# ПРО УТВОРЕННЯ ВАРІАНТІВ РОЗБИТТЯ В ЗАДАЧІ КЛАСТЕРИЗАЦІЇ

*Надія ТИМОФІЄВА*
*МННЦ ІТіС НАНУ та МОН України 03022, Київ, просп. Ак. Глушкова, 40,*
*тел.: (044) 266 53 52, факс (044) 266 1570*

## Abstract

A new method for a narrowing of sphere of a search of an optimum decision in a clustering problem is proposed. A constant parameters for any problem are revealed, and a conformity with natural laws of changing objective function values depending on them are established. One of them is a finite sequence (a variant of partitioning) which is generated a combination of elements of combinatorial matrix. The rules of formation this sequence are defined. A conformity with natural laws of changing objective function values depending on these variants of partitioing is analised. It is proved that variants of partitioning in a clustering problem consists of subsets. It is shown, that the variants of partitioning for which the objective function take eithher maximum or minimum, can be contained in different subsets or into one of them.

## 1. Вступ

Відомі методи знаходження оптимального розв'язку в задачах комбінаторної оптимізації шляхом відтинання неефективних їх варіантів полягають в аналізі залежності значення цільової функції від вхідних даних. Але в цих задачах клас цільової функції залежить від порядку комбінаторних конфігурацій (аргумента) у їх множині. В задачах кластеризації цей порядок установлюється для підмножин, на які, з використанням незалежних від вхідних даних параметрів, можна розділити множину ізоморфних розбиттів. З урахуванням вхідних даних оговорена множина упорядковується утвореними підмножинами так, що для одержаного порядку визначається деяка закономірність зміни значень функції цілі.

В статті визначаються правила утворення варіантів розбиття з елементів комбінаторної матриці, якою задаються вхідні дані. Доведено, що множина варіантів розбиття в задачі кластеризації розділяється на підмножини. Відповідно на такі ж підмножини розділяється і множина ізоморфних розбиттів. На наведених прикладах показано, що варіанти, для яких цільова функція набуває або найбільшого або найменшого значень можуть знаходитися у різних таких підмножинах або в одній їз них. Використання цієї властивості дозволяє звужувати область пошуку оптимального розв'язку в цих задачах.

## 2. Основні положення

Уточнимо деякі поняття. Розбиттям $n$-елементної множини $A=\{a_1,...,a_n\}$ на $\eta$ підмножин (блоків) назвемо множину підмножин $\rho^k=(\rho_1^k,...,\rho_{\eta^k}^k)$ таку, що $\rho_1^k\cup...\cup\rho_{\eta^k}^k=A$, $\rho_s^k\neq\varnothing$, $\rho_t^k\cap\rho_s^k=\varnothing$, $t\neq s$, $t,s\in\{1,...,\eta^k\}$, $\eta^k\in\{1,...,n\}$ – кількість $\eta^k$ в $\rho^k$. Підмножина $\rho_s^k=(a_1,...,a_{\xi_s^k})$, $a_r\in A$, $t\in\{1,...,\xi_s^k\}$, може мати від $1$ до $n$ елементів ($\xi_s^k\in\{1,...,n\}$). Верхній індекс $k$ в $\rho^k$ позначає порядковий номер розбиття у множині усіх можливих розбиттів $\Theta$, $k\in\{1,...,q\}$, $q$ – кількість елементів у $\Theta$.

Перенумеруємо усі елементи в $\rho^k$ незалежно від їх належності до підмножин $\rho_s^k$ від $1$ до $n$ і розглянемо $\rho^k$ як перестановку $\omega^k$ елементів $1,...,n$. В залежності від умови задачі розбиття $\rho^k$ позначимо як $\rho^k=\omega^k=(\omega_1^k,...,\omega_n^k)$ або як $\rho^k=(\rho_1^k,...,\rho_{\eta^k}^k)$, а елементи будь-якої підмножини $\rho_s^k$ – як елементи $\omega_t^k$ перестановки $\omega^k$.

