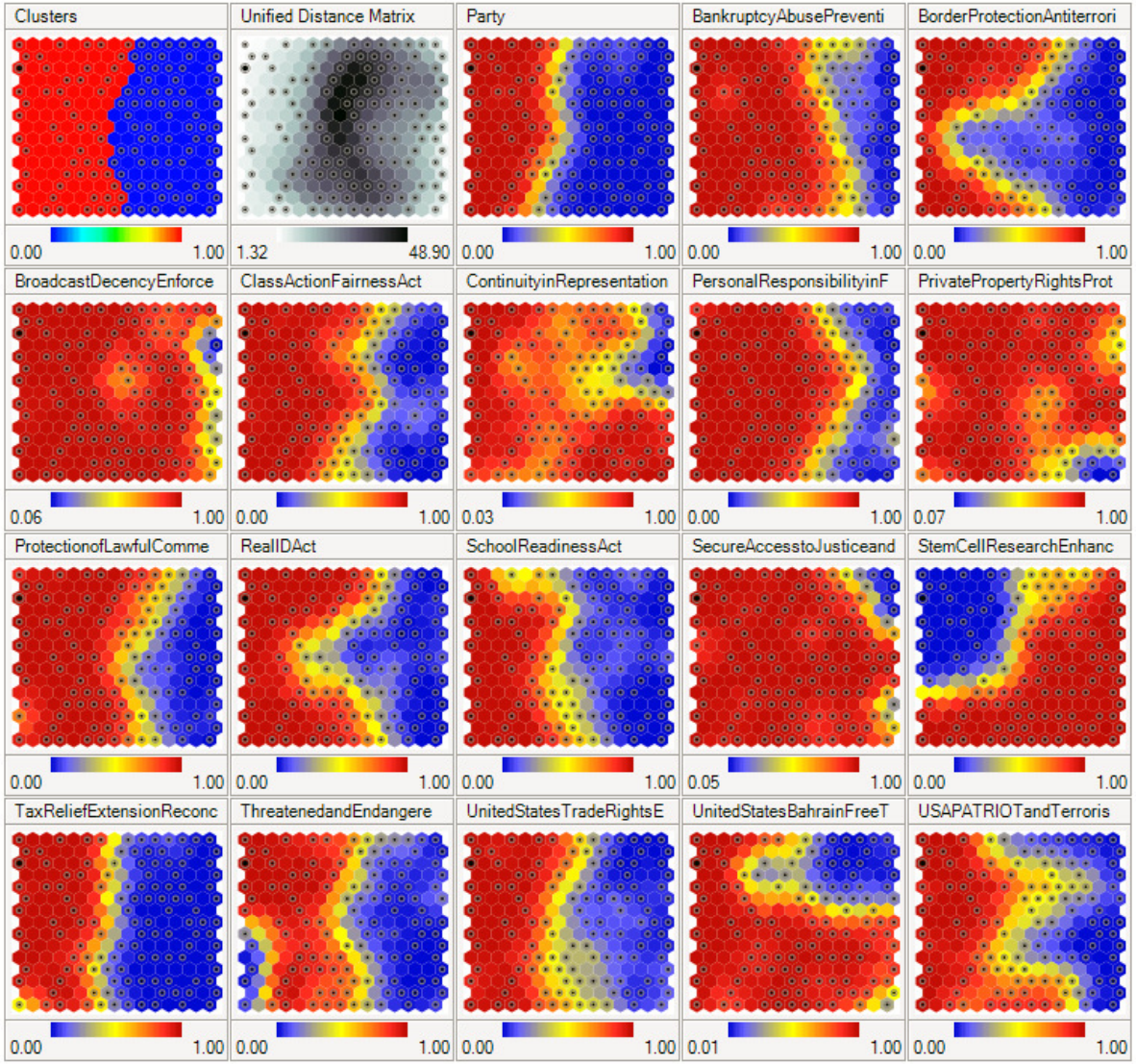
Интерпретация карт Кохонена

Два типа графиков — цветных карт $M \times H$:

- Цвет узла (m, h) — локальная плотность в точке (m, h) — среднее расстояние до k ближайших точек выборки;
- По одной карте на каждый признак:
  цвет узла (m, h) — значение j-й компоненты вектора $w_{m,h}$.

