Введение в кластеризацию

А. А. Бояров, А. А. Сенов

СПбГУ
математико–механический факультет

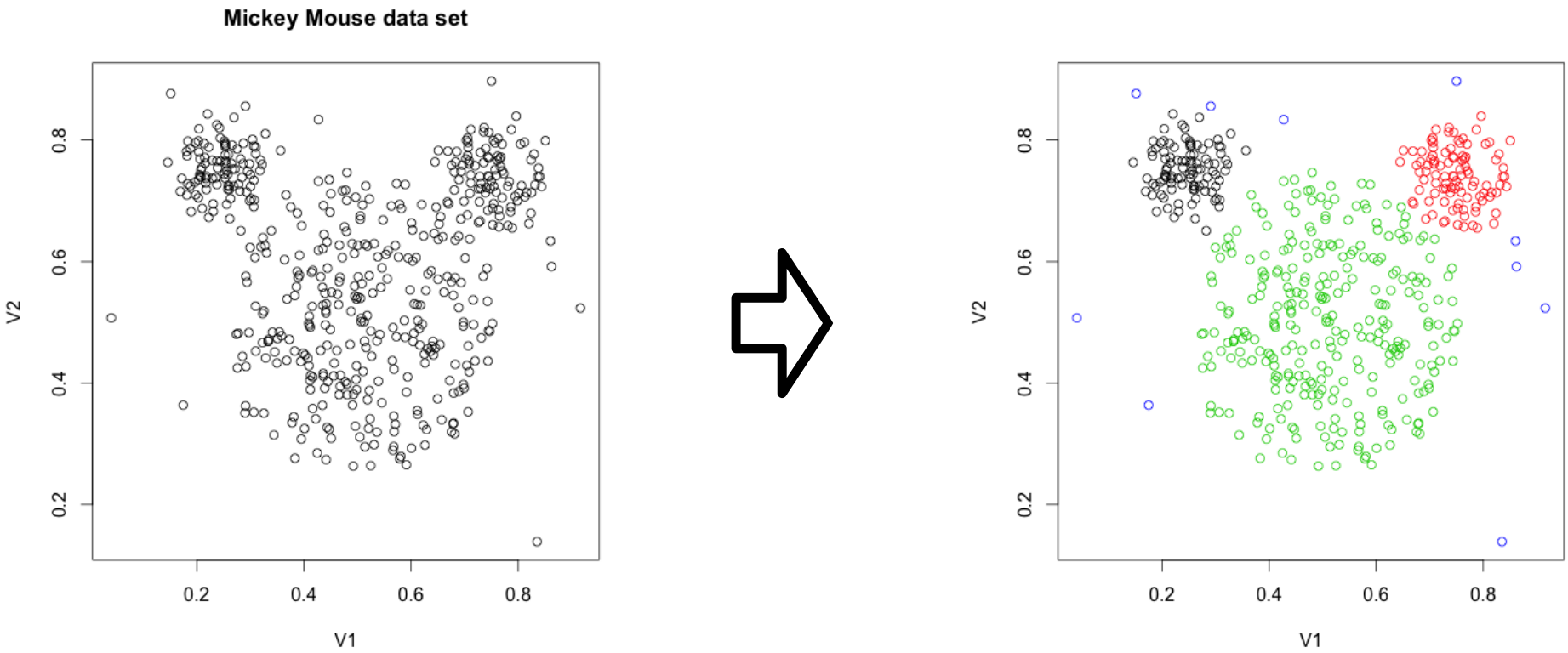
12 марта 2014

Определение

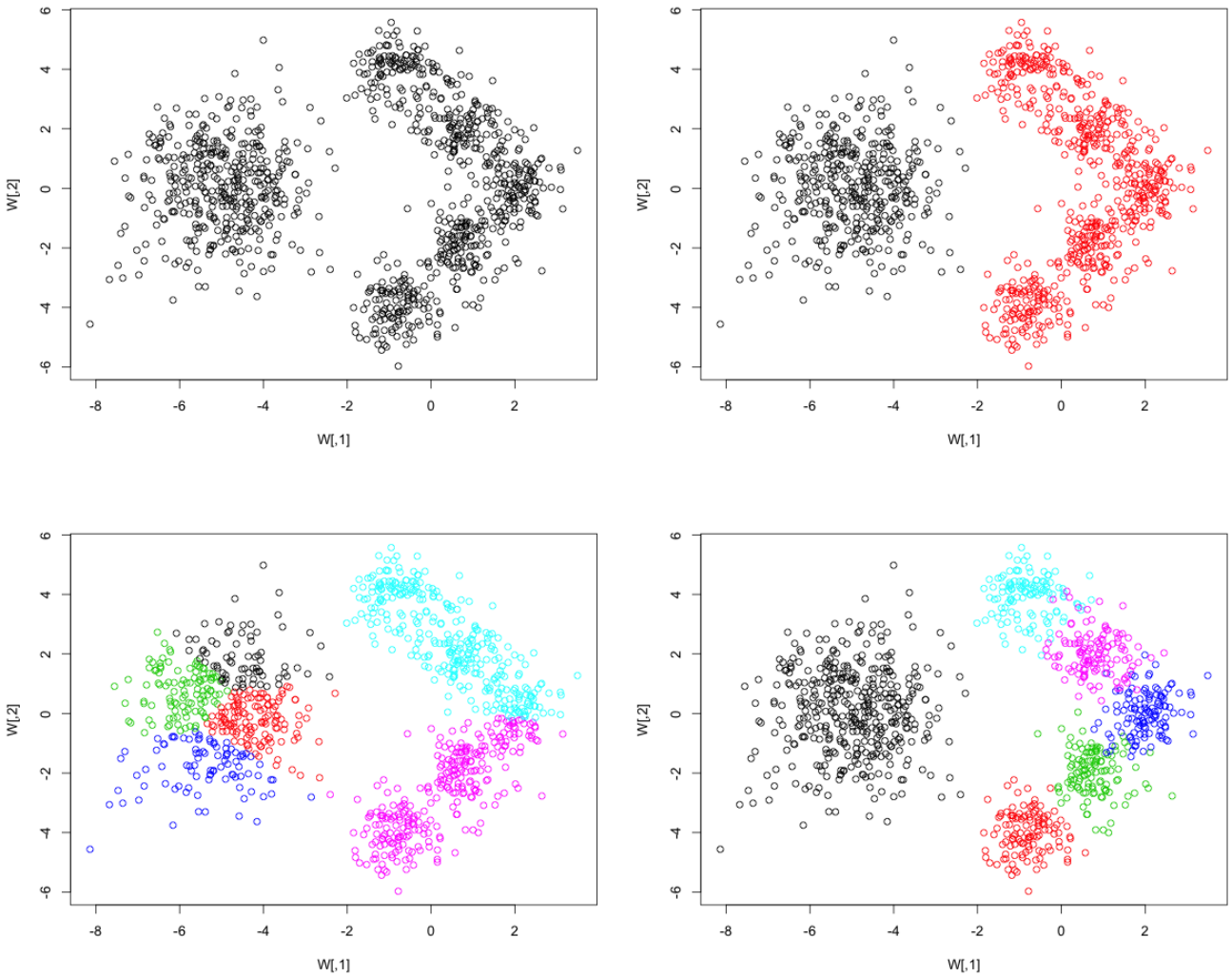
Кластеризация — задача разбиения множества объектов на группы (кластера) таким образом, что бы объекты из разных групп отличались друг от друга значительнее чем объекты из одной группы.