Два розбиття $\rho^k$ і $\rho^i$ назвемо ізоморфними, якщо $\eta^k=\eta^i$, і для будь-якої підмножини $\rho_t^k\subset\rho^k$ знайдеться підмножина $\rho_s^i\subset\rho^i$, для якої $\xi_t^k=\xi_s^i$. Підмножину ізоморфних розбиттів позначимо $\Theta_{\eta^k}\subset\Theta$. **П**ерше розбиття (перестановку) у підмножині $\Theta_{\eta^k}\subset\Theta$ позначимо

$$\rho^1=((1,...,\xi_1^1),(\xi_1^1+1,\xi_1^1+2,...\ ...,\xi_1^1+\xi_2^1),...,(\sum_{j=1}^{\eta^1-1}\xi_j^1+1,...,n)). \qquad (1)$$

Розбиття $\rho^k$ за кількістю підмножин і кількістю в них елементів розділяються на чотири типи [1]. До першого типу відносяться $\rho^k$, кількість елементів у всіх підмножинах якого – різна. Кількість елементів у підмножинах $\rho_j^k$ розбиття другого типу однакова. В розбиття третього типу входять дві і більше підмножини, які містять один елемент. Хоча б одна підмножина повинна містити більше ніж один елемент. В розбиття четвертого типу входять дві і більше підмножини, кількість елементів у яких однакова. З них одна підмножина повинна мати порівнянно з іншими найбільше елементів.

В подальшому розглянемо зміну значень цільової функції в задачі кластеризації для підмножини ізоморфних розбиттів. Для неї комбінаторною матрицею може бути будь яка із заданих. Покладемо, що вхідні дані задані комбінаторною симетричною матрицею $Q(\rho^k)$, де $g_{rp}(\rho^k) \in Q(\rho^k)$ – кількість зв'язків між заданими елементами $a_r, a_p \in A$ базової множини $A = \{a_1, ..., a_n\}$. Елементи симетричної (0,1)-матриці $C$ $c_{rp} = 1$, якщо $a_r, a_p$ знаходяться в одній підмножині $\rho_j^k \subset \rho^k$ і $c_{rp} = 0$ в іншому випадку. Послідовність наддіагональних елементів матриці $Q(\rho^1)$ задамо числовою функцією $f(j)|_1^m$, а матриці $C$ – функцією $\varphi(j)|_1^m$, де $m = \frac{n(n-1)}{2}$. Комбінаторною функцією $\beta(f(j), \rho^k)|_1^m$ задамо послідовність $f(j)|_1^m$, яка змінюється в залежності від розбиття $\rho^k \in \Theta$. Цільова функція набуде вигляду

$$F(\rho^k) = \sum_{j=1}^{m} \beta_j (f(j), \rho^k) \varphi(j). \qquad (2)$$

## 3. Утворення варіантів розбиття

Оскільки значення функції $\varphi(j)|_1^m$ $\varphi(j) \in \{0,1\}$, то не всі елементи з матриці $Q(\rho^k)$ для розбиття $\rho^k \in \Theta$ приймають участь у формуванні виразу (2) Розглянемо послідовність, яка містить лише ті елементи з $Q(\rho^k)$, по яких для $\rho^k \in \Theta$ оцінюється результат розв'язку задачі (2).

Уведемо множину $u(\rho^k, l)|_1^z = (u_1(\rho^k, 1), ..., u_z(\rho^k, z))$, у якій значення $u_l(\rho^k, l) = \beta_j(f(j), \rho^k)\varphi(j)$, якщо $\varphi(j) = 1$, де $z$ – кількість одиниць у $\varphi(j)|_1^m$. Назвемо $u(\rho^k, l)|_1^z$ варіантом розбиття. Множину варіантів розбиття позначимо $H_u = \left(u(\rho^k, j)|_1^z = (u_1(\rho^k, 1), ..., u_z(\rho^k, z)), k = \overline{1, d}\right)$, $d$ – їх кількість.