Пример:
Задача UCI house-votes (US Congress voting patterns)
Объекты — конгрессмены;
Признаки — вопросы, выносившиеся на голосование;
Есть целевой признак {демократ,республиканец}.

Интерпретация карт Кохонена (пример)



Достоинства и недостатки карт Кохонена

Достоинства:

- Возможность визуального анализа многомерных данных.

Недостатки:

- **Субъективность.** Карта зависит не только от кластерной структуры данных, но и...
  — от свойств сглаживающего ядра;
  — от (случайной) инициализации;
  — от (случайного) выбора x_i в ходе итераций.
- **Искажения.** Близкие объекты исходного пространства могут переходить в далёкие точки на карте, и наоборот.

Рекомендуется только для разведочного анализа данных.

Резюме по сетям Кохонена

- Сеть Кохонена решает задачу кластеризации.
- Основные стратегии — мягкая WTM и жёсткая WTA.
- Мягкая конкуренция ускоряет сходимость.
- Карта Кохонена используется для визуализации многомерных данных, разведочного анализа данных, интерпретации кластеров по признакам.
- Карта Кохонена может быть субъективной и искажённой

Агломеративная иерархическая кластеризация

Алгоритм Ланса-Уильямса [1967]

сначала все кластеры одноэлементные:
$t := 1; \quad C_t = \{\{x_1\}, \ldots, \{x_\ell\}\};$
$R(\{x_i\}, \{x_j\}) := \rho(x_i, x_j);$
для всех $t = 2, \ldots, \ell$ (t — номер итерации):
  найти в C_{t-1} два ближайших кластера:
  $(U, V) := \arg\min_{U \neq V} R(U, V);$
  $R_t := R(U, V);$
  слить их в один кластер:
  $W := U \cup V;$
  $C_t := C_{t-1} \cup \{W\} \setminus \{U, V\};$
  для всех $S \in C_t$
    вычислить $R(W, S)$ по формуле Ланса-Уильямса;

Формула Ланса-Уильямса

Как определить расстояние $R(W,S)$
между кластерами $W = U \cup V$ и S,
зная расстояния $R(U,S)$, $R(V,S)$, $R(U,V)$?

Формула, обобщающая большинство разумных способов определить это расстояние [Ланс, Уильямс, 1967]:

$$\begin{aligned} R(U \cup V, S) = {} & \alpha_U \cdot R(U,S) + \\ & + \alpha_V \cdot R(V,S) + \\ & + \beta \cdot R(U,V) + \\ & + \gamma \cdot \big|R(U,S) - R(V,S)\big|, \end{aligned}$$

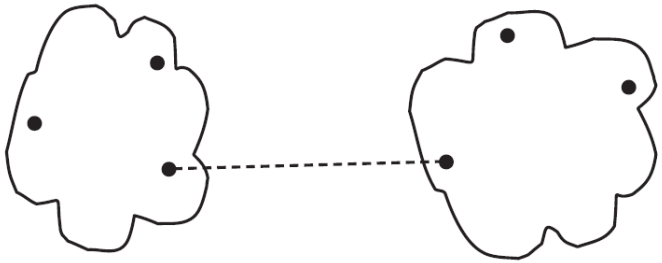
где α_U, α_V, β, γ — числовые параметры.

Частные случаи формулы Ланса-Уильямса

1. Расстояние ближнего соседа:

$$R^{б}(W,S) = \min_{w\in W, s\in S} \rho(w,s);$$

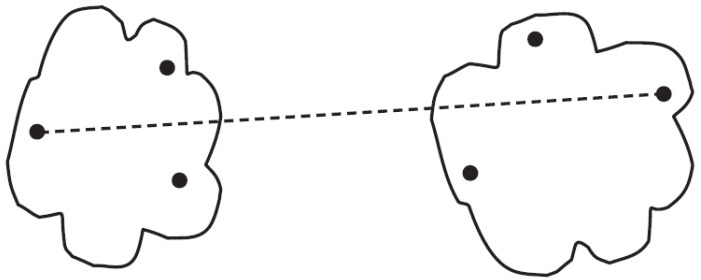
$$\alpha_U = \alpha_V = \tfrac{1}{2},\ \ \beta = 0,\ \ \gamma = -\tfrac{1}{2}.$$



