Кластерізація багатовимірних та багатопараметрових зображень "без вчителя" на основі генетичного алгоритму

О. М. Ахметшин. Е. В. Пирогов

Дніпропетровський Національний Університет, Україна, 49050,
м. Дніпропетровськ, пер. Науковий 13
Електронна пошта : akhm@mail.dsu.dp.ua

Abstract

In article offered new unsupervised clastering method multimeasure and multiparameter data based on genetic algorithm. The model is capable to an estimation of amount clasters in data. The article consist of the theoretical description of the offered model and her parameters, examples of clastering multiparameter NMR medical image and multimeasure image of earth surface.

Анотація

В роботі запропоновано новий метод кластерізації "без вчителя" багатовимірних та багатопараметрових даних оснований на генетичному алгоритмі. Отримана модель здатна до оцінки кількості кластерів в даних. Робота складається з теоретичного опису запропонованої моделі, її параметрів та прикладів кластерізвції багатопараметрового ЯМР медичного зображення і багатовимірного зображення земної поверхні.

1. Вступ

Серед методів клатерізації особливе місце займають алгоритми, що дають змогу оцінити кількість кластерів в даних. Оцінка кількості кластерів в даних, що аналізуються для багатьох прикладних задач розпізнання є вирішальним для висновків про належність об'єктів до однієї групи.

Актуальними задачами, які потребують алгоритмів самоорганізуючоїся кластерізації є задачі медичної діагностики, прикладні задачі економіки, дистанційні задачі геофізики та ин.

Але при великій потребі в алгоритмах даного типу їх кількість досі мала. Це пояснюється складністю побудови моделі кластерізації в якій будуть одночасно покращуватись параметри кластерів відносно міжкластерної відстані та внутрішньокластернї щільності і при цьому буде варіюватись кількість кластерів з метою знаходження глобального оптимуму. Прикладом моделі для кластерізації без вчителя є самоорганізуючі карти Кохонена [2].

В роботі представлено новий метод кластерізації "без вчителя" на основі генетичного алгоритму.

2. Метод

Вибором моделі генетичного алгоритму для вирішення задачі самоорганізуючоїся кластерізації стало те, що генетичні алгоритми – евристичні алгоритми що здатні знаходити глобальний максимум функції. Евристична модель генетичного алгоритму ґрунтується на твердженні про те, що не вигідно отримувати якісно покращенні результати за допомогою поганих початкових даних.

Нехай необхідно кластеризовати багатовимірну вибірку $D = \{d_1 .. d_n\}$ де n – кількість даних в r-мірній виборці.

Для розробки алгоритму кластерізації "без вчителя" на основі моделі генетичного алгоритму потрібно ввести наступні поняття :

1. Популяція – множина реалізацій розбиття даних на кроці ітераційного процесу.
$$Q = \{g_1 .. g_m\} \quad (1)$$
2. Ген – реалізація розбиття даних на кластери.
$$g_i = L_1 .. L_{k\_i} \quad (2)$$

k_i – кількість кластерів в даній реалізації. L_i – індекс елемента даних, що вважається центром кластера.

За допомогою генетичного алгоритму буде виконуватись ітераційний пошук оптимального розбиття на кластери. На кожному кроці ітерації оцінюються отримані розбиття і для подальшого розвитку популяції покращуються ті розбиття, що є кращими в відношенні середньої міжклатерної відстані та середньої внутрішньокластерної щільності. Отримані таким чином нові реалізації розбиття будуть заміняти ті розбиття, що оцінені як погані на даному кроці ітераційного процесу.

Генетичний алгоритм складається з операторів селекції, схрещування та мутації [1].

Селекцією називається вибір двох генів з популяції для проведення операції схрещування. При цьому імовірність того, що ген буде вибрано обраховується за формулою :

$$P(g_i) = \frac{\sqrt{P_i^{dist\,2} + P_i^{dens\,2}}}{\sum_{i=1}^{m} \sqrt{P_i^{dist\,2} + P_i^{dens\,2}}} \quad (3)$$

P_i^{dist} - імовірність того, що ген буде вибрано завдяки кращій середній міжклатерній відстані.

$$P_i^{dist} = \frac{\sum_{l=1}^{k_i}\sum_{j=1}^{k_i} dist(d_{L_j^{g_i}}, d_{L_l^{g_i}})}{k_i - 1} \quad (4)$$

$dist(d_{L_j^{g_i}}, d_{L_l^{g_i}})$ - відстань між двома кластерами.

