

# Нечеткая кластеризация потоков данных методом d-FuzzyStream

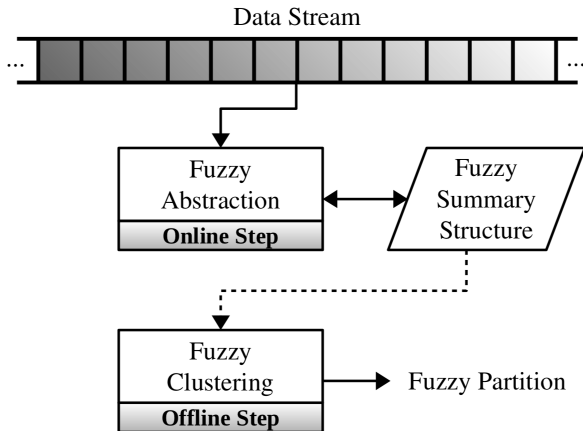
Поздняков Виталий

Высшая школа экономики

Декабрь 2019

# Методология FOOF (Fuzzy Online-Offline Framework)

- ▶ Используется для работы с потоком данных
- ▶ Нечеткая версия OOF (Online-Offline Framework)



- ▶ Первый этап — микрокластезация (d-FuzzyStreram)
- ▶ Второй этап — макрокластеризация (Weighted Fuzzy C-Means)

# Определения

Нечеткий микрокластер (Fuzzy Micro-Cluster, FMiC) задается вектором

$$FMiC = (\overline{CF}, SSD, N, t, M)$$

- ▶  $\overline{CF}_i = \sum \mu_{ij} x_j$  — линейная взвешенная сумма наблюдений
- ▶  $SSD_i = \sum \mu_{ij}^m d(x_j, c_i)^2$  — квадратичная взвешенная сумма расстояний до наблюдений
- ▶  $N_i$  — количество наблюдений
- ▶  $t_i$  — дата последнего наблюдения
- ▶  $M_i = \sum_{x_j \in C_i} \mu_{ij}$  — сумма степеней принадлежности

Важные свойства: **инкрементность, аддитивность**

# Определения

Тогда можно выразить

- ▶  $c = \overline{CF}/M$  — центроид микрокластера
- ▶  $dp = \sqrt{\frac{SSD}{N}}$  — нечеткое рассеивание (fuzzy dispersion), отражает радиус микрокластера
- ▶  $FR_{ij} = \frac{dp_i + dp_j}{d(c_i, c_j)}$  — матрица близости микрокластеров  $i$  и  $j$ .  
Чем больше значение, тем ближе микрокластеры
- ▶  $\tau \in [0, +\infty]$  — пороговое значение близости, при котором кластеры  $i$  и  $j$  объединяются в один. При  $\tau < 1$  границы кластеров не пересекаются

