

Введение в кластеризацию

А. А. Бояров, А. А. Сенов

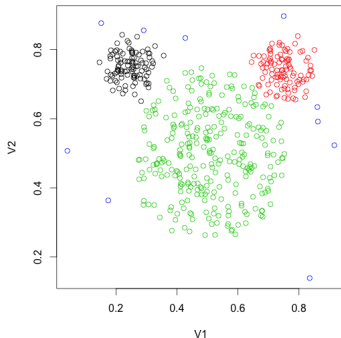
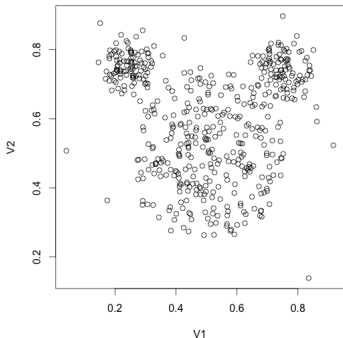
СПбГУ
математико–механический факультет

12 марта 2014

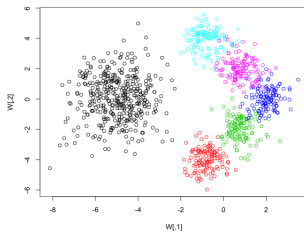
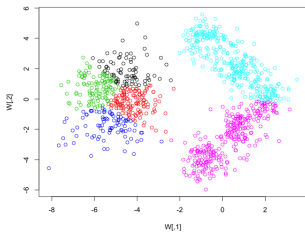
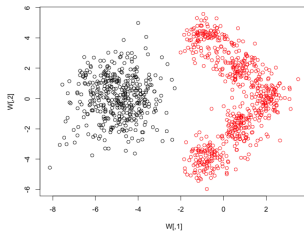
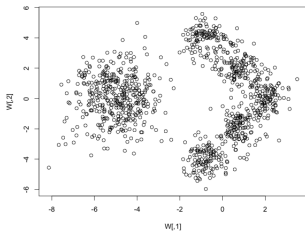
Кластеризация — задача разбиения множества объектов на группы (кластера) таким образом, что бы объекты из разных групп отличались друг от друга значительно больше чем объекты из одной группы.

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

Mickey Mouse data set



Пример



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

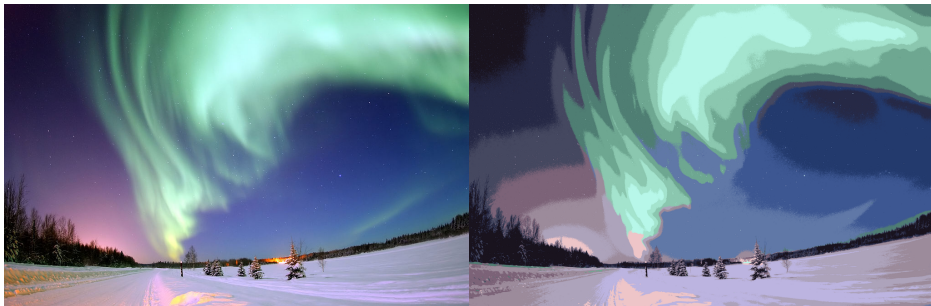


Figure: Сжатие изображения с помощью k-means ($k = 16$)



(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

The screenshot shows a search engine interface with a search bar containing 'data mining' and a 'Search' button. Below the search bar, there are navigation links for 'Web', 'Bing', 'News', 'Images', 'Wiki', 'Jobs', 'PubMed', and 'RR PUT'. The search results are displayed in a list format, with the top 77 results shown. The first result is 'Data mining - Wikipedia, the free encyclopedia', followed by 'Data Mining: What is Data Mining?' and 'Data Mining'.

Search results for **data mining**:

- Data mining - Wikipedia, the free encyclopedia**
Data mining (the analysis step of the "Knowledge Discover
http://en.wikipedia.org/wiki/Data_mining [Bing, Bieldo, Gt
- Data Mining: What is Data Mining?**
Generally, **data mining** (sometimes called **data** or knowle
<http://www.anderson.ucla.edu/faculty/jason.frand/teacher/t>
- Data Mining**
Data mining is a powerful new technology with great poter
<http://www.eco.utexas.edu/~norman/BUS.FOR/course.mat/>

The screenshot shows the Clusty search engine interface with a search bar containing 'mars' and a 'Search' button. Below the search bar, there are navigation links for 'web', 'news', 'images', 'blogs', 'wikipedia', 'jobs', and 'more'. The search results are displayed in a list format, with the top 584 results shown. The first result is 'The Rise and Fall of Ziggy Stardust and the Spiders from Mars', followed by 'Study pinpoints source of Mars meteorites' and 'Violent splinter group mars peace deal with Pakistan Taliban'.

