2013. – Львів, Вид-во Нац. ун-ту "Львівська політехніка", 2013, P. 220–221. 2. Fisher T. An overview of current approaches to mashup generation / T. Fischer, F. Bakalov, A. Nauerz // Proceedings of the International Workshop on Knowledge Services and Mashups, 2009, P. 157-158. 3. About IFTTT [Електронний ресурс] / IFTTT Inc. // Режим доступу: URL: https://ifttt.com/wtf. 4. Сповіщення. Слідкуйте за новим цікавим вмістом в Інтернеті [Електронний ресурс] / Google Inc. // Режим доступу: URL: https://www.google.com/alerts. 5. Automate your Dropbox [Електронний ресурс] / Wappwolf Inc. // Режим доступу: URL: http://wappwolf.com/dropboxautomator. 6. Giusy L. Mashups for data integration: An analysis. / L. Giusy, H. Hakim // Technical Report UNSW-CSE-TR-0810, 2008, P. 68–69. 7. Кушнірецька І. І. Визначення структури і змісту вхідних інформаційних ресурсів для роботи Мешап-системи / І. І. Кушнірецька, О. І. Кушнірецька, А. Ю. Берко // Технологический аудит и резервы производства. – 2014. – № 6/3(20). – С. 4–9. 8. ISO 15926-7: Data integration, sharing, exchange, and hand-over between computer systems. Part 7: Implementation methods for the integration of distributed systems: Template methodology, 2007.

УДК 681.513

І. А. Лур'є[1], В. В. Осипенко[2], В. І. Литвиненко[1], М. А.Таиф[1], Н. В. Корніловська[1]
[1]Херсонський національний технічний університет,
[2]Національний університет біоресурсів і природокористування України

ГІБРИДИЗАЦІЯ АЛГОРИТМУ ІНДУКТИВНОГО КЛАСТЕР-АНАЛІЗУ З ВИКОРИСТАННЯМ ОЦІНКИ ЩІЛЬНОСТІ РОЗПОДІЛУ ДАНИХ

Запропоновано нову техніку кластеризації, в основу якої покладено два методи: щільнісний алгоритм DBSCAN та індуктивний алгоритм об'єктивної кластеризації. Експерементально доведено, що комбінацією двох цих методів дозволяє вирішити проблему розпізнавання кластерів різної нелінійної форми та значно підвищити точність при розпізнаванні складних об'єктів.

Ключові слова: щільнісний метод кластеризації DBSCAN, об'єктивний алгоритм кластеризації, індуктивні методи самоорганізації моделей, МГУА, гібридні методи кластеризації.

In this article proposed a new clustering technique, which is based on two methods: density algorithm DBSCAN and inductive objective clustering algorithm. Experimentally proved that the combination of two these methods can solve the problem of recognition of clusters of different nonlinear form, and greatly increase the accuracy in the detection of complex objects.

Key words: Density-based spatial clustering, DBSCAN, objective clustering algorithm, inductive methods of self-organization models, GMDH, hybrid clustering methods.

Вступ

Задача кластеризації – окремий випадок задачі навчання без вчителя, що зводиться до розбиття наявної множини об'єктів даних на підмножини так, щоб елементи однієї підмножини істотно відрізнялися деяким набором властивостей від елементів всіх інших підмножин. Існує багато різних алгоритмів кластеризації. Деякі з них поділяють множину на заздалегідь відому кількість кластерів, деякі автоматично вибирають кількість кластерів. Алгоритм DBSCAN (Density Based Spatial Clustering of Applications with Noise) – щільнісний алгоритм для кластеризації

просторових даних з присутністю шуму) із автоматичним вибором кількості кластерів. Він оснований на припущенні, що щільність точок всередині кластерів більша, ніж поза кластерами. Цей алгоритм дає змогу знаходити кластери довільної форми. Алгоритм запропонували Мартін Эстер, Ганс-Пітер Кригель і колеги для вирішення проблеми розбиття даних (спочатку просторових) на кластери довільної форми [1]. Більшість алгоритмів створюють кластери, що за формою близькі до сферичних, бо мінімізують відстань об'єктів до центру кластера. Автори DBSCAN експериментально показали, що цей алгоритм здатний розпізнати кластери різної форми. Ідея, що покладена в основу цього алгоритму, полягає в тому, що всередині кожного кластера спостерігається типова щільність точок (об'єктів), що помітно вища ніж щільність зовні кластера, а також щільність в областях із шумом нижча ніж щільність кожного із кластерів. З іншого боку, індуктивні методи кластеризації [2] дозволяють при неточних зашумлених даних та коротких вибірках, використовуючи мінімум обраного квадратичного критерію, знаходити нефізичну модель (вирішальне правило), точність якої менша ніж структура повної фізичної моделі. Перебір множини моделей-кандидатів за зовнішніми критеріями є необхідним тільки для нефізичних моделей. За малих дисперсій завад доцільно використовувати внутрішні критерії перебору. Із збільшенням завад доцільно переходити до непараметричних алгоритмів. Застосування індуктивних методів кластеризації доцільне тому, що вони майже завжди гарантують знаходження оптимальної кількості кластерів, яке адекватне рівню шуму у вибірці даних.

Основна ідея цієї роботи полягає в тому, щоб об'єднати щільнісний алгоритму DBSCAN, який дозволяє розпізнавати кластери різної форми та індуктивний алгоритм кластеризації, який дозволить значно підвищити точність при розпізнаванні складних об'єктів.

Запропоновано нову технологію кластеризації, в основу якої покладено два методи: щільнісний алгоритм DBSCAN та індуктивний алгоритм об'єктивної кластеризації. Комбінуючи ці методи, можна вирішити деякі з перерахованих вище проблем з доволі високим результатом.

Мета роботи – розроблення методологічних засад побудови гібридних індуктивних алгоритмів кластер-аналізу для виділення (кластеризації) об'єктів із складними нелінійними формами з високими характеристиками точності розпізнавання та роздільної здатності.

Для досягнення мети роботи при розв'язанні прикладних задач опрацювання даних запропоновано використати індуктивний алгоритм кластеризації з вбудованим алгоритмом DBSCAN, в якому розв'язок, що наближений до глобального мінімуму, отримують послідовним використанням DBSCAN операцій.

