УДК 519.237.8:519.254:004.9

О. Г. Байбуз, М. Г. Сидорова, О. В. Лапец

Дніпропетровський національний університет імені Олеся Гончара

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ КЛАСТЕРНОГО АНАЛІЗУ РЕЗУЛЬТАТІВ ПСИХОЛОГІЧНОГО ТЕСТУ В УМОВАХ НЕВИЗНАЧЕНОСТІ

Запропоновано інформаційну технологію кластерного аналізу результатів міні-мульт тесту для визначення психологічних особливостей хворих на артеріальну гіпертензію. Технологія забезпечує підтримку прийняття рішень дослідником щодо вибору найякіснішого розв'язку в умовах невизначеності.

Ключові слова: *кластерний аналіз, інформаційна технологія, оцінка якості, міні-мульт тест.*

Предложена информационная технология кластерного анализа результатов мини-мульт теста для определения психологических особенностей больных артериальной гипертензией. Технология обеспечивает поддержку принятия решений исследователем по выбору качественного решения в условиях неопределенности.

Ключевые слова: *кластерный анализ, информационная технология, оценка качества, мини-мульт тест.*

The information technology of cluster analysis of mini-mult test results to determine the psychological characteristics of patients with arterial hypertension has proposed. Technology provides support for decision-making by researcher at the choice of quality solutions in the face of uncertainty.

Keywords: *cluster analysis, information technology, quality evaluation, mini-mult test.*

Вступ. Застосування кластерного аналізу є корисним у різноманітних предметних галузях, у тому числі і в медицині, і в психології, оскільки методи кластеризації дають змогу поділити сукупність об'єктів на однорідні за певним формальним критерієм подібності групи (кластери). Основною властивістю цих груп є те, що об'єкти, які належать одному кластеру, подібніші між собою, ніж об'єкти з різних кластерів. Таку класифікацію можна виконувати одночасно за досить великою кількістю ознак. Можна сформулювати такі цілі кластеризації:

- Розуміння даних шляхом виявлення кластерної структури.

Розбиття вибірки на групи схожих об’єктів дозволяє застосовувати до кожного кластера свій метод аналізу при подальшій обробці даних і прийнятті рішень.

- Стиснення даних. Якщо початкова вибірка надмірно велика, то можна скоротити її, залишивши по одному найбільш типовому представнику від кожного кластера.
- Виявлення новизни. Виділяються нетипові об’єкти, які не вдається приєднати ні до одного з кластерів.
- Формулювання та перевірка гіпотез на основі дослідження даних.

Аналіз літературних даних і постановка проблеми. Задачам кластерного аналізу приділено багато уваги [1–4]. Існують різні підходи і напрями досліджень, розроблено безліч методів та алгоритмів, багато дослідників розглядали дану тематику у своїх наукових роботах. Проте і досі існують питання, які не знайшли свого повного розв’язку (оцінка якості результатів, вибір методу та оптимальної кількості кластерів, визначення подібності об’єктів тощо). Якщо немає жодної експертної інформації щодо отримуваних кластерів та їх кількості, то досліднику необхідно приймати низку рішень в умовах невизначеності. Оскільки випадковий необґрунтований вибір методу та параметрів може призвести до того, що отримане розбиття буде зовсім відмінним від природної, притаманної досліджуваним даним, кластерної структури, актуальною задачею є розробка інформаційних технологій, що дозволять оцінювати якість отримуваних угруповань залежно від різних налаштувань параметрів та зменшити ризик вибору невідповідного розв’язку.

Постановка задачі. В роботі поставлено за мету провести кластерний аналіз даних, які є результатом дослідження когнітивних функцій хворих на артеріальну гіпертензію II і III стадії (з давністю перенесеного інсульту не менше 6 місяців). Було обстежено 46 пацієнтів, серед яких 24 чоловіки й 22 жінки, середній вік яких склав 46,7 року, тривалість артеріальної гіпертензії – 4,6 року. Для визначення психологічних особливостей хворих було використано міні-мульт тест, який являє собою скорочений варіант MMPI тесту та має 11 шкал (з них 3 – оцінні: відвертості, вірогідності, корекції, та 8 базисних: іпохондрії, депресії, істерії, психопатії, паранойяльності, психастенії, шизоїдності, гіпоманії), дійсні числа.

