УДК 519.816:519.237.8

О.М. Мацуга, Т.Г. Ємел'яненко

Дніпропетровський національний університет ім. О. Гончара

ПІДТРИМКА ПРИЙНЯТТЯ РІШЕНЬ ПІД ЧАС КЛАСТЕРНОГО АНАЛІЗУ МЕДИЧНИХ ДАНИХ

Запропоновано використовувати методи теорії підтримки прийняття рішень під час вибору методів кластеризації при проведенні аналізу медичних даних.

Вступ. Розв'язання задачі кластерного аналізу за медичними показниками може бути виконано набором методів. Різні методи кластерного аналізу розбивають множину пацієнтів на кластери. Виникає задача підтримки прийняття рішень на етапі формування рекомендацій дослідникові з метою вибору того чи іншого методу для проведення кластерного аналізу.

Аналіз останніх досліджень і публікацій. Методи кластерного аналізу висвітлено у роботах С.А. Айвазяна [1], И.Д. Манделя [2], Н.Г. Загоруйко [3], М. Жамбю [4] та інших. Як найбільш уживані слід виділити методи агломеративно-ієрархічні, *K*-середніх та Forel.

Ідея агломеративно-ієрархічних методів [1–4] полягає в такому. Спочатку кожен об'єкт розглядається як окремий кластер. Далі знаходять два найбільш близько розташовані кластери S_i та S_j і об'єднують їх в один S_{i+j}. Процес об'єднання продовжується доки всі об'єкти не утворять один кластер. Під час об'єднання виникає потреба в обчисленні відстані від нового кластеру S_{i+j} до всіх інших. Вона може бути підрахована за формулою Ланса-Уільямса:

$$d\left(S_{i+j}, S_m\right) = \alpha_i d\left(S_i, S_m\right) + \alpha_j d\left(S_j, S_m\right) + \\ +\beta d\left(S_i, S_j\right) + \gamma \left| d\left(S_i, S_m\right) - d\left(S_j, S_m\right) \right|,$$

де $d(\square,\square)$ – функція відстані; S_m ($m \neq i, j$) – кластер, відстань до якого обчислюється; α_i, α_j, β, γ – числові параметри.

Сполучення параметрів α_i, α_j, β, γ відповідають різним способам обчислення відстані між кластерами та породжують інші види агломеративно-ієрархічних методів:

1) найближчого сусіда (або одного зв'язку):

$$\alpha_i = 0,5\ ;\ \alpha_j = 0,5\ ;\ \beta = 0\ ;\ \gamma = -0,5\ ;$$

2) найвіддаленішого сусіда (або повного зв'язку):

$$\alpha_i = 0,5\ ;\ \alpha_j = 0,5\ ;\ \beta = 0\ ;\ \gamma = 0,5\ ;$$

3) середнього зв'язку:

$$\alpha_i = \frac{n_i}{n_i + n_j}\ ;\ \alpha_j = \frac{n_j}{n_i + n_j}\ ;\ \beta = 0\ ;\ \gamma = 0\ ,$$

де n_i , n_j – кількість об'єктів у кластерах S_i та S_j , які об'єднують;

4) медіанного зв'язку:

$$\alpha_i = 0,5\ ;\ \alpha_j = 0,5\ ;\ \beta = 0\ ;\ \gamma = 0\ ;$$

5) центроїдний:

$$\alpha_i = \frac{n_i}{n_i + n_j}\ ;\ \alpha_j = \frac{n_j}{n_i + n_j}\ ;\ \beta = -\frac{n_i n_j}{n_i + n_j}\ ;\ \gamma = 0\ ;$$

6) Варда:

$$\alpha_i = \frac{n_m + n_i}{n_m + n_i + n_j}\ ;\quad \alpha_j = \frac{n_m + n_j}{n_m + n_i + n_j}\ ;\quad \beta = -\frac{n_m}{n_m + n_i + n_j}\ ;\quad \gamma = 0\ ,$$

де n_m – кількість об'єктів у кластері S_m , відстань до якого обчислюють.

Слід відзначити, що кластеризація, одержана центроїдним методом, має інверсії та не має властивості редуктивності [4], що дозволяє говорити про «погане» розбиття. Тому даний метод у роботі не використовувався.

