УДК 519.254

О. П. Приставка, М. Г. Сидорова

Дніпропетровський національний університет імені Олеся Гончара

ПІДТРИМКА ПРИЙНЯТТЯ РІШЕНЬ У ЗАДАЧАХ КЛАСТЕРНОГО АНАЛІЗУ

Запропонована інформаційна технологія підтримки прийняття рішень щодо вибору найкращого розбиття в умовах невизначеності кластерного аналізу.

Ключові слова: *кластерний аналіз, підтримка прийняття рішень, оцінка якості кластеризації, інформаційна технологія.*

Предложена информационная технология поддержки принятия решений при выборе лучшего разбиения в условиях неопределенности кластерного анализа.

Ключевые слова: *кластерный анализ, поддержка принятия решений, оценка качества кластеризации, информационная технология.*

Proposed information technology to decision making support in choosing the best partition in the face of uncertainty of the cluster analysis.

Key words: *cluster analysis, decision making support, assessment of the quality of clustering, information technology.*

Постановка проблеми. Кластеризація – об'єднання в групи схожих об'єктів – є одним з фундаментальних завдань у галузі аналізу даних і Data Mining. Дана задача стала предметом широкого дослідження, оскільки список прикладних областей, де вона застосовується, досить широкий: сегментація зображень, маркетинг, медицина, соціальні науки, аналіз текстів і багато інших.

Алгоритми кластерного аналізу 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 наступні цілі кластеризації:

– Розуміння даних шляхом виявлення кластерної структури. Розбиття вибірки на групи схожих об'єктів дозволяє спростити

подальшу обробку даних і прийняття рішень, застосовуючи до кожного кластера свій метод аналізу (стратегія «розділяй і володарюй»).

– 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.

– Представлення та перевірка гіпотез на основі дослідження даних.

– Може використовуватися в якості попереднього кроку обробки для інших алгоритмів, наприклад, алгоритмів класифікації, які будуть спиратися на виявлені заздалегідь кластери.

Кластеризація може бути знайдена під різними іменами в різних контекстах, таких як самонавчання (в розпізнаванні), чисельна таксономія (в біології, екології), типологія (в області соціальних наук) та розбиття (в теорії графів).

Задачам кластерного аналізу приділено багато уваги. Існують різні підходи і напрямки досліджень, розроблено безліч методів та алгоритмів, багато дослідників присвятили свої наукові роботи даній тематиці. Проте і досі існують питання, які не знайшли свого повного розв’язку .

Одним з найактуальніших питань кластерного аналізу є оцінювання результатів та пошук розбиття, що найкраще відповідає структурі даних. У більшості задач кластеризації дослідники зіштовхуються з проблемою вибору оптимального числа кластерів, що відповідає природі досліджуваних об’єктів.

Для розв’язання таких задач у літературі на даний момент існує велика кількість функціоналів та індексів якості, що дозволяють у кількісному вигляді оцінювати відповідність вихідного розбиття природній структурі даних, а також порівнювати результати отримані різними методами або при різних значеннях параметрів. Визначення функціоналів якості, головним чином, ґрунтується на таких критеріях як компактність та відокремленість кластерів, але все ж таки до кожного з них закладено різні поняття кластера та однорідності, тому вони досить часто демонструють зовсім різні результати, обираючи різні розбиття як найкращі. І перед дослідником знову постає питання, який критерій якості обрати. Тому виникає потреба розроки нового підходу, який би міг враховувати результати різних функціоналів якості одночасно та забезпечувати більш точну оцінку результатів. **Аналіз досліджень і публікацій.** Над методами кластерного аналізу активно працюють багато дослідників. Їх основними завданнями є подолання недоліків методів, що існують на сьогоднішній день, розробка нових алгоритмів, оцінка якості, візуаліза-

ція та інтерпретація результатів, пошук рішень основних проблем у даній галузі.