Наддіагональні елементи матриці $Q(\omega^1)$ задамо підмножинами $S_p$, $p \in \{1, ..., n-1\}$. Підмножина $S_1$ містить $n-1$ елементів першого рядка. Множина $S_2$ містить $n-2$ значення, які належать другому рядку і т.д. Підмножина $S_{n-1}$ складається з одного $m$–го значення.

Нижче сформулюємо леми і теореми, які із-за обмеження на об'єм наводяться без доведення.

**Лема 1.** Утворення варіантів розбиття $u(\rho^k, l)|_1^z \in H_u$ проводиться вибиранням з кожної множини $S_1, ..., S_{n-1}$ від 1 до $\gamma$ елементів, або не вибирається ні один елемент, $\gamma$ – кількість елементів у множині $S_p$.

Доведення очевидне.

**Лема 2.** Множина $H_u$ складається з підмножин $K_r$, $r = \overline{1, q^*}$. В $K_1$ входять варіанти розбиття, які містять елементи матриці $Q(\rho^k)$, починаючи з адреса 1 і більше, в $K_2$ – входять елементи, починаючи з адреса 2 і т. д., а в $K_{q^*}$ – з найменшим номером адреса $q^*$, де $q^* \in \{n-1, ..., 2n-2\}$.

Нескладно замітити, якщо у $\rho^k \in \Theta$ $\eta^k = 2$ причому $\xi_1^k = n-1$, а $\xi_2^k = 1$, то кількість підмножин у $H_u$ дорівнює три: $K_1$, $K_2$ і $K_n$. Якщо $\eta^k \geq 2$, а $\xi_1^k \geq \xi_2^k \geq ... \geq \xi_{\eta^k}^k$, причому $\xi_p^k \neq 1$, то кількість підмножин у $H_u$ дорівнює $n-1$. Для $\eta^k = 2$ і $\xi_1^k = \xi_2^k$ кількість підмножин дорівнює $q^* = n - \frac{n-2}{2}$.

## 4. Цільова функція в залежності від варіантів розбиття

Розглянемо задачі розбиття для різних типів $\rho^k$, у яких комбінаторна функція апроксимується монотонними, опуклими і вгнутими функціями.

**Теорема 1.** Якщо функція $\beta(f(j), \rho^1)|_1^m$ апроксимується монотонно неспадною, то найбільшого значення функція (2) для першого типу $\rho^k$ набуває для розбиття (1), для якого

$\xi^i_1 < \xi^i_2 < ... < \xi^i_{\eta^i}$, а найменшого – для розбиття (1), для якого $\xi^k_1 > \xi^k_2 > ... > \xi^k_{\eta^k}$. В обох випадках $u(\rho^i, l)|^z_1 \in K_1$. Якщо $\eta^k = 2$ і $\xi^k_{\eta^k} = 1$, то варіант розбиття для $F(\rho^k)$ найбільшого належить підмножині $K_{q^*}$. Якщо $\xi^i_1 = \xi^i_2 = ... = \xi^i_{\eta^i - 1}$, а $\xi^k_{\eta^k} = 1$, то найбільшого значення функція (2) набуває для $\rho^k$, для якого $u(\rho^k, l)|^z_1 \in K_\zeta$, а найменшого, якщо $u(\rho^k, l)|^z_1 \in K_{q^*}$, де $\zeta = \left\lfloor \frac{n-1}{2} \right\rfloor$. Для другого типу варіант розбиття для $F(\rho^k)$ найбільшого належить підмножині $K_1$, а найменшого – підмножині $K_{q^*}$.

