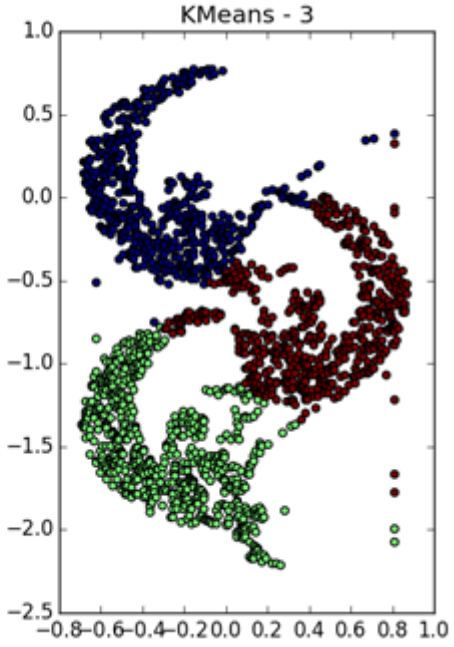
Задача кластеризации данных



Группа людей, действуя совместно,
может свершить такое,
о чем поодиночке они не могли бы и мечтать.

Франклин Рузвельт

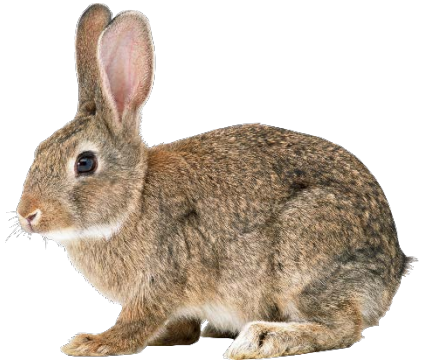
Содержание

- Основные концепции
- Разделительная кластеризация
- Иерархическая кластеризация
- Плотностная кластеризация
- **Нечеткая кластеризация**
- Меры качества кластеризации

Четкая *vs.* нечеткая кластеризация

A

B



C

D

E

Четкая кластеризация

Objects	C_1	C_2
A	1	0
B	1	0
C	0	1
D	0	1
E	0	1

Нечеткая кластеризация

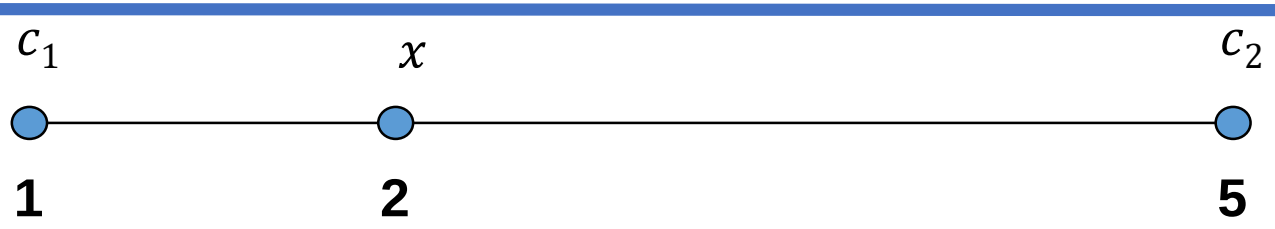
Objects	$P(C_1)$	$P(C_2)$
A	0.90	0.10
B	0.80	0.20
C	0.15	0.85
D	0.30	0.70
E	0.25	0.75

Четкая *vs.* нечеткая кластеризация

- $SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} w_{ij}\, dist(x_i, c_j)^2$
- w_{ij} – вес факта $x_i \in c_j$, $\sum_{j=1}^{k} w_{ij} = 1$
- Способы минимизации SSE
  - Фиксировать c_j и найти w_{ij}
  - Фиксировать w_{ij} и вычислить c_j
- $w_{ij} \in \{0,1\}$

- $SSE = \sum_{j=1}^{k} \sum_{i=1}^{n} w_{ij}^{m}\, dist(x_i, c_j)^2$
- w_{ij} – вес факта $x_i \in c_j$, $\sum_{j=1}^{k} w_{ij} = 1$
- $m > 1$ – «размытость» кластеров (обычно $m = 2$)
- Способы минимизации SSE
  - Фиксировать c_j и найти w_{ij}
  - Фиксировать w_{ij} и вычислить c_j
- $w_{ij} \in [0,1]$

