АЛГОРИТМИ СІМЕЙСТВА FOREL (ФОРмальний ЕЛемент)

Надія І. Недашківська n.nedashkivska@gmail.com

Алгоритм FOREL (ФОРмальний ЕЛемент)

Дано: Множина об'єктів $I = \{i_1, i_2, \ldots, i_n\}$

$i_j = \{x_1, x_2, \ldots, x_m\}$ m ознак $x_h \in R$

Ідея. В один кластер мають потрапити об'єкти, "близькі" до деякого "центрального" об'єкту.

$\rho_k = \sum_j \rho(c_k, i_j)$ - сума внутрішньокластерних відстаней для k-го кластеру, $j = 1, \ldots, n_k$

$F = \sum_k \rho_k$ - критерій розбиття на кластери (мінімізувати)

Будуються кластери сферичної форми.

Кількість кластерів залежить від радіуса сфер.

Алгоритм FOREL (ФОРмальний ЕЛемент)

1) Задаємо R_0 – мінімальний радіус, який охоплює всі n об'єктів.

2) Зменшуємо радіус сфер: $0.9R_0$. Центр – випадковим чином.

Шукаємо точки, відстань до яких є меншою за заданий радіус, і розраховується центр мас цих "внутрішніх" точок.

Центр сфери переноситься в центр мас і знову розраховуються внутрішні точки, поки центр не стабілізується.

Сфера "зупиняється" в області локального максимуму щільності точок у просторі ознак.

Алгоритм FOREL (ФОРмальний ЕЛемент)

- Ініціалізуємо множ. некластеризованих точок $U:=I$, задаємо R
- Поки у вибірці є некластеризовані точки, тобто $U \neq \varnothing$:
  - випадковим чином вибрати $i_0 \in U$
  - повторювати:
    - утворити кластер – сферу з центром i_0 і радіусом R

$$C_0 := \{i_k \in I \mid d(i_k, i_0) \leq R\}$$

    - помістити центр сфери в центр мас кластера

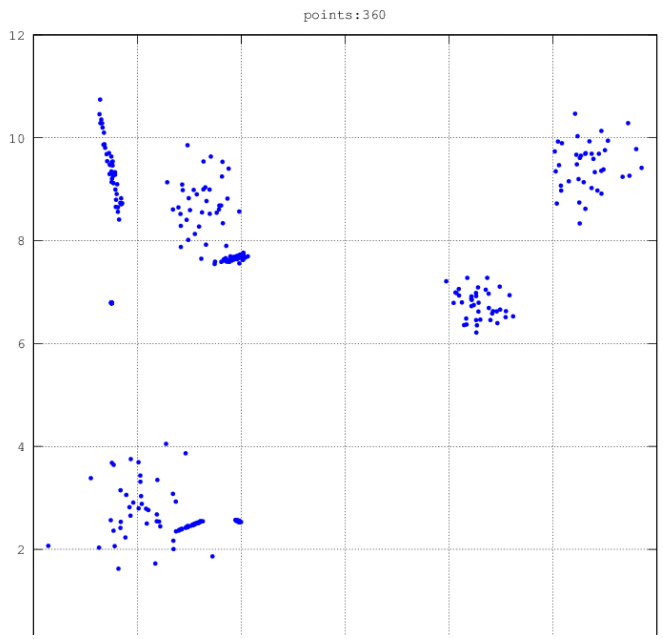
$$i_0 := \frac{1}{|C_0|} \sum_{i_k \in C_0} i_k$$