Определение

Кластеризация — задача разбиения множества объектов на группы (кластера) таким образом, что бы объекты из разных групп отличались друг от друга значительнее чем объекты из одной группы.



Пример



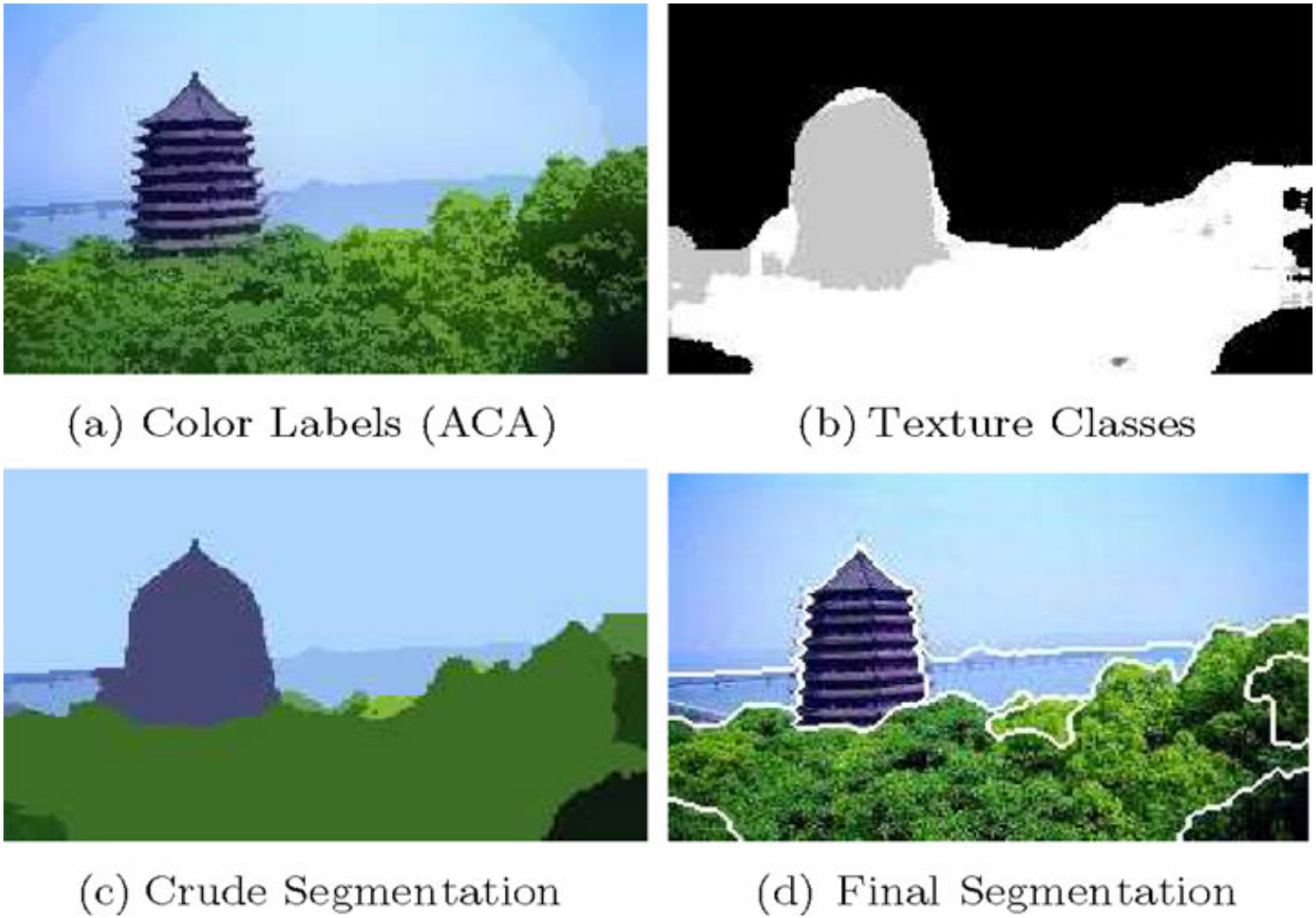
Применения

- Обработка изображений
  сжатие, сегментация
- Рекомендательные системы
  модели поведения, категоризация товаров
- Поисковые сервисы
  структуризация выдачи, выделение тем
- Медицина
  анализ изображений
- Биология
  категоризация видов, определение происхождения, анализ ДНК