Постановка задачи

Формальна постановка задачі кластеризації виглядає так: нехай X – множина об'єктів, Y – множина номерів (імен, міток) кластерів. Задано також функцію відстані між об'єктами $r(x, x')$. Потрібно розбити вибірку на підмножини, що не перетинаються (кластери), так, щоб кожен кластер складався з об'єктів, близьких за метрикою r, а об'єкти різних кластерів істотно відрізнялися. При цьому кожному об'єкту $x_i \in X^m$ відповідає номер кластера y_i. При цьому алгоритм кластеризації можна роглядати як функцію $a: X \to Y$, яка будь-якому об'єкту $x \in Y$ ставить у відповідність номер кластера $y \in Y$. Множина Y у деяких випадках відома заздалегідь, однак частіше ставиться задача визначення оптимальної кількісті кластерів за тим чи іншим критерієм якості кластеризації [3].

Методологічні основи індуктивного моделювання складних систем та індуктивного алгоритму кластерізації

Вперше ідею методу групового урахування аргументів подав О. Г. Івахненко у 1968 році [4]. Як зазначено в [5, 6], МГУА – це потужна інформаційна технологія розв'язання задач структурно-параметричної ідентифікації моделей складних об'єктів або моделювання за експериментальними даними в умовах невизначеності, а також задач інтелектуального аналізу даних. В подальшому метод групового урахування аргументів трансформувався від "евристичної самоорганізації" [7, 8], "індуктивної самоорганізації моделей складних систем" [21] до "індуктивного моделювання

складних систем". Серед основних принципів індуктивного моделювання складних систем виділяються такі [7, 8, 9]: принцип самоорганізації; принцип зовнішнього доповнення і принцип свободи вибору рішень.

Принцип самоорганізації моделей ґрунтується на індуктивному підході моделювання складних систем, відкидає шлях розширення й ускладнення моделі, збільшення вихідного обсягу інформації про об'єкт та постулює існування оптимальної, що обмежена за масштабами області моделювання однієї моделі оптимальної складності. Її можна синтезувати за допомогою самоорганізації, тобто перебору багатьох моделей-претендентів за доцільно вибраними зовнішніми критеріями селекції моделей. Оптимізація моделі за деяким ансамблем критеріїв визначає досяжні за заданого рівня шумів та обсягу спостережень результати моделювання.

Принцип зовнішнього доповнення пов'язаний з теоремою Геделя, яка говорить, що "... тільки зовнішні критерії, засновані на новій інформації, дозволяють синтезувати справжню модель об'єкта, що прихована в зашумлених даних". Інакше кажучи, відповідно до цього принципу, лише зовнішні критерії (тобто розраховані на основі "свіжих" даних, які не використовувалися для синтезу моделі) при збільшенні складності моделі проходять через їхні мінімуми. Застосування цього принципу реалізується через поділ вихідної таблиці даних на дві частини: A і B.

Принцип свободи вибору рішень: відповідно до нього для кожного покоління (або ряду селекції моделі) існує деякий мінімум комбінацій, які називаються свободою вибору і забезпечують збіжність багаторядних селекцій до моделі оптимальної складності. Принцип свободи вибору рішень та покрокова (багаторядна) процедура прийняття рішення вперше реалізовані в персептроні. Персептрон складається з декількох рядів зв'язків. Після кожного ряду зв'язків є спеціальний пристрій, що пропускає в наступний ряд найімовірніші рішення. На останньому пристрої приймається єдине і остаточне рішення. Інакше кажучи, цілеспрямовано вибираючи моделі для визначення моделі оптимальної складності відповідно до викладених принципів, необхідно дотримуватися таких правил: для кожного покоління (або ряду селекції) моделей існує деякий мінімум комбінацій, який називається свободою вибору. Занадто велика кількість поколінь призводить до індуциту (інформаційна матриця стає погано зумовленою). Що складніше завдання селекції, то більше потрібно поколінь для отримання моделі оптимальної складності. Свобода вибору забезпечується тим, що на кожен наступний ряд селекції передається не одне рішення, а декілька кращих, які відібрані на останньому ряді. Д. Габор сформулював цей принцип так: приймати рішення в певний момент часу необхідно так, щоб у наступний момент часу, коли виникне необхідність у черговому рішенні, зберігалася б свобода вибору рішень [10].

Ці принципи покладено в основу технології розв'язання задач індуктивного синтезу моделей за експериментальними даними. Найзагальнішу постановку задачі індуктивного синтезу моделей за експериментальними даними або структурно-параметричної ідентифікації подано у працях [5, 6; 11, 12]. Згідно з цими роботами така постановка зводиться до пошуку екстремуму деякого критерію CR на множині різних моделей $\Im$:

$$f^* = \arg\min CR(f). \qquad (1)$$

Оскільки (1) не є завершеним формулюванням задачі, його необхідно додатково визначити, зокрема:

- задати відому з аналізу експерименту апріорну або експертну інформацію про вид, характер та обсяг початкової інформації;
- вказати клас базисних функцій, з яких повинна формуватися множина $\Im$;
- визначити спосіб генерації моделей f;
- задати метод оцінювання параметрів;
- задати критерій $CR(f)$ порівняння моделей та вказати метод його мінімізації.

Отже, технологія розв'язування такої задачі складається з таких основних етапів [6]:

1) подання вибірки даних експерименту, апріорної та експертної інформації та поділ таблиці даних щонайменше на дві підвибірки, які не перетинаються (питанням поділу присвячені роботи [13], [14] та ін.);

2) визначення класу базисних функцій;
3) генерування різних структур моделей у вибраному класі;
4) оцінювання параметрів згенерованих структур і формування множини $\Im$;
5) мінімізація заданого критерію $CR(f)$ і вибір оптимальної моделі f^*;
6) перевірка адекватності отриманої оптимальної моделі;
7) ухвалення рішення про завершення процесу моделювання.

Із загальних позицій, задача ідентифікації полягає у формуванні за вибіркою даних експерименту певної множини $\Im$ моделей різної структури вигляду [6]:

$$\hat{y}_f = f\left(X, \hat{q}_f\right) \tag{2}$$

і знаходженні оптимальної моделі за умовою

$$f^* = \arg\min_{f\in\Im} CR\left(y, f\left(X, \hat{q}_f\right)\right), \tag{3}$$

причому оцінки параметрів в (2) для кожної моделі $f \in \Im$ є розв'язком ще однієї задачі знаходження екстремуму [6, 11]:

$$\hat{q}_f = \arg\min_{q_f\in R^{s_f}} QR\left(y, X, q, s_f\right). \tag{4}$$

У термінах індуктивного моделювання складних систем s_f називається складністю моделі f і дорівнює кількості ненульових компонентів у (3); QR – критерій якості розв'язання задачі параметричної ідентифікації для кожної синтезованої моделі, що генерується в задачі структурної ідентифікації [6], [11]. Невизначеності, які містяться в (3), (4), можна також зарахувати до класу концептуальних, що впливає на якість розв'язку. Ці невизначеності поділяються на дві групи, тобто такі, що стосуються вихідних даних, і такі, які стосуються технології моделювання.