Необхідно розробити інформаційну технологію, що дозволить зрозуміти, яка кластерна структура притаманна досліджуваним даним,

проаналізувати отримані кластери, виявити схожість властивостей об'єктів аналізу, візуалізувати отримані результати для подальшої їх інтерпретації.

Основний матеріал. Для проведення кластерного аналізу результатів психологічного тесту при відсутності будь-якої допоміжної експертної інформації щодо отримуваних кластерів було розроблено інформаційну технологію, що складається з п'яти основних етапів.

1. *Попередня обробка даних*. Для зведення досліджуваних ознак до єдиного масштабу пропонується стандартизація, що обчислюється за формулою: $x'_{ij} = \frac{x_{ij} - \bar{x}_j}{\hat{\sigma}_j}, i = \overline{1,n}, j = \overline{1,p}$, де $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, $\hat{\sigma}_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$. Після такого перетворення усі ознаки будуть мати однаковий вплив на результат кластеризації.

Важливим є проаналізувати інформативність досліджуваних ознак, з цією метою реалізовано метод апроксимації матриці відстаней та метод «гойдалки». Було виявлено, що усі 11 ознак є інформативними та мають бути розглянуті у подальшому аналізі.

2. *Визначення наявності кластерної структури.* Пропонується застосування критерію Дуда – Харта:

$$F_{DH} = w_2 / w_1,$$

де w_2 – сума квадратів внутрішньокластерних відстаней у випадку, коли дані розподілені на 2 кластери, w_1 – сума квадратів внутрішньокластерних відстаней у випадку одного кластеру.

Гіпотеза існування єдиного кластеру однорідних даних відхиляється, якщо значення критерію менше критичного значення, обчисленого за формулою:

$$F_{кр} = 1 - \frac{2}{\pi p} - u_{(1-\alpha)}\sqrt{2(1 - 8/(\pi^2 p))/(np)},$$

де $u_{(1-\alpha)}$ – квантиль стандартного нормального розподілу рівня $(1-\alpha)$.

Отримане значення критерію свідчить про те, що досліджувані дані не є однорідними та мають кластерну структуру.

3. *Застосування методів кластерного аналізу*. У загальному вигляді задача кластерного аналізу полягає у визначенні розбиття $G = \{g_1, g_2, \ldots, g_K\}$ на кластери множини $X = \{x_{ij}\}, i = \overline{1,N}, j = \overline{1,p}$, де

N – кількість об'єктів аналізу, p – кількість досліджуваних ознак, x_{ij} – значення j-ї ознаки і-го об'єкта, $g_i = \{x_l\}$, $i = \overline{1,K}$, $x_l = \{x_{lj}\}$, $l = \overline{1,N_i}$, $j = \overline{1,p}$, N_i – кількість об'єктів у і-му кластері, $\sum_{i=1}^{K} N_i = N$, $\bigcup_{i=1}^{K} g_i = X$, $g_i \bigcap g_j = \varnothing, i \neq j$.

Розглянемо декілька алгоритмів кластерного аналізу з тих, що склали ядро розробленого авторами програмного забезпечення.

Алгоритм агломеративної ієрархічної кластеризації.

1. Кожен об'єкт $x_i, i = \overline{1,N}$ вважаємо окремим кластером g_i, $N_i = 1$. Обираємо метрику та обчислюємо матрицю $D = \{d_{ij}\}, i, j = \overline{1,N}$, де d_{ij} – відстань між x_i та x_j.

2. У матриці відстаней D знаходимо мінімальний елемент d_{ij} і кластери g_i та g_j об'єднуємо $g_{i+j} = g_i \cup g_j$, $N_{i+j} = N_i + N_j$.

3. З матриці D вилучаємо відстані від g_i та g_j до інших кластерів та додаємо відстані, що відповідають новому кластеру g_{i+j}.

Для обчислення відстані між кластерами використовується загальна формула Ланса – Уільямса:

$$d(g_{i+j}, g_h) = \alpha_1 d(g_i, g_h) + \alpha_2 d(g_j, g_h) + \beta d(g_i, g_j) + \gamma \left| d(g_i, g_h) - d(g_j, g_h) \right|$$

Задаючи відповідні значення параметрів α_l, α_m, β, γ, отримаємо різні види агломеративних ієрархічних методів [4].