Применения. Сжатие изображений

Figure: Сжатие изображения с помощью k-means (k = 16)

Применения. Сегментация изображений



(a) Color Labels (ACA)

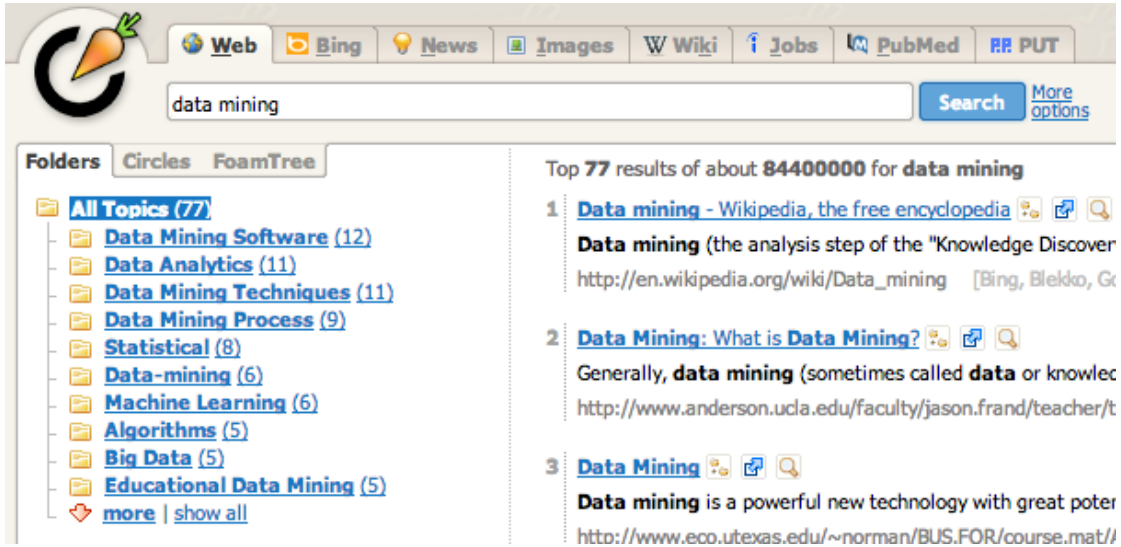
(b) Texture Classes

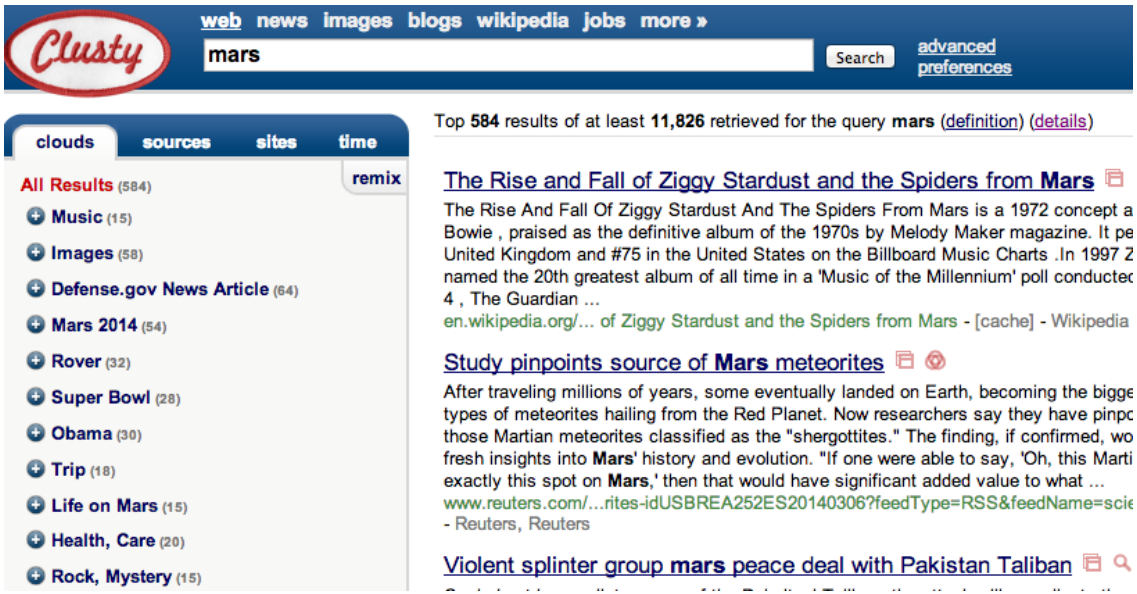
(c) Crude Segmentation

(d) Final Segmentation

Figure: Сегментация изображения на основе содержания (content-based)