Опис методів кластерного аналізу можна знайти в багатьох джерелах, наприклад у фундаментальних роботах І. Д. Манделя [1], С. А. Айвазяна [2], М. Жамбю [3], Н. Г. Загоруйко [4]. У [5] висвітлені основні поняття, цілі, задачі, проблеми і труднощі, а також переваги і сильні сторони кластерного аналізу. Подано огляд існуючих напрямів та підходів кластеризації, розглянуто найпопулярніші методи з детальним описом кожного з них. У роботі також представлені приклади застосування технік кластерного аналізу в таких галузях, як сегментування зображень, розпізнавання об‘єктів та символів, пошук документів та інформації, data mining.

Існує три підходи дослідження точності кластеризації. Перший ґрунтується на зовнішніх критеріях, тобто оцінюються результати алгоритму кластеризації на основі заздалегідь визначеної структури, яка накладається на набір даних і відображає наші припущення. Другий підхід заснований на внутрішніх критеріях. Можна оцінювати результати алгоритму кластеризації в термінах величин, які пов'язані з векторами даних (наприклад, матриці близькості). Третій підхід використовує відносні критерії. Основна ідея полягає в оцінці структури кластеризації, порівнюючи її з іншими кластеризаційними схемами, отриманими іншими алгоритмами або з різними значеннями параметрів. У [6; 7] розглянуто кожен з трьох підходів, представлені посилання на літературу, а також розглянуто фундаментальні концепції в даній області.

У [8] наведено порівняльну характеристику та проведено ряд експериментів з метою визначення точності тридцяти індексів якості представлених у літературі. У [9] автори пропонують новий індекс якості кластеризації. Проведено оцінку надійності представленого індексу як теоретично, так і практично, та порівняння з іншими відомими в літературі індексами. Також наведено огляд близьких за тематикою робіт. Близько 50 функціоналів якості наведено в [1]. У [10] представлена система інтелектуального аналізу даних, яка дозволяє застосування алгоритмів кластеризації та проведення оцінки якості отриманих рішень. Методи визначення оптимальної кількості кластерів розглянуто в [11; 12, 13].

Постановка задачі. Розробити інформаційну технологію кластерного аналізу, яка б здійснювала підтримку прийняття рішень при виборі найкращого розбиття. Застосувати даний підхід до кластеризації даних медичного обстеження хворих на серцеву недостатність.

Основний матеріал. Для розв‘язання поставленої задачі пропонується інформаційна технологія, яка складається з наступних етапів:

1. Проводимо попередню обробку даних: відбір інформативних ознак методом «Гойдалки» та стандартизацію [14].

2. Отримуємо розбиття різними методами або при різних значеннях параметрів, та розглядаємо їх в якості альтернатив.

3. Для кожної альтернативи обчислюємо значення будь-якої комбінації наступних функціоналів якості, які вважаємо експертами:

– Сума внутрішньокластерних дисперсій за всіма ознаками:

$$F_1(C) = \sum_{i=1}^{K}\sum_{j=1}^{p}\left(\frac{1}{n_i - 1}\sum_{h=1}^{n_i}\left(x_{hj}^{(i)} - \bar{x}_j^{(i)}\right)^2\right) \to \min.$$

– Сума квадратів відстаней до центрів кластерів: $F_2(C) = \sum_{i=1}^{K}\sum_{X_j \in C_i} d^2(X_j^{(i)}, M_i) \to \min$, де $M_i = (\bar{x}_1^{(i)}, \bar{x}_2^{(i)}, ..., \bar{x}_p^{(i)})$ – центр кластера C_i.