Повторюємо кроки 2–3, доки не отримаємо необхідну кількість кластерів або усі об'єкти не будуть об'єднані в один кластер для побудови дендрограми (рис. 2). На основі візуального аналізу дендрограми можна зробити висновок, що кластерна структура досліджуваних даних складається з двох груп. Це підтверджує і критерій різниці між рівнями дендрограми (рис. 1).

Рисунок 1 – Дослідження кількості кластерів на основі дендрограми

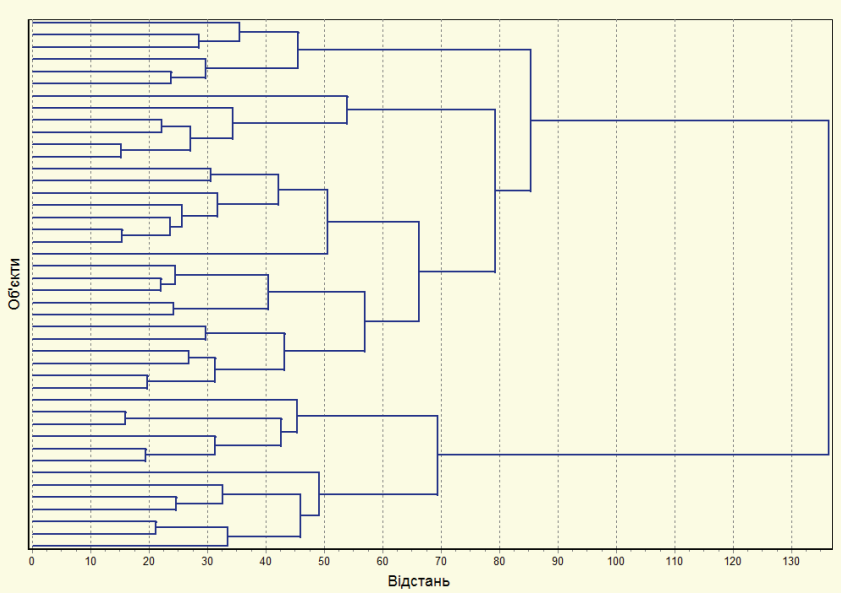

Рисунок 2 – Результат ієрархічної кластеризації у вигляді дендрограми

Методи нечіткої кластеризації [5–6] дозволяють одному і тому самому об'єкту належати одночасно кільком (або навіть усім) кластерам, але з різним ступенем приналежності. Таким чином, можна проаналізувати ступінь відокремленості кластерів один від одного та виявити їх перетини.

Алгоритм Уіндхема. Знаходитмо розв’язок оптимізаційної задачі

$$Q(P,V)=\sum_{l=1}^{K}\sum_{i=1}^{N}\sum_{j=1}^{N}\mu_{li}^{2}\nu_{lj}^{2}d(x_i,x_j)\to\min$$

у такому вигляді:

$$P^{*}=\arg\min_{P}\left\{\begin{matrix}Q(P,V):P=(M^{1},...,M^{K}),\\ M^{l}=\left[(\mu_{l1,...,}\mu_{lN}),(\nu_{l1,...,}\nu_{lN})\right]\end{matrix}\right\},$$

$$0\le\mu_{li}\le 1,\quad 0\le\nu_{lj}\le 1,\quad \sum_{l=1}^{K}\mu_{li}=1,\quad \sum_{j=1}^{N}\nu_{li}=1,\quad \sum_{i=1}^{N}\mu_{li}>0,\quad \sum_{l=1}^{K}\nu_{lj}>0,$$

$i,j=1,...,N$, $l=1,...,K$.

1. Випадковим чином задаємо матрицю розбиття $P_{K\times N}\in\left[\mu_{li}\right]$, $0\le\mu_{li}\le 1$, $\sum_{i=1}^{N}\mu_{li}>0$ та отримуємо початкове розбиття $P_{(0)}=(M_{(0)}^{1},...,M_{(0)}^{K})$ на K нечітких кластерів.

2. Покладаємо значення i = 1.