Применения. Структуризация результатов поиска





Применения. Визуализация и анализ

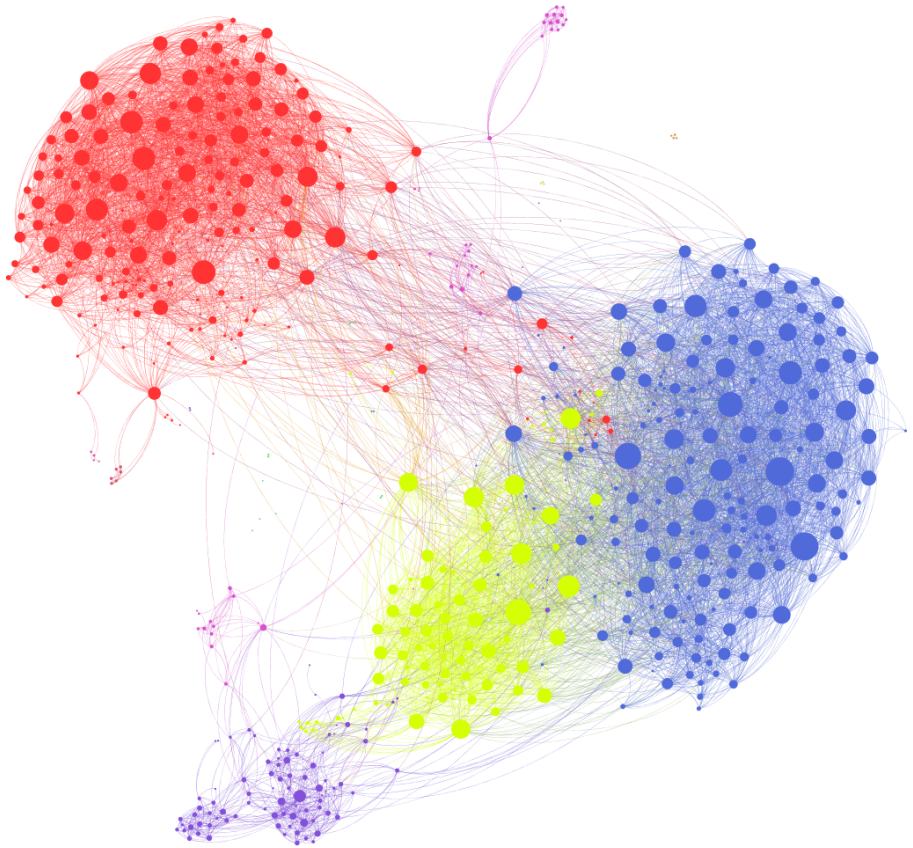


Figure: Взаимосвязь друзей в facebook.com (griffsgraphs.com)

Применения. Визуализация и анализ

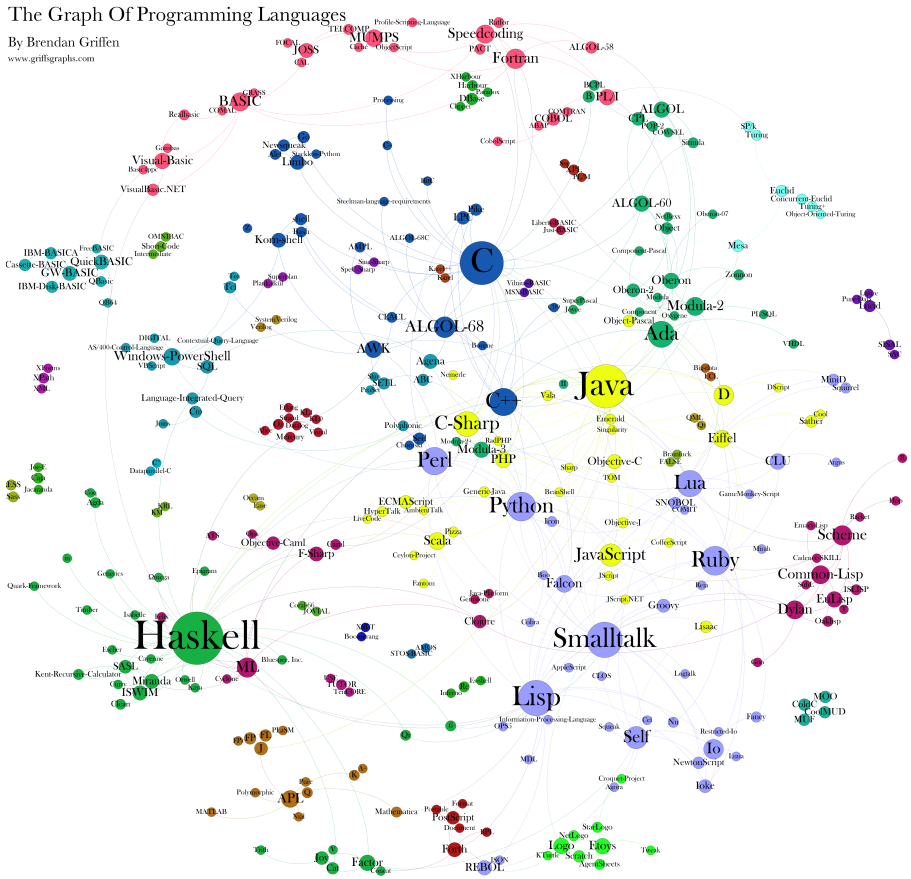


Figure: Взаимосвязь ЯП на основе данных dbpedia.org (griffsgraphs.com)