Дані експерименту зазвичай містять такі види невизначеності:

1) структурна невизначеність: неповне знання зв'язків вхід-вихід, що не дає змоги однозначно задати структуру моделі (2);
2) стохастична невизначеність: невідомий характер та рівні шуму у даних;
3) інформаційна невизначеність або якість вибірки.

Невизначеності, що стосуються процедури моделювання, є такими:

1) функціональна невизначеність проявляється у неможливості точного виборі того або іншого базисного набору функцій;
2) параметрична невизначеність – стосується вибору критерію QR і методу розв'язання задачі параметричної ідентифікації (4);
3) критеріальна невизначеність – стосується вибору критерію CR і методу розв'язання основної задачі структурної ідентифікації (3).

У роботі [15] зазначено, що погляд на кластеризацію як на модель дозволяє перенести в теорію кластерного аналізу основні поняття й принципи теорії самоорганізації моделей на основі методу групового урахування аргументів (МГУА) [16, 17]. Самоорганізацією кластеризації називається їх перебір з метою вибору оптимальної кластеризації. Що більша неточність даних, то простіша оптимальна кластеризація (складність вимірюється числом кластерів і числом ознак). В алгоритмах об'єктивного кластерного аналізу (ОКК) кластери утворюються за внутрішнім критерієм (що складніше, то точніше), а оптимальні їх кількість і склад ансамблю ознак визначаються за зовнішнім критерієм (утворює мінімум в області недоускладненої кластеризації, що є оптимальною для заданого рівня дисперсії завад) [15]. Перебір варіантів кластеризації реалізує алгоритм ОКК [18]. Побудова ієрархічного дерева кластеризації впорядковує та скорочує перебір, причому оптимальна за критерієм кластеризація не втрачається [19]. Фізичну кластеризацію знаходять за критерієм балансу кластеризації. Для обчислення критерію вибірку даних поділяють

на дві рівні частини. На кожній підвибірці будується дерево кластеризації, і на кожному кроці розраховується критерій балансу при однаковому числі кластерів. Критерій вимагає знайти кластеризацію, при якій збігатимуться як число, так і координати центрів (середніх точок) кластерів, що відповідають один одному [20]:

$$BL = \frac{1}{MK}\sum_{j=1}^{M}\sum_{i=1}^{K}(x_{oA} - x_{oB})^2 \to \min, \tag{5}$$

де K – число кластерів на даному кроці побудови дерева; M – число координат; x_{oA} – координати центрів кластерів, що побудовані на частині A; x_{oB} – координати центрів кластерів, що побудовані на частині B.

У роботах [21–23] автор запропонував вдосконалений метод індуктивного кластерного аналізу, який можна застосувати як самостійний інструмент при розв'язанні задач прогнозування поведінки складних систем в умовах невизначеності, так і для розв'язання задач кластеризації у широкому сенсі, тобто паралельного виконання кластеризації та вибору підмножин інформативних ознак. З позицій загальної постановки задачі кластерного аналізу в широкому розумінні, тобто вирішення комплексного завдання з: (1) побудови оптимальної кластеризації та з (2) одночасного конструювання оптимального ансамблю інформативних ознак таку задачу формально можна сформулювати так.

Нехай загальний масив вхідних даних має вигляд:

$$\overset{\#}{X} = \left(x_{0j} \,\mathbb{M}\, x_{ij} \in X\right),\ j = \overline{1,m}, i = \overline{1,n}\ , \tag{6}$$

де $\left\{x_{0j},\ j = 1,...,m\right\}$ – значення цільових ознак заданих об'єктів, наприклад рейтингів m експертів у задачі виділення гомогенних груп експертів; $x_{ij} \in X$ $(i = 1,...,n,\ \ j = 1,...,m)$ – простір ознак.

Необхідно:

1) синтезувати підмножину $\left\{x_h^*\right\} = X^* \subset X$, $h = 1,...,n^*, n^* \le n$ із зазначених ознак, найкращу за заданим критерієм оптимальності та яка давала б змогу:

2) класифікувати всіх учасників експертних змагань (у вказаній задачі) на $k < m$, $k = 1,...,\ \mathrm{K}$ однорідних груп за результатами відбіркової кваліфікаційної сесії та:

3) відібрати єдину групу (ЕКВР) k^* із $m^* < m$ експертів, що відповідають вимогам та проблематиці конкретного проекту за заданими критеріями.

Метод відбору експертів до ЕКВР можна подати так.

Крок 1. Попереднє формування групи експертів $w_k \in \Omega$, для яких їхній рейтинг x_0 є не нижчим за прийнятий для проекту рівень.

Крок 2. Поділ 6 [22] на дві непересічні підмножини Ω^A й Ω^B при цьому: $\Omega^A \cup \Omega^B = \Omega,\ \Omega^A \cap \Omega^B = \varnothing$, де $w_{ij} \in \Omega, i = 1,...,n, j = 1,...,m$ – відповідає опису j-го експерта, який зарахований до участі в конкурсних змаганнях. Сформована таким чином загальна матриця даних $\overset{\#}{X}$ матиме такий вигляд:

$$\overset{\#}{X} = \left[\left(x_{0j} \,\mathbb{M}\, X\right)^A \mathbf{N} \left(x_{0j} \,\mathbb{M}\, X\right)^B\right],$$
$$j = 1,...,m^A = m^B,\ \ m^A + m^B = m. \tag{7}$$

Крок 3. Налаштування однієї з відомих процедур кластеризації (наприклад, класичного алгоритму *k-means* або іншого) та кластеризація об'єктів $w_k \in \Omega$ за допомогою вибраного алгоритму незалежно на підмножинах Ω^A і Ω^B в просторі $\overset{\#}{X}$ за однією із схем багаторядних

алгоритмів індуктивного моделювання складних систем (ІМСС) з індуктивним нарощуванням кількості ознак в їх ансамблях. При цьому може застосовуватися простий критерій вигляду:

$$D^2(\hat{m}) = \sum_{k=1}^{K} (\hat{m}_k^A - \hat{m}_k^B)^2 \ . \tag{8}$$

Правило зупинки: індуктивна процедура добігає кінця за умови:

$$D^2(\hat{m})_s \le D^2(\hat{m})_{s+1} \ , \tag{9}$$

де s – ряд селекції в термінах ІМСС.