P_i^{dens} - імовірність того, що ген буде вибрано завдяки кращій середній внутріщньокластерній щільності. При цьому для оцінки щільності в кластері використовується середньоквадратичне відхилення даних, що належать кластеру відносно центра кластера :

$$P_i^{dens} = \frac{1 - P_i^{std}}{\sum_{j=1}^{M} 1 - P_j^{std}} \quad (5)$$

Обраховується як обернена імовірності вибору гена відносно середньоквадратичного відхилення в кластері.

$$P_i^{dens} = \frac{\sum_{j=1}^{k_i} std(d_{L_j^{g_i}}, U_{L_j^{g_i}})}{k_i - 1} \quad (6)$$

$std(d_{L_j^{g_i}}, U_{L_j^{g_i}})$ -середньквадраничне відхилення обраховане для $U_{L_j^{g_i}}$ – множини даних, що входить до кластера, відносно центра кластера $d_{L_j^{g_i}}$

Схрещування – операція, за допомогою якої знаходять нові реалізації розбиття на кластери. Операція проводиться над генами, що були вибрані за допомогою оператора селекції. Нехай вибрано два гена $\mathbf{g^1}$ та $\mathbf{g^2}$ виберемо один з цих генів як головний. Тобто такий, кластери якого будуть переноситись до нового гена незмінними. Нехай це $\mathbf{g^1}$

Обробка головного гена. Спочатку проводять операцію видалення з розбиття таких кластерів, відсутність яких покращить середню міжкластерну відстань в реалізації та середню внутрішньокластеру щільність. Для цього обраховують обернені імовірності видалення кластерів з гена g^1 :

$$P_i^{g^1} = \frac{\sqrt{P(g^1)_i^{dist^2} + P(g^1)_i^{dens^2}}}{\sum_{i=1}^{m}\sqrt{P(g^1)_i^{dist^2} + P(g^1)_i^{dens^2}}} \quad (7)$$

де $P(g^1)_i^{dist} = \frac{\sum_{j=1}^{k_{g^1}} dist(d_{L_i^{g^1}}, d_{L_j^{g^1}})}{\sum_{k=1}^{k_{g^1}}\sum_{j=1}^{k_{g^1}} dist(d_{L_j^{g^1}}, d_{L_k^{g^1}})}$ (8)

$$P(g^1)_i^{dens} = \frac{std(d_{L_i^{g^1}}, U_{L_i^{g^1}})}{\sum_{k=1}^{k_{g^1}} std(d_{L_k^{g^1}}, U_{L_k^{g^1}})} \quad (9)$$

Кластери, що потрібно видалити із розбиття задовольняють умові :

$$P_i^{g^1} < w / k_{g^1} \quad (10)$$

де w – **коефіцієнт жорсткості видалення. w→1.**

Після видалення з головного гена поганих кластерів формується **новий ген** $\mathbf{g^*}$, до якого входять всі кластери, що залишились в головному.

Синтез нових кластерів. Синтез нових кластерів проводять на основі кластерів, що вже є в новому гені та кластерів, що є в **допоміжному гені** $\mathbf{g^2}$.

Для того, щоб встановити, з яких саме кластерів необхідно формувати нові кластери знаходять матрицю схрещування :

$$B = \begin{bmatrix} b_{11}..k_{1k_{g^*}} \\ b_{k_{g^2}1}..k_{k_{g^2}k_{g^*}} \end{bmatrix} \quad (11)$$

де :

$$b_{ij} = \begin{cases} 0 - \textit{немає схрещування} \\ 1 - \textit{схрещуються } i-\textit{й кластер з нового гена та } j-\textit{й кластер з допоміжного} \end{cases} \quad (12)$$

Ця матриця формується за допомогою ймовірностей схрещування :

$$P_{ij}^{cross} = P_i^{g^*} * P_j^{g^2} \quad (13)$$

де $P_i^{g^*}$ та $P_i^{g^2}$ обраховуються за формулою (7).