Цели кластеризации

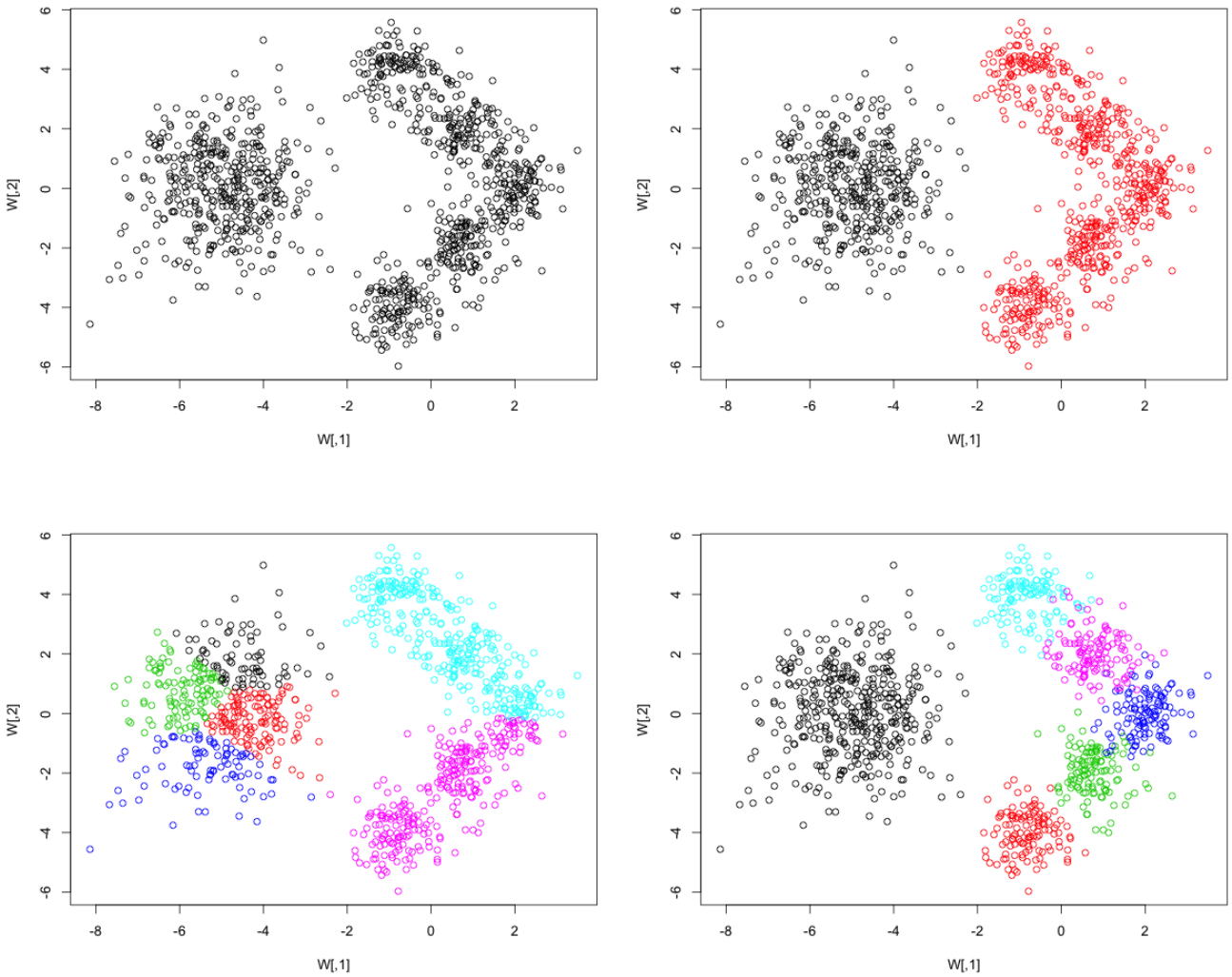
- Описание данных
  — результаты кластеризации интересны сами по себе
- Сжатие данных
  — результаты кластеризации вспомогательны
- Определение аномалий
  — определить объекты, не подходящие ни одному из кластеров

Неоднозначность кластеризации

Решение задачи кластеризации неоднозначно:

- количество кластеров k, как правило, неизвестно
- мера различия q между объектами зачастую неопределена
- множество критериев качества кластеризации
- множество алгоритмов кластеризации
- множество теоретических подходов

Пример



Формальности

Дано: $\mathbb{W}$ — выборка размера $N = |\mathbb{W}|$, $k \leq N$ — число кластеров.
Дополнительно: $q : \mathbb{W} \times \mathbb{W} \to \mathbb{R}$ — мера различия на $\mathbb{W}$
(больше $q \rightsquigarrow$ больше различие).

Результат алгоритма кластеризации — функция $\gamma_{\chi^k} : \mathbb{W} \to \{1, \ldots, K\}$

$$\chi^k(\mathbb{W}) = \{\mathbb{W}_1, \ldots, \mathbb{W}_k\}, \quad \bigcup_{i=1}^{k} \mathbb{W}_i = \mathbb{W}$$

$$\gamma_{\chi^k}(w) = i : w \in \mathbb{W}_i.$$

Хотелось бы

$$q(w_1, w_2) < q(w_3, w_4) \qquad \forall w_{1,2,3} \in \mathbb{W}_i, w_4 \in \mathbb{W}_j, \quad i \neq j$$

Разновидности кластеризации

1. По типу входных данных
   - Матрица различий между объектами
   - Множество объектов и мера различия
2. По структуре кластеров
   - Плоская
   - Иерархическая
3. По виду принадлежности к кластерам
   - Четкая
   - Нечеткая

Виды входных данных

- Множество объектов и метрика
  Имеем $\mathbb{W} = \{w_i\}_{i=1}^N$ и меру различия $q : \mathbb{W} \times \mathbb{W} \to \mathbb{R}$
  Обычно, q приходится выбирать самостоятельно.

- Матрица различий между объектами.
  Вместо $\mathbb{W} = \{w_i\}_{i=1}^N$ и метрики имеем $\mathbf{Q} = \{q_{i,j}\}_{i,j=1}^N$
  — *матрицу близости*, где $d_{i,j}$ — близость между i и j объектами.

Если надо, можем посчитать $\mathbf{Q}$ $(q_{i,j} = q(w_i, w_j))$.

Меры различия

Меры различия для $\mathbb{W} \subset \mathbb{R}^d$ (количественные данные)

- $q(w, w') = \sqrt{\sum_{i=1}^{d}(w_i - w'_i)^2}$ — расстояние Евклида
- $q(w, w') = \max_i(w_i - w'_i)$ — расстояние Чебышева
- $q(w, w') = \sum_{i=1}^{d} |w_i - w'_i|$ — расстояние городских кварталов
- $q(w, w') = ||w - w'||_p$ — расстояние в ℓ_p
- $q(w, w') = \frac{1-r(w,w')}{2}$, $r(w, w')$ — коэффициент корреляции Пирсона
- $q(w, w') = \sqrt{(w - \overline{w})^T \Sigma_{\mathbb{W}}^{-1}(w - \overline{w})}$ — расстояние Махалонобиса
  $\overline{w}$ — среднее, а $\Sigma_{\mathbb{W}}$ — ковариационная матрица векторов из $\mathbb{W}$

Меры различия

Меры различия для $\mathbb{W} \subset \{0,1\}^d$ (качественные данные)

- $q(w, w') = 1 - \frac{w \cdot w'}{|w|^2 + |w|^2 - w \cdot w'}$ — расстояние Танимото (Tanimoto/Jaccard distance) $w \cdot w'$ — "побитовое" умножение
- $q(w, w') = \sum_i p_i(w_i, w'_i)$ — расстояние на основе штрафов
  $p_i(w_i, w'_i)$ — функция штрафа за несоответсвие в i-ом признаке
  $(p_i(a, a) = 0,\ p_i(a, b) \geq 0)$
  p_i может задаваться исследователем, а может быть посчитана из данных
- Статья о сравнении метрик для качественных (категориальных) данных

Выбор меры различия не менее важен чем выбор алгоритма.