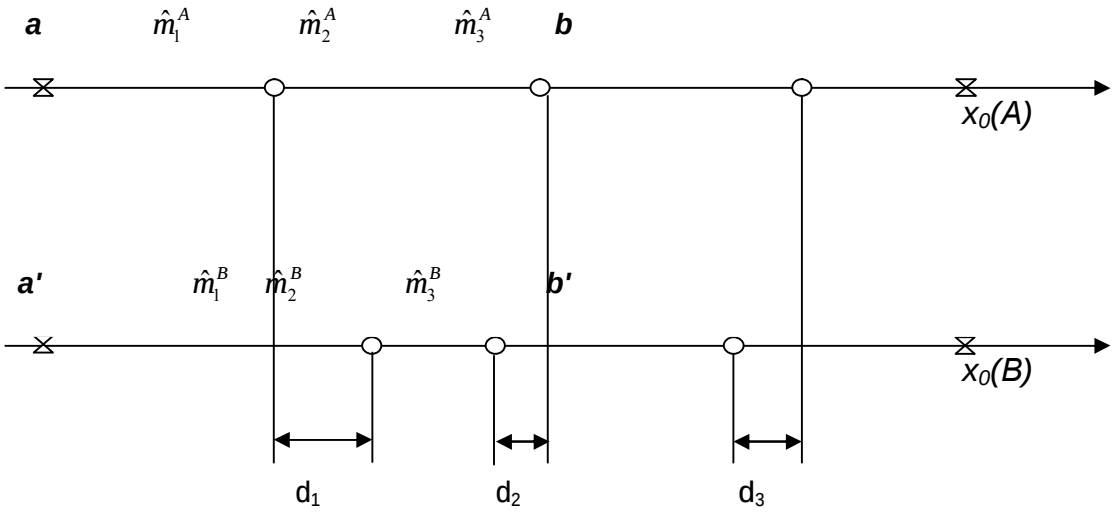

Рис. 1. До принципу роботи критерію несуперечності кластеризацій для $n_k = 3$

При цьому фіксується кількість виділених кластерів $k^{*(A)} = k^{*(B)} = K^*$ і підпростір ознак:

$$\{x_l^*\} = X^*, \ l = 1,...,n^*, \ n^* \le n \ , \tag{10}$$

Критерій [10] вимагає, щоб сума квадратів відхилень між центрами "колективів" експертів (кластерів), що згруповані за рейтинговим показником (по осі цільової ознаки x_0) та встановлені незалежно на підмножинах Ω^A і Ω^B, була мінімальною.

Існує багато методів, що використовуються в задачах кластеризації. Проте процеси, які зустрічаються на практиці, є значно складнішими, ніж ті, на які орієнтовані розроблені методи. Крім того, дані можуть містити неоднорідні спостереження та невідомі взаємозв'язки між змінними. Все це призводить до того, що немає однозначних відповідей на питання вибору якнайкращого методу кластеризації. Індуктивні алгоритми МГУА дають можливість автоматично знаходити взаємозалежності в даних, вибирати оптимальну структуру кластерів та налаштовувати (оцінювати) параметри. Тому перспективним є розвиток методів індуктивної кластеризації, оскільки вони мають значні переваги. Однак, крім переваг вони мають і деякі обмеження та недоліки, що вимагає пошуку шляхів їх вдосконалення.

Щільнісний алгоритм кластеризації просторових даних з присутністю шуму DBSCAN

Як було сказано вище, алгоритм DBSCAN являє собою щільнісний алгоритм, призначений для кластеризації просторових даних, в яких присутній шум. Він був запропонований в роботі [23] для вирішення проблеми розбиття спочатку просторових даних на кластери довільної форми. Більшість алгоритмів, які виробляють плоске розбиття, створюють кластери, які за формою близькі до сферичних, оскільки мінімізують відстань документів до центру кластера [24].

Автори DBSCAN експериментально показали, що їхній алгоритм здатний розпізнати кластери різної форми. Основна ідея, покладена в основу алгоритму, полягає в тому, що всередині кожного кластера спостерігається типова щільність точок (об'єктів), яка помітно вища, ніж щільність зовні кластера, а також щільність в областях з шумом нижча за щільності будь-якого з кластерів. Тобто, для кожної точки кластера її сусідство заданого радіуса повинно містити не менше деякої кількості точок, яке задається граничним значенням [25].

В основу цього алгоритму покладено кілька визначень [25]:

- e -околом об'єкта називається окіл радіуса e деякого об'єкта;
- кореневим об'єктом називається об'єкт, e -окіл якого містить не менше деякого мінімального числа MinPts об'єктів;
- об'єкт p безпосередньо щільнісно досяжний з об'єкта q, якщо p знаходиться в e -околі q і q є кореневим об'єктом;
- об'єкт p щільнісно досяжний з об'єкта q при заданих e і параметра MinPts, якщо існує послідовність об'єктів $p, \mathbf{K}, p$, де $p = q$ p = q і $p = p$, така, що $p+1$ безпосередньо щільнісно досяжний з p, $1 \le i \le n$;
- об'єкт p щільнісно з'єднаний з об'єктом q при заданих e і MinPts, якщо існує об'єкт o такий, що p і q щільнісно досяжні з o.

Для пошуку кластерів алгоритм DBSCAN перевіряє e -окіл кожного об'єкта. Якщо e -окіл об'єкта p містить більше точок, ніж MinPts, то створюється новий кластер з кореневим об'єктом p. Потім DBSCAN ітеративно збирає об'єкти безпосередньо щільнісно досяжні з кореневих об'єктів, які можуть привести до об'єднання декількох щільнісно досяжних кластерів. Процес завершується, коли до жодного кластера не можна додати жодного нового об'єкта [26].

Хоча алгоритм DBSCAN не вимагає заздалегідь визначати кількість кластерів, може виникнути потреба у вказівках значень параметрів e і MinPts, які безпосередньо впливають на результат кластеризації. Оптимальні значення цих параметрів складно визначити, особливо для багатовимірних просторів даних. Крім того, розподіл даних у таких просторах часто несиметричний, що не дозволяє використовувати для їх кластеризації глобальні параметри щільності.