Після створення матриці схрещування для кожної пари кластерів активних для цієї матриці проводять схрещування за наступною схемою :

1. Не хай схрещуються два кластери задані індексами L_1 та L_2 тоді їх бінарне перетворення :

$$L_1 \rightarrow LB_1\{lb_1^1 .. lb_{b\_max}^1\}$$
$$L_2 \rightarrow LB_2\{lb_1^1 .. lb_{b\_max}^1\} \quad (14)$$

Де $b\_max = \log_2(n)$

2. Формують бінарну матрицю допустимих нащадків :

$$LB = \begin{bmatrix} lb_1^1, lb_2^2, lb_3^2, .., lb_{b\_max}^2 \\ lb_1^1, lb_2^1, lb_3^2, .., lb_{b\_max}^2 \\ ..... \\ lb_1^1, lb_2^1, lb_3^1, .., lb_{b\_max}^2 \end{bmatrix} \xrightarrow{int} LI = \begin{bmatrix} l_1 \\ ..... \\ l_{b\_max-1} \end{bmatrix} \quad (15)$$

З матриці нащадків вибирають такого нащадка, який знаходиться як надалі від батьків. Тобто до нової реалізації розбиття вноситься такий кластер, що знаходиться як надалі від кластерів, що його утворили.

3. Після отримання нового кластеру L^* його вносять до g^*

Сформований ген g^* потрібно внести до популяції – для цього з популяції виноситься ген з найгіршими характеристиками міжкластерної відстані та внутрішньокластерної щільності. Імовірність того, що ген буде видалено з популяції обернена імовірності того, що ген буде брати участь в схрещуванні. Тобто :

$$\overline{P}(g_i) = \frac{\sqrt{\overline{P}_i^{dist\,2} + \overline{P}_i^{dens\,2}}}{\sum_{i=1}^{m} \sqrt{\overline{P}_i^{dist\,2} + \overline{P}_i^{dens\,2}}} \qquad (16)$$

де

$$\overline{P}_i^{dist} = \frac{1 - P_i^{dist}}{\sum_{j=1}^{M} 1 - P_j^{dist}} \qquad (17)$$

$$\overline{P}_i^{dens\,2} = \frac{1 - P_i^{dens}}{\sum_{j=1}^{M} 1 - P_j^{sens}} \qquad (18)$$

Оператор мутації використовується для того, щоб знайдені в процесі еволюціонування популяції розбиття прагнули до глобального оптимуму. Тобто оператор мутації – цілеспрямоване внесення в популяцію випадкової інформації з метою виведення системи із локальних оптимумів. В залежності від наявних даних імовірність мутації можливо варіювати з метою отримання більш швидкої збіжності ітераційного процесу.

4. Параметри моделі.

Вхідними параметрами моделі є :

1. **Розмір популяції**. Цей параметр відповідає за швидкість сходження еволюційного процесу до стаціонарного стану. Встановлення його дуже малим буде гальмувати процес завдяки невеликій кількості реалізацій, що будуть порівнюватись між собою. Але вибір великої популяції може гальмувати час виконання ітерацій.
2. **Коефіцієнт жорсткості видалення** – відповідає за видалення "поганих" кластерів на етапі синтезу нового гена. $W \in (0,1)$ Є регулювальником сходження алгоритму. При $W=1$ як і при $W \to 0$ алгоритм буде розходитись, бо в першому випадку кількість кластерів в новому гені буде постійно зменшуватись, в другому – кількість кластерів буде постійно зростати.
3. **Імовірність мутації** – відповідає за виведення популяції з локального мінімуму. Дуже велике значення цього коефіцієнта буде лише погіршувати вже знайдені розбиття і може привести до розходження алгоритму, дуже мале значення призведе до того, що процес пошуку буде затримано в локальному оптимумі, і навіть до того, що локальний оптимум буде вибрано, як глобальний. Для більшості задач імовірність мутації 10% така ж як для природної мутації.

Важливим аспектом є встановлення критерію зупинки алгоритму. Такий критерій може бути отримано за допомогою аналізу стабільності прогресу середньої міжкластерної відстані та середнього внутрішньомластерної щільності.

5. Використання метода для кластерізації багатовимірних та багатопараметрових зображень.