– Відношення середньої внутрішньокластерної і середньої міжкластерної відстаней: $F_3(C) = \dfrac{\widetilde{F}(C)}{\widetilde{\widetilde{F}}(C)} \to \min$, де

$$\widetilde{F}(C) = \frac{1}{\sum_{i=1}^{K}\frac{n_i(n_i-1)}{2}}\sum_{i=1}^{K}\sum_{j=1}^{n_i-1}\sum_{h=j+1}^{n_i} d(X_j^{(i)}, X_h^{(i)}),$$

$$\widetilde{\widetilde{F}}(C) = \frac{1}{\prod_{i=1}^{K} n_i}\sum_{i=1}^{K-1}\sum_{j=1}^{n_i}\left(\sum_{m=i+1}^{K}\sum_{h=1}^{n_m} d(X_j^{(i)}, X_h^{(m)})\right).$$

– Сума внутрішньокластерних відстаней:

$$F_4(C) = \sum_{l=1}^{K}\sum_{i=1}^{n_l-1}\sum_{j=i+1}^{n_l} d(X_i^{(l)}, X_j^{(l)}) \to \min.$$

4. Отримані результати представляємо у вигляді матриці $X = \{x_{ij}; i = \overline{1,n}, j = \overline{1,m}\}$, де n – кількість методів, m – кількість експертів, x_{ij} – оцінка, яку поставив j-й експерт і-й альтернативі.

5. Зводимо виставлені оцінки до єдиного масштабу: $x_{ij} = \dfrac{x_{ij}}{\sum_{i=1}^{n} x_{ij}}$.

6. Застосовуємо один з наступних методів прийняття рішень [14]:

– процедура Борда

Для кожного експерта виконуємо впорядкування альтернатив у порядку спадання їх адекватності. Обчислюємо колективну оцінку якості кожного варіанта як суму рангових місць. Найкращим результатом вважається той, що буде мати найменшу оцінку.

– плюралітарна процедура

Оцінки кожного експерта впорядковуються. Для кожної альтернативи обчислюється колективна оцінка, що дорівнює кількості експертів, які поставили її на перше місце. Найкращою вважається альтернатива з максимальною оцінкою.

– множинний аналіз

Оцінка адекватності альтернатив проводиться за рекурентною процедурою:

1) Задаємо крок $t = 1,$ та $k_j = \frac{1}{m}.$

2) Обчислюємо оцінки альтернатив на t-му кроці $x_i^t = \sum_{j=1}^{m} x_{ij} k_j^{t-1}, i = \overline{1,n}$.

3) Обчислюємо $\lambda^t = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} x_i^t, t = 1,2,...$

4) Збільшуємо $t : t = t + 1$. Обчислюємо значення компетентності експертів на t-му кроці $k_j^t = \frac{1}{\lambda^t} \sum_{i=1}^{n} x_{ij} x_i^t, \sum_{j=1}^{m} k_j^t = 1, j = \overline{1,m}$.

5) Повторюємо пункти 2–4, доки процес не зійдеться з деякою заданою точністю ε. Доведено, що процес є збіжним. Найкращим вважається результат з мінімальною оцінкою. Даний метод дозволяє також оцінити узгодженість експертів на основі дисперсійного коефіцієнта конкордації.

Проілюструємо можливі випадки застосування запропонованої технології на даних медичного обстеження хворих на хронічну серцеву недостатність. Дані отримані за допомогою допплер-ехокардіографії в Українському державному науково-дослідному інституті медико-соціальних проблем інвалідності.

Вибір методу. Проведемо розподілення хворих на 5 груп (кластерів), що відповідають стадіям захворювання, різними ієрархічними методами: найближчого сусіда, найвіддаленішого сусіда, середнього зв‘язку, центрального зв‘язку, Варда. За допомогою запропонованої технології визначимо, який метод краще реалізує дану задачу. Оцінки функціона-

лів якості наведено у таблиці 1, результати методів прийняття рішень – у таблиці 2.

