УДК 519.254

О. П. Приставка, доктор технічних наук, професор
кафедри математичного забезпечення ЕОМ
Дніпропетровського національного університету
ім. Олеся Гончара
М. Г. Сидорова, аспірантка Дніпропетровського
національного університету ім. Олеся Гончара

ІНТЕГРОВАНА ТЕХНОЛОГІЯ КЛАСТЕРНОГО АНАЛІЗУ

Запропоновано інформаційну технологію автоматизованої обробки та аналізу даних, що реалізує обчислювальні схеми кластеризації, класифікації, підтримки прийняття рішень і становить ядро розробленого програмного забезпечення "MedISA".

Предложена информационная технология автоматизированной обработки и анализа данных, которая реализует вычислительные схемы кластеризации, классификации, поддержки принятия решений и составляет ядро разработанного программного обеспечения "MedISA".

Proposed information technology for automated data processing and analysis, which implements the computational schemes of clustering, classification, decision support and is the core of the developed software "MedISA".

Ключові слова. Інформаційна технологія, кластеризація, класифікація, підтримка прийняття рішень.

Вступ. Останнім часом у зв'язку з удосконаленням технологій запису і зберігання даних спостерігається тенденція накопичення великої кількості інформації. Виникає завдання обробки наборів даних значних обсягів з метою виявлення прихованих у них знань, закономірностей, властивостей, тенденцій, кращого розуміння структури. Це призводить до появи такого поняття, як Data Mining, або інтелектуальний аналіз [1–3]. Data mining – це мультидисциплінарна галузь, що виникла і розвивається на базі таких наук, як прикладна статистика, розпізнавання образів, штучний інтелект, теорія баз даних та ін. Головна особливість Data Mining – це поєднання фундаментального математичного апарату і останніх досягнень у сфері інформаційних технологій. Найчастіше до завдань Data Mining зараховують такі: кластеризація, класифікація, прогнозування, виявлення асоціацій і послідовностей, візуалізація.

Кластеризація – об'єднання в групи схожих об'єктів – одне з фундаментальних завдань у галузі аналізу даних і Data Mining. Список прикладних галузей, де вона застосовується, досить широкий: сегментація зображень, маркетинг, медицина, аналіз текстів і багато інших. На сучасному етапі кластеризація часто виступає першим кроком в аналізі даних.

Завдання класифікації полягає в побудові моделі за деяким принципом на основі множини об'єктів навчальної вибірки, які мають подібні класифікаційні ознаки, та зарахування нових спостережень до одного з наперед заданих класів. Задачам кластерного аналізу та класифікації приділено багато уваги [4–12]. Існують різні підходи і напрямки досліджень, розроблено безліч методів та алгоритмів, багато дослідників присвятили свої наукові праці даній тематиці. Проте й досі існують питання, які не знайшли свого повного висвітлення. До таких питань належать оцінка якості отриманих результатів, вибір оптимальної кількості кластерів, а також пошук нових, більш формалізованих методів кластеризації.

Таким чином, швидке зростання обсягу інформації в науці та бізнесі, а також потреба аналізу та визначення корисної інформації на основі величезних наборів даних призводить до необхідності розробки інформаційних систем та програмних засобів, що дозволять виконувати такі завдання.

Постановка завдання. Мета роботи – створити систему інтелектуального аналізу, що реалізує задачі кластеризації, класифікації, візуалізації, обробки та аналізу інформації, забезпечує підтримку прийняття рішень; запропонувати інформаційну технологію та розробити обчислювальні схеми методів, що забезпечать математичну основу і становитимуть ядро даної системи; продемонструвати роботу програми на наборах реальних даних у сфері медицини.

Результати дослідження. Щоб виконати поставлене завдання, запропоновано інформаційну технологію та розроблено систему автоматизованої обробки даних MedISA, яка реалізує цю технологію. Програмне забезпечення створено в середовищі Delphi 7.0, взаємодіє з користувачем у діалоговому режимі. Дає змогу завантажувати текстові файли у форматі *.txt і таблиці баз даних формату *.dbf. Поряд із кластеризацією та класифікацією система дозволяє проводити первинний статистичний та імовірнісний аналізи для кращого розуміння природи досліджуваних даних.

Перетворення досліджуваних даних. Перед проведенням кластерного аналізу слід звернути увагу на такі питання.

1. Досить часто об'єкти в досліджуваних наборах даних характеризуються великою кількістю ознак, причому деякі ознаки не несуть значної інформації, а лише збільшують розмірність даних. Це ускладнює роботу алгоритмів, виникає потреба у великому обсязі пам'яті та значній кількості машинного часу, що робить деякі методи непридатними для таких наборів даних. Для розв'язання цієї проблеми система пропонує два варіанти зменшення розмірності даних:

− користувач може сам обирати ознаки, спираючись на знання предмета, власні припущення та спостереження;

− реалізовано ітераційний метод вибору інформативних ознак "Гойдалки". Даний метод дозволяє впорядкувати не лише ознаки за їх інформативністю, але й об'єкти [13].