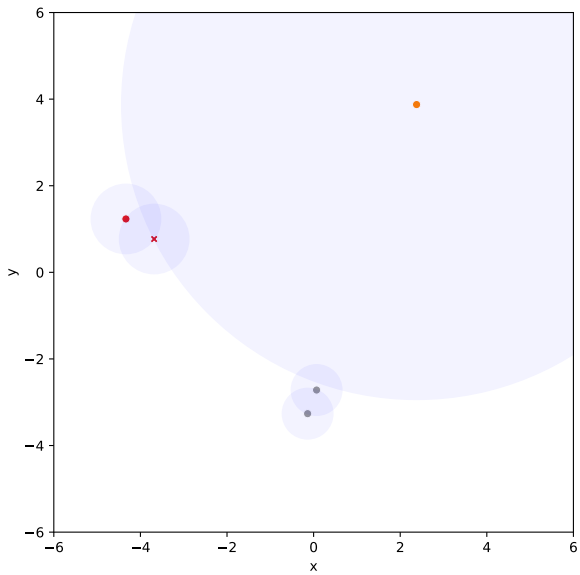
## Алгоритм d-FuzzyStream

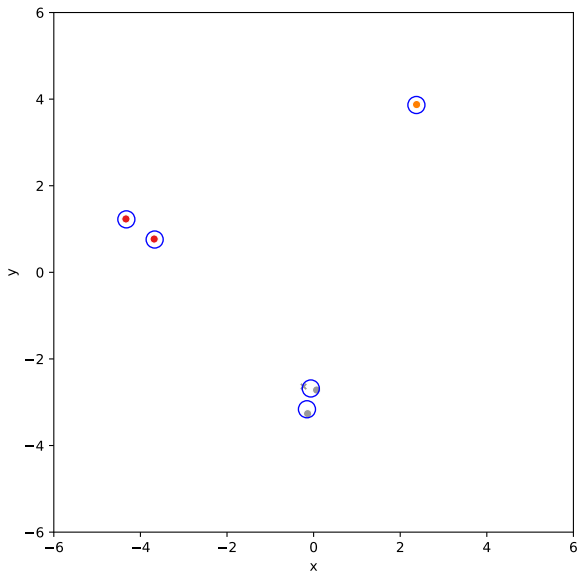
Входные параметры:  $minFMiC$ ,  $maxFMiC$  — минимальное и максимальное количество микрокластеров,  $\tau$  — порог объединения микрокластеров,  $m$  — коэффициент размытия

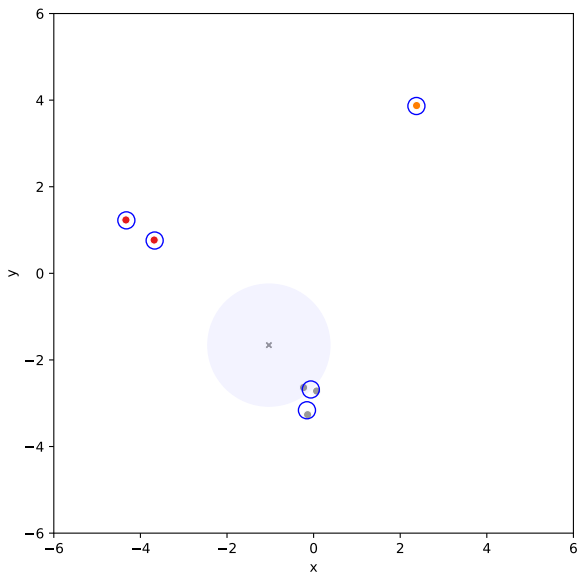
1. Пока количество микрокластеров меньше минимального, создаем новые микрокластеры под каждое наблюдение
2. Если новое наблюдение попадает в радиус хотя бы одного микрокластера, то оно инкрементируется в параметры этих микрокластеров со степенью принадлежности

$$\mu_{ik} = 1 / \sum_j \left( \frac{d(x_k, v_i)^2}{d(x_k, v_j)^2} \right)^{\frac{1}{m-1}}$$

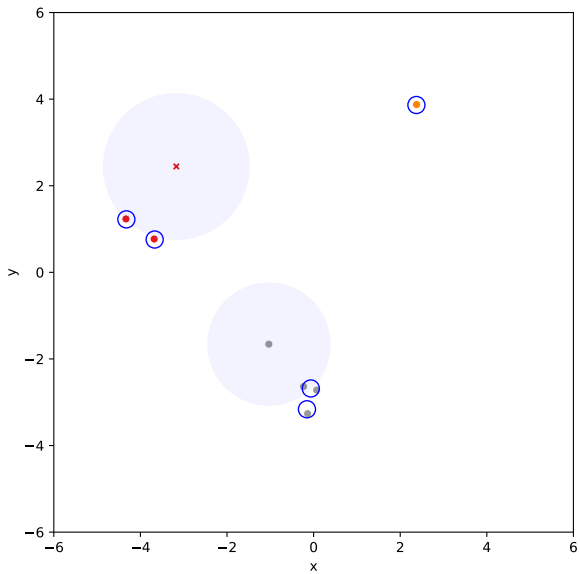
3. Если новое наблюдение не попадает в радиус микрокластера, то создаем под него новый микрокластер
4. При превышении максимального количества кластеров удаляется самый старый
5. При превышении порога близости кластеры объединяются