Робота алгоритму DBSCAN зводиться до наступного.

Вхід: множина об’єктів S, *Eps* і *MinPt.*

Об’єкт може бути в одному із трьох станів:

1. Не зазначений.
2. Зазначений, що не є внутрішнім об’єктом ніякого кластера.
3. Віднесений до деякого кластера.

Крок 1. Встановити для всіх елементів множини S прапор “не зазначений”. Присвоїти поточному кластеру C_j нульовий номер, $j = 0$. Множині шумових точок Noise = 0.

Крок 2. Для кожного $s_i \in S$ такого, що прапор (s_i) = “не зазначений”, виконати:

Крок 3. Прапор (s_i) := “зазначений”;

Крок 4. $N_i = N_{Eps}(s_i) = \{q \in S | dist(s_i, q) \le Eps\}$

Крок 5. Якщо $|s_i| < MinPt$, то

$Noise = Noise + \{s_i\}$

Інакше номер наступного кластера $j = j + 1$;

EXPANDCLUSTER $(s_i, N_i, C_j, Eps, MinPt)$;

Вихід: множина кластерів $C = (C_j)$.

EXPANDCLUSTER

Вхід: поточний об'єкт s_i, його – сусідство N_i, поточний кластер C_j і $Eps, MinPt$.

Крок 1. $C_j = C_j + \{s_i\}$;

Крок 2. Для всіх крапок $s_k \in N_i$:

Крок 3. Якщо прапор (s_k) = "не зазначений", то

Крок 4. Прапор (s_k) = "зазначений";

Крок 5 . $N_{ik} = N_{Eps}(s_k)$;

Крок 6. Якщо $|N_{ik}| \geq MinPt$, то $N_i = N_i + N_{ik}$;

Крок 7. Якщо $p : s_k \in C_p, p = \overline{1,(C)}\}$, те $C_j = C_j + \{s_k\}$;

Вихід: кластер C_j .

Як показують дослідження [27], розглянутий алгоритм кластеризації володіє рядом переваг, які дають можливість використовувати цей метод для роботи з кластерами різної природи (форми). Застосування даного алгоритму дозволяє працювати з вибірками великого обсягу і дає можливість працювати з *n*-мірними об'єктами (це об'єкти, кількість атрибутів яких понад 3, за умови адекватного вибору функції для розрахунку відстані (у загальному випадку можна використовувати метрику Міньковського). Проте, істотним недоліком є достатньо трудомістка процедура визначення необхідних параметрів для коректної роботи алгоритму. Детальніше переваги і недоліки алгоритму DBSCAN наведено в табл. 1.

Таблиця 1

Переваги і недоліки алгоритму DBSCAN

Переваги	Недоліки
DBSCAN може знаходити кластери довільної форми	DBSCAN є не зовсім детермінованим алгоритмом: межові точки, які досяжні з більш ніж одного кластера, можуть бути частиною іншого кластера, залежно від порядку опрацювання даних.
DBSCAN стійкий до шуму і викидів, тобто всі викиди виносяться в окремий кластер	Якість роботи DBSCAN залежить від використовуваної міри відстані. Найчастіше використовується евклідова відстань. Але для багатовимірних даних цей показник може виявитися майже марним через так зване "прокляття розмірності" (важко знайти відповідне значення для ε). Цей ефект, однак, також присутній в будь-яких інших алгоритмах, основаних на евклідовій відстані.
Не вимагає апріорного задання кількості кластерів, на відміну від алгоритму К-середніх	DBSCAN не може скопіювати дані, які мають велику різницю у щільності, так як комбінації minPts ε не можуть бути обрані відповідним чином для всіх кластерів.
Використовує лише два параметри, в основному не чутливі до впорядкованості точок в базі даних	Якщо дані і масштаб не дуже добре зрозумілі, вибирати значущу відстань порогу ε може бути дуже важко.
Дозволяє працювати з вибірками даних великої розмірності	Істотним недоліком є доволі трудомістка процедура визначення необхідних параметрів для коректної процедури алгоритму.
Визначення параметрів minPts і ε дозволяють працювати з *n*-мірними об'єктами за умови адекватного вибору функції для розрахунку відстані	

Розроблення гібридного алгоритму

Як показали результати досліджень [28], застосування однорідних методів для вирішення складних проблем далеко не завжди приводить до успіху. У гібридній архітектурі, що поєднує кілька парадигм, ефективність одного підходу може компенсувати слабкість іншого. Комбінуючи різні підходи, можна уникнути недоліків, що властиві кожному окремо. Гібридні алгоритми, як правило, складаються з різних компонентів, об'єднаних в інтересах досягнення поставлених цілей. Інтеграцією та гібридизацією різних методів та інформаційних технологій можна вирішувати складні завдання, які неможливо вирішити на основі будь-яких окремих методів або технологій. При цьому у разі інтеграції різнорідних інформаційних технологій слід очікувати синергетичних ефектів вищого порядку, ніж при об'єднання різних моделей у межах однієї технології.