2. Якщо ознаки об'єктів вимірюються в різних одиницях або сильно відрізняються за значеннями, доцільно звести їх до єдиного масштабу. Найбільше вживаною з таких процедур є стандартизація. Програма

обчислюють нові значення за формулою: $x'_{ij} = \frac{x_{ij} - \bar{x}_j}{\hat{\sigma}_j}, i = \overline{1,n}, j = \overline{1,p}$, де $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, $\hat{\sigma}_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2}$.

Кластеризація. Розроблено обчислювальні схеми алгоритмів, що ґрунтуються на методах ієрархічної [4–8], швидкої ієрархічної [5, 7] та розділової кластеризації [4,6–8]. Запропоновано три метрики відстаней: евклідова, манхеттенська, Чебишева.

Ієрархічні методи дозволяють отримати наочне уявлення про структуру всієї досліджуваної сукупності об'єктів у вигляді дендрограми і демонструють процес об'єднання об'єктів у кластери. На основі класичного агломеративного методу ієрархічної кластеризації, який широко представлений у літературі [4–8], розроблено такий алгоритм.

Ієрархічний агломеративний метод.

1. Нехай обрано метрику й обчислено матрицю відстаней між об'єктами $D = \{d_{ij}\}, i, j = \overline{1,n}$, де $d_{ij} = d(i,j)$.

2. Кожен об'єкт вважають окремим кластером. У матриці відстаней D знаходять мінімальний елемент d_{ij} і кластери C_i та C_j об'єднують в один кластер $C_{i+j} = C_i \cup C_j$.

3. Змінюють матрицю D шляхом вилучення відстаней від C_l та C_h до інших кластерів і додавання відстаней до кластера C_{i+j}.

4. Далі на кожному кроці процедуру повторюють, тобто знаходять мінімальний елемент у перетвореній матриці відстаней і відповідні кластери об'єднують. Процедуру слід повторювати N–1 раз, доки всі об'єкти не будуть об'єднані.

Для обчислення відстані між кластерами існує загальна формула Ланса–Вільямса: $d(C_{h+m}, S_k) = \alpha_l d(C_l, C_k) + \alpha_m d(C_m, C_k) + \beta d(C_l, C_m) + \gamma |d(C_l, C_k) - d(C_m, C_k)|$. Задаючи різні значення параметрів α_l, α_m, β, γ, отримаємо різні види агломеративних ієрархічних методів:

1. $\alpha_l = \frac{1}{2}; \alpha_m = \frac{1}{2}; \beta = 0; \gamma = -\frac{1}{2}$ – одиничного зв'язку (найближчого сусіда);
2. $\alpha_l = \frac{1}{2}; \alpha_m = \frac{1}{2}; \beta = 0; \gamma = \frac{1}{2}$ – повного зв'язку (найвіддаленішого сусіда);
3. $\alpha_l = \frac{n_l}{n_l + n_m}; \alpha_m = \frac{n_m}{n_l + n_m}; \beta = 0; \gamma = 0$ – середнього зв'язку;
4. $\alpha_l = \frac{n_l}{n_l + n_m}; \alpha_m = \frac{n_m}{n_l + n_m}; \beta = -\frac{n_l n_m}{(n_l + n_m)^2}; \gamma = 0$ – між центрами;
5. $\alpha_l = \frac{n_k + n_l}{n_k + n_l + n_m}; \alpha_k = \frac{n_k + n_m}{n_k + n_l + n_m}; \beta = -\frac{n_k}{n_k + n_l + n_m}; \gamma = 0$ – Варда.

Дослідження алгоритмів ієрархічної класифікації приводять до висновків, що найбільш трудомістка операція – пошук пари найближчих кластерів на кожному кроці. Це обмежує використання алгоритму для вибірок великого обсягу. Тому було розроблено обчислювальні схеми швидких ієрархічних методів. Ідея полягає в тому, щоб перебирати лише найбільш близькі пари. Тобто для економії пам'яті і зменшення необхідного числа порівнянь при пошуку найменшої відстані слід виключити з розгляду ті відстані, які не впливають на виконання обчислень. Результати швидких методів ієрархічної кластеризації повністю збігаються з результатами класичних ієрархічних методів. Відмінність полягає лише у швидкості роботи. На основі методів, описаних у працях [5, 7], запропоновано такі обчислювальні схеми.

Швидкі ієрархічні методи кластеризації.

1. Нехай обрано метрику відстані й обчислено матрицю відстаней $D = \{d_{ij}\}, i, j = \overline{1,n}$, де

Кожен об'єкт вважають окремим кластером.