Метод *K*-середніх існує за двома варіантами: Мак-Кіна та Болла-Холла, що відрізняються між собою порядком стабілізації [1; 2]. У дослідженнях використано варіант Болла-Холла. Випадковим чином обираються *K* центрів кластерів, кожен об'єкт приєднується до кластера, чий центр ближчий, після чого центри кластерів перераховуються як центри мас. Тоді об'єкти знову перерозподіляються за найближчими кластерами, центри переобчислюються і т. д. Ітераційний процес закінчується, коли центри кластерів стабілізуються.

Алгоритм Forel [3] дозволяє виділити кластери простої сферичної форми. В основі лежить базовий алгоритм. Задаючись фіксованим радіусом *R*, поміщають у випадковому чині обрану точку множини об'єктів – центр сфери обраного радіусу. Знаходять координати центру мас точок, що потрапили всередину сфери, та переносять центр сфери до цієї точки. Коли сфера стабілізується, вважають точки, що потрапили до неї, одним кластером, вилучають їх із множини, і переходять до пошуку

нового кластера за аналогічним алгоритмом. Для одержання заданої кількості кластерів базовий алгоритм реалізується при змінних значеннях радіуса.

Результати застосування кожного методу можна представити у вигляді розбиття $S = \{S_1, S_2, \ldots, S_K\}$ на K кластерів

$$S_l = \left\{ X_1^{(l)}, X_2^{(l)}, \ldots, X_{n_l}^{(l)} \right\} = \begin{pmatrix} x_{11}^{(l)} & x_{12}^{(l)} & \cdots & x_{1q}^{(l)} \\ x_{21}^{(l)} & x_{22}^{(l)} & \cdots & x_{2q}^{(l)} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n_l 1}^{(l)} & x_{n_l 2}^{(l)} & \cdots & x_{1q}^{(l)} \end{pmatrix};$$

n_l – кількість об'єктів у кластері S_l; q – розмірність даних; $X_i^{(l)} = \left(x_{i1}^{(l)}, x_{i2}^{(l)}, \ldots, x_{iq}^{(l)} \right)$ – i-й об'єкт кластера S_l; $x_{ih}^{(l)}$ – значення h-ї ознаки об'єкта $X_i^{(l)}$ ($h = \overline{1,q}$).

Для оцінки якості розбиття S відомо близько 50 функціоналів якості [1, 2]. Найбільш поширені такі:

1. Сума квадратів відстаней до центрів кластерів, яка повинна бути мінімальною:

$$Q_1(S) = \sum_{l=1}^{K} \sum_{i=1}^{n_l} d^2 \left(X_i^{(l)}, \overline{X}_l \right),$$

де $\overline{X}_l = \left(\overline{x}_1^{(l)}, \overline{x}_2^{(l)}, \ldots, \overline{x}_q^{(l)} \right)$ – центр кластера S_l.

2. Сума попарних внутрішньокластерних відстаней, що також має бути мінімальною:

$$Q_2(S) = \sum_{l=1}^{K} \sum_{i=1}^{n_l - 1} \sum_{j=i+1}^{n_l} d \left(X_i^{(l)}, X_j^{(l)} \right).$$

Зручність даного функціонала в тому, що його мінімізація автоматично забезпечує максимізацію суми міжкластерних відстаней.

3. Загальна внутрішньокластерна дисперсія має бути мінімальною:

$$Q_3(S) = \det \left(\sum_{l=1}^{K} n_l V_l \right),$$

де $V = \left\| v_{ij}^{(l)}, i, j = \overline{1,q} \right\|$ – матриця коваріацій кластера S_l, елементи якої обраховуються за формулою

$$v_{ij}^{(l)} = \frac{1}{n_l}\sum_{h=1}^{n_l}\left(x_{hi}^{(l)} - \overline{x}_i^{(l)}\right)\left(x_{hj}^{(l)} - \overline{x}_j^{(l)}\right).$$

4. Відношення функціоналів, що повинно бути мінімальним:

$$Q_4(S) = \frac{Q_4'(S)}{Q_4''(S)},$$

де $Q_4'(S)$ – середня внутрішньокластерна відстань

$$Q_4'(S) = \frac{1}{\sum_{l=1}^{K}\frac{n_l(n_l-1)}{2}}\cdot\sum_{l=1}^{K}\sum_{i=1}^{n_l-1}\sum_{j=i+1}^{n_l} d\left(X_i^{(l)}, X_j^{(l)}\right);$$

$Q_4''(S)$ – середня міжкластерна відстань

$$Q_4''(S) = \frac{1}{\prod_{l=1}^{K} n_l}\cdot\sum_{l=1}^{K-1}\sum_{i=1}^{n_l}\left(\sum_{m=l+1}^{K}\sum_{j=1}^{n_m} d\left(X_i^{(l)}, X_j^{(l)}\right)\right).$$

Оскільки мінімізація $Q_4'(S)$ не гарантує максимізації $Q_4''(S)$, щоб урахувати як внутрішньокластерну, так і міжкластерну відстань застосовується відношення у вигляді $Q_4(S)$.