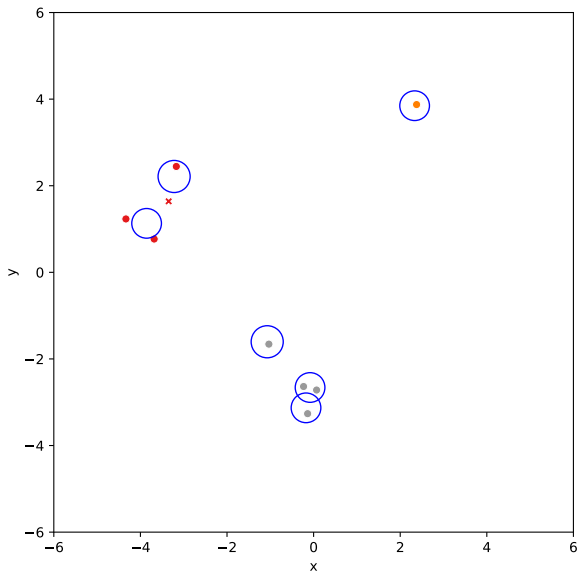


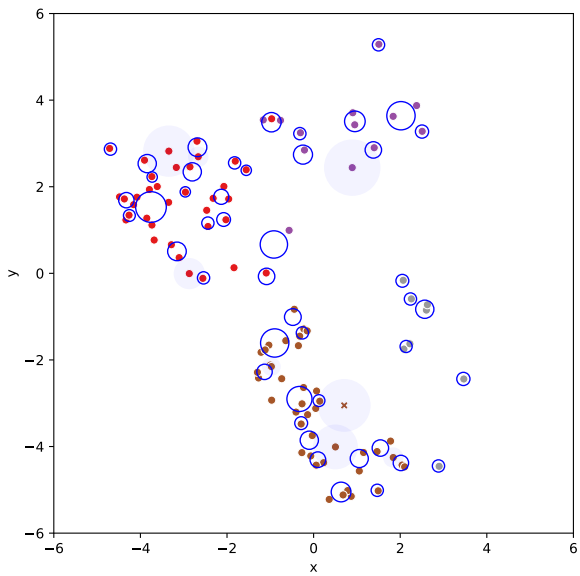






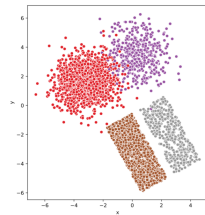
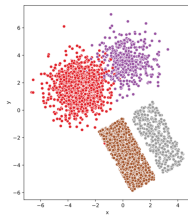
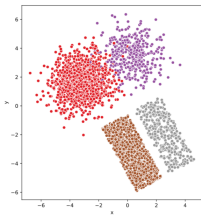




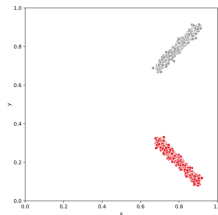
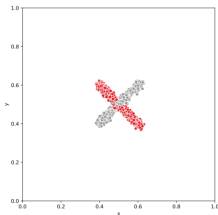
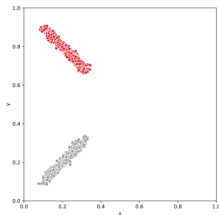


# Датасеты для тестирования

- ▶ Сгенерированные 10000 наблюдений (3 временных среза)



- ▶ Сгенерированные 11000 наблюдений (3 временных среза)



# Метрики качества

- ▶ *Creations* — количество созданных микрокластеров
- ▶ *Removals* — количество удаленных микрокластеров
- ▶ *Absorptions* — количество попаданий в радиус микрокластера
- ▶ *Merges* — количество объединений микрокластеров
- ▶  $Purity = \frac{1}{N} \sum_i \max_j |C_i \cap T_j|$  — мера «чистоты» кластеризации, где  $T_j$  — размеченный заранее класс наблюдения

# Результаты тестирования

**Датасет #1 (1000 - 2000 - 3000)**

#	purity	creations	removals	merges	absorptions	seconds
0	0.194	523	91	332	623	29
1	0.129	824	451	274	182	75
2	0.127	821	387	335	295	68
<b>Среднее</b>	<b>0.15</b>	<b>722.67</b>	<b>309.67</b>	<b>313.67</b>	<b>366.67</b>	<b>57.33</b>

**Датасет #2 (1000 - 2000 - 3000)**

#	purity	creations	removals	merges	absorptions	seconds
0	0.194	523	91	332	623	29
1	0.129	824	451	274	182	75
2	0.127	821	387	335	295	68
<b>Среднее</b>	<b>0.15</b>	<b>722.67</b>	<b>309.67</b>	<b>313.67</b>	<b>366.67</b>	<b>57.33</b>

## Литература

- ▶ A Framework for Clustering Evolving Data Streams, 2003, Charu C. Aggarwal, Jiawei Han, Jianyong Wang, Philip S. Yu
- ▶ FuzzStream: Fuzzy Data Stream Clustering Based on the Online-Offline Framework, 2017, Priscilla de Abreu Lopes and Heloisa de Arruda Camargo
- ▶ d-FuzzStream: A Dispersion-Based Fuzzy Data Stream Clustering, 2018, Leonardo Schick, Priscilla de Abreu Lopes and Heloisa de Arruda Camargo
- ▶ Merging Clusters in Summary Structures for Data Stream Mining based on Fuzzy Similarity Measures, 2019, Leonardo Schick, Priscilla de Abreu Lopes and Heloisa A. Camargo