2. Обирається параметр δ. Залежно від способу обчислення δ існують різні методи швидкої кластеризації.

Метод 1. Як δ обирають середнє значення матриці відстаней.

Метод 2. Задаються параметри n_1 та n_2. Якщо кількість кластерів не перевищує поріг n_1, то δ не обирають, а працюють з матрицею D. Інакше обирається випадковим чином n_2 відстаней, і δ прирівнюється до найменшої з них. Параметри n_1 та n_2 впливають лише на час виконання алгоритму, а не на результат кластеризації. Як початковий вибір можна запропонувати $n_1 = n_2 = 20$.

3. Формується множина пар $P(\delta) = \{(C_i, C_m) : d(C_i, C_m) \le \delta\}$. Далі алгоритм збігається з алгоритмом класичної агломеративної кластеризації, тільки замість матриці D використовується множина $P(\delta)$.

4. Коли всі такі пари будуть вичерпані, параметр δ збільшується і формується нова скорочена множина пар. Так триває до повного злиття всіх об'єктів в один кластер.

Метод К-середніх є найбільш вживаним алгоритмом розділової кластеризації. Існує багато модифікацій цього методу, найвідоміші варіанти Болла–Холла [6–8] і Мак-Кіна [4, 6–8], на основі яких і розроблено обчислювальні схеми.

Метод К-середніх Болла–Холла:

1. Задається число K – кількість кластерів.

2. Серед усієї множини об'єктів обирається *K* точок одним з таких способів:
- перші *K* об'єктів;
- випадкові *K* об'єктів;
- найвіддаленіші *K* об'єктів.

Кожна з цих *K*точок вважається центром окремого кластера.

3. Далі кожен об'єкт зараховуємо до кластера, центр якого найближчий. Тобто для кожної точки $X_i, i=\overline{1,n}$ обчислюємо відстані $d(X_i, M_j), j=\overline{1,K}$ та обираємо номер того кластера, де буде досягатися мінімум.

4. Після того, як усі об'єкти розподілено, перераховуються центри кластерів $M_i=(\bar{x}_{i1}, \bar{x}_{i2}, ..., \bar{x}_{ip}), i=\overline{1,K}$, де $\bar{x}_{ij}=\frac{1}{n_i}\sum_{k=1}^{n_i} x_{kj}, j=\overline{1,p}$.

5. Повторюємо кроки 3–4, доки центри не стабілізуються або кількість ітерацій не перевищить задану.

Метод К-середніх Мак-Кіна:

Даний варіант методу відрізняється від попереднього тільки тим, що центри кластерів перераховуються кожного разу після зарахування точки до найближчого кластера.

Підтримка прийняття рішень. Одне з найактуальніших питань кластерного аналізу – оцінювання результатів і пошук розбиття, яке найкраще відповідає структурі даних. У більшості задач кластеризації дослідники стикаються з проблемою вибору оптимальної кількості кластерів, що відповідає природі досліджуваних об'єктів.

Для розв'язання таких задач у літературі існує велика кількість функціоналів та індексів якості, що дозволяють у кількісному вигляді оцінювати відповідність вихідного розбиття та природної структури даних, а також порівнювати результати, отримані різними методами або при різних значеннях параметрів [6, 9, 10]. Визначення функціоналів якості ґрунтується переважно 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. Ми розробили технологію для застосування методів теорії підтримки прийняття рішень, згідно з якою можна враховувати результати різних функціоналів якості одночасно. Такий підхід має забезпечити більш точну оцінку результатів.

Наша обчислювальна схема складається з кількох етапів.

1. Отримуємо різні розбиття та розглядаємо їх як альтернативи.

2. Для кожної альтернативи обчислюємо значення наступних функціоналів якості, які вважаємо експертами:

− сума внутрішньокластерних дисперсій за всіма ознаками

$$Q_1(C)=\sum_{i=1}^{K}\sum_{j=1}^{p}\left(\frac{1}{n_i-1}\sum_{k=1}^{n_i}\left(x_{kj}^{(i)}-\bar{x}_j^{(i)}\right)^2\right)\to min, \quad (1)$$

− сума квадратів відстаней до центрів кластерів

$$Q_2(C)=\sum_{i=1}^{K}\sum_{X_j^{(i)}\in C_i} d^2(X_j^{(i)}, M_i)\to min, \quad (2)$$

де $M_i=(\bar{x}_1^{(i)}, \bar{x}_2^{(i)}, ..., \bar{x}_p^{(i)})$ – центр кластера C_i;

− відношення середньої внутрішньокластерної та середньої міжкластерної відстаней.