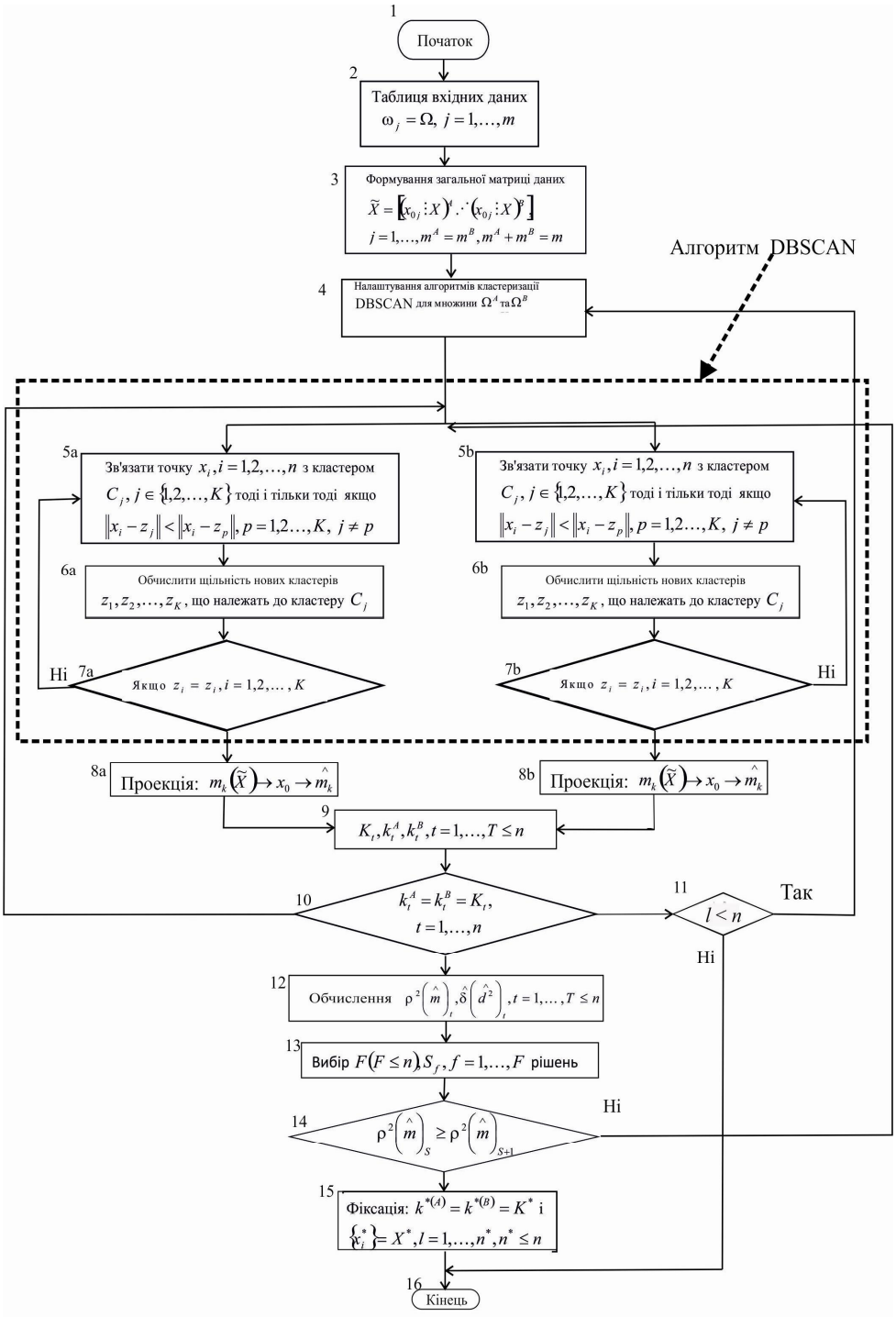

Рис. 2. Схема індуктивного методу кластер-аналізу із застосуванням алгоритму DBSCAN

У загальному вигляді процедура зводиться до такого.

Крок 1. Поділ початкової таблиці даних на дві частини *A i B* (Ω^A і Ω^B) за вимогами методології індуктивного моделювання складних систем. Підготовлена загальна матриця даних $\overset{\#}{X}$ матиме такий умовний вигляд (припустимо, що m – парне):

$$\overset{\#}{X} = \left[\left(x_{0j} \mathbb{M} X \right)^A \mathbf{N} \left(x_{0j} \mathbb{M} X \right)^B \right], \quad j = 1, ..., m^A = m^B, \; m^A + m^B = m. \tag{11}$$

Крок 2. Налаштування процедури кластеризації за DBSCAN алгоритмом.

Крок 3. Кластеризація об'єктів $w_k \in \Omega$ за допомогою вибраного і уже налаштованого алгоритму незалежно на підмножинах Ω^A і Ω^B в просторі X за однією з класичних схем алгоритмів МГУА з індуктивним нарощуванням кількості ознак в їх ансамблях. Багаторядна індуктивна процедура кластеризації може бути такою.

1-й ряд селекції:

1.1) кластеризація об'єктів на підмножинах Ω^A і Ω^B за ансамблями $\{x_i\}$, $i = 1, ..., n$;

1.2) проектування центрів отриманих кластерів на вісь x_0;

1.3) для кластеризацій, в яких виконується умова $k_t^A = k_t^B = K_t$ (t – поточний номер кластеризації, $k_t^{(\cdot)}$ – кількість кластерів в t–й кластеризації), обчислюються значення критерію оптимальності $r^2(\hat{m})$.

2-й ряд селекції:

2.1) кластеризація об'єктів на підмножинах Ω^A і Ω^B за ансамблями $\{x_i, x_j\}$, $i, j = 1, ..., n$, $i \neq j$;

2.2) виконуються п.п. (1.2) – (1.3) і за критерієм якості кластеризації відбираються F $(F \leq n)$ кращих кластеризацій S_f та відповідних ансамблів ознак X_f, $f = 1, ..., F$.

3-й і наступні ряди селекції:

3.1) кластеризація об'єктів на підмножинах Ω^A і Ω^B за ансамблями $\{X_f, x_l\}$, $f = 1, ..., F$, $l = 1, ..., n$ за умови, що ознака з індексом l не присутня в уже створених ансамблях X_f.

3.2) виконується п. (2.2).

Правило зупинки: індуктивна процедура зупиняється за умови:

$$D^2(\hat{m})_s \leq D^2(\hat{m})_{s+1}, \tag{12}$$

де s – ряд селекції в термінах МГУА. При цьому фіксується значення $k^{*(A)} = k^{*(B)} = K^*$, $K^* \leq m/2$ і підпростір інформативних ознак $\{x_l^*\} = X^*$, $l = 1, ..., n^*$, $n^* \leq n$, а $D^2(\hat{m})_s$ – значення системного критерію якості кластеризації (8), який задається залежно від розв'язуваної задачі та природи даних.

Отже, у результаті розв'язання задачі кластеризації в широкому сенсі маємо:

– синтезовану підмножину $\{x_h^*\} = X^* \subset X$, $h = 1, ..., n^*, n^* \leq n$ із усіх заданих з експерименту ознак, що є найкращою за заданим критерієм оптимальності і яка дозволяє:

– класифікувати всі об'єкти з Ω на $k < m$, $k = 1, ..., K$ однорідних груп.

Експерименти

Для проведення експериментів було відібрано тестові дані з бази даних обчислювальної школи Східнофінського університету [29], що мають складну двовимірну просторову форму.