**Теорема 2.** Якщо комбінаторна функція апроксимується монотонно незростаючою, то найбільшого значення функція (2) для першого типу $\rho^k$ набуває для розбиття (1), для якого $\xi^i_1 > \xi^i_2 > ... > \xi^i_{\eta^i}$, а найменшого – для розбиття (1), для якого $\xi^k_1 < \xi^k_2 < ... < \xi^k_{\eta^k}$. В обох випадках $u(\rho^i, l)|^z_1 \in K_1$. Якщо $\eta^k = 2$ і $\xi^k_{\eta^k} = 1$, то варіант розбиття для $F(\rho^k)$ найменшого належить підмножині $K_{q^*}$. Якщо $\xi^i_1 = \xi^i_2 = ... = \xi^i_{\eta^i - 1}$, а $\xi^i_{\eta^i} = 1$ то найбільшого значення функція (2) набуває для $\rho^i$, якщо $u(\rho^i, l)|^z_1 \in K_\zeta$, а найменшого – для $\rho^k$ якщо $u(\rho^k, l)|^z_1 \in K_{q^*}$. Для другого типу варіант розбиття для $F(\rho^k)$ найбільшого належить підмножині $K_{q^*}$, а найменшого – підмножині $K_1$.

**Теорема 3.** Якщо $\beta(f(j), \rho^1)|^m_1$ апроксимується опуклою функцією, то цільова функція (2) для першого типу розбиттів набуває найбільшого значення для $\rho^i$, для якого $u(\rho^i, l)|^z_1 \in K_{q^*}$, а найменшого – для $\rho^k$, для якого $u(\rho^k, l)|^z_1 \in K_1$. Якщо $\eta^k = 2$ і $\xi^k_{\eta^k} = 1$, то варіант розбиття для $F(\rho^k)$ найбільшого належить підмножині $K_{q^*}$, а найменшого – підмножині $K_\zeta$. Для другого типу $\rho^k$ варіант розбиття для $F(\rho^k)$ найбільшого належить підмножині $K_1$, а найменшого – підмножині $K_{q^*}$.

**Теорема 4**. Якщо $\beta(f(j), \rho^1)|^m_1$ апроксимується вгнутою функцією, то цільова функція (2) для першого типу розбиттів набуває найбільшого значення для $\rho^i$, для якого $u(\rho^i, l)|^z_1 \in K_1$, а найменшого – для розбиття $\rho^k$, для якого $u(\rho^k, l)|^z_1 \in K_{q^*}$. Якщо $\eta^k = 2$ і $\xi^k_{\eta^k} = 1$, то варіант розбиття для $F(\rho^k)$ найменшого належить підмножині $K_{q^*}$, а найбільшого – підмножині $K_\zeta$. Для другого типу $\rho^k$ варіант розбиття для $F(\rho^k)$ найменшого належить підмножині $K_1$, а найбільшого – підмножині $K_{q^*}$.

У табл. 1а, 2а, 3а наведені підмножини варіантів розбиття. У першій колонці знаходиться множина $K_r$, до якої відноситься варіант розбиття, у другій – розбиття $\rho^k$, у третій – варіант розбиття $u(\rho^k, l)|^z_1$. У табл. 1б, 2б, 3б для тих же параметрів наведено значення цільової функції для різних задач, вхідні дані в яких задано комбінаторними функціями. У першій колонці знаходиться множина $K_r$, до якої відноситься варіант розбиття, у другій – розбиття $\rho^k$, у третій – значення цільової функції, якщо $\beta(f(j), \rho^k)|^m_1 = (1,...,15)$, у четвертій – значення цільової функції, якщо $\beta(f(j), \rho^k)|^m_1 = (15,...,1)$, у п'ятій – значення цільової функції, якщо $\beta(f(j), \rho^k)|^m_1 = (1, 2, 3, ..., 8, ..., 3, 2, 1)$, у шостій – значення цільової функції, якщо $\beta(f(j), \rho^k)|^m_1 = (8, 7, 6, ..., 1, ..., 6, 7, 8)$.

