УДК 519.161

К.т.н., доцент Зорін Ю.М., студент Подольський С.В.

**Національний технічний університету України
«Київський політехнічний інститут»**

ГЕНЕТИЧНИЙ АЛГОРИТМ КЛАСТЕРИЗАЦІЇ МНОЖИНИ ВІДПОВІДНО ДО ЗАДАНОГО КРИТЕРІЮ

Abstract

Yuri Zorin, Assoc. Prof., PhD; Sergey Podolsky, student

Genetic algorithm of set clustering with accordance to given criterion

The paper describes an alternative approach to genetic algorithm of cluster analysis implementation. The problem of Data Mining and clustering is studied and discussed. Genome representation and genetic operators of crossover, mutation and fitness-function evaluation are proposed. The analysis of efficiency of the genetic algorithm is fulfilled. The prospects for further research and modification are proposed.

Вступ

Задача кластеризації (об’єднання в групи схожих об’єктів) – одна з фундаментальних задач аналізу даних (Data Mining), яка відноситься до таких, що самонавчаються і не мають чіткого результату вирішення. Дана задача формулюється як метод редукції (ущільнення) деякої множини даних в більш компактну класифікацію об’єктів. Кластерний аналіз широко застосовується для сегментації зображень, маркетингових досліджень, прогнозування тощо [1]. До популярних алгоритмів кластеризації відносять Expectation-Maximization, k-середніх (k-means, fuzzy c-means) та інші [2].

У загальному вигляді задача кластеризації формулюється наступним чином. Нехай задана множина спостережень X та скінченна вибірка об’єктів

$$X^m = \{x_1, x_2, \dots, x_m,\} \subset X.$$

Необхідно розбити вибірку X^m на підмножини, що не перетинаються, – кластери $S_1, S_2, \dots, S_K$ таким чином, щоб забезпечити мінімум (екстремум) деякого критерію P (функціонала якості), тобто:

$$S = \{S_1, S_2, \dots, S_K\} : F(S) \to \min_S P \left(\max_S P\right).$$

Алгоритм кластеризації – це функція $\alpha: X^m \to S$, яка будь-якому об'єкту x_i ставить у відповідність кластер $S_j \in S$.

Постановка задачі

Метою роботи є розробка генетичного алгоритму (ГА) кластеризації заданої вибірки (скінченної множини) об'єктів, де критерій кластеризації задається у вигляді цільової функції ГА та основних генетичних операцій для його реалізації.

Генетичний алгоритм кластеризації заданої вибірки

Для здійснення кластеризації ГА працює поетапно. На кожному етапі ГА групує і повертає один кластер. На першому етапі з заданої вихідної вибірки об'єктів ГА виділяє перший кластер, оцінка розміру якого задається у вигляді діапазону – верхньої та нижньої границі. Після завершення роботи першого етапу ГА із множини об'єктів, що залишилися некластеризованими, аналогічним методом виділяється другий кластер. Таким чином, ГА застосовується до вибірки до тих пір, поки вона не буде вичерпана повністю. Такий підхід введено з метою зменшення розміру геному і, відповідно, складності та часу виконання генетичних операцій.

Кодування геному й обчислення пристосованості

Кожна хромосома репрезентує кластер і містить посилання на всі об'єкти кластеру. На кожному етапі ГА створюється початкова популяція випадково згенерованих варіантів кластерів. Оскільки розмір кластеру не обов'язково є сталим, то розмір кожного кластеру також генерується випадково у межах заданого діапазону. Фактично діапазон задає рекомендацію для ГА, якого розміру кластер необхідно повернути. Максимальний розмір кластеру дорівнює потужності множини об'єктів, що залишилася некластеризованою на поточному етапі роботи алгоритму. Оскільки визначення пристосованості передбачає обчислення міри компактності кластеру відповідно до метрики (способу оцінки подібності) [3], то мінімальний розмір кластеру дорівнює двом. Альтернативним параметром роботи алгоритму є кількість кластерів, яку необхідно отримати. При цьому розміри кластерів розраховуються і вважаються однаковими, тобто сталими.