$$Q_3(C)=\frac{\tilde{Q}(C)}{\tilde{\tilde{Q}}(C)}\to min, \quad (3)$$

де $\tilde{Q}(C)=\frac{1}{\sum_{i=1}^{K}\frac{n_i(n_i-1)}{2}}\sum_{i=1}^{K}\sum_{j=1}^{n_i-1}\sum_{k=j+1}^{n_i} d(X_j^{(i)}, X_k^{(i)})$, $\tilde{\tilde{Q}}(C)=\frac{1}{\prod_{i=1}^{K} n_i}\sum_{i=1}^{K-1}\sum_{j=1}^{n_i}\left(\sum_{m=i+1}^{K}\sum_{k=1}^{n_m} d(X_j^{(i)}, X_k^{(m)})\right)$;

− сума внутрішньокластерних відстаней

$$Q_4(C)=\sum_{i=1}^{K}\sum_{j=1}^{n_i-1}\sum_{k=j+1}^{n_i} d(X_j^{(i)}, X_k^{(i)})\to min. \quad (4)$$

3. Отримані результати подаємо у вигляді матриці $X=\{x_{ij}; i=\overline{1,n}, j=\overline{1,m}\}$, де *n* – кількість методів, *m* – кількість експертів, x_{ij} зведена до одиничної шкали оцінка, яку поставив *j*-й експерт *i*-й альтернативі.

4. Застосовуємо один з таких методів теорії прийняття рішень [13].

• *Процедура Борда*

Для кожного експерта виконуємо впорядкування альтернатив за спаданням 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 $\lambda^t = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} x_i^t, t = 1, 2, \ldots$

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.

Даний метод дозволяє також оцінити узгодженість експертів на основі дисперсійного коефіцієнта конкордації.

Класифікація. Запропоновано інформаційну технологію контрольованого машинного навчання, розроблено обчислювальні схеми восьми методів класифікації об'єктів [11, 12]. Якість моделі перевіряється шляхом підрахунку ймовірності помилки. Модель, для якої це значення буде найменшим, вважається найкращою та використовується для класифікації нових об'єктів. Структуру технології наведено на рис. 1.

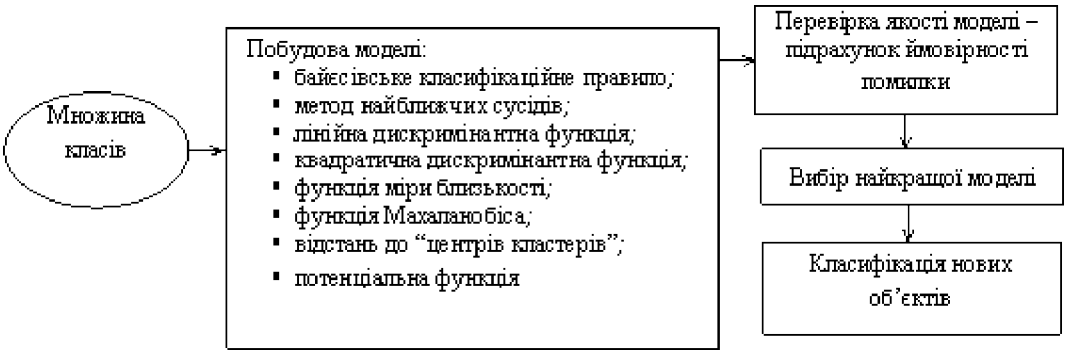

Рис. 1. Структура інформаційної технології класифікації

Метод найближчих сусідів.

1.Задаємо кількість сусідів k.

2.Для кожного об'єкта X знаходимо його k найближчих сусідів. Об'єкт X_i називається найближчим сусідом об'єкта X, якщо $d(X_i, X) = \min_i d(X_i, X), i = 1, 2, \ldots, N$.

3.Об'єкт X зараховується до того класу, до якого належить його найближчий сусід (у випадку методу найближчого сусіда) або більшість з його K сусідів (у випадку методу K-найближчих сусідів).

Байєсівське класифікаційне правило.

1. Обчислюємо байєсівські визначальні функції, які мають вигляд: $B_i(X) = P(\Pi_i) \cdot f(X | \Pi_i)$, де $P(\Pi_i)$ – апріорна ймовірність появи класу Π_i, $P(\Pi_i) = n_i / N$, де n_i – кількість навчальних об'єктів класу Π_i, $f(X | \Pi_i)$ – щільність розподілу класу [ ] в p-вимірному ознаковому просторі. Для нормального розподілу, наприклад, $f(X | \Pi_i) = \frac{1}{(2\pi)^{p/2} \Sigma_i^{p/2}} \exp\{-\frac{1}{2}(X - \mu_i)^T \Sigma_i^{-1} (X - \mu_i)\}$, де $\mu_i = \frac{1}{n_i} \sum_{x_j \in \Pi_i} X_j$ – вектор середніх класу [ ] [ ] – вибіркова коваріаційна матриця класу [ ]

2. Об'єкт X зараховуємо до класу Π_i, якщо [ ] $i, j = 1, 2, \ldots, K$.