5. Відношення функціоналів, що має бути мінімальним:

$$Q_5(S) = \frac{Q_5'(S)}{Q_5''(C)},$$

де $Q_5'(S)$ – сума середніх внутрішньокластерних відстаней до центрів

$$Q_5'(S) = \sum_{l=1}^{K}\frac{1}{n_l}\sum_{i=1}^{n_l} d^2\left(X_i^{(l)}, \overline{X}_l\right);$$

$Q_5''(S)$ – сума міжкластерних відстаней:

$$Q_5''(S) = \sum_{l=1}^{K} d^2\left(\overline{X}_l, \overline{X}\right),$$

де $\overline{X}$ – центр мас усієї множини об'єктів.

Наведені методи кластерного аналізу приводять до різних варіантів розбиття заданої множини об'єктів. Більш того, навіть один і той же самий метод може давати різні результати, як наприклад, *K*-середніх та Forel, що залежать від початкової ініціалізації центрів кластерів. Не вирішують даної проблеми і функціонали якості , оскільки до кожного з них закладено різні

поняття кластера та однорідності, що призводить до неоднозначних результатів. Тому виникає задача вибору методу кластерного аналізу на етапі формування рекомендацій досліднику. Для її вирішення запропоновано використовувати методи теорії підтримки прийняття рішень.

Постановка задачі. Задано результати обстеження пацієнтів, хворих на артеріальну гіпертензію (АГ), для яких замірялися такі показники: систолічний артеріальний тиск (АТ), фракція викиду лівого шлуночка (ЛШ), товщина задньої стінки ЛШ та психологічні показники за тестом MMPI – іпохондрія, депресія, істерія. Дані представлено у вигляді матриці дійсних чисел $X = \left(X_1, X_2, \ldots X_n\right) = \left\{x_{ih}, i = \overline{1,n}, h = \overline{1,q}\right\}$, де n – кількість пацієнтів; q – кількість показників, які замірювалися ($q = 5$); $X_i = \left(x_{i1}, \ldots, x_{i5}\right)$ – результати обстеження i-го пацієнта; x_{ih} – значення j-го показника, заміряні в i-го пацієнта. Дані зібрані на базі Кримського республіканського НДІ фізичних методів лікування та медичної кліматології ім. І.М. Сєченова.

Будемо вважати, що відстань між пацієнтами обчислюється за евклідовою метрикою. Необхідно обрати один з методів кластерного аналізу, за яким можна розбити хворих на кластери.

Задачу вибору методу кластерного аналізу сформулюємо як задачу підтримки прийняття рішень під час оцінювання альтернатив за множиною критеріїв. Альтернативи $A_j \in A,\ j = \overline{1, C_a}$, де C_a – їхня кількість, оцінюються за множиною критеріїв $K = \left\{k_p, p = \overline{1, C_k}\right\}$, де C_k – кількість критеріїв, у припущенні, що для кожної альтернативи $A_j \in A$ задано кортеж оцінок $r_j = \left\{r_{jp}; p = \overline{1, C_k}\right\}$ даної альтернативи за всіма критеріями, тобто задана матриця $r = \left\{r_{jp}; j = \overline{1, C_a}, p = \overline{1, C_k}\right\}$. Необхідно знайти кількісні оцінки відносної корисності, тобто сформувати множину оцінок альтернатив $\left\{o_j, j = \overline{1, C_a}\right\}$. Ними в даному випадку, виступають методи кластерного аналізу, а критеріями – функціонали якості методів.

Основний матеріал. Для розв'язання поставленої задачі пропонується обчислювальна схема, яка складається з таких етапів:

1. Стандартизація вихідних даних за умови, що показники вимірюються в різних одиницях:

$$\tilde{x}_{ih} = \frac{x_{ih} - \overline{x}_h}{\hat{\sigma}_h},\ i = \overline{1,n},\ h = \overline{1,q}$$

де $\overline{x}_h = \frac{1}{n}\sum_{y=1}^{n} x_{yh}$; $\hat{\sigma}_h = \frac{1}{n}\sum_{y=1}^{n}\left(x_{yh} - \overline{x}_h\right)^2$.