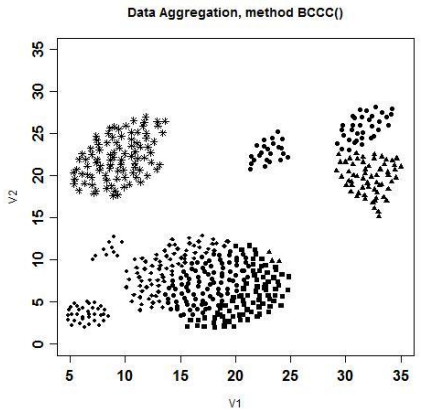

Рис. 3. Дані Aggregation: класів – 7; розмірність – 2; кількість екземплярів – 788

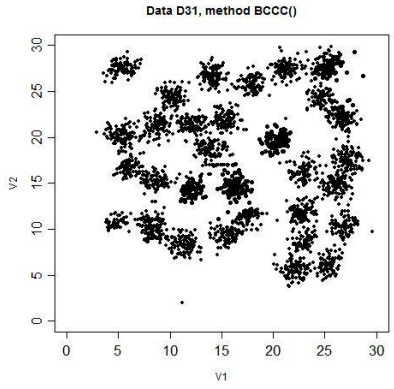

Рис. 4. Дані D31: класів – 31; розмірність – 2; кількість екземплярів – 3100

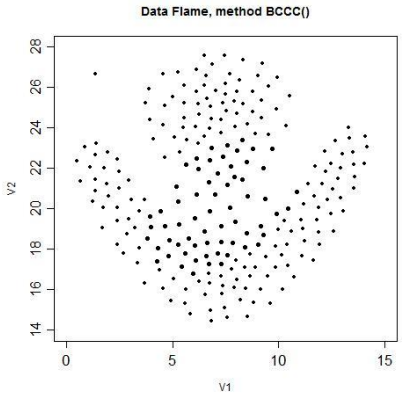

Рис. 5. Данні Flame: класів – 2; розмірність – 2; кількість екземплярів – 240

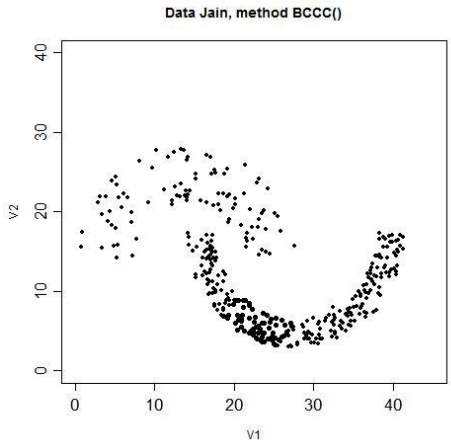

Рис. 6. Дані Jain: класів – 2; розмірність – 2; кількість екземплярів – 373

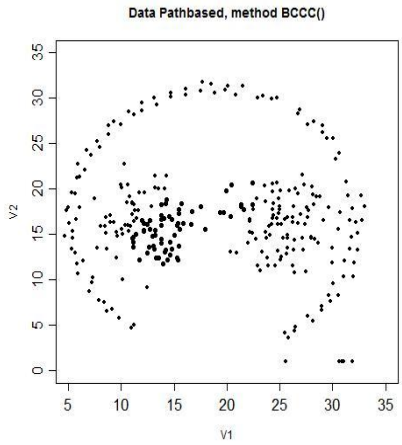

Рис. 7. Дані Pathbased: класів – 3; розмірність – 2; кількість екземплярів – 300

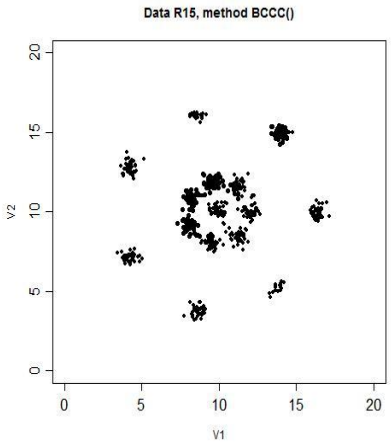

Рис. 8. Дані R15: класів – 15; розмірність – 2; кількість екземплярів – 600

Для порівняльних досліджень було використано алгоритми кластеризації з пакета WEKA. Результати наведено у табл. 2.

Таблиця 2

Результати порівняльних експериментів (відсоток правильно розпізнаних даних) кластеризації

Дані	Індуктивний DBSCAN	DBSCAN	K-середніх	EM -алгоритм	COWEB
Aggregation	98,5	91,0	85,5	88,1	90,1
D31	99,1	94,3	81,3	85,6	89,2
Flame	93,7	88,4	78,2	80,3	85,4
Jain	99,0	95,0	76,0	78,8	92,4
Pathbased	98,2	96,0	80,5	80,2	89,1
R15	100,0	96,8	89,0	93,1	91,3

Висновки

Вивчений та реалізований щільнісний алгоритм кластеризації просторових даних із присутністю шуму DBSCAN [1]. Основна перевага цього алгоритму полягає в тому, що він підходить для виділення кластерів довільної форми. Обчислювальна складність алгоритму $O(n^2)$. Основний його недолік – це можливість виникнення проблеми, якщо щільність різних кластерів значно відрізняється. Представляє інтерес використання алгоритмів кластеризації, що враховують динамічну природу масиву даних.