Кластеризация: плоская и иерархическая

- *Плоская кластеризация* — кластера равнозначны
  1. K-means
  2. EM-algorithm
  3. Spectral clustering
  4. Mean-shift
  5. ...
- *Иерархическая кластеризация* — новые кластера последовательно строятся их уже найденых, образуя иерархическую (древовидную) структуру
  1. CURE
  2. Brown clustering
  3. OPTICS
  4. ...

Плоская структура кластеров: $\chi^k(\mathbb{W}) = \{\mathbb{W}_1, \ldots, \mathbb{W}_k\}$

Иерархическая структура кластеров: $\chi^k(\mathbb{W}) = \{\mathbb{W}_1^{(l)}, \ldots, \mathbb{W}_{k^{(l)}}^{(l)}\}_{l=1}^{L}$

Иерархическая кластеризация. Иерархия кластеров

$$w_1, w_2, \ldots, w_{i-1}, w_i, \ldots, w_{j-1}, w_j, \ldots, w_{l-1}, w_l, \ldots, w_{s-1}, w_s, \ldots, w_{N-1}, w_N$$

$$\mathbb{w}_1^{(1)}, \quad \ldots \quad , \mathbb{w}_i^{(1)}, \quad \ldots \quad , \mathbb{w}_j^{(1)}, \quad \ldots \quad , \mathbb{w}_s^{(1)}, \quad \ldots \quad , \mathbb{w}_{k^{(1)}}^{(1)}$$

$$\mathbb{w}_1^{(2)}, \quad \ldots \quad , \mathbb{w}_i^{(2)}, \quad \ldots \quad , \mathbb{w}_j^{(2)}, \quad \ldots \quad , \mathbb{w}_{k^{(2)}}^{(2)}$$

$$\ldots$$

$$\mathbb{w}_1^{(l)}, \quad \ldots \quad , \mathbb{w}_i^{(l)}, \quad \ldots \quad , \mathbb{w}_{k^{(l)}}^{(l)}$$