Багатовимірні та багатопараметрові зображення використовуються в багатьох галузях прикладної науки. Це можуть бути медичні мультизображення чи мультизображення земної поверхні. Аналіз таких зображень з метою виділення інформації про наявність об'єктів різних типів – важлива задача діагностики.

Було оброблено багатопараметрове ЯМР медичне зображення, що складалось з трьох гомографічних зображень (рис 1). Для кластерізації було використано метрику Евкліда. Отримане кластерізоване зображення виявило наявність 49 кластерів (рис 2.а). Ітераційний алгоритм ініціювався за допомогою популяції з 10 реалізацій. Центри кластерів в початкових реалізаціях було задано випадково, максимальна кількість кластерів в початковій реалізації 20. Імовірність мутації 10%. Коефіцієнт жорсткості видалення в операторі схрещування 0.9. Зупинка ітераційного процесу проводилась за допомогою аналізу усталеності прогресу міжкластерної відстані та прогресу внутрішньокластерної щільності (рис 2.б рис 2.в). Отримано 588 ітерацій.

Також було кластерізовано багатовимірне зображення земної поверхні яке мало 3 параметри (рис 3.а). Клатерізоване зображення містить 32 кластери (рис 3.б). Розмір популяції містив 10 реалізацій. Початкова популяція формувалась випадково з максимальною кількістю кластерів в реалізації 10. Імовірність мутації 10%. Коефіцієнт жорсткості видалення в операторі схрещування 0.9. Ітераційний процес стабілізувався на 303 ітерації.

Висновки.

Використання генетичного алгоритму як базового для моделі самоорганізуючоїся кластерізації – принціпово новий підхід до вирішення проблеми побудови алгоритмів кластерізації "без вчителя".

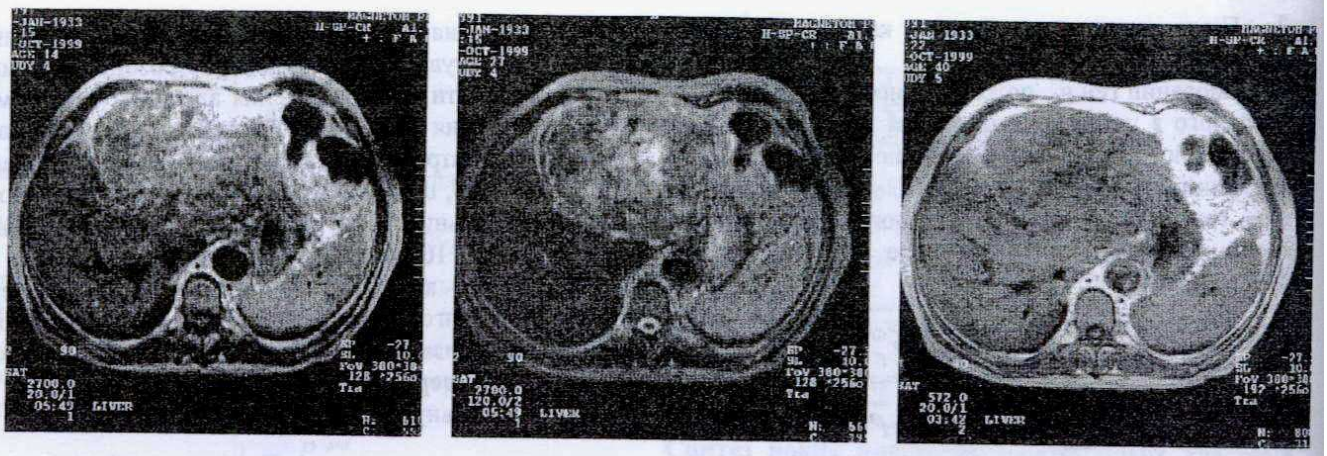

Рис. 1 Багатопараметрове ЯМР медичне зображення.