Таблиця 1

Оцінки якості ієрархічних методів кластерного аналізу

Ієрархічні методи	Функціонали якості			
	F_1	F_2	F_3	F_4
найближчого сусіда	0,282	0,239	0,001	0,289
найвіддаленішого сусіда	0,195	0,176	0,080	0,132
середнього зв‘язку	0,151	0,200	0,002	0,240
центрального зв‘язку	0,183	0,232	0,001	0,283
Варда	0,190	0,153	0,920	0,060

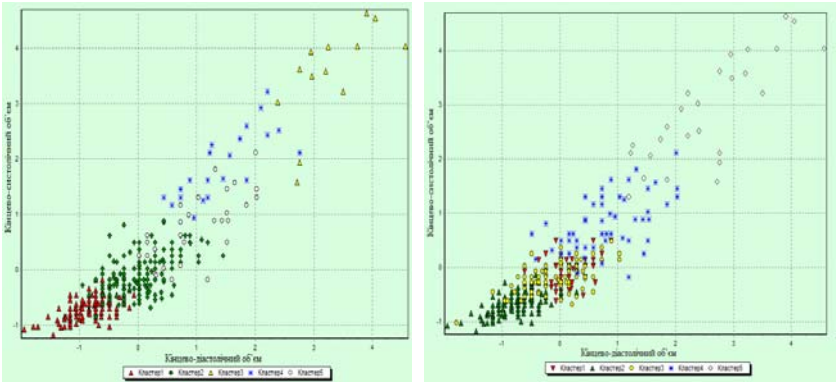
Очевидно, що користуючись лише таблицею значень функціоналів якості, досить складно виділити якийсь метод як найкращий.

Таблиця 2

Результати методів прийняття рішень

Ієрархічні методи	Множинний аналіз	Процедура Борда	Плюралітарна процедура
найближчого сусіда	0,140	16	1
найвіддал. сусіда	0,128	12	0
середнього зв‘язку	0,100	10	1
між центрами	0,119	12	0
Варда	0,513	10	2

За отриманими результатами найкращими слід вважати розбиття, отримані ієрархічними методами середнього зв‘язку та Варда. Також при виборі методу важливу роль відіграє візуальний аналіз (рис 1).



а *б*

Рис. 1. Візуальний аналіз на основі діаграми розбиття: а – метод середнього зв‘язку, б – метод Варда

Визначення параметрів методу. Один і той самий метод кластерного аналізу може демонструвати різні результати залежно від обраних вхідних параметрів. Тому важливо вибрати саме такі параметри, що найкраще підходять для кластеризації відповідних даних. Застосуємо метод К-середніх при різних варіантах вибору початкових центрів. Результати функціоналів якості наведено в таблиці 3.

Таблиця 3

Оцінки якості методу К-середніх при різних варіантах вибору початкових центрів

Вибір початкових центрів	Функціонали якості			
	F_1	F_2	F_3	F_4
перші	0,330	0,329	0,239	0,320
випадкові	0,331	0,332	0,386	0,322
найвіддаленіші	0,339	0,338	0,375	0,358

Усі функціонали якості однозначно віддали перевагу вибору перших точок як початкових центрів.

Визначення оптимальної кількості кластерів. У більшості задач кластерного аналізу немає ніякої початкової інформації щодо кількості кластерів присутній у структурі даних. Тому дана задача є досить актуальною та потребує розробки методів її вирішення.

Для ієрархічних методів кількість кластерів не є вхідним параметром та може бути визначена шляхом аналізу дендрограми (рис.2).

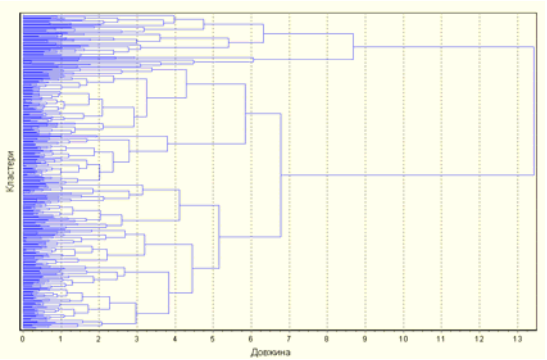


Рис. 2. Результати ієрархічної кластеризації медичних даних у вигляді дендрограми

Для неієрархічних методів кількість кластерів має бути визначена заздалегідь. Застосуємо запропоновану технологію і до цієї задачі. Проведемо кластеризацію медичних даних методом К-середніх у варіанті Мак-Кіна для числа кластерів від 2 до 10. Для кожного розбиття обчислюємо значення функціоналів якості. До отриманої матриці застосовуємо множинний аналіз, результати якого наведено в таблиці 4.