Четкая *vs.* нечеткая кластеризация

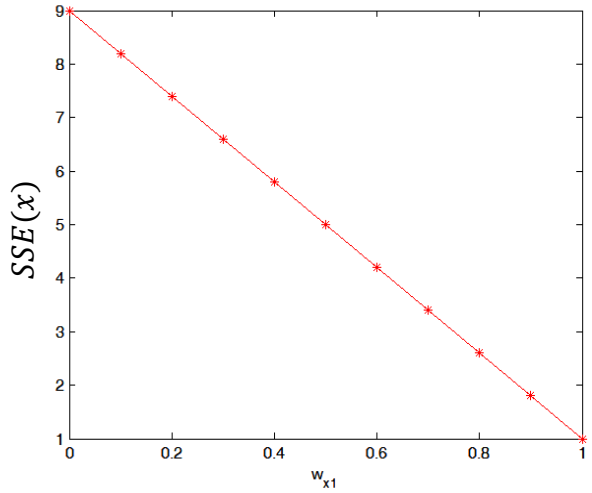


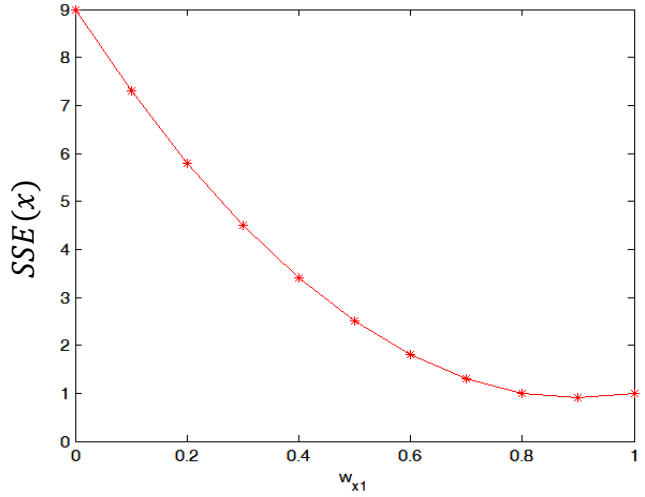
$$SSE(x) = w_{x1}(2-1)^2 + w_{x2}(5-2)^2 =$$
$$= w_{x1} + 9w_{x2}$$

$SSE(x) \to \min$ при $w_{x1} = 1, w_{x2} = 0$

$$SSE(x) = w_{x1}^2(2-1)^2 + w_{x2}^2(5-2)^2 =$$
$$= w_{x1}^2 + 9w_{x2}^2$$

$SSE(x) \to \min$ при $w_{x1} = 0.9, w_{x2} = 0.1$





Алгоритм Fuzzy *c*-Means

- k – количество кластеров
- $X = \{x_1, x_2, \dots, x_n\}$ – множество d-мерных точек
- $C \in \mathbb{R}^{k \times d}$ – *матрица центроидов*
  - c_j – центр j-го кластера (d-мерный вектор)
- $U \in \mathbb{R}^{n \times k}$ – *матрица принадлежности*
  - $0 \le u_{ij} \le 1$ – степень принадлежности (расстояние) между точкой x_i и центроидом c_j
- Минимизируемая *целевая функция*
  - $J_{FCM}(X, k, m) = \sum_{j=1}^{k} \sum_{i=1}^{n} u_{ij}^m \, dist(x_i, c_j)^2$
  - $m > 1$ – размытость