Search results for **mars**:

Top 584 results of at least 11,826 retrieved for the query **mars** (definition) (details)

- The Rise and Fall of Ziggy Stardust and the Spiders from Mars**
The Rise And Fall Of Ziggy Stardust And The Spiders From Mars is a 1972 concept a Bowie , praised as the definitive album of the 1970s by Melody Maker magazine. It pe United Kingdom and #75 in the United States on the Billboard Music Charts . In 1997 Z named the 20th greatest album of all time in a "Music of the Millennium" poll conducted 4 , The Guardian ...
[en.wikipedia.org/...of_Ziggy_Stardust_and_the_Spiders_from_Mars_-\[cache\]-_Wikipedia](http://en.wikipedia.org/...of_Ziggy_Stardust_and_the_Spiders_from_Mars_-[cache]-_Wikipedia)
- Study pinpoints source of Mars meteorites**
After traveling millions of years, some eventually landed on Earth, becoming the bigge types of meteorites hailing from the Red Planet. Now researchers say they have pincp those Martian meteorites classified as the "shergottites." The finding, if confirmed, wo fresh insights into Mars' history and evolution. "If one were able to say, 'Oh, this Marti exactly this spot on Mars,' then that would have significant added value to what ...
www.reuters.com/...rites-idUSBREA252ES20140306?feedType=RSS&feedName=scit - Reuters, Reuters
- Violent splinter group mars peace deal with Pakistan Taliban**

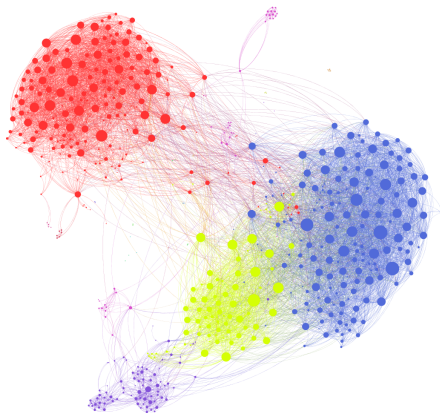


Figure: Взаимосвязь друзей в facebook.com (griffsgraphs.com)

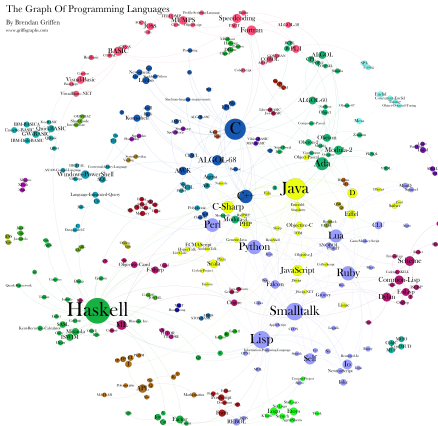


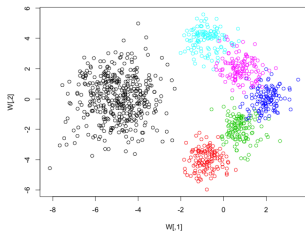
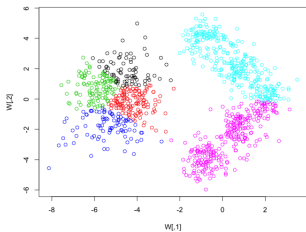
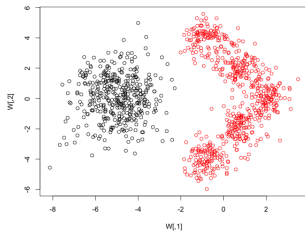
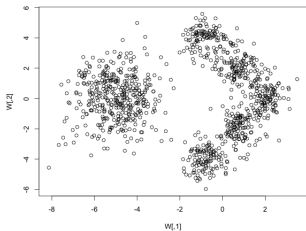
Figure: Взаимосвязь ЯП на основе данных dbpedia.org (griffioengraphs.com)

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

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



Дано: \mathbb{W} — выборка размера $N = |\mathbb{W}|$, $k \leq N$ — число кластеров.

Дополнительно: $q : \mathbb{W} \times \mathbb{W} \rightarrow \mathbb{R}$ — мера различия на \mathbb{W}

(больше $q \rightsquigarrow$ больше различие).

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

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

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

Хотелось бы

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

- 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

- Множество объектов и метрика

Имеем $\mathbb{W} = \{w_i\}_{i=1}^N$ и меру различия $q : \mathbb{W} \times \mathbb{W} \rightarrow \mathbb{R}$

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 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
- $q(w, w') = \sqrt{(w - \bar{w})^T \Sigma_{\mathbb{W}}^{-1} (w - \bar{w})}$ — расстояние Махалонобиса
 \bar{w} — среднее, а $\Sigma_{\mathbb{W}}$ — ковариационная матрица векторов из \mathbb{W}

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 несоответствие 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.

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

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

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

Иерархическая кластеризация. Иерархия кластеров

$W_1, W_2, \dots, W_{i-1}, W_i, \dots, W_{j-1}, W_j, \dots, W_{l-1}, W_l, \dots, W_{s-1}, W_s, \dots, W_{N-1}, W_N$

$W_1^{(1)}, \dots, W_i^{(1)}, \dots, W_j^{(1)}, \dots, W_s^{(1)}, \dots, W_{k^{(1)}}^{(1)}$

$W_1^{(2)}, \dots, W_i^{(2)}, \dots, W_j^{(2)}, \dots, W_{k^{(2)}}^{(2)}$

...