Непараметричні алгоритми розпізнавання за мірою схожості з еталонними класами.

1. Обчислюємо спеціальну функцію – міру схожості [ ] між об'єктом X і класом [ ] Залежно від виду функції схожості можна виділити різні методи. Розглянемо такі функції схожості:

– *функція міри близькості* [ ]

– *відстань до центра* де – евклідова відстань, – вектор середніх класу

– *потенціальна функція* де

– *відстань Махаланобіса* де і відповідно вектор середніх і коваріаційна матриця класу

2. Об'єкт *X* належить до класу Π_i, якщо для функцій і для функції .

Лінійна дискримінантна функція для нормальних розподілів з рівними коваріаційними матрицями.

1. Для кожного класу обчислюємо функцію де

2. Об'єкт *X* належить до класу Π_i, якщо

Квадратична дискримінантна функція для нормальних розподілів з різними коваріаційними матрицями.

1. Для кожного класу обчислюємо функцію

2. Об'єкт *X* належить до класу якщо $i, j = 1, 2, \ldots, K$.

Візуалізація та інтерпретація результатів. У системах Data Mining візуалізація розглядається як досить важливий спосіб розвідувального аналізу даних, вона має як самостійну цінність для кращого розуміння природи досліджуваних даних, так і використовується для полегшення оцінки та інтерпретації результатів кластеризації, відіграє значну роль у підтримці прийняття рішень.

Система MedISA має широкий спектр засобів зображення даних. Це графіки й таблиці, гістограми і дендрограми, різного роду діаграми, в тому числі діаграми розсіювання кластерів, списки й текстові коментарі, розроблена процедура картографічної візуалізації значень градації досліджуваних показників на території України.

Результати роботи системи розглянемо на двох наборах даних. Перший з них містить інформацію медичного обстеження хворих на серцеву недостатність, складається з 394 об'єктів, що характеризуються сімома ознаками (рис. 2–9). Другий набір даних містить показники первинної інвалідності дорослого населення України в розрізі адміністративних територій за класами та формами захворювань. Налічує 27 об'єктів, що відповідають адміністративним територіям і розглядаються у п'ятивимірному ознаковому просторі (рис. 10–11). В обох випадках попередньо проведено стандартизацію даних.

Головна перевага ієрархічних методів кластеризації полягає у їх здатності давати наочне уявлення про структуру досліджуваних даних у вигляді дендрограми (рис. 2). Дендрограма – це графік, який являє собою вкладене групування об'єктів, що змінюється на різних рівнях ієрархії та наочно демонструє послідовність об'єднання елементів аналізу.

Для того щоб отримати розбиття об'єктів на класи ієрархічним методом, потрібно задати кількість кластерів, на якій слід припинити процес об'єднання. Лінія на дендрограмі (рис. 2) показує рівень, на якому отримуємо 5 кластерів. На рис. 3 наведено діаграму розсіювання, що є двовимірним зрізом простору ознак. Кожен об'єкт розглядається як точка в геометричному просторі. Координатні осі відповідають обраним користувачем ознакам. Різними позначками зображені об'єкти різних кластерів. Діаграма розсіювання також застосовується для зображення кластерів, отриманих неієрархічними методами.