2. Проведення кластерного аналізу за вихідними (або стандартизованими) даними сьома методами. Як результат – одержання семи варіантів розбиття множини пацієнтів: $S^{(1)}$ – ієрархічним методом найближчого сусіда, $S^{(2)}$ – найвіддаленішого сусіда, $S^{(3)}$ – середнього зв'язку, $S^{(4)}$ – медіанного зв'язку, $S^{(5)}$ – Варда, $S^{(6)}$ – методом *K*-середніх та $S^{(7)}$ – алгоритмом Forel.

3. Обчислення значень п'яти функціоналів якості, наведених вище, для кожного з варіантів розбиття, з метою їхнього використання в якості експертних оцінок $r_{jp} = Q_p\left(S^{(j)}\right),\ j = \overline{1,7},\ p = \overline{1,5}$.

4. Вибір методу кластерного аналізу на основі оцінки альтернатив та узгодженості експертів. Альтернативи можуть оцінюватися за рекурентною процедурою, що наведена в [5]

$$x_j^t = \sum_{p=1}^{C_k} r_{jp} k_p^{t-1},\ j = \overline{1,C_a},$$

$$\lambda^t = \sum_{j=1}^{C_a}\sum_{p=1}^{C_k} r_{jp} x_j^t,\ t = 1,2,.,..$$

$$k_p^t = \frac{1}{\lambda^t}\sum_{j=1}^{C_a} r_{jp} x_j^t,\ p = \overline{1,C_k},$$

$$\sum_{p=1}^{C_k} k_p^t = 1,$$

k_p^t – оцінка адекватності альтернативи на t -му кроці.

Тоді, $x^t = \frac{1}{\lambda^{t-1}} B x^{t-1}$, $k^t = \frac{1}{\lambda^t} U k^{t-1}$, $t=1,2,\ldots,$

де B – матриця розмірності $C_a \times C_a$, U – матриця розмірності $C_k \times C_k$, $B = rr'$, $U = r'r$, $r = \left\{r_{jp}; j = \overline{1,C_a}, p = \overline{1,C_k}\right\}$.

Збіжність обчислювального процесу випливає зі збіжності до власних

векторів матриць B та U, тобто, до максимальних власних чисел цих матриць. Матриці B і U повинні бути додатньо визначеними. Додатня визначеність матриць є наслідком додатності $r_{jp}, j = \overline{1, C_a}, p = \overline{1, C_k}$.

Оцінка узгодженості [5] виконується за допомогою дисперсійного коефіцієнта конкордації. Для цього матрицю $r = \left\{ r_{jp}; j = \overline{1, C_a}, p = \overline{1, C_k} \right\}$ переформовують у матрицю рангів $\left\{ R_{jp} \right\}$, на підставі якої обчислюють характеристики:

$$g_j = \sum_{p=1}^{C_k} R_{jp}, j = \overline{1, C_a}; \qquad D = \frac{1}{C_a - 1} \sum_{j=1}^{C_a} (g_j - \overline{g}),$$

де $\overline{g} = \frac{1}{C_a} \sum_{j=1}^{C_a} g_j$. Тоді $W = \frac{D}{D_{\max}}$,

де W – дисперсійний коефіцієнт конкордації $0 < W < 1$. Значущість W, а отже, перевірка гіпотези H_0 : $W = 0$ при $C_a > 6$ відбувається на підставі статистичної характеристики для незв'язаних рангів $\chi^2 = WC_k(C_a - 1)$, або для зв'язаних рангів

$$\chi^2 = \frac{12 \sum_{j=1}^{C_a} \left(g_j - \overline{g} \right)^2}{C_k C_a (C_a + 1) - \frac{1}{C_a - 1} \sum_{p=1}^{C_k} T_p},$$

де $T_p = \sum_{i=1}^{H_p} \left(h_i^3 - h_i \right)$, H_p – кількість груп рівних рангів у p-му ранжируванні, h_i – кількість рівних рангів в i-й групі зв'язаних рангів при ранжируванні за p. Обидві статистичні характеристики мають χ^2 – розподіл з $\nu = C_a - 1$ кількістю степенів вільності.