x	$x_{i,1}$	…	$x_{i,d}$
1			
…			
n			

c	$c_{j,1}$	…	$c_{j,d}$
1			
…			
k			

u	1	…	k
1			
…			
n			

Алгоритм Fuzzy *c*-Means

Input: X, k, m, ε

Output: U

$s := 0, U^{(0)} := (u_{ij})$ {initialization}

repeat

{computation of new centroids' coordinates}

Compute $C^{(s)} := (c_j)$ using formula

where $u_{ij} \in U^{(s)}$

{update matrixes values}

Compute $U^{(s)}$ and $U^{(s+1)}$ using formula

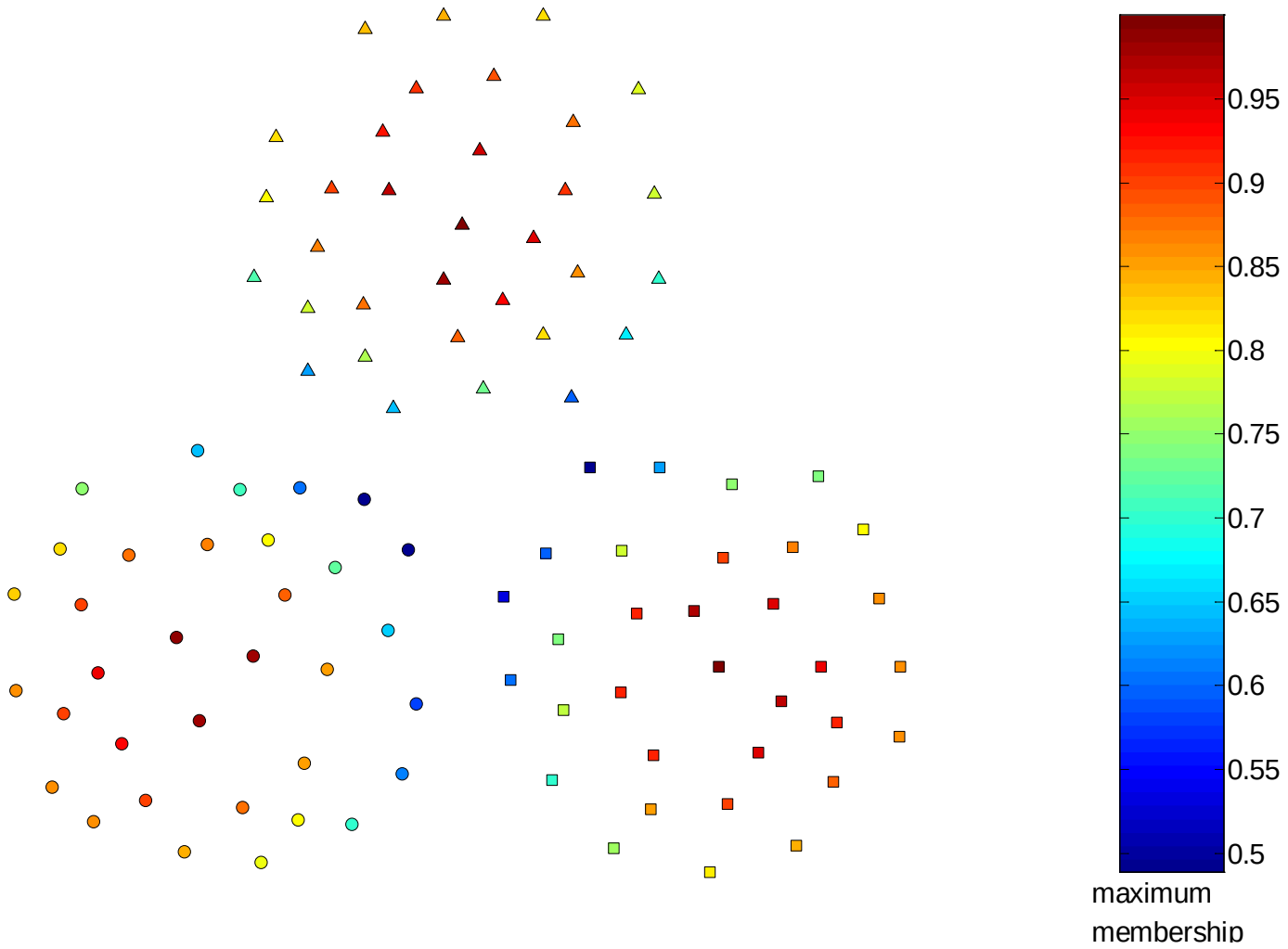
$s := s + 1$

until $\max_{ij}\{|u_{ij}^{(s)} - u_{ij}^{(s-1)}|\} \geqslant \varepsilon$

$$\forall j, l \quad c_{jl} = \frac{\sum_{i=1}^{n} u_{ij}^{m} \cdot x_{il}}{\sum_{i=1}^{n} u_{ij}^{m}}$$

$$u_{ij} = \sum_{t=1}^{k} \left(\frac{\rho(x_i, c_j)}{\rho(x_i, c_t)} \right)^{\frac{2}{1-m}}$$