3. Покладаємо $P_{(i)}=P_{(0)}$ і обчислюємо матрицю прототипів $V_{(0)}$ таким чином:

$$\nu_{li} = \frac{1 \Big/ \sum_{i=1}^{n} \mu_{li}^2 d(x_i, x_j)}{\sum_{j=1}^{n} (1 \Big/ \sum_{i=1}^{n} \mu_{li}^2 d(x_i, x_j))}, \tag{1.1}$$

$i, j = 1,..., N$, $l = 1,..., K$.

4. Покладаємо $V_{(i+1)} = V_{(i)}$ і обчислюємо матрицю розбиття $P_{(i)}$ таким чином:

$$\mu_{li} = \frac{1 \Big/ \sum_{j=1}^{n} \nu_{lj}^2 d(x_i, x_j)}{\sum_{k=1}^{c} (1 \Big/ \sum_{j=1}^{n} \nu_{kj}^2 d(x_i, x_j))}, \tag{1.2}$$

$i, j = 1,..., N$, $l = 1,..., K$.

5. Обчислюємо матрицю прототипів $V_{(i)}$ на основі $P_{(i)}$ відповідно до співвідношення (1.1).

Якщо $Q(P_i, V_i) - Q(P_{i-1}, V_{i-1}) < \varepsilon$, то $P^* = P_{(i)}$, $V^* = V_{(i)}$, інакше i = i +1 і переходимо на крок 4.

Аналізуючи графік приналежності, який наведено на рис/ 3 (чим ближче до 1 значення функції приналежності об’єкта, тим з більшою вірогідністю можна віднести його відповідному кластеру), можна зробити висновок, що в цілому об’єкти чітко розподіляються на 2 кластера, і лише стосовно 2–3 об’єктів рішення про віднесення їх певному кластеру є менш однозначним.

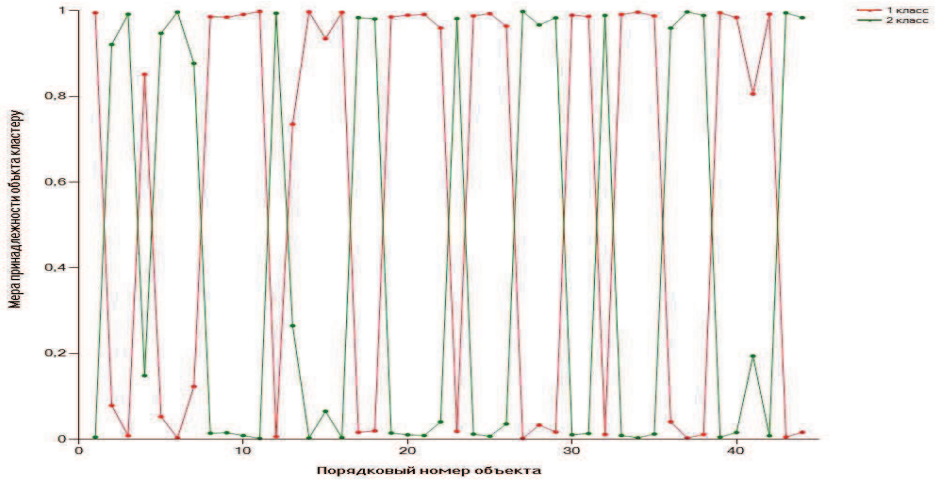

Рисунок 3 – Графік функції приналежності об’єктів дослідження нечітким кластерам

На рис. 4 результати кластеризації представлені у вигляді діаграми розсіювання (об'єкти дослідження зображуються точками у N-вимірному просторі та проектуються на площину). Візуально можна виділити 2 кластери з не досить чіткою межею.