2. Расстояние дальнего соседа:

$$R^{д}(W,S) = \max_{w\in W, s\in S} \rho(w,s);$$

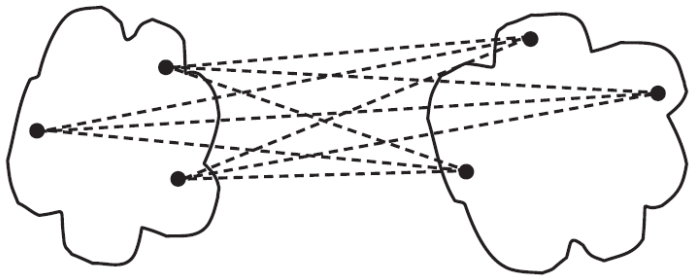
$$\alpha_U = \alpha_V = \tfrac{1}{2},\ \ \beta = 0,\ \ \gamma = \tfrac{1}{2}.$$



3. Групповое среднее расстояние:

$$R^{г}(W,S) = \frac{1}{|W||S|} \sum_{w\in W} \sum_{s\in S} \rho(w,s);$$

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



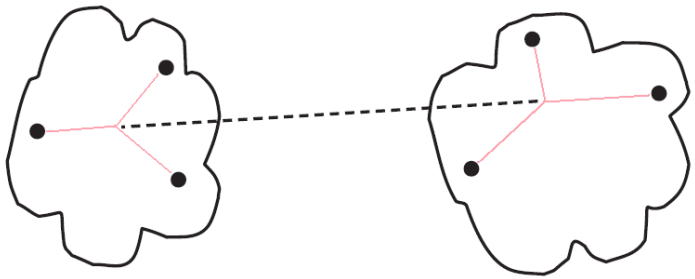
Частные случаи формулы Ланса-Уильямса

4. Расстояние между центрами:

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

$$\alpha_U = \frac{|U|}{|W|},\ \ \alpha_V = \frac{|V|}{|W|},$$

$$\beta = -\alpha_U\alpha_V,\ \ \gamma = 0.$$



5. Расстояние Уорда:

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

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

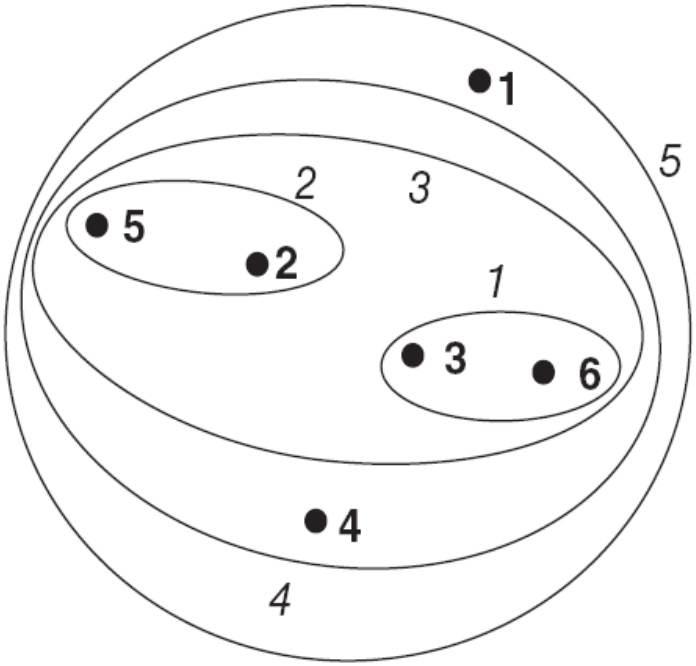
Проблема выбора

Какая функция расстояния лучше?