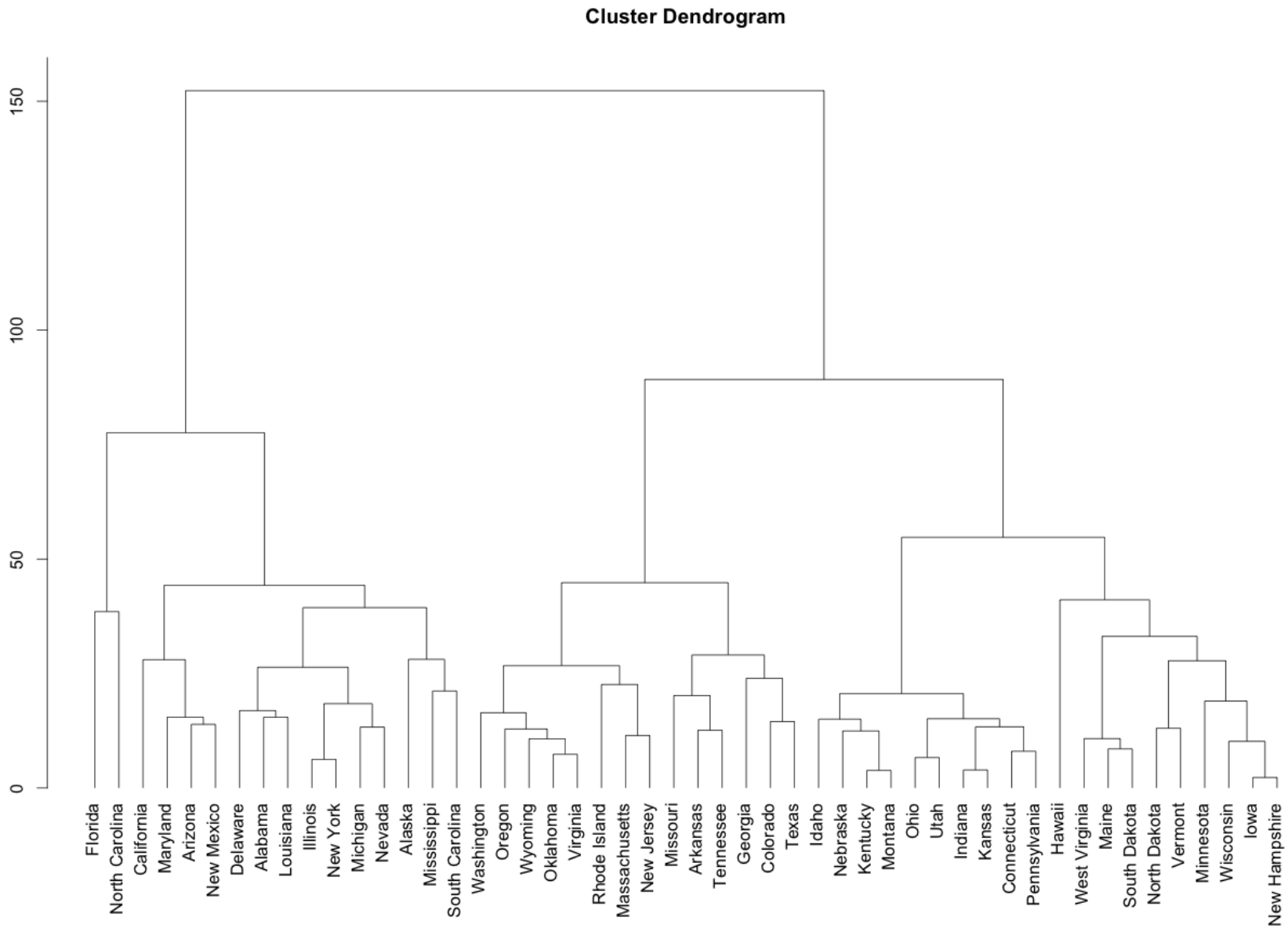
$$\ldots$$

$$\mathbb{w}_1^{(L)}$$

$$N = k^{(0)} > k^{(1)} > \ldots > k^{(L)} = 1$$

$$\bigcup_{i=1}^{k^{(l)}} \mathbb{w}_i^{(l)} = \mathbb{W} \quad \forall 1 \leq l \leq L$$

Иерархическая кластеризации. Дендрограма



Кластеризация: четкая и нечеткая

- *Четкая кластеризация*
  — каждый объект $w \in \mathbb{W}$ принадлежит строго к одному кластеру из $\chi^k(\mathbb{W})$

$$\gamma_{\chi^k}(w) = i : w \in \mathbb{W}_i.$$

- *Нечеткая (fuzzy) кластеризация*
  — каждый объект $w \in \mathbb{W}$ принадлежит каждому кластеру из $\chi^k(\mathbb{W})$ в *определенной степени*

$$\gamma_{\mathit{fuzzy}\,\chi^k}(w) = f \in \mathbb{R}^k : \quad f_i \geq 0, \sum_i f_i = 1.$$

K-means

Открыл Lloyd (1957, Bell Labs) для дискретизации аналоговых сигналов, опубликовал Forgy (1965).

- Разумная цель

$$F(\chi^k) = F(\mathbb{W}_1, \ldots, \mathbb{W}_k, m_1, \ldots, m_k) = \sum_{\mathbb{W}_i \in \chi_k} \sum_{w \in \mathbb{W}_i} q(w, m_i) \longrightarrow \min_{\chi^k}$$

- Сложная (NP-hard) задача, но есть эффективный алгоритм оптимизации
- Общая идея:
  0. Задать начальное разбиение на кластера
  1. Найти среднее каждого кластера — *центроиды*
  2. Каждый объект определить в кластер с ближайшим центроидом
  3. Повторить пункты 1–2

K-means. Алгоритм

Дано:

- $\mathbb{W} \subset \mathbb{R}^d$ — данные для кластеризации
- k — количество кластеров
- $m_1^{(0)}, \ldots, m_k^{(0)} \subset \mathbb{R}^d$ — начальные центроиды
- $q(\cdot, \cdot)$ — мера различия (расстояние)
- *критерий остановки*

Алгоритм

0. $t \leftarrow 0$
1. $\mathbb{W}_i^{(t)} \leftarrow \left\{ w \in \mathbb{W} : q\left(w, m_i^{(t)}\right) \leq q\left(w, m_j^{(t)}\right) \quad \forall 1 \leq j \leq k \right\}$
2. $m_i^{(t)} \leftarrow \frac{1}{|\mathbb{W}_i^{(t)}|} \sum_{w \in \mathbb{W}_i^{(t)}} w; \quad t \leftarrow t + 1$
3. Шаги 1–3 повторяются пока не выполнен *критерий остановки*

K-means. Сходимость

Основная идея: чередование assignment (expactation) step и update (maximization) step

$$\text{Assignment step.} \quad \sum_{\mathbb{W}_i \in \chi_k} \sum_{w \in \mathbb{W}_i} q(w, m_i) \longrightarrow \min_{\mathbb{W}_1, \ldots, \mathbb{W}_k}$$

$$\text{Update step.} \quad \sum_{\mathbb{W}_i \in \chi_k} \sum_{w \in \mathbb{W}_i} q(w, m_i) \longrightarrow \min_{m_1, \ldots, m_k}$$

Theorem

Для каждой итерации t алгоритма $k-means$ верно неравенство:

$$F\left(\left\{\mathbb{W}_i^{(t)}\right\}, \left\{m_i^{(t)}\right\}\right) \geq F\left(\left\{\mathbb{W}_i^{(t+1)}\right\}, \left\{m_i^{(t+1)}\right\}\right)$$

$$F\left(\left\{\mathbb{W}_i^{(t)}\right\}, \left\{m_i^{(t)}\right\}\right) \xrightarrow{t \to \infty} = \min F\left(\left\{\mathbb{W}_i^{(t)}\right\}, \left\{m_i^{(t)}\right\}\right)?$$

K-means. Проблемы и решения

1. Сходимость к локальному минимуму
2. Результат сильно зависит от начальных кластеров
3. Необходимо указывать количество кластеров
4. Неустойчив к выбросам в данных
5. Ищет “шарообразные” кластеры
6. “Жесткая” принадлежность к кластерам

Что можно сделать:

1. Запускать много раз — выбрать лучший
2. Запускать много раз, или эксперементировать с алгоритмами выбора начальных кластеров: $Lloyd - Forgy$, $Random\ partitions$, $k - means++$, $preclustering$, ...
3. Эксперементировать с k: попробовать несколько и выбрать лучший
4. Использовать более устойчивый вариант: $k - medians$
5. Преобразовать $\mathbb{W}$ или q
6. Использовать нечеткую вариацию: $c - means$

Инициализация кластеров. Основные алгоритмы

Lloyd–Forgy (default)

1. $m_1, \ldots, m_k = \textit{random from } \mathbb{W}$

Random partitions

0. $\mathbb{W}_1 = \emptyset, \ldots, \mathbb{W}_k = \emptyset$
1. foreach $w \in \mathbb{W}$: $\quad \mathbb{W}_{\textit{rand}(k)} \leftarrow w$