1. Ester, Martin; Kriegel, Hans-Peter; Sander, Jörg; Xu, Xiaowei (1996). Simoudis, Evangelos; Han, Jiawei; ayyad, Usama M., eds. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96). AAAI Press. pp. 226–231. ISBN 1-57735-004-9.CiteSeerX: 10.1.1.71.1980. 2. Ивахненко А. Г. Объективная кластеризация на основе теории самоорганизации моделей // Автоматика. – 1987. – № 5. – С. 6–15. 3. Загоруйко Н. Г. Когнитивный анализ данных / Рос. акад. наук, Сиб. отд-ние, Ин-т математики им. С. Л. Соболева. – Новосибирск : Гео, 2013. – 186 с. 4. Ивахненко А. Г. Метод групового урахування аргументів – конкурент методу стохастичної апроксимації // Автоматика. – 1968. – № 3. – С. 58–72. 5.Степашко В. С. Елементи теорії індуктивного моделювання / Стан та перспективи розвитку інформатики в Україні: монографія / Колектив авторів. – К.: Наукова думка, 2010. – 1008 с. – С. 471–486. 216. 6. Степашко В. С. Самоорганизация прогнозирующих моделей сложных процессов и систем. – ХУ Всероссийская научно-техническая конференция – Нейроинформатика – 2013: Лекции по нейроинформатике / Ю. В. Тюменцев – отв. ред. – М.: НИЯУ МИФИ, 2013. – 320 с. – С. 150–170. 7. Ivakhnenko A. G. Heuristic Self-Organization in Problems of Engineering Cybernetics. – Automatica. – № 6. – 1970. – P. 207–219. 8. Ивахненко А. Г. Системы эвристической самоорганизации в технической кибернетике. – К.: Техніка, 1971. – 392 с. 9. Ивахненко А. Г. Помехоустойчивость моделирования / А. Г. Ивахненко, В. С. Степашко. – К.: Наукова думка. – 1985. – 216 с. 10. Габор Д. Перспективы планирования // Автоматика. 1972. – №2. – С. 16–22. 11. 217. Степашко В. С. Теоретические аспекты МГУА как метода индуктивного моделирования / В. С. Степашко // УСиМ. – 2003. – №2. – 31–38. 12. Степашко В. С. Метод критических дисперсий как аналитический апарат теории індуктивного моделирования / В. С. Степашко // Проблемы управления и информатики. – 2008. – №2. – С. 8–26. 13. Висоцький В. М. Про найкращий поділ вхідних даних в алгоритмах МГУА / В. М. Висоцький // Автоматика. – 1976. – № 3. – С. 71–74. 14. Юрачковский Ю. П. Оптимальное разбиение исходной выборки данных на обучающую и проверочную последовательности на основе анализа функции распределения критерия / Ю. П. Юрачковский, А. Н. Горшков // Автоматика. — 1980. – № 2. – С. 51–59. 15.Сарычева Л. В. Объективный кластерный анализ данных на основе МГУА // Проблемы управления и информатики. – 2008. – № 2. – С. 86–104. 16. Ивахненко А. Г. Объективная кластеризация на основе теории самоорганизации моделей // Автоматика. – 1987. – № 5. – С. 6–15. 17. Ивахненко А. Г. Алгоритмы метода группового учета аргументов (МГУА) при непрерывных и бинарных признаках / Препр. Ин-т кибернетики им. В. М. Глушкова. – К., 1992. – 49 с. 18. Madala H. R. and Ivakhnenko A. G. Inductive Learning Algorithms for Complex Systems Modeling. CRC Press Inc., Boca Raton, 1994. 19. Zholnarsky A. A. Agglomerative Cluster Analysis Procedures for Multidimensional Objects: A Test for Convergence. Pattern Recognition and Image Analysis, 1992, vol. 25, no.4, pp. 389–390. 20. Ivakhnenko A. G. and Ivakhnenko G. A. The Review of Problems Solvable by Algorithms of the Group Method of Data Handling (GMDH). Pattern Recognition and Image Analysis, 1995, vol. 5, no.4. P. 527–535. 21. Осипенко В. В. Индуктивный алгоритм кластер-анализа в инструментарии системных информационно-аналитических исследований / В. В. Осипенко // Управляющие системы и машины, №2, 2013. – С. 26–32. 22. Wojcik W. The use of inductive clustering algorithms for forming expert groups in large-scale innovation projects / W. Wojcik, V. Osypenko, V. Lytvynenko // Electronika. Instytut Electroniki i Technik Informacyjnych Politechnika Lubelska. – 2013. – No. 8. – P. 45–49. 23.Осипенко В. В., Литвиненко В. І., Лурьє І. А. Комбіноване використання

індуктивного алгоритму кластеризації та алгоритму k-середніх // Інтелектуальні системи прийняття рішень і проблеми обчислювального інтелекту: зб. наук. праць за матеріалами міжнар. наук. конф. – Херсон: ХНТУ, 2014. – С. 309–311. 24. Ester M., Kriegel H.-P., Sander J. and Xu X. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, OR, 226-231. 25. Henrik Bäcklund, Anders Hedblom, Niklas Neijman A Density-Based Spatial Clustering of Application with Noise. – Linköpings Universitet. – 2011. 26. http://habrahabr.ru/post/164417. 27. Чапланов А. П., Чапланова Е. Б. Кластеризация объектов с помощью алгоритма DBSCAN / Системи обробки інформації. – 2006. – Вип. 9. – С. 82–84. 28. Литвиненко В. И. Гибридные искусственные иммунные системы и мягкие вычисления (обзор) / Індуктивне моделювання складних систем: зб. наук. пр. – К.: МННЦ ІТС НАН та МОН України, 2009. – Вип. 1. – С. 114–130. 29. https://cs.joensuu.fi/sipu/datasets/.

УДК 004.652.4+004.827

Н. І. Мельникова
Національний університет "Львівська політехніка",
кафедра "Інформаційні системи та мережі"

ОСОБЛИВОСТІ ОПРАЦЮВАННЯ МЕДИЧНОЇ ІНФОРМАЦІЇ ДЛЯ СИСТЕМ ПІДТРИМКИ ПРИЙНЯТТЯ ЛІКУВАЛЬНИХ РІШЕНЬ

Проаналізовано основні підходи до опрацювання персоналізованої медичної інформації для прийняття лікувальних рішень і як засіб її реалізації запропоновано архітектуру системи підтримки прийняття лікувальних рішень, проаналізовано отримані результати. Окреслено основні етапи аналізу медичної інформації засобами системи підтримки прийняття лікарських рішень, що дають змогу декомпонувати керівні процеси і описують відношення між керівними потоками та деталізують послідовність використання методів представлення даних у системі.

Ключові слова: система підтримки прийняття рішень, лікувальна експертна система, архітектура інформаційної технології.

The article has analyzed the main approaches to the processing of personalized medical information for medical decision making and a tool of implementing it has proposed the system of support medical decision-making, the architecture of this system and the analysis of the results have been done. Here the main stages of analysis of medical information by tools of medical decision support solutions are presented. That enables the decomposition of control process and describes the relationship between management of streams and detalisation of sequence of used methods of presenting data in the system.

Key words: support decision making system, treatment expert system, architecture of information technology.

Вступ

Сьогодні розвиток інформаційних технологій в медицині вражає своїми швидкими темпами, а саме впровадження нанотехнологій для реабілітації та лікування хворих; інтелектуальних агентів для діагностування захворювань; систем підтримки прийняття рішень, що підвищують об'єктивність дій у процесах діагностики чи вибору лікування хворих з різними патологіями. Все це підвищує якість надання медичної допомоги, але залишається ще ряд завдань, вирішення яких сприяло б оптимізації оцінювання та опрацювання персоналізованих медичних даних для процесу підтримки прийняття лікувальних рішень.