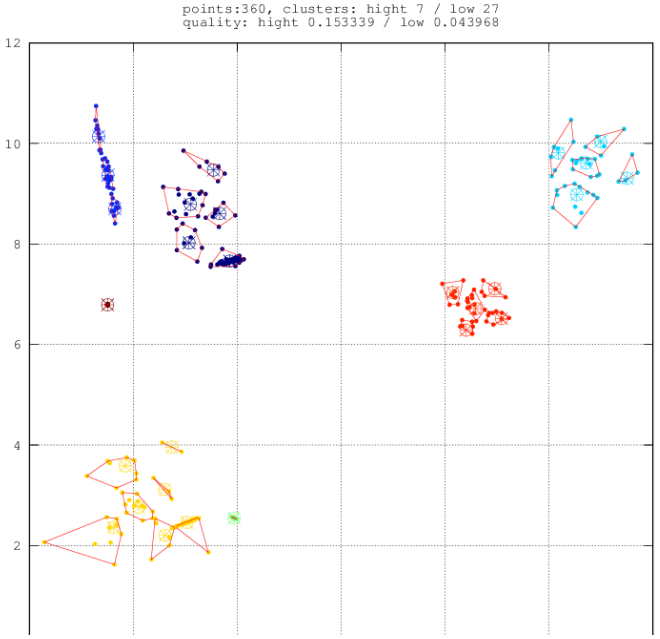
  - поки центр i_0 не стабілізується
  - відмітити всі точки C_0 як кластеризовані: $U := U \setminus C_0$

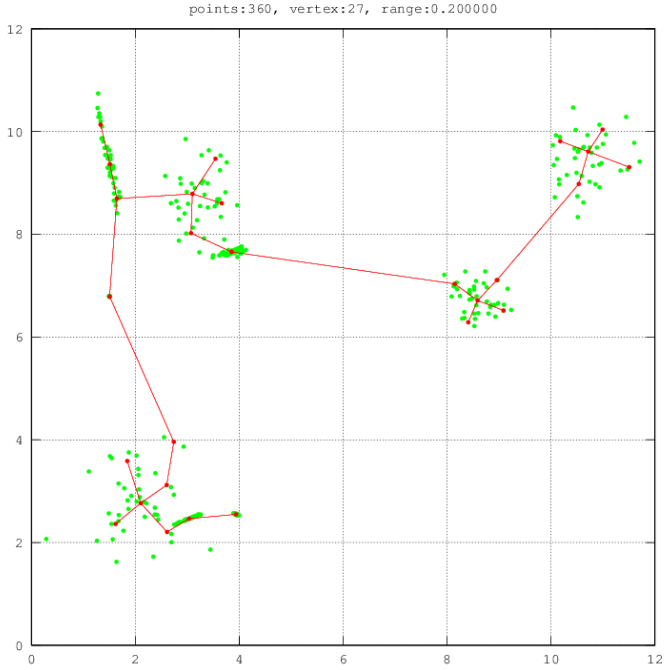
Алгоритм FOREL (ФОРмальний ЕЛемент)

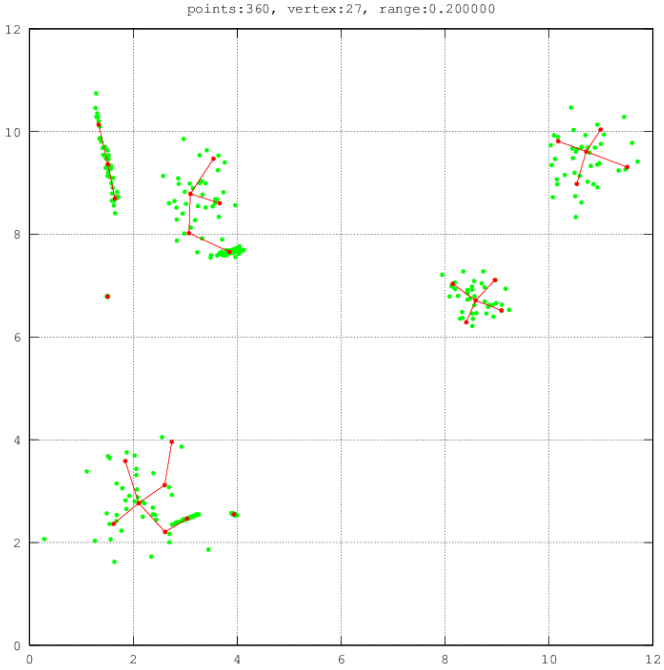
- Ініціалізуємо множину некластеризованих точок $U:=I$.
- Поки у вибірці є некластеризовані точки:
- …
- повторювати:
- …
- …
- поки …
- відмітити всі точки C_0 як кластеризовані: $U:=U \setminus C_0$
- Застосувати алгоритм мінімального покриваючого дерева до множини центрів всіх знайдених кластерів.
- Кожний об’єкт $i_k \in I$ приписати кластеру з найближчим центром.