Таблиця 1

а) $n = 6$, $\eta^k = 2$, $\xi^k_1 = 5$, $\xi^k_2 = 1$

| 1 | 2 | 3 |
|---|---|---|
| $K_1$ | (1, 2, 3, 4, 5), (6) | 1,2,3,4,6,7,8,10,11,13 |
| $K_1$ | (6, 1, 2, 4, 5), (3) | 1,3,4,5,7,8,9,13,14,15 |
| $K_1$ | (6, 1, 2, 3, 5), (4) | 1,2,4,5,6,8,9,11,12,15 |
| $K_1$ | (6, 1, 2, 3, 4), (5) | 1,2,3,5,6,7,9,10,12,14 |
| $K_2$ | (6, 1, 3, 4, 5), (2) | 2,3,4,5,10,11,12,13,14,15 |
| $K_6$ | (6, 2, 3, 4, 5), (1) | 6,7,8,9,10,11,12,13,14,15, |

б)

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| $K_1$ | (1, 2, 3, 4, 5), (6) | 65 | 95 | 45 | 45 |
| $K_1$ | (6, 1, 2, 4, 5), (3) | 79 | 81 | 41 | 49 |
| $K_1$ | (6, 1, 2, 3, 5), (4) | 73 | 87 | 43 | 47 |
| $K_1$ | (6, 1, 2, 3, 4), (5) | 69 | 91 | 43 | 48 |
| $K_2$ | (6, 1, 3, 4, 5), (2) | 89 | 71 | 35 | 55 |
| $K_6$ | (6, 2, 3, 4, 5), (1) | 105 | 55 | 49 | 41 |

Таблиця 2

а) $n = 6,\ \eta^k = 2,\quad \xi_1^k = 3,\quad \xi_2^k = 3$

| 1 | 2 | 3 |
|---|---|---|
| $K_1$ | (1, 2, 3), (4, 5, 6) | 1, 2, 6, 13, 14, 15 |
| $K_1$ | (1, 2, 4), (3, 5, 6) | 1, 3, 7, 11, 12, 15 |
| $K_1$ | (1, 2, 5), (3, 4, 6) | 1, 4, 8, 10, 12, 14 |
| $K_1$ | (1, 2, 6), (3, 4, 5) | 1, 5, 9, 10, 11, 13 |
| $K_2$ | (1, 3, 4), (2, 5, 6) | 2, 3, 8, 9, 10, 15 |
| $K_2$ | (1, 3, 5), (2, 4, 6) | 2, 4, 7, 9, 11, 14 |
| $K_2$ | (1, 3, 6), (2, 4, 5) | 2, 5, 7, 8, 12, 13 |
| $K_3$ | (2, 3, 5), (1, 4, 6) | 3, 5, 6, 8, 11, 14 |
| $K_3$ | (2, 3, 6), (1, 4, 5) | 3, 4, 6, 9, 12, 13 |
| $K_4$ | (2, 3, 4), (1, 5, 6) | 4, 5, 6, 7, 10, 15 |

б)

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| $K_1$ | (1, 2, 3), (4, 5, 6) | 51 | 45 | 15 | 39 |
| $K_1$ | (1, 2, 4), (3, 5, 6) | 49 | 47 | 21 | 33 |
| $K_1$ | (1, 2, 5), (3, 4, 6) | 49 | 47 | 25 | 29 |
| $K_1$ | (1, 2, 6), (3, 4, 5) | 49 | 47 | 27 | 27 |
| $K_2$ | (1, 3, 4), (2, 5, 6) | 47 | 49 | 27 | 27 |
| $K_2$ | (1, 3, 5), (2, 4, 6) | 47 | 49 | 27 | 27 |
| $K_2$ | (1, 3, 6), (2, 4, 5) | 47 | 49 | 29 | 25 |
| $K_3$ | (2, 3, 6), (1, 4, 5) | 47 | 49 | 27 | 27 |
| $K_3$ | (2, 3, 5), (1, 4, 6) | 47 | 49 | 29 | 25 |
| $K_4$ | (2, 3, 4), (1, 5, 6) | 47 | 49 | 29 | 25 |