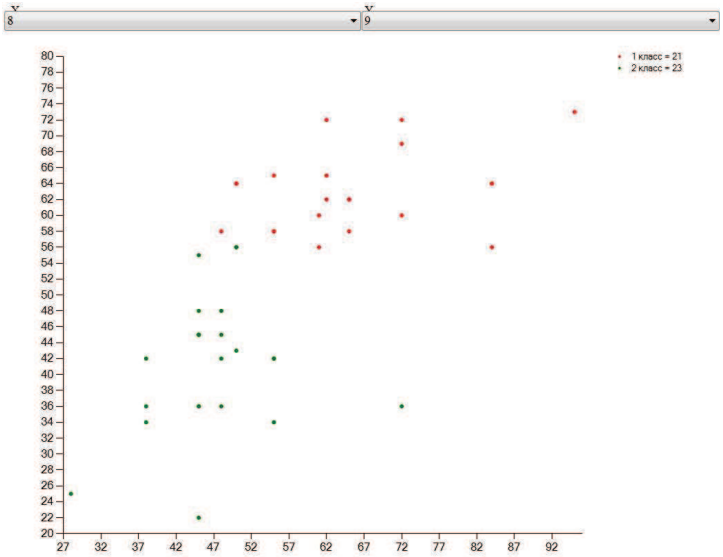

Рисунок 4 – Результат кластеризації у вигляді діаграми розсіювання

4. *Оцінювання якості, оптимальної кількості кластерів та вибір результуючого розв'язку.* На попередньому етапі ми отримали набір розв'язків $G = \{G_1, G_2, \ldots, G_n\}$, $G_i = \{g_1^{(i)}, g_2^{(i)}, \ldots, g_{K_i}^{(i)}\}$, серед яких необхідно виявити такий, що має найвищу якість. Оцінювання здійснюється за допомогою функціоналів якості, таких як сума внутрішньокластерних дисперсій за усіма ознаками, сума внутрішньокластерних відстаней, відношення внутрішньокластерних та міжкластерних відстаней; індексів Данна, Беждека – Данна, Калінського й Гарабача, а також на основі багатокритеріальних оцінок, отриманих із застосуванням колективних методів прийняття рішень [7–8].

Іноді декілька результатів кластеризації можуть представляти різні, але еквівалентні за якістю угруповання. У цьому випадку, замість того, щоб обирати один з розв'язків, пропонується використовувати ансамблевий підхід і отримати колективне результуюче рішення [9–10].

5. *Візуалізація і аналіз результатів.* Для аналізу та інтерпретації результатів розроблена система пропонує обчислення статистичних характеристик кожного кластера, а також широкий спектр засобів візуалізації. Було виявлено, що середні значення усіх ознак одного кластеру менші, ніж другого.

Висновки. У роботі запропоновано інформаційну технологію кластерного аналізу результатів міні-мульт тесту для визначення

психологічних особливостей хворих на артеріальну гіпертензію. Технологія забезпечує підтримку прийняття рішень дослідником щодо вибору найякіснішого розв'язку в умовах невизначеності. Досліджено кластерну структуру даних, виявлено два кластери, проаналізовано нечіткість приналежності об'єктів кожній з груп, а також статистичні характеристики усіх ознак за кожним кластером.

Бібліографічні посилання

1. Berkhin P. A survey of clustering data mining techniques // Grouping Multidimensional Data. 2006. P. 25–71.

2. Миркин Б. Г. Методы кластер-анализа для поддержки принятия решений: обзор. М. 2011. 88 с.

3. Бериков В.С., Лбов Г. С. Современные тенденции в кластерном анализе. Всероссийский конкурсный отбор обзорно-аналитических статей по приоритетному направлению «Информационно-телекоммуникационные системы». 2008. 26 с.

4. Мандель И. Д. Кластерный анализ. М. 1988. 176 с.

5. Вятченин Д. А. Нечеткие методы автоматической классификации: монография. Мн. 2004. 219 с.

6. Oliveira J. V., Pedrycz W. Advances in fuzzy clustering and its applications. Chichester. 2007. 435 p.

7. Pascual D., Pla F., Sanchez J. S. Cluster validation using information stability measures // Pattern Recognition Letters. 2010. Vol. 31. P. 454–461.

8. Приставка О. П., Сидорова М.Г. Підтримка прийняття рішень в задачах кластерного аналізу // Актуальні проблеми автоматизації та інформаційних технологій : зб. наук. праць. Дніпропетровськ. 2011. Т. 15. С.117–125.

9. Сидорова М. Г. Застосування ансамблів алгоритмів для підвищення стійкості результатів кластеризації. // Актуальні проблеми автоматизації та інформаційних технологій : зб. наук. праць. Дніпропетровськ. 2013. Т. 17. С. 22 –29.

10. Pestunov I. A., Berikov V. B., Kulikova E. A. Rylov Ensemble of Clustering Algorithms for Large Datasets. // Optoelectronics, Instrumentation and Data Processing. 2011. Vol. 47. No. 3. P. 245–252.

Надійшла до редколегії 28.11.16.