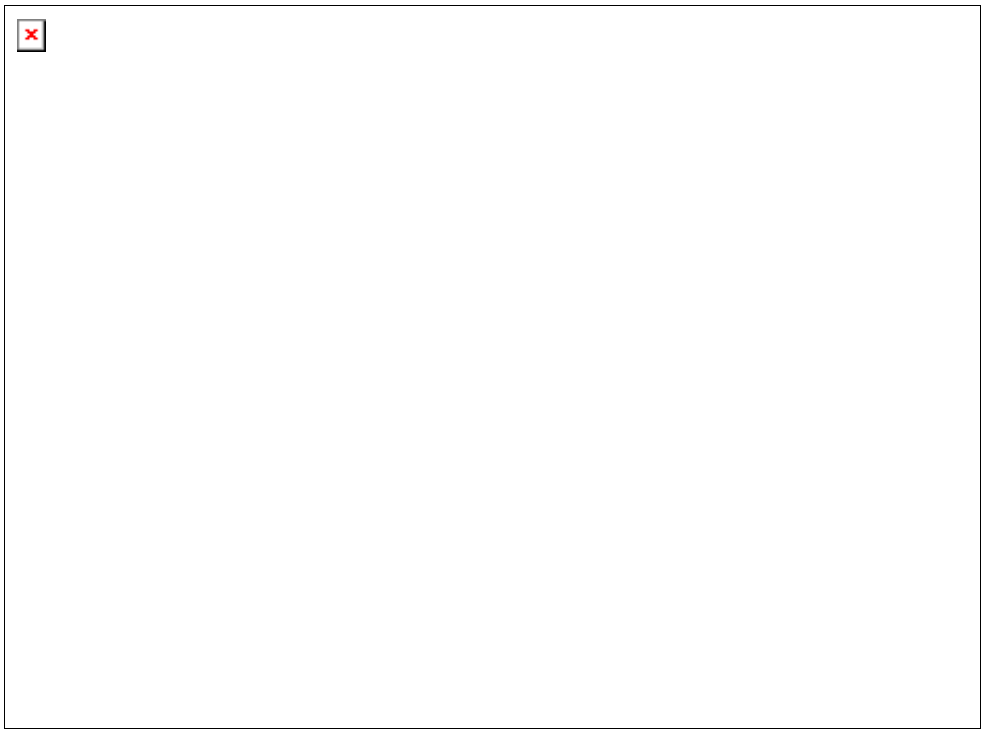

Рис. 2. Дендрограма, що відображає результати ієрархічного методу повного зв’язку

Рис. 3. Діаграма розсіювання

Для вибору методу, що дає найкраще розбиття, проведемо кластеризацію різними алгоритмами і скористаємося методами прийняття рішень. На вкладці “Функціонали якості” програми (рис. 4) відображено таблицю з оцінками.

Рис. 4. Вкладка “Функціонали якості” системи MedISA

Як бачимо, визначитися з вибором, ґрунтуючись лише на оцінках функціоналів якості, досить складно, тому застосуємо методи прийняття рішень. Результати множинного аналізу наведені на рис. 5, а на рис. 6 маємо висновки процедур Борда та плюралітарної.

Рис. 5. Результати множинного аналізу

У результаті множинного аналізу було обрано ієрархічний метод повного зв’язку (віддаленого сусіда). Зроблено висновок про неузгодженість оцінок експертів.

Найкращим методом за процедурами Борда та плюралітарної є метод *K*-середніх Болла–Холла при виборі перших *k* точок як початкових центрів.

1. Процедура Борда

Метод	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Оцінка	30	27	21	22	12	19	13	24	20	23

* Чим менше оцінка, тим краще метод

1. Плюралітарна процедура

Метод	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10
Оцінка	1	0	0	0	2	1	1	0	0	0

* Чим більше оцінка, тим краще метод

Рис. 6. Результати процедур Борда і плюралітарної

Для аналізу отриманих кластерів запропоновано провести первинний статистичний та імовірнісний аналіз [14]. Для кожної ознаки з будь-якого класу можна обчислити статистичні характеристики (рис. 7), відновити нормальний та сплайн-нормальний (з 1 і 2 вузлами склеювання) розподіли (рис. 8), за допомогою критеріїв згоди Пірсона та Колмогорова перевірити, чи відповідає відтворений розподіл реальним даним, визначити границі норми.

	Точкові оцінки	Середньокв-ні відхилення	Довірчі інтервали
Середнє арифметичне	74,211024	0,469940	(73,289756; 75,132291)
Середнє квадратичне	3,875228	0,332298	(3,223794; 4,526663)
Коефіцієнт варіації Пірсона	0,052219	0,004490	(0,043417; 0,061021)
Коефіцієнт асиметрії	0,313926	0,284308	(-0,243429; 0,871281)
Коефіцієнт ексцесу	-0,070725	0,532471	(-1,114577; 0,973126)
Коефіцієнт контрексцесу	3,760210	0,079281	(3,604787; 3,915633)

Рис. 7. Характеристики обчислені для ознаки “Фракція викиду” другого кластера

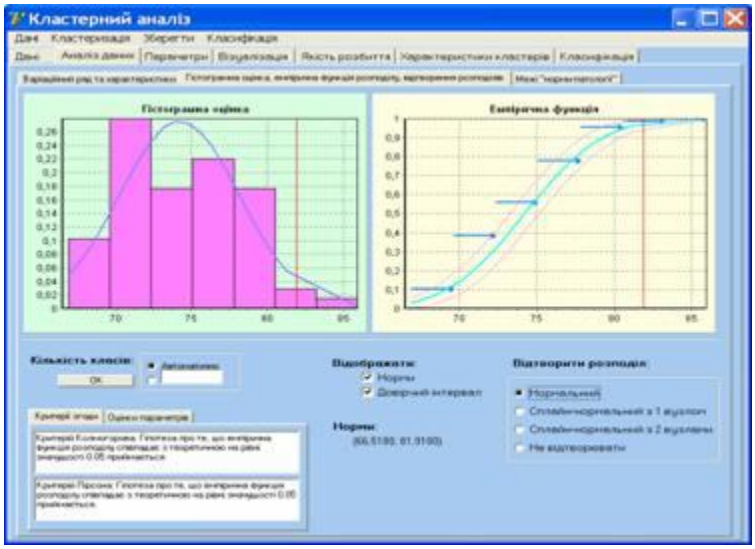

Рис. 8. Результати імовірнісного аналізу для ознаки “Фракція викиду” другого кластера

Також є можливість провести порівняльний аналіз кластерів за однією з характеристик, наприклад за середнім значенням ознак, у табличному і графічному вигляді (рис. 9).