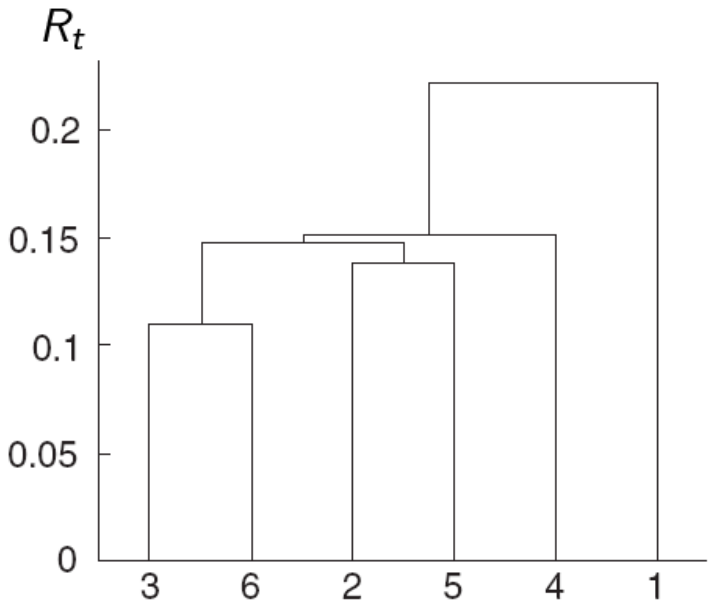
Визуализация кластерной структуры

1. Расстояние ближнего соседа:

Диаграмма вложения



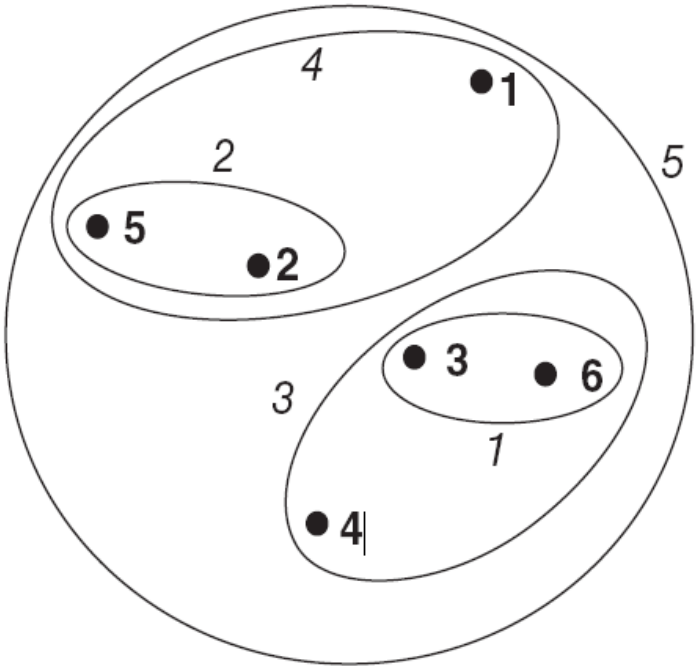
Дендрограмма



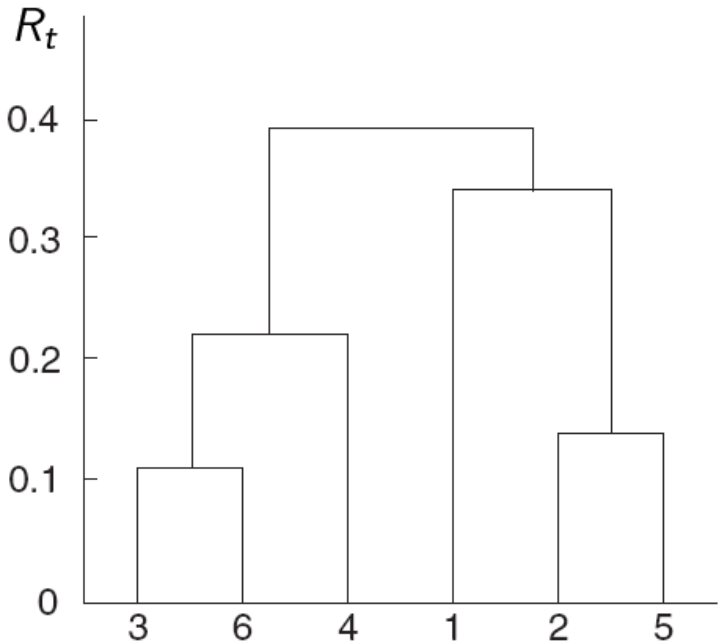
Визуализация кластерной структуры

2. Расстояние дальнего соседа:

Диаграмма вложения



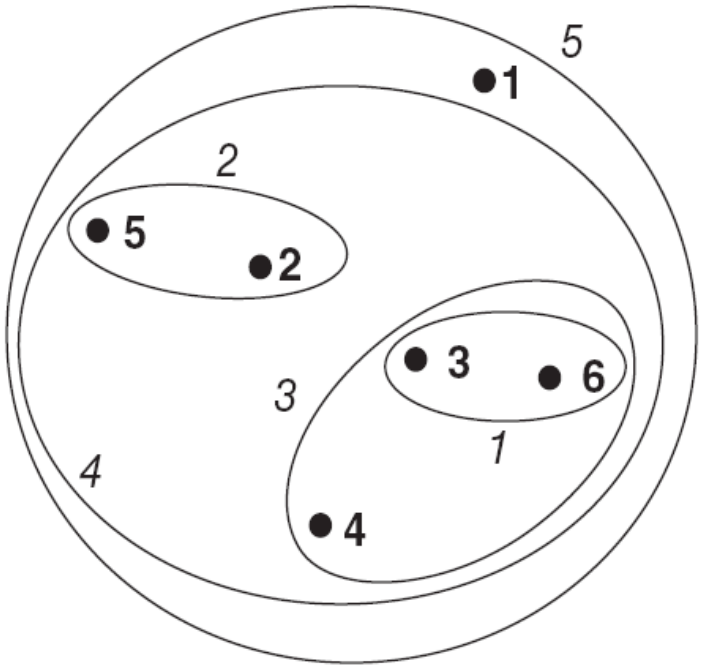
Дендрограмма



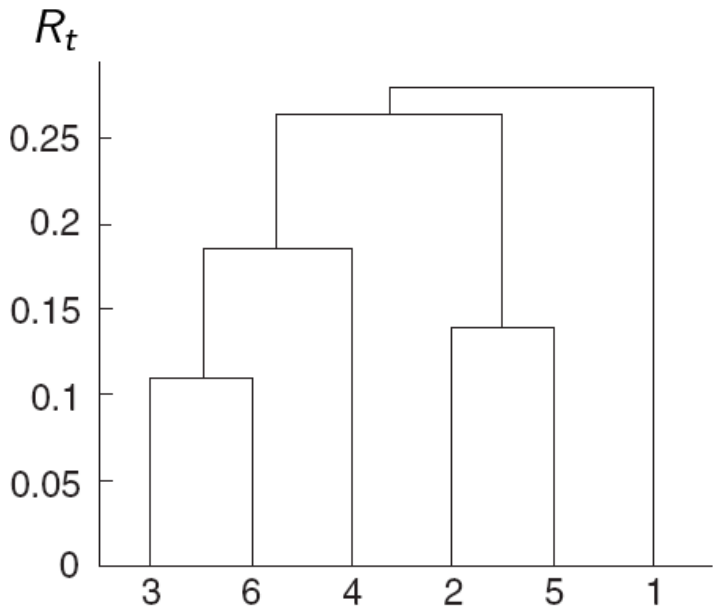
Визуализация кластерной структуры

3. Групповое среднее расстояние:

Диаграмма вложения



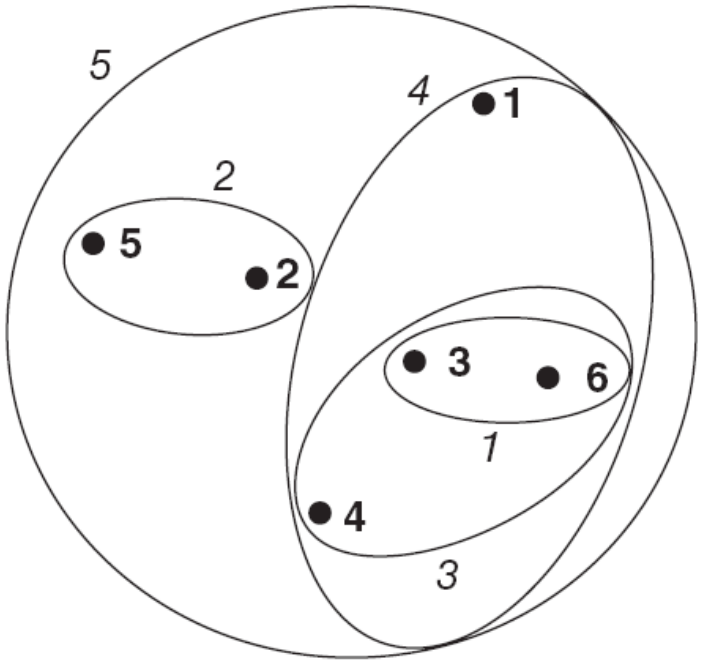
Дендрограмма



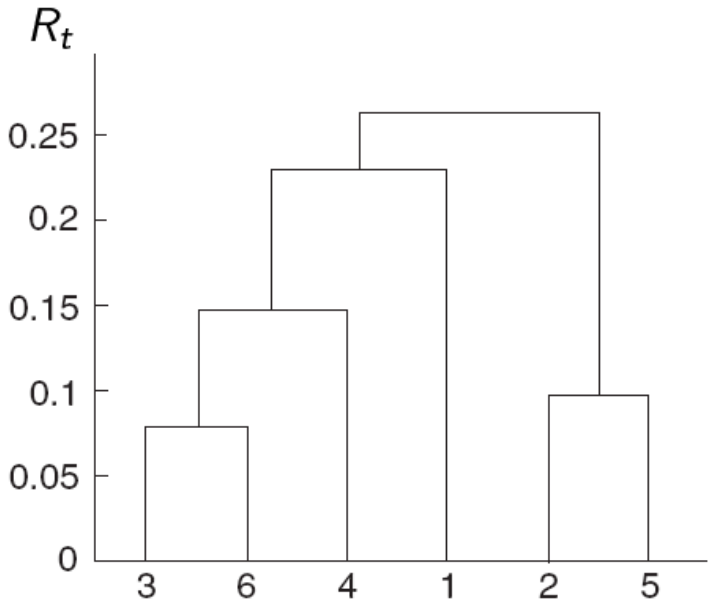
Визуализация кластерной структуры

5. Расстояние Уорда:

Диаграмма вложения



Дендрограмма



Свойство монотонности

Определение

Кластеризация *монотонна*, если при каждом слиянии расстояние между объединяемыми кластерами только увеличивается: $R_2 \leqslant R_3 \leqslant \ldots \leqslant R_\ell$.

Теорема (Миллиган, 1979)

Кластеризация монотонна, если выполняются условия

$$\alpha_U \geqslant 0, \quad \alpha_V \geqslant 0, \quad \alpha_U + \alpha_V + \beta \geqslant 1, \quad \min\{\alpha_U, \alpha_V\} + \gamma \geqslant 0.$$

Если кластеризация монотонна, то дендрограмма не имеет самопересечений.

$R^{ц}$ не монотонно; $R^{б}$, $R^{д}$, $R^{г}$, $R^{у}$ — монотонны.

Свойства сжатия и растяжения

Определение

Кластеризация *сжимающая*, если $R_t \leqslant \rho(\mu_U, \mu_V), \ \forall t$.
Кластеризация *растягивающая*, если $R_t \geqslant \rho(\mu_U, \mu_V), \ \forall t$.
Иначе кластеризация *сохраняет метрику пространства*.

Свойство **растяжения** наиболее желательно, так как оно способствует более чёткому отделению кластеров.

$R^{\text{б}}$ — сжимающее; $R^{\text{д}}$, $R^{\text{у}}$ — растягивающие;
$R^{\text{г}}$, $R^{\text{ц}}$ — сохраняющие.

