УДК 519.254

Система «MedISA» кластеризації даних

О.П. Приставка[1], М.Г. Сидорова[1]

Abstract – **Proposed the information technology of cluster analysis, which is a part of the developed software, "MedISA" for automatic data processing.**

Keywords – **Cluster analysis, hierarchical and partitioning methods, K-means, decision-making, visualization.**

I. Вступ

Мета роботи полягала в розробці інформаційної технології та програмного забезпечення кластерного аналізу даних, представлених у вигляді матриці дійсних чисел $X = \{x_{ij}; i = \overline{1, N}, j = \overline{1, p}\}$, де N – кількість об‘єктів, кожен об‘єкт характеризується набором з p ознак, x_{ij} – значення j-ї ознаки, що спостерігається в і-го об‘єкта, дійсне число.

II. Ядро системи «MedISA»

В роботі запропоновано інформаційну технологію кластерного аналізу, що входить до складу розробленої в середовищі Delphi 7.0 автоматизованої системи «MedISA» обробки медичних даних, ядро якої складають:

- Процедури роботи з базою даних, формування локальних баз даних за запитами користувача, стандартизація даних.
- Обробка даних на основі первинного статистичного аналізу та імовірнісний статистичний аналіз, задача якого є відновлення функцій розподілу.
- Алгоритми кластерного аналізу. Реалізовано ієрархічні методи, що представлені класичним агломеративним та двома швидкими алгоритмами, а також ітеративний метод К-середніх у варіантах Болла-Холла та Мак-Кіна. Система дозволяє обирати метрику відстані між кластерами (ближнього сусіда, дальнього сусіда, середня, між центрами, відстань Варда) та об‘єктами (евклідова, манхеттенська, Чебишева). Вибір початкових центрів представлено трьома способами: перші к точок, найвіддаленіші та випадкові.[1,2]
- Вибір оптимальної кількості кластерів здійснюється на основі індексу Calinski-Harabasz та методу, що базується на різниці між рівнями об‘єднання при ієрархічній кластеризації.[3]
- Оцінка якості кластеризації та вибір найадекватнішого методу кластерного аналізу для досліджуваної сукупності об‘єктів на основі функціоналів якості, множинного аналізу та колективних методів прийняття рішень (процедури Борда та плюралітарна).[4]
- Для аналізу та інтерпретації отриманих результатів запропоновано обчислення статистичних характеристик у кожному з кластерів, а також потужний блок візуалізації даних, що містить різноманітні графіки, діаграми розсіювання, дендрограми, таблиці та текстові коментарі.
- Реалізовані наступні методи класифікації: байесовське класифікаційне правило, метод найближчих сусідів, лінійна дискримінантна функція, квадратична дискримінантна функція, методи, що ґрунтуються на функції міри близькості, функції Махаланобіса, відстані до «центрів кластерів», потенціальній функції.

Система має багатовіконний інтерфейс та взаємодіє з користувачем у діалоговому режимі. Робота програми починається з завантаження даних, що мають бути кількісними та міститися у текстовому файлі або dbf-форматі. Далі користувач має змогу провести стандартизацію даних, обрати метод кластеризації, задати необхідні вхідні параметри, оцінити якість розбиття та провести аналіз отриманих результатів за допомогою статистичних характеристик та засобів візуалізації. Розбиття вихідної вибірки на кластери, одержані різними методами або при різних значеннях параметрів, можуть значно відрізнятися. Тому вибір алгоритму та параметрів для отримання найкращого результату є досить важливим та складним. В системі велика увага приділена підтримці прийняття рішень, що здійснюється на основі функціоналів якості та колективних методів вибору. Програма дозволяє порівнювати якість різних методів при застосуванні їх до вибірки досліджуваних об‘єктів, підбирати найоптимальніші вхідні параметри до кожного метода, а також визначати можливу кількість кластерів.

III. Практична реалізація

Запропонована система знайшла практичну реалізацію на медичних даних, що є показниками первинної інвалідності дорослого й працездатного населення в Україні та в розрізі адміністративних територій; за класами та формами захворювань. Задача полягала у розподіленні адміністративних територій України на три групи в залежності від рівня інвалідності за такими показниками: гостра ревматична гарячка та хронічні ревматичні хвороби серця; гіпертонічна хвороба; ішемічна хвороба серця; цереброваскулярні хвороби; хвороби артерій, артеріол та вен.

Для кластеризації в якості оптимального методу було обрано ієрархічний метод з обчисленням відстані Варда між кластерами та манхеттенською метрикою.

На Рис. 1 представлена діаграма розсіювання, де різними позначками представлені об‘єкти, що потрапили до різних кластерів.