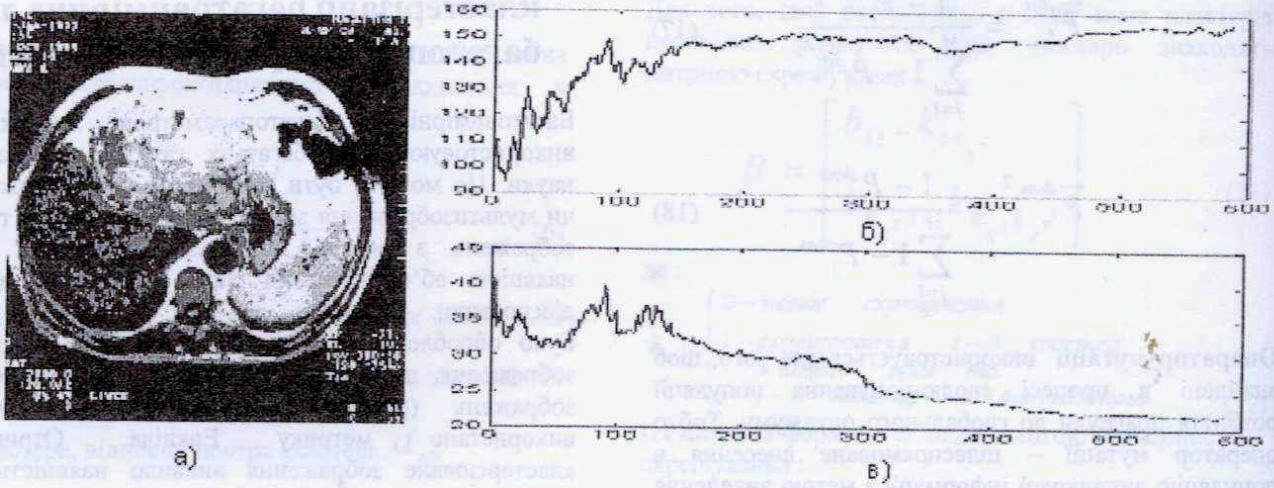

Рис. 2 Результат кластерізації багатопараметрового медичного зображення : а) нове кластері зоване зображення (49 кластерів); б) прогрес середньої міжкластерної відстані; в) прогрес середнього внутрішньокластерного середньоквадратичного відхилення.

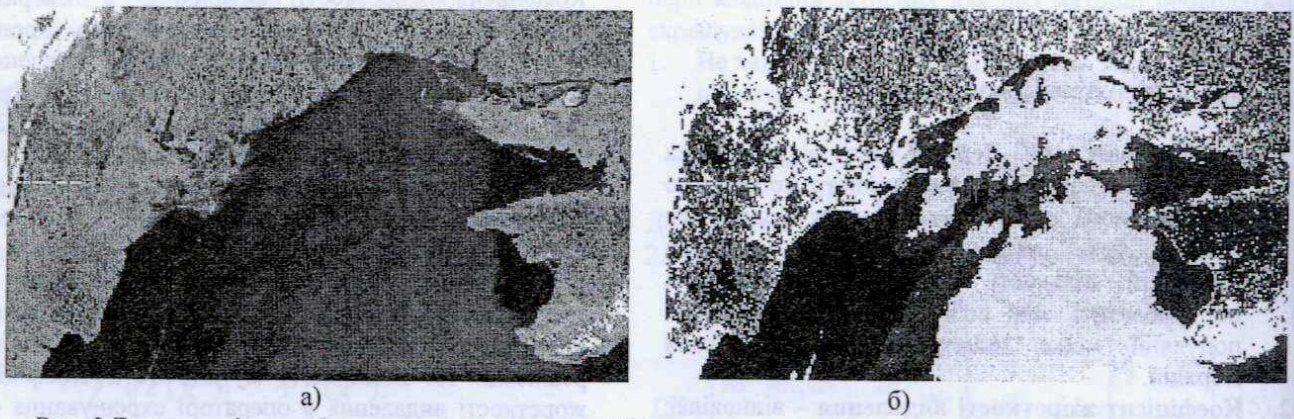

Рис. 3 Багатовимірне зображення земної поверхні (а) та нове кластері зоване зображення (б) 32 кластера.

В роботі розроблено алгоритм кластерізації багатовимірних даних "без вчителя", що базується на генетичному алгоритмі та виконана його комп'ютерна реалізація для медичних зображень та зображень земної поверхні.

Література

1. Goldberg D. E. Genetic algorithms in search, optimization and machine learning. Addison Wesleg, MA. 1989.
2. Kohonen D. Self-organizing maps. NY. Springer-Verlag, 1995