Запропоновану обчислювальну схема застосовано до даних обстеження хворих на АГ, для яких під час медичного та психологічного обстеження замірялись п'ять показників: систолічний АТ, фракція викиду ЛШ, товщина задньої стінки ЛШ, іпохондрія, депресія, істерія. Одиниці виміру всіх показників різні, тому попередньо була здійснена стандартизація.

Під час проведення кластерного аналізу сімома методами задано кількість кластерів $K = 6$. Матрицю оцінок методів за функціоналами якості наведено в таблиці 1.

Таблиця 1

Оцінки якості методів кластерного аналізу

	Функціонали якості				
Метод	Q_1	Q_2	Q_3	Q_4	Q_5
ієрархічний найближчого сусіда	1,000	1,000	1,000	0,782	0,087
ієрархічний найдальшого сусіда	0,586	0,161	0,175	0,847	0,926
ієрархічний простого середнього	0,724	0,244	0,179	1,000	0,797
ієрархічний середнього зв'язку	0,935	0,807	0,561	0,917	0,433
ієрархічний Варда	0,505	0,094	0,077	0,804	1,000
K-середніх	0,485	0,091	0,071	0,807	0,795
Forel	0,845	0,644	0,496	0,847	0,706

Оцінка якості методів кластерного аналізу за рекурентною процедурою дала таке ранжирування методів: *K*-середніх (оцінка альтернативи 0,109, чим нижче вага, тим краще результат кластеризації), ієрархічний Варда (0,118), ієрархічний найдальшого сусіда (0,128), ієрархічний простого середнього (0,144), Forel (0,162), ієрархічний середнього зв'язку (0,169), ієрархічний найближчого сусіда (0,170). Оцінка узгодженості методів кластеризації дала такі результати: коефіцієнт конкордації дорівнює 0,420, він є значущим з рівнем помилки першого роду $\alpha = 0{,}1$.

Отже, як найкраще розбиття можна рекомендувати результати методу *K*-середніх. Середні величини показників, за якими проведена кластеризація, (табл. 2) дали змогу охарактеризувати клініко-функціональні та психологічні особливості хворих кожного кластера.

Кластери суттєво різняться за показниками систолічного АТ, іпохондрії, депресії та істерії. Перший кластер утворили пацієнти, в яких зазначені показники знаходяться в допустимих межах. У пацієнтів другого та третього кластерів спостерігається м'який ступінь підвищення систолічного АТ при АГ, а для третього – ще й підвищення за шкалою іпохондрії. У пацієнтів четвертого кластера додатково значно перевищують норму значення всіх психологічних показників. До п'ятого кластера ввійшли хворі з помірним

ступенем підвищення систолічного АТ, але допустимими значеннями психологічних показників. Хворі шостого кластера відрізняються від них тяжким ступенем підвищення систолічного АТ.

Таблиця 2

Середні величини показників

Показник\Кластер	1	2	3	4	5	6
САТ	139	146	147	141	164	196
Фракція викиду ЛШ	64	62	62	64	62	62
Товщина задньої стінки ЛШ	1,2	1,2	1,3	1,3	1,4	1,3
Іпохондрія	65	54	70	93	55	69
Депресія	54	47	62	78	46	59
Істерія	64	48	64	75	52	62

Одержане розбиття пацієнтів не лише за клініко-функціональними показниками, але й з урахуванням психологічних особливостей, може забезпечити диференційований підхід до внутрішньонозологічної діагностики АГ.

Висновки. Таким чином, запропоновано обчислювальну схему кластерного аналізу медичних даних, яка забезпечує підтримку прийняття рішень на етапі формування рекомендацій досліднику з метою вибору того чи іншого методу. Запропоновану схему апробовано на даних клініко-інструментального та психологічного обстеження хворих на артеріальну гіпертензію.

Бібліографічні посилання

1. **Айвазян С.А.** Классификация многомерных наблюдений / С.А. Айвазян, З.И. Бежаева, О.В. Староверов. – М., 1974. – 240 с.
2. **Мандель И.Д.** Кластерный анализ / И.Д. Мандель. – М., 1988. – 176 с.
3. **Загоруйко Н.Г.** Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск, 1999. – 270 с.
4. **Жамбю М.** Иерархический кластер-анализ и соответствия / М. Жамбю. – М., 1988. – 279 с.
5. **Емельяненко Т.Г.** Принятие решений в системах мониторинга / Т.Г. Емельяненко, А.В. Зберовский, А.Ф. Приставка, Б.Е. Собко. – Д., 2005. – 224 с.

Надійшла до редколегії 03.06.08