Алгоритм FOREL (ФОРмальний Елемент)









Алгоритм FOREL (ФОРмальний ЕЛемент): переваги

- підвищується ефективність алгоритму МПД – центрів кластерів (сфер) набагато менше ніж початкових об'єктів;

- можливість описувати кластери довільної геометричної форми – варіюючи R, отримуємо кластеризації різного ступеня детальності. Якщо кластери близькі за формою до шарів, то R беруть великим. Для описання кластерів більш складної структури необхідно зменшувати R.

Алгоритм FOREL (ФОРмальний ЕЛемент): недоліки

- чутливість до вибору початк. наближення i_0 кожного кластеру

Для уникнення генерують декілька (напр., 10 - 20) кластеризацій.

Вибирається кластеризація, яка доставляє найкраще значення заданому функціоналу якості (напр. $F = \sum_k \sum_j \rho(c_k, i_j)$).

Алгоритм FOREL- 2

- Отримати розбиття із заданою кількістю кластерів k.

- Радіус сфери на кожній ітерації зменшується, напр. вдвічі.

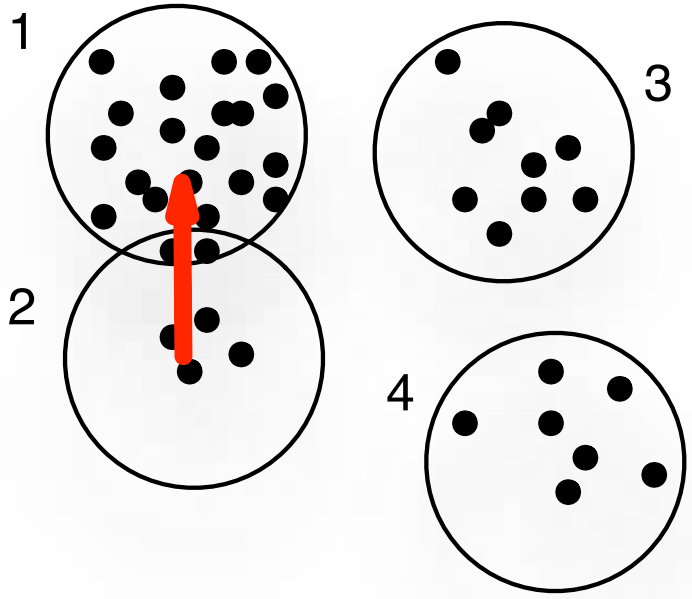
- Функціонал якості кластеризації:

$$F = f(g_i)\sum_{k=1}^{g}\rho_k \qquad f(g_i) = \begin{cases} 1, & if\ g_i = g \\ \infty, & if\ g_i \neq g \end{cases}$$

Найкращому варіанту кластеризації відповідає мінімальне значення F.

Алгоритм FOREL- 3

Використовується коли дані наряду з декількома локальними згустками точок мають ще одиночні точки або невеликі їх скупчення, які випадково розкидані у просторі між згустками.



1, 3, 4 - “стійкі” кластери

2 – випадковий, “нестійкий”

Причини появи випадкових кластерів:

- помилки в даних,
- невдалий вибір радіусу сфер.

