ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

УДК 004.825

к.т.н. **Волосюк Ю.В.** (ЄУ, м. Миколаїв)

АНАЛІЗ АЛГОРИТМІВ КЛАСТЕРИЗАЦІЇ ДЛЯ ЗАДАЧ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ

В статі представлена загальна характеристика процедури кластерного аналізу. Наведено огляд існуючих підходів до вирішення задачі кластеризації та математичних методів кластерного аналізу даних.

Описані етапи процесу кластеризації, розглянуті питання вибору міри відстані і ваг для класифікуючих властивостей об'єктів. Проведено класифікацію та аналіз існуючих алгоритмів кластерного аналізу, розглянуті переваги та недоліки цих алгоритмів. Обґрунтовано доцільність використання карт Кохонена в методиках кластеризації з метою дослідження наявності чи відсутності кластерної структури в даних, числа кластерів, законів сумісного розподілу ознак, залежностей тощо. Надано порівняльну таблицю алгоритмів.

Зроблено висновок щодо необхідності подальшого розроблення простих в реалізації алгоритмів, які потребують мінімальної кількості початкових параметрів, дозволяють проводити багатоваріантний аналіз і дають задовільні результати.

Ключові слова*: *кластерний аналіз, функція відстані, пошук центрів, алгоритм порогової величини, алгоритм k-середніх.

Постановка проблеми. Стрімке посилення потоків та зростання об'ємів інформації в навколишньому світі спонукає сучасні технології до актуального завдання підвищення ефективності пошуку необхідної інформації в глобальному інформаційному просторі. Це завдання вимагає дослідження та розроблення методів і алгоритмів розподілу інформаційних моделей об'єктів на певні групи і класи. Завдання такого роду виникають в таких сучасних інформаційних технологіях як *Data Mining, Text Mining, Web Mining*, розпізнавання образів, машинне навчання. В цьому напряму науковий інтерес представляють методи кластеризації. Важливість цих методів полягає в тому, що вони дозволяють виділити групи інформаційних об'єктів, близьких за певними ознаками, без будь-якої попередньої інформації про розподіл інформаційних об'єктів на групи.

Кластеризація в *Data Mining* здобуває цінність за умови, коли вона виступає одним з етапів аналізу даних, побудови закінченого аналітичного розв'язку. Аналітику частіше легше виділити групи схожих об'єктів, вивчити їхні особливості й побудувати для кожної групи окрему модель, чим створювати одну загальну модель для всіх даних, що висуває проблему аналізу алгоритмів кластеризації для задач інтелектуального аналізу даних в розряд актуальних.

Аналіз останніх досліджень і публікацій. Найважливіший внесок у розвиток методів і алгоритмів кластеризації внесли такі вчені як: С. А. Айвазян в частині розроблення класифікації багатовимірних спостережень, В. М. Бухштабер в процесі розроблення ряду методів автоматичної класифікації на основі математичних моделей, І. С. Єнюков – методів кластеризації об'єктів із категоризаційними ознаками, Л. Д. Мешалкін – локальних методів класифікації, І. Д. Мандель – в процесі дослідження ряду функцій оцінки якості кластеризації [2]. Цими науковцями розроблено ряд методів і алгоритмів кластеризації та класифікації багатовимірних інформаційних моделей об'єктів.

Формулювання цілей статті. Метою статті є опис математичних методів кластерного аналізу даних та етапів процесу кластеризації, аналіз існуючих алгоритмів проведення кластерного аналізу, розгляд переваг та недоліків алгоритмів кластеризації.

Виклад основного матеріалу дослідження. Кластеризація (або кластерний аналіз) – представляє собою сукупність математичних методів, призначених для формування відносно

«віддалених» один від одного груп «близьких» між собою об’єктів за інформацією про відстані або зв’язки (міри відстаней) між ними.

Завдання кластеризації відноситься до статистичної обробки, а також до широкого класу завдань навчання без вчителя. Головна відмінність кластеризації від класифікації полягає в тому, що перелік груп чітко не заданий і визначається в процесі роботи алгоритму.

Застосування кластерного аналізу в загальному виді зводиться до наступних етапів [1]:

1. Ідентифікація вибірки об'єктів для кластеризації.
2. Визначення множини змінних, за якими планується проводити оцінку об'єктів у вибірці. При необхідності – нормалізація значень змінних.
3. Обчислення значень тієї або іншої міри схожості між об'єктами.
4. Застосування одного з методів кластерного аналізу для створення груп подібних об'єктів (кластерів).
5. Перевірка вірогідності результатів кластерного розв'язку.