	KDR	KSR	KDO	KSO	FV	IMMLG	SDLA
Кластер 1	6,7576	5,3956	237,3396	142,4192	40,0010	254,5433	41,4800
Кластер 2	4,9125	2,7812	113,8676	29,3341	74,2110	141,2003	29,7500
Кластер 3	5,6521	4,0305	157,9013	72,1506	53,9895	191,9074	41,9836
Кластер 4	4,4081	2,6900	88,6419	27,2978	69,3293	98,4825	32,8261
Кластер 5	5,1345	3,3374	126,2936	45,5822	63,5843	142,7693	28,8137

Рис. 9. Порівняльна характеристика кластерів за середніми значеннями показників

Система MedISA містить процедуру картографічної візуалізації кластеризації у випадку, коли об’єктами набору даних виступають адміністративні території України. На рис. 10 наведено цифрову карту України, що автоматично будується засобами розробленого програмного забезпечення. Різний колір областей відповідає різним кластерам.

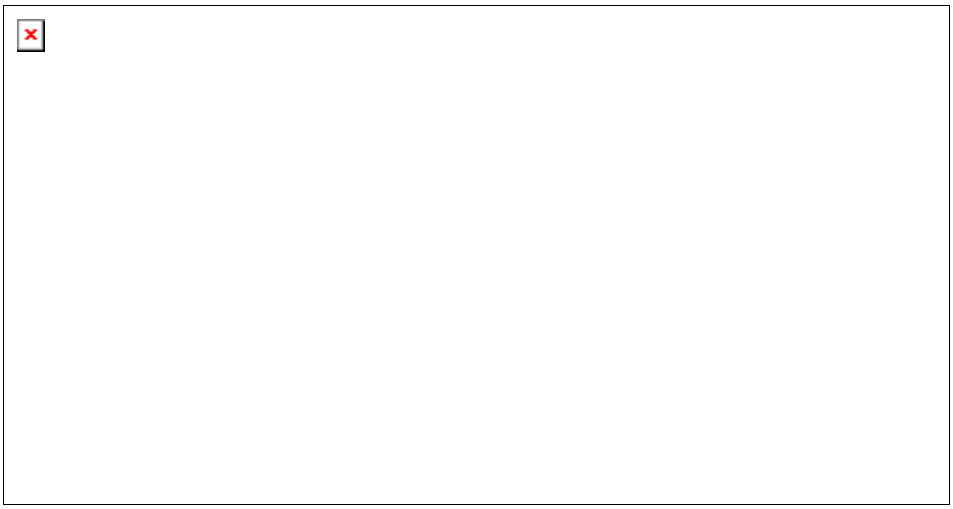

Рис. 10. Карта України з нанесенням результатів кластеризації

Програма реалізує картографічну візуалізацію значень градації медичних показників на території України за допомогою поліноміальних сплайнів від двох змінних на основі *B*-сплайнів, близьких до інтерполяційних у середньому, що дозволяє відображати дані в інтуїтивно-зрозумілому вигляді для полегшення їх аналізу, інтерпретації та розуміння [15]. На рис. 11 наведено результати картографічної візуалізації показника інвалідності (на 10 тис. населення) за ішемічною хворобою серця для дорослого населення на 2010 р.