Алгоритм FOREL- 3

Вхід: Множина об'єктів $I = \{i_1, i_2, \ldots, i_n\}$

$S = \mathrm{FOREL}(R, I)$ R – заданий радіус

FOREL-3: - Використовується той самий радіус сфер R

- Початкові точки – центри кластерів, отримані в S

- Формування нового кластеру робиться за участю всіх n точок (З базового алгоритму FOREL вилучаємо етап $U := U \setminus C_0$)

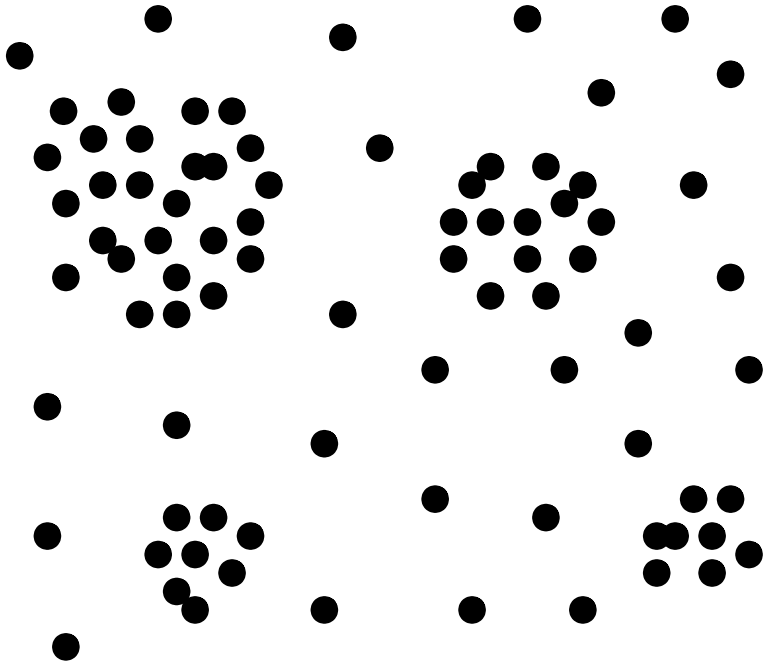
- Функціонал якості кластеризації:

$$F = \sum_{k=1}^{g} n_k f(k) \qquad f(k) = \begin{cases} 1, & if\ k-th\ cluster\ is\ stable \\ 0, & if\ k-th\ cluster\ is\ unstable \end{cases}$$

Результат: Множина стійких кластерів

Алгоритм FOREL- 4

Використовується коли є згустки різного розміру на рівномір-ному фоні.



Етапи алгоритму:

- Шукаються потенційні центри майбутніх кластерів,
- Перевіряється, чи є ці центри центрами стійких кластерів.

Алгоритм FOREL- 4

Етап 1:

Поки множина некластеризованих точок не є порожньою:

1.1. Базовий FOREL з достатньо великим радіусом R

n_j – кількість внутрішніх точок отриманого кластеру

1.2. Якщо $n_j \geq d$:

- центр кластеру заноситься в список претендентів на центр “стійкого” кластеру

Інакше: - список претендентів не змінюється

1.3. Внутрішні точки кластеру виключаються з подальшого розгляду.

Алгоритм FOREL- 4

Етап 2: Розглядається вся множина з n об'єктів

2.1. Список претендентів впорядковується за спаданням n_j

2.2. Поки не кінець списку претендентів:

- в центр кластеру-претендента поміщається сфера радіуса R
- $n'_j := n_j$ – кількість внутрішніх точок кластеру, $R_j^{\min} := R$
- Поки швидкість зменшення n'_j є малою:
  - значення $R_j^{\min}$ зменшується
  - розраховується n'_j – кількість внутрішніх точок кластеру,

2.3. Вибирається g кластерів з найбільшими значеннями n'_j, де g – задана загальна кількість кластерів.

Це рівнозначно максимізації $F = \sum_{j=1}^{g} n'_j$

Вибір кількості кластерів g

- Якщо кластери слугують для подальшого машинного використання, то можна вибирати великі значення g

- Якщо кластери в подальшому будуть використовуватися людиною, то $g=7\pm2$ (число Міллера)

Для алгоритмів сімейства FOREL