$W_1^{(l)}, \dots, W_i^{(l)}, \dots, W_{k^{(l)}}^{(l)}$

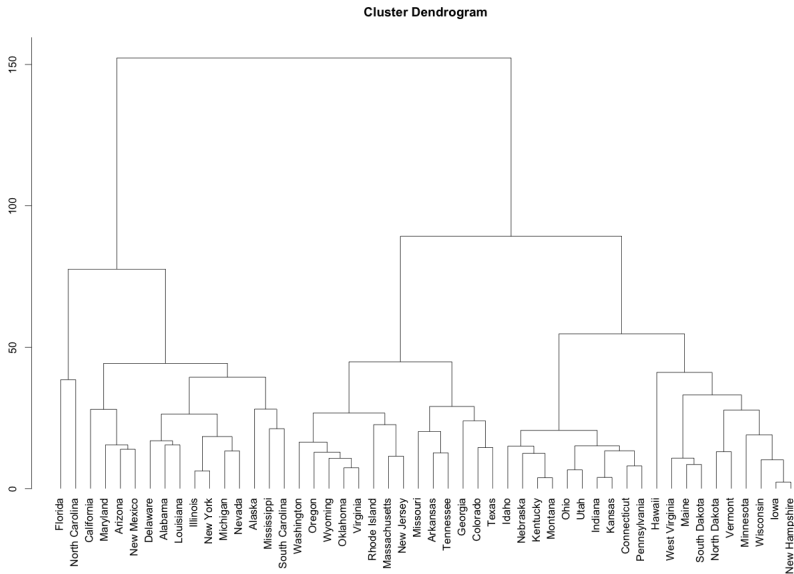
...

$W_1^{(L)}$

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

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

Иерархическая кластеризации. Дендрограмма



- 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

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

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

- Разумная цель

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

- 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:

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

Алгоритм

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

K-means. Сходимость

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

$$\text{Assignment step. } \sum_{\mathbb{W}_i \in \mathcal{X}_k} \sum_{w \in \mathbb{W}_i} q(w, m_i) \longrightarrow \min_{\mathbb{W}_1, \dots, \mathbb{W}_k}$$

$$\text{Update step. } \sum_{\mathbb{W}_i \in \mathcal{X}_k} \sum_{w \in \mathbb{W}_i} q(w, m_i) \longrightarrow \min_{m_1, \dots, m_k}$$

Theorem

Для каждой итерации t алгоритма k – means верно неравенство:

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

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

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 экспериментировать с алгоритмами выбора начальных кластеров: *Lloyd – Forgy*, *Random partitions*, *k – means ++*, *preclustering*, ...
- 3 Экспериментировать с k : попробовать несколько и выбрать лучший
- 4 Использовать более устойчивый вариант: *k – medians*
- 5 Преобразовать W или q
- 6 Использовать нечеткую вариацию: *c – means*

Lloyd–Forgy (default)

- 1 $m_1, \dots, m_k = \text{random from } \mathbb{W}$

Random partitions

- 0 $\mathbb{W}_1 = \emptyset, \dots, \mathbb{W}_k = \emptyset$
- 1 foreach $w \in \mathbb{W}$: $\mathbb{W}_{\text{rand}(k)} \leftarrow w$

K-means++

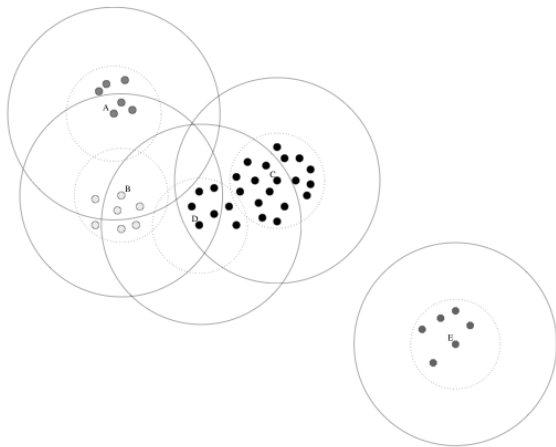
- 0 $j = 1; m_j = \text{random from } \mathbb{W}$
- 1 $D(w_i) = \min_{m \in \{m_1, \dots, m_j\}} q(w_i, m) \quad \forall w_i \in \mathbb{W}$
- 2 $m_{j+1} = w_i$ с вероятностью $p_i = \frac{D(w_i)}{\sum_{w \in \mathbb{W}} D(w)}$; $j = j + 1$

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



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

Assignment step. $\sum_{w \in \mathbb{W}} \sum_{i=1}^k q(w, m_i) \gamma_{\text{fuzzy}}(w)_i \rightarrow \min_{\gamma_{\text{fuzzy}}}$

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

Update step. $\sum_{w \in \mathbb{W}} \sum_{i=1}^k q(w, m_i) \gamma_{\text{fuzzy}}(w)_i \rightarrow \min_{m_1, \dots, m_k}$

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