Таблиця 3

а) $n = 5,\ \eta^k = 3,\quad \xi_1^k = 2,\quad \xi_2^k = 2,\quad \xi_2^k = 1$

| 1 | 2 | 3 |
|---|---|---|
| $K_1$ | (1, 2), (3, 4), (5) | 1, 8 |
| $K_1$ | (2, 1), (5,3), (4) | 1, 9 |
| $K_1$ | (4, 5), (1,2), (3) | 1, 10 |
| $K_2$ | (3, 1), (2, 4), (5) | 2, 6 |
| $K_2$ | (2, 5), (1,3), (4) | 2, 7 |
| $K_2$ | (5,4), (3,1), (2) | 2, 10 |
| $K_3$ | (3, 2), (1, 4), (5) | 3, 5 |
| $K_3$ | (1,4), (5,2), (3) | 3, 7 |
| $K_3$ | (5,3), (4,1), (2) | 3, 9 |

Продовження таблиці 3а)

| 1 | 2 | 3 |
|---|---|---|
| $K_4$ | (5, 1), (2,3), (4) | 4, 5 |
| $K_4$ | (1, 5), (4,2), (3) | 4, 6 |
| $K_4$ | (3,4), (5,1), (2) | 4, 8 |
| $K_5$ | (2,3), (4,5), (1) | 5, 10 |
| $K_6$ | (4,2), (3,5), (1) | 6 9 |
| $K_7$ | (4,3), (2,5), (1) | 7, 8 |

б)

| 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| $K_1$ | (1, 2), (3, 4), (5) | 9 | 13 | 4 | 8 |
| $K_1$ | (2, 1), (5,3), (4) | 10 | 12 | 3 | 9 |
| $K_1$ | (4, 5), (1,2), (3) | 11 | 11 | 2 | 10 |
| $K_2$ | (3, 1), (2, 4), (5) | 8 | 14 | 7 | 5 |
| $K_2$ | (2, 5), (1,3), (4) | 9 | 13 | 6 | 6 |
| $K_2$ | (5,4), (3,1), (2) | 12 | 10 | 3 | 9 |
| $K_3$ | (3, 2), (1, 4), (5) | 8 | 14 | 8 | 4 |
| $K_3$ | (1,4), (5,2), (3) | 10 | 12 | 7 | 5 |
| $K_3$ | (5,3), (4,1), (2) | 12 | 10 | 5 | 7 |
| $K_4$ | (5, 1), (2,3), (4) | 9 | 13 | 9 | 3 |
| $K_4$ | (1, 5), (4,2), (3) | 10 | 12 | 9 | 3 |
| $K_4$ | (3,4), (5,1), (2) | 12 | 10 | 7 | 5 |
| $K_5$ | (2,3), (4,5), (1) | 15 | 7 | 6 | 6 |
| $K_6$ | (4,2), (3,5), (1) | 15 | 7 | 7 | 5 |
| $K_7$ | (4,3), (2,5), (1) | 15 | 7 | 7 | 5 |

## 5. Висновки

Для розв'язання задачі кластеризації великих розмірностей, як правило, використовуються методи випадкового пошуку (метод Монте-Карло) або наближені алгоритми послідовного типу, в яких використано, наприклад, метод "найближчого сусіда". Пошук оптимального розв'язку цими методами може проводитися не по всіх розглянутих підмножинах, а лише в одній із них. Тому для задач, для яких максимальне і мінімальне значення цільової функції знаходиться в різних підмножинах (теореми 3,4) з їх допомогою можна знайти розбиття, для якого функція цілі набуває не найменшого значення, а найбільшого і навпаки. При розробці алгоритмів на основі цих методів необхідно проводити пошук оптимального розв'язку по всіх підмножинах. В цьому випадку складність задачі дорівнює $n^2$.

## Література

1. Тимофеева Н.К. О некоторых свойствах разбиений множества на подмножества// УСиМ. – 2002. – N 5.– С. 6–23.