[1] Дніпропетровський національний університет імені Олеся Гончара, пр. Гагаріна, 72, Дніпропетровськ, 49050, УКРАЇНА, E-mail: Sidorova.M.G@gmail.com

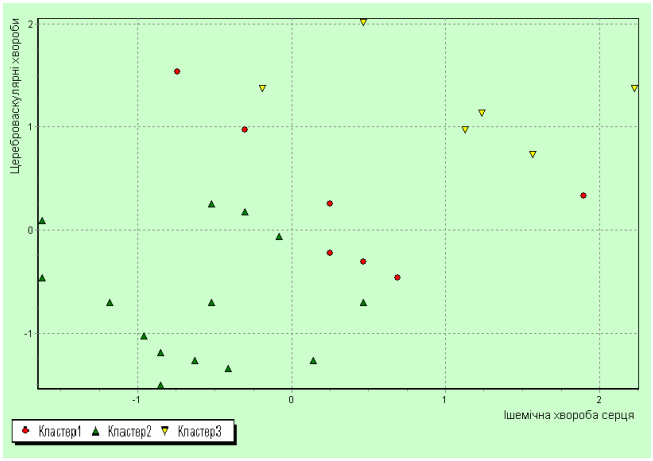

Рис.1. Діаграма розсіювання отриманих кластерів

Процес об‘єднання у кластери та ієрархію об‘єктів можна представити у вигляді дендрограми на Рис. 2.

На Рис. 3 наведено цифрову карту України, що автоматично будується засобами розробленого програмного забезпечення при використанні бібліотеки OpenGL. Різний колір областей відповідає різним кластерам.

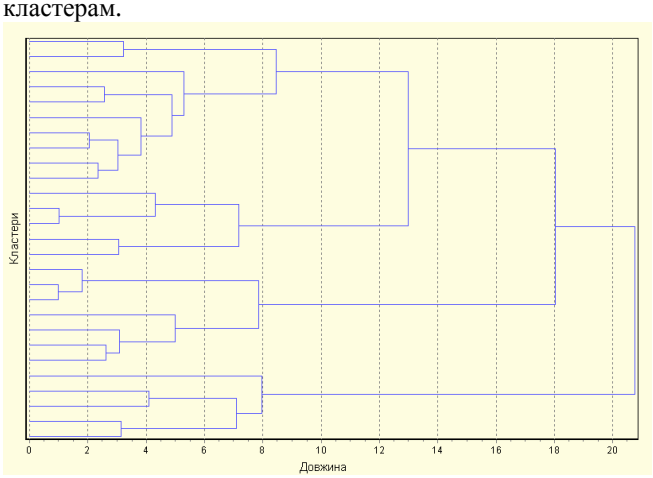

Рис.2. Дендрограма – результат ієрархічного агломеративного методу

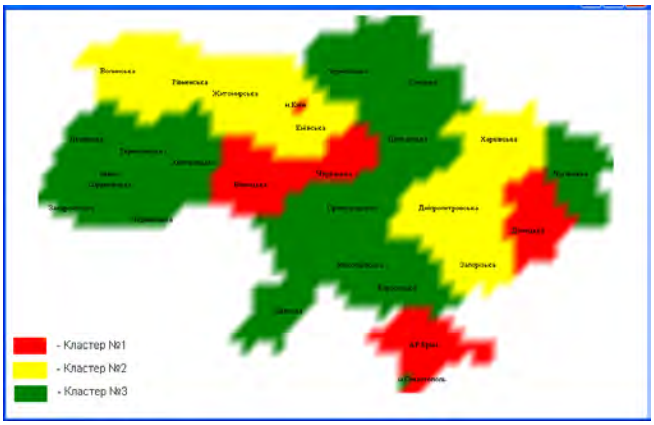

Рис.3. Карта України з нанесенням результатів кластеризації

Таким чином, перший кластер складають Вінницька, Черкаська, Донецька області, АР Крим та м. Київ; до другого кластеру увійшли Волинська, Рівненська, Житомирська, Київська, Харківська, Дніпропетровська та Запорізька; третій кластер об‘єднує адміністративні одиниці, що залишилися.

Для оцінки отриманих кластерів, розглянемо середні показники для кожної ознаки у кожній з груп (Таб. I).

ТАБЛИЦЯ 1

СЕРЕДНІ ПОКАЗНИКИ

ОЗНАКА	КЛАСТЕР №1	КЛАСТЕР №2	КЛАСТЕР №3
Ревматичні хвороби серця	0.50	0.24	0.26
Гіпертонічна хвороба	0.80	0.26	0.49
Ішемічна хвороба серця	4.12	3.03	4.19
Цереброваскулярні хвороби	4.86	3.54	5.33
Хвороби артерій, артеріол та вен	1.02	0.79	1.15

Таким чином, бачимо, що другий кластер має найнижчі середні показники інвалідності для всіх ознак. Перший кластер характеризується найвищими середніми показниками для ознак «Ревматичні хвороби серця» та «Гіпертонічна хвороба», а третій – для ознак «Ішемічна хвороба серця», «Цереброваскулярні хвороби» та «Хвороби артерій, артеріол та вен».

III. ВИСНОВОК

У роботі запропонована інформаційна технологія кластеризації, а також програмне забезпечення «MedISA», що містить дану технологію та дозволяє у зручній формі проводити кластерний аналіз, оцінювати якість розбиття, пропонує методи підтримки прийняття рішень та полегшення інтерпретації отриманих результатів. Проведено практичну реалізацію системи на медичних даних та наведено отримані результати. Система також може бути застосована до інших наборів даних у різних галузях науки.

СПИСОК ПОСИЛАНЬ

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

[2] Жамбю М. Иерархический кластер-анализ и соответствия / М. Жамбю. – М., 1988. – 279 с.

[3] Milligan G.W. and Cooper M.C. An examination of procedures for determining the number of clusters in a data set, Psychometrika, 1985, 50, 159-179p.

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