Проблема повышения эффективности алгоритма

Проблема эффективности:

- самая трудоёмкая операция в алгоритме Ланса-Уильямса — — поиск ближайших кластеров — $O(\ell^2)$ операций:

$$\text{шаг 3:} \quad (U, V) := \arg\min_{U \neq V} R(U, V).$$

- значит, построение всего дерева — $O(\ell^3)$ операций.

Идея повышения эффективности:

- перебирать лишь наиболее близкие пары:

$$\text{шаг 3:} \quad (U, V) := \arg\min_{R(U,V) \leqslant \delta} R(U, V).$$

- периодически увеличивать параметр δ.

Быстрый (редуктивный) алгоритм Ланса-Уильямса

1: сначала все кластеры одноэлементные:
$t := 1; \quad C_t = \{\{x_1\}, \ldots, \{x_\ell\}\};$
$R(\{x_i\}, \{x_j\}) := \rho(x_i, x_j);$
2: выбрать начальное значение параметра δ;
$P(\delta) := \{(U, V) \mid U, V \in C_t,\ R(U, V) \leqslant \delta\};$
3: **для всех** $t = 2, \ldots, \ell$ (t — номер итерации):
4: **если** $P(\delta) = \varnothing$ **то** увеличить δ так, чтобы $P(\delta) \neq \varnothing$;
5: $(U, V) := \underset{(U,V) \in P(\delta)}{\arg\min} R(U, V);$
$R_t := R(U, V);$
6: $C_t := C_{t-1} \cup \{W\} \setminus \{U, V\};$
7: **для всех** $S \in C_t$
8: вычислить $R(W, S)$ по формуле Ланса-Уильямса;
9: **если** $R(W, S) \leqslant \delta$ **то** $P(\delta) := P(\delta) \cup \{(W, S)\};$

Свойство редуктивности

Всегда ли быстрый алгоритм строит ту же кластеризацию?

Определение (Брюинош, 1978)

Расстояние R называется *редуктивным*, если
для любого $\delta > 0$ и любых δ-близких кластеров $R(U, V) \leqslant \delta$
объединение δ-окрестностей U и V
содержит δ-окрестность объединения $W = U \cup V$:

$$\{S\colon R(U \cup V, S) < \delta\} \subseteq \{S\colon R(S, U) < \delta\} \cup \{S\colon R(S, V) < \delta\}.$$

Теорема

Если расстояние R редуктивно, то быстрый алгоритм приводит к той же кластеризации, что и исходный алгоритм.

Свойство редуктивности

Теорема (Диде и Моро, 1984)

Расстояние R является редуктивным, если

$$\alpha_U \geqslant 0,\ \alpha_V \geqslant 0,\ \alpha_U+\alpha_V+\min\{\beta,0\} \geqslant 1,\ \min\{\alpha_U,\alpha_V\}+\gamma \geqslant 0.$$

Сравните с условием монотонности (теорема Миллигана):

$$\alpha_U \geqslant 0,\ \ \alpha_V \geqslant 0,\ \ \alpha_U+\alpha_V+\beta \geqslant 1,\ \ \min\{\alpha_U,\alpha_V\}+\gamma \geqslant 0.$$

Утверждение

Всякое редуктивное расстояние является монотонным.

$R^{ц}$ не редуктивное; $R^{б}$, $R^{д}$, $R^{г}$, $R^{у}$ — редуктивные.

Рекомендации и выводы

Стратегия выбора параметра δ на шагах 2 и 4:

- Если $|C_t| \leqslant n_1$, то $P(\delta) := \{(U, V)\colon U, V \in C_t\}$.
- Иначе выбрать n_2 случайных расстояний $R(U, V)$; $\delta :=$ минимальное из них;
- n_1, n_2 влияют только на скорость, но не на результат кластеризации; сначала можно положить $n_1 = n_2 = 20$.

Общие рекомендации по иерархической кластеризации:

- лучше пользоваться $R^{\text{у}}$ — расстоянием Уорда;
- лучше пользоваться быстрым алгоритмом;
- определение числа кластеров — по максимуму $|R_{t+1} - R_t|$, тогда результирующее множество кластеров $:= C_t$.