Після одержання й аналізу результатів можливе коригування обраної метрики й методу кластеризації до одержання оптимального результату.

Міри відстаней. Вузловим моментом в кластерному аналізі даних вважається вибір метрики (або міри близькості об’єктів). Подібність або розходження між об’єктами класифікації встановлюється в залежності від обраної метричної відстані між ними. Проблема подібності полягає не в простому віднесенні об'єктів до тих або інших класів, а в тому, що такий розподіл повинен задовольняти критеріям наукового знання. Кількісне визначення подібності опирається на поняття метрики. При такому підході об'єкти представляються крапками в багатомірному координатному просторі, причому подібності й відмінності між ними визначаються із метричних відстаней. Розмірність простору визначається числом змінних, що описують об'єкт.

При класифікації використовуються різні міри відстані [3; 4]. В процесі розгляду методичних підходів визначення «подібності» об’єктів з’ясовано, що для початку потрібно скласти вектор характеристик для кожного об'єкта (як правило, це набір числових значень, наприклад, ріст-вага людини). Однак існують також алгоритми, що працюють із якісними (тобто категорійними) характеристиками. Після того, як визначено вектор характеристик, доцільно провести нормалізацію, з метою отримання однакового внеску усіма компонентами при розрахунках «відстаней». У процесі нормалізації всі значення приводяться до деякого діапазону, наприклад, [-1,-1] або [0,1]. Нарешті, для кожної пари об'єктів визначається «відстань» між ними (ступінь подібності). Серед сукупності метрик виділимо наступні:

1. Евклідова відстань. Найпоширеніша функція відстані. Являє собою геометричну відстань в багатомірному просторі:

$$\rho(X,Y)=\sqrt{\sum_{i}^{n}(x_i-y_i)^2}.$$

З геометричної точки зору, евклідова міра відстані може виявитися безглуздою, якщо ознаки визначені в різних одиницях. Для коригування ситуації, вдаються до нормування кожної ознаки. Застосування евклідової відстані виправдано в наступних випадках [4]:

а) властивості (ознаки) об’єкта однорідні за фізичним змістом і однаково важливі для класифікації;
б) ознаковий простір збігається з геометричним простором.

2. Квадрат евклідової відстані. Застосовується для додання більшої ваги об’єктам, що максимально віддалені один від одного. Ця відстань обчислюється наступним чином:

$$\rho(X,Y)=\sum_{i}^{n}(x_i-y_i)^2.$$

3. Відстань міських кварталів (манхеттенська відстань). Ця відстань є середньою різниць по координатах. У більшості випадків ця міра відстані приводить до таких самих результатів, як і для звичайної евклідової відстані. Однак для цієї міри вплив окремих більших різниць (викидів) зменшується через те, що вони не зводяться у квадрат. Формула для розрахунків манхеттенської відстані наступна:

$$\rho(X,Y) = \sum_{i}^{n} |x_i - y_i|.$$

4. Відстань Чебишева. Ця відстань може виявитися корисною, коли потрібно визначити два об'єкти як «різні», якщо вони різняться за будь-якою однією координатою. Відстань Чебишева обчислюється за формулою:

$$\rho(\mathrm{X},\mathrm{Y}) = \max(|x_i - y_i|).$$

5. Степенева відстань. Застосовується у випадку, коли необхідно збільшити або зменшити вагу, що відноситься до розмірності, для якої відповідні об'єкти значно відрізняються. Степенева відстань обчислюється за наступною формулою:

$$\rho(\mathrm{X},\mathrm{Y}) = \sqrt[r]{\sum_{i}^{n} (x_i - y_i)^p},$$

де r і p – параметри, обумовлені користувачем. Параметр p відповідальний за поступове зважування різниць по окремих координатах, параметр r відповідальний за прогресивне зважування більших відстаней між об'єктами. Якщо обидва параметри – r і p рівні двом, то ця відстань збігається з відстанню Евкліда.

Вибір метрики у повній мірі залежить від дослідника, оскільки результати кластеризації можуть суттєво відрізнятися при використанні різних мір. Таким чином, вибір міри відстані і ваг для класифікуючих властивостей – дуже важливий етап, тому що від цих процедур залежать склад і кількість формованих класів, а також ступінь подібності об’єктів всередині класів.

Класифікація алгоритмів. На рис. 1 представлена класифікація алгоритмів та методів кластерного аналізу. Сутність ієрархічних агломеративних методів полягає у тому, що на першому кроці кожний об’єкт вибірки розглядається як окремий кластер. Процес об’єднання кластерів відбувається послідовно: на підставі матриці відстаней або матриці подібності