Таблиця 4

Результати множинного аналізу при дослідженні оптимальної кількості кластерів

k	2	3	4	5	6	7	8	9	10
Оцінка	0,050	0,037	0,033	0,031	0,032	0,032	0,033	0,070	0,682

Результати показали, що оптимальне число кластерів для даної структури даних дорівнює 5.

Висновки. Таким чином, запропонована інформаційна технологія підтримки прийняття рішень при виборі найкращого розбиття в умовах невизначеності кластерного аналізу, яка дозволяє враховувати результати різних функціоналів якості одночасно тим самим забезпечує більш точну оцінку результатів. Розроблено програмне забезпечення, до складу якого входить запропонована технологія. Продемонстровано практичну реалізацію на даних медичного обстеження хворих на серцеву недостатність.

Бібліографічні посилання

1. **Мандель И. Д.** Кластерный анализ / И. Д. Мандель. – М., 1988. – 176 с.
2. **Айвазян С. А.** Классификация многомерных наблюдений / С. А. Айвазян, З. И. Бежаева, О. В. Староверов. – М., 1974. – 240 с.

3. **Жамбю М.** Иерархический кластер-анализ и соответствия / М. Жамбю. – М., 1988. – 279 с.
4. **Загоруйко Н. Г.** Прикладные методы анализа данных и знаний / Н. Г. Загоруйко. – Новосибирск, 1999. – 270 с.
5. **Jain A. K.** Data Clustering: A Review//A.K. Jain, M.N. Murty , P.J. Flunn//ACM Computing Surveys, Vol. 31, №. 3, September 1999. – P.265-323.
6. **Halkidi M.** On Clustering Validation Techniques / M. Halkidi, Y. Batistakis, M. Vazirgiannis // Journal of Intelligent Information Systems. – 2001. 17:2/3, – P. 107–145.
7. **Гусарова Л.** Проверка обоснованности кластерного решения / Л. Гусарова, И. Яцкив // Proceedings of International Conference Rel-Stat'03, – Vol. 2.
8. **Milligan G.** An examination of procedures for determining the number of clusters in a data set/ G. Milligan, M. Cooper// Psychometrika – Vol. 50, №. 2, June 1985.– P. 159–179.
9. **Halkidi M.** Clustering validity assessment:Finding the optimal partitioning of a data set / M. Halkidi, M.Vazirgiannis // Proceedings of ICDM, 2001. – P. 187-194.
10. **Bolshakova N.** An Integrated Tool for Microarray Data Clustering and Cluster Validity Assessment/ N. Bolshakova, F. Azuaje, P. Cunningham //Bioinformatics Advance Access published – 2004, December 17, P. 133-137.
11. **Шалымов Д. С.** Алгоритмы устойчивой кластеризации на основе индексных функций и функций устойчивости / Д. С. Шалымов // Стохастическая оптимизация в информатике. – 2008. –Вып. 4. – С. 236-248.
12. **И. Яцкив** Методы определения количества кластеров при классификации без обучения / И. Яцкив, Л. Гусарова // Transport and Telecommunication, – 2003, Vol.4, № 1.
13. **C. Sugar** Finding the number of clusters in a data set: An information theoretic approach / C. Sugar, G. James // Journal of the American Statistical Association 2003; 98(463): 750-763.
14. **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 15.06.11