УДК 004.9:519.23

# ДО ПИТАННЯ МІЖЦЕНТРОВИХ ВІДХИЛЕНЬ В ЗАДАЧАХ БІКЛАСТЕРИЗАЦІЇ СКЛАДНИХ ОБ'ЄКТІВ

В.В. Осипенко, доктор технічних наук, доцент
*Київський національний університет технологій та дизайну*

Ключові слова: кластер-аналіз, бікластеризація, критерій, ознака, об'єкт.

У традиційних постановках задач кластерного аналізу "без учителя" мають місце відомі некоректності, для подолання яких необхідно застосовувати певну апріорну інформацію разом із евристичними припущеннями стосовно наявної вибірки об'єктів. Звідси виникає проблема конструювання спеціальних критеріїв оптимальності при синтезі кластеризацій та вибору підпростору інформативних ознак.

З метою регуляризації некоректностей «класичного» кластер-аналізу, проблему розглядатимемо в межах постановки задачі кластеризації у широкому сенсі або бікластеризації [1], яка почасти зустрічається в економіці, екології, енергетиці, медицині, біології і у багатьох інших прикладних напрямках.

## Постановка задачі бікластеризації

Нехай задано загальний масив вхідних даних в такому виді:

$$\tilde{X} = \left( x_{0j} : x_{ij} \in X \right),\ j = \overline{1,m}, i = \overline{1,n}\ , \qquad (1)$$

де $x_{0j} \in (x_{01}, x_{02}, ..., x_{0m})$ – вектор цільових ознак, $X$ – матриця вхідних ознак. Тобто, кожен об'єкт $\omega_j \in \Omega$ описується як $\omega_j = \left( x_{0j} : x_{ij} \in X \right), i = \overline{1,n}$ .

Необхідно:

1) синтезувати підмножину $\left\{ x_\eta^* \right\} = X^* \subset X$ , $\eta = 1, ..., n^*, n^* \le n$ із наявних ознак, найкращу за заданим критерієм оптимальності та яка дозволила б:
2) класифікувати всі об'єкти з $\Omega$ на $k < m$, $k = 1, ..., K$ однорідних груп.

## Індуктивний кластер-аналіз з вбудованим критерієм найменших міжцентрових відхилень

Відомо, що серед основних характеристик *k*-го кластера (для зручності будемо розглядати евклідів простір $\mathbb{R}^N$) є його центр маси в просторі ознак *X*:

$$\bar{m}_k(X) = \left\{ \frac{1}{r_k} \sum_{l=1}^{r_k} x_{li},\ i = 1,...,n \right\},\ x_i \in X \quad , \tag{2}$$

Використаємо цільову ознаку, як регуляризуючий елемент і обчислимо центр *k*-го кластера лише за значеннями $x_0$ об'єктів $\omega_j^k$ в *k*-му кластері, тобто уже в евклідовому просторі $\square^1$.

Вираз (2) набуде вигляду:

$$\bar{m}_k(x_0) = m_k = \frac{1}{r_k} \sum_{l=1}^{r_k} x_{0l} \,. \tag{3}$$

Як і в індуктивному моделюванні [2] вхідна множина $\omega_j \in \Omega,\ j = \overline{1,m}$ тут ділиться на дві підмножини $\Omega^A$ і $\Omega^B$ такі, що: $\Omega^A \cup \Omega^B = \Omega,\ \Omega^A \cap \Omega^B = \varnothing$.

Нехай на підмножинах $\Omega^A$ і $\Omega^B$ по одній з процедур досягнуто кластеризації [3] $s_t^A \in S^A$ і $s_t^B \in S^B$ з однаковими кількостями кластерів $k_t{}^A = k_t{}^B = K_t$ ($t$ – номер кластеризації, що відповідає деякому підпростору ознак $X_t \subset X$ ) в евклідовому підпросторі ознак $X_t \subset X$ і нехай для всіх $K_t$ кластерів із $s_t^A$ і $s_t^B$ обчислені їх центри $\dot{m}_k^A$ і $\dot{m}_k^B$, $k = 1,...,K_t$, по осі $x_0$.

Тоді критерій оптимальності регуляризованої бікластеризації можна записати в загальному нормалізованому вигляді як:

$$\rho^2(\dot{m}) = \sum_{k=1}^{K} (\dot{m}_k^A - \dot{m}_k^B)^2 \ \Big/ \ \sum_{k=1}^{K} (\dot{m}_k^A + \dot{m}_k^B)^2 \to \min \,. \tag{4}$$

Критерій (4) в індуктивному кластер-аналізі нами було названо критерієм найменших міжцентрових відхилень. При цьому міжцентрові відхилення відносяться до різних підгруп і, звичайно, до різних кластерних груп, або підмножин. Проте, і це важливо підкреслити, що обидві підгрупи належать одному й тому ж експерименту, тобто вважається і це очевидно, що об'єкти відібрано із статистично однаковими характеристиками.

Список використаних джерел

1. V.V. Osypenko, "Two approaches to solving the problem clusterization in the broad sense from the position of inductive modeling", Bulletin NUBiP of Ukraine. Ser. Energy and Automation , no.1. pp. 83-97, 2014.

2. A.G. Ivakhnenko, "Polynomial theory of complex system", IEEE Transaction on Systems, Man and Cybernetics, vol. SMC-1, no. 4, pp. 364–378, 1971.

3. Brian S. Everitt, Sabine Landau, Morven Leese, Daniel Stahl. "Cluster Analysis", 5th Edition, John Wiley and sons. Ltd., 352 p., 2012.