Для тестування роботи алгоритму у якості міри подібності була застосована евклідова відстань, яка обчислюється за формулою

$$D(x,y) = \sqrt{\sum_{i}^{m}(x_i - y_i)^2},$$

де x, y – пара об'єктів у m-вимірному просторі; x_i, y_i – їх відповідні проекції. Для реалізації даного критерію було розроблено функцію обчислення пристосованості, яка повертає середнє значення відстані об'єктів кластеру до його центру тяжіння:

$$F(S) = \frac{(N_s)^{\alpha}}{\sum_{j}^{N_s} D(x_j, C_s)},$$

де N_s – розмір кластеру S, $D(x_j, C_s)$ – евклідова відстань j-го об'єкта x_j кластеру до центру тяжіння C_s кластеру, радіус-вектор якого обчислюється як фізичний центр маси:

$$\overrightarrow{C_s} = \frac{\sum_{j}^{N_s} \overrightarrow{x_j}}{N_s}.$$

Коефіцієнт α введено у зв'язку з тим, що розмір кластеру може бути не заданий. Фактично дана фітнес-функція характеризує обернену щільність об'єктів кластеру, тому для двох кластерів з різним розміром і однаковою щільністю значення пристосованості буде однаковим. Але в такому випадку ГА буде схильним до групування меншої кількості об'єктів з такою ж щільністю, які знаходяться найближче один до одного: двох, трьох, або в залежності від заданого мінімального розміру кластеру. У такому разі необхідно «заохочувати» більші за розміром кластери такої ж щільності, що було здійснено введенням нелінійної залежності фітнес-функції від розміру кластеру. Експериментальним шляхом було встановлено, що оптимальне значення даного коефіцієнту $\alpha = 1{,}1 \div 1{,}5$ в залежності від очікуваного розміру кластеру та його однорідності. У випадку, якщо розмір кластеру задано конкретно або в межах вузького діапазону, доцільно залишити $\alpha = 1$.

Обчислювальна складність такої фітнес-функції становить $O(2 \cdot N_s)$.

Генетичні операції

Оскільки геном подається у вигляді множини об'єктів, порядок розташування яких не має значення, то доцільним є кросовер на основі операцій над множинами. Алгоритм виконання кросоверу наступний:

1) знайти перетин $I = A \cap B$ двох множин-батьків A, B;
2) знайти різниці множин $R_A = A/I,\ R_B = B/I$;
3) розбити випадковим чином множини R_A і R_B на пари однакових за розміром множин $R1_A$, $R2_A$ та $R1_B$, $R2_B$ відповідно;
4) створити двох нащадків $U1$ та $U2$, виконавши об'єднання множин $U1 = I \cup R1_A \cup R2_B$ та $U2 = I \cup R1_B \cup R2_A$.

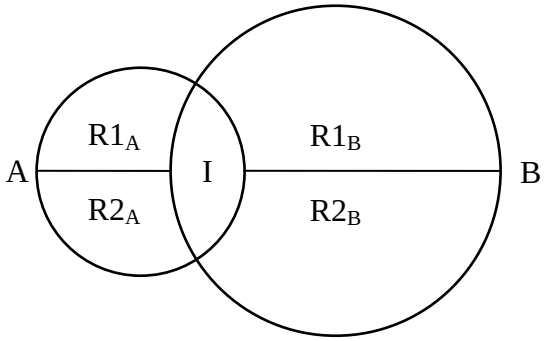

Рис. 1. Виконання кросоверу

Дана реалізація кросоверу дозволяє отримати нащадків, розміри яких варіюються також в межах меншого та більшого розмірів обох батьків (Рис. 1).

Операція мутації може виконувати додавання (конкатенацію), видалення або зміну одного випадкового об'єкту поточного кластеру з урахуванням можливості конфліктів – хромосома не може містити пари однакових елементів.

Висновки

Розроблений ГА показав ефективні результати кластеризації при кількості об'єктів 300-800. При цьому було помічено, що чим більш точно задано розмір кластерів, тим якісніше виконується кластеризація. Зміна функції пристосованості ГА дозволяє задавати нові критерії кластеризації, а отже і вирішувати нові задачі. В подальшому планується адаптація та застосування ГА для інших критеріїв з метою вирішення задач декомпозиції графів – задачі розбиття графа на фіксоване число підграфів заданих порядків, розбиття графа на максимальні сильно зв'язні підграфи.

Література

1. *Jain A.K., Murty M.N., Flynn P.J.* Data Clustering: A Review // ACM Computing Surveys, Vol. 31, No. 3, September 1999. – P. 265.
2. *Fung G.* A Comprehensive Overview of Basic Clustering Algorithms. – IEEE, June, 2001 – Citeseer. – P. 6.
3. *Hathaway R.J., Bezdek J.C.* Optimization of Clustering Criteria by Reformulation // IEEE Transactions on Fuzzy Systems, Vol. 3, № 2, May 1995. – P. 242.