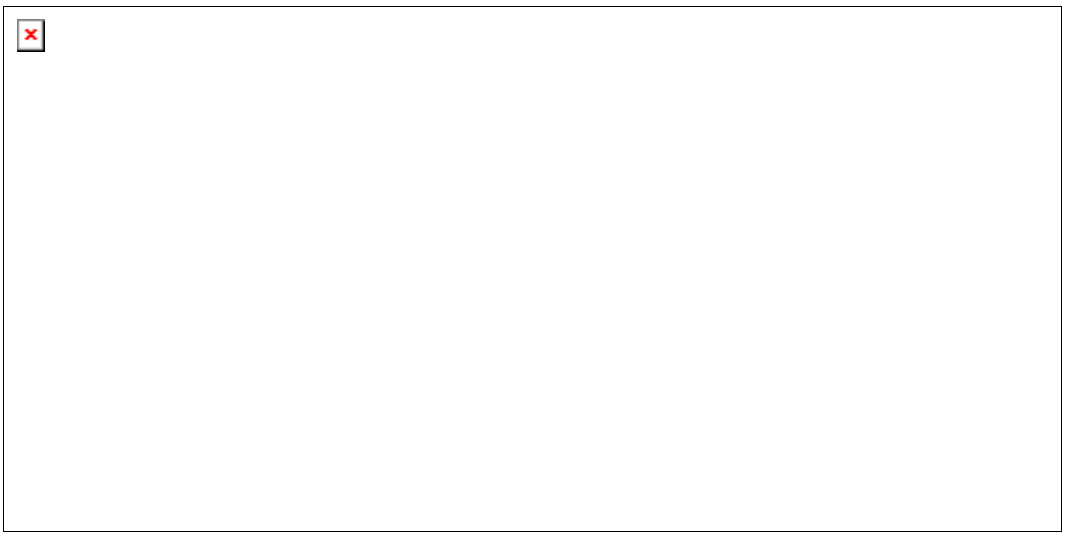

Рис. 11. Результати картографічної візуалізації

Висновки. Розроблено інформаційну технологію проведення кластерного аналізу, ядро якої становлять обчислювальні схеми на основі ієрархічного агломеративного методу, швидких ієрархічних методів, методу *K*-середніх у варіантах Болла–Холла та Мак-Кіна. При цьому запропоновано 3 типи метрик: евклідова, манхеттенська, Чебишева. Для ієрархічної кластеризації реалізовано 5 типів міжкластерних відстаней: близького сусіда (одиничного зв'язку), далекого сусіда (повного зв'язку), середня, між центрами та відстань Ворда. Запропоновано технологію підтримки прийняття рішень у задачах кластеризації, яка ґрунтується на функціоналах якості, множинному аналізі, процедурах Борда та плюралітарній. Розроблено програмне забезпечення MedISA, в якому реалізовано всі запропоновані процедури, а також обчислювальні схеми класифікації, первинного статистичного та імовірнісного аналізу, засоби візуалізації. Продемонстровано результати роботи програми на наборах реальних даних у галузі медицини. Ця система може бути використана і в інших галузях науки і техніки.

Література

1. Дюк В. Data Mining : учебный курс / В. Дюк, А. Самойленко. – СПб. : Питер, 2001. – 368 с.

2. Чубукова И. А. Data Mining : учебное пособие / И. А. Чубукова. – М. : Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – 382 с.

3. Пятецкий-Шапиро Г. Data Mining и перегрузка информацией / Г. Пятецкий-Шапиро // Анализ данных и процессов / А. А. Барсегян, М. С. Куприянов, И. И. Холод и др. – 3-е изд., перераб. и доп. – СПб. : БХВ-Петербург, 2009. – 512 с.

4. Айвазян С. А. Классификация многомерных наблюдений / Айвазян С. А., Бежаева З. И., Староверов О. В. – М., 1974. – 240 с.

5. Жамбю М. Иерархический кластер-анализ и соответствия / Жамбю М. – М., 1988. – 279 с.

6. Мандель И. Д. Кластерный анализ / Мандель И. Д. – М., 1988. – 176 с.

7. Воронцов К. В. Алгоритмы кластеризации и многомерного шкалирования : курс лекций / Воронцов К. В. – М. : Изд-во МГУ, 2007.

8. Jain A. K. Data Clustering: A Review / A. K. Jain, M. N. Murty, P. J. Flunn // ACM Computing Surveys. – September 1999. – Vol. 31, – No. 3. – P. 265–323.

9. Halkidi M. On Clustering Validation Techniques/ M. Halkidi, Y. Batistakis, M. Vazirgiannis // Journal of Intelligent Information Systems. – 2001. – 17:2/3. – P. 107–145.

10. Milligan G. An examination of procedures for determining the number of clusters in a data set / G. Milligan, M. Cooper // Psychometrika. – June 1985. – Vol. 50. – No. 2. – P. 159–179.

11. Бусыгин Б. Н. Распознавание образов при геолого-географизическом прогнозировании / Б. Н. Бусыгин, Л. В. Мирошниченко. – Днепропетровск: Изд-во ДГУ, 1991. – 168 с.

12. Kotsiantis S. B. Supervised Machine Learning: A Review of Classification Techniques // Informatica. – 2007. – V. 31. – P. 249–268.

13. Емельяненко Т. Г. Принятие решений в системах мониторинга / Т. Г. Емельяненко, А. В. Зберовский, А. Ф. Приставка, Б. Е. Собко. – Д., 2005. – 224 с.

14. Приставка О. П. Статистичний аналіз в АСОД: Відтворення розподілів. Критерії однорідності / Приставка О. П., Приставка П. О., Смирнов С. О. – Д. : РВВ ДДУ, 2000. – 112 с.

15. Приставка П. О. Поліноміальні сплайни при обробці даних : монографія / Приставка П. О.– Д. : Вид-во Дніпропетр. ун-ту, 2004. – 236 с.