Алгоритм Fuzzy c-Means: пример работы



Алгоритм Fuzzy c-Means: пример работы

S Name	I GiveBirth	I NumWings	I CanFly	I LiveInWater	I NumLegs	I NumHands	I NumHorns	D cluster_0	D cluster_1	D cluster_2	D cluster_3	D cluster_4	D cluster_5
Mermaid	-1	0	0	1	0	2	0						
Whale	1	0	0	1	0	0	0						
Dolphin	1	0	0	1	0	0	0						
Python	0	0	0	0	0	0	0						
Salmon	0	0	0	1	0	0	0						
Leopard_shark	1	0	0	1	0	0	0						
Eel	0	0	0	1	0	0	0						
Unicorn	1	0	0	0	4	0	1						
Pegasus	1	2	1	0	4	0	0						
Bat	1	2	1	0	4	0	0						
Cat	1	0	0	0	4	0	0						
Porcupine	1	0	0	0	4	0	0						
Tiger	1	0	0	0	4	0	0						
Horse	1	0	0	0	4	0	0						
Cow	1	0	0	0	4	0	2						
Antelope	1	0	0	0	4	0	2						
Sphinx	-1	0	0	0	4	0	0						
Platypus	0	0	0	0	4	0	0						
Frog	0	0	0	1	4	0	0						
Komodo	0	0	0	0	4	0	0						
Turtle	0	0	0	1	4	0	0						
Salamander	0	0	0	1	4	0	0						
Gila_monster	0	0	0	0	4	0	0						
Devil	-1	0	1	0	2	2	2						
Centaur	1	0	0	0	4	2	0						
Human	1	0	0	0	2	2	0						
Monkey	1	0	0	0	2	2	0						
Dragon	0	2	1	0	2	0	0						
Angel	-1	2	1	0	2	2	0						
Pigeon	0	2	1	0	2	0	0						
Penguin	0	2	0	1	2	0	0						
Owl	0	2	1	0	2	0	0						
Eagle	0	2	1	0	2	0	0						
Kiwi	0	2	0	0	2	0	0						

За и против Fuzzy c-Means

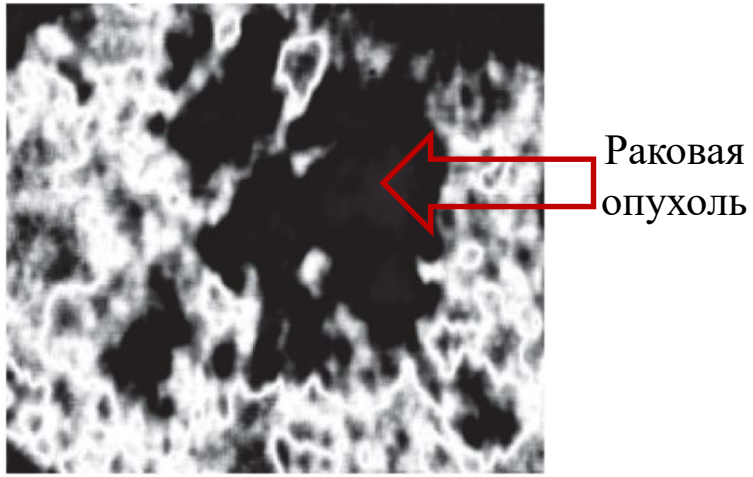
- Достоинства

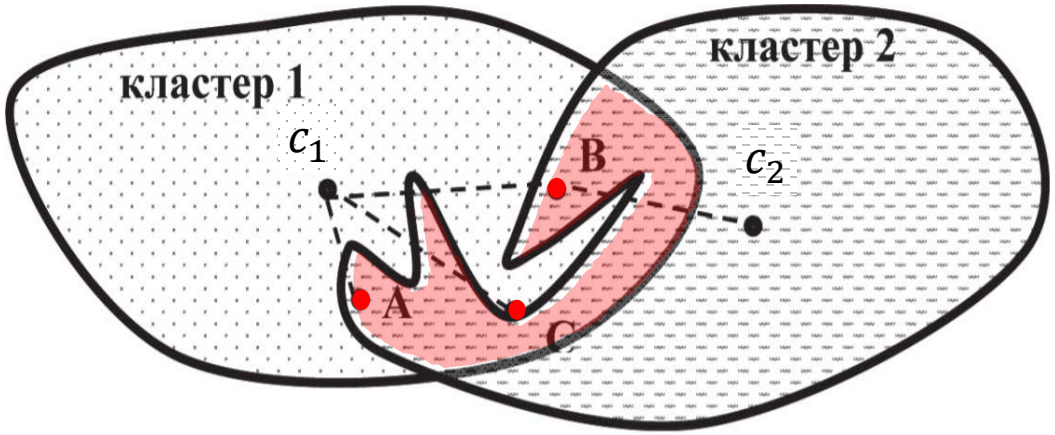
  Дает лучшие результаты для перекрывающихся кластеров и сравнительно лучшие результаты, чем k-means

- Недостатки
  - Количество кластеров необходимо задавать
  - Меньшее значение ε улучшает результаты, но ценой большего количества итераций

Применение нечеткой кластеризации

- Сегментация радиологических изображений





- Восстановление пропущенных данных
  - Нечеткая кластеризация точек без пропущенных координат
  - Создание прототипов: замена пропущенных координат соответствующими координатами центроидов
  - Выбор прототипа с минимальным расстоянием до центроида

Литература

- Tan P.-N., Steinbach M., Karpatne A., Kumar V. Introduction to Data Mining. 2nd Edition. Pearson, 2019. 839 p. ISBN: 978-0-13-312890-1

  8.2.1 Fuzzy Clustering, pp. 621-626