K-means++

0. $j = 1$; $m_j = \textit{random from } \mathbb{W}$
1. $D(w_i) = \min\limits_{m \in \{m_1, \ldots, m_j\}} q(w_i, m) \quad \forall w_i \in \mathbb{W}$
2. $m_{j+1} = w_i$ с вероятностью $p_i = \frac{D(w_i)}{\sum_{w \in \mathbb{W}} D(w)}$; $j = j + 1$

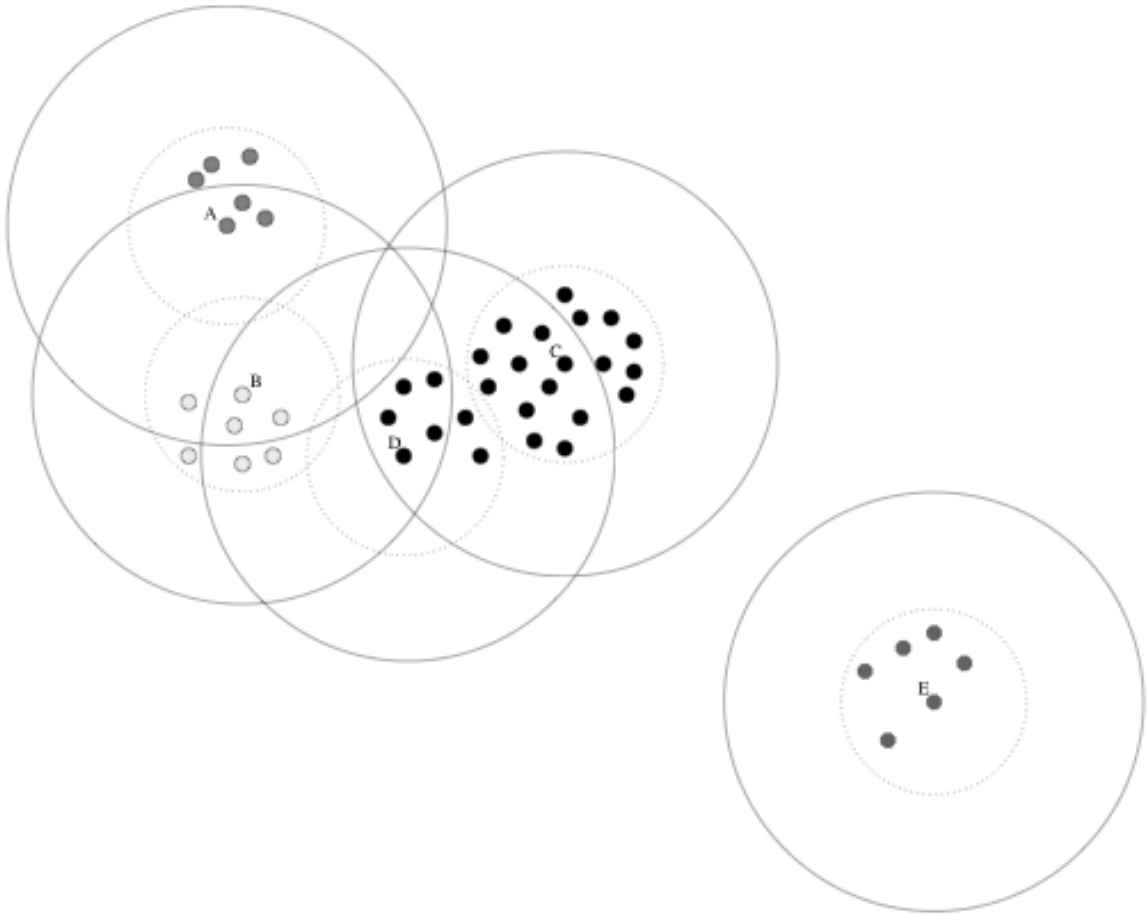
Инициализация кластеров. Прекластеризация

Идея: найти начальные кластера алгоритмом попроще

Canopy clustering

0. Выбрать $0 < T2 < T1$, $j = 1$
1. $m_j =$ случайная точка из $\mathbb{W}$
2. $\mathbb{W}_j = \{w \in \mathbb{W} : q(w, m_j) < T1\}$
3. $\mathbb{W} = \mathbb{W} \setminus \{w \in \mathbb{W} : q(w, m_j) < T2\}$

Инициализация кластеров. Canopy clustering



C-means

$$F\left(\left\{\mathbb{W}_i^{(t)}\right\}, \left\{m_i^{(t)}\right\}\right) = \sum_{\mathbb{W}_i \in \chi_k} \sum_{w \in \mathbb{W}_i} q(w, m_i) = \sum_{w \in \mathbb{W}} \sum_{i=1}^{k} q(w, m_i) \gamma_{\chi^k}(w)_i$$

Assignment step.

$$\sum_{w \in \mathbb{W}} \sum_{i=1}^{k} q(w, m_i) \gamma_{\textbf{fuzzy}}(w)_i \longrightarrow \min_{\gamma_{\textbf{fuzzy}}}$$

$$\gamma_{\textbf{fuzzy}}(w)_i = \left(\sum_{j=1}^{k} \left(\frac{q(w, m_i)}{q(w, m_j)} \right)^{\frac{2}{m-1}} \right)^{-m} \quad \forall w \in \mathbb{W}, i = 1..k$$

Update step.

$$\sum_{w \in \mathbb{W}} \sum_{i=1}^{k} q(w, m_i) \gamma_{\textbf{fuzzy}}(w)_i \longrightarrow \min_{m_1, \ldots, m_k}$$

$$m_i = \sum_{w \in \mathbb{W}} \gamma_{\textbf{fuzzy}}(w)_i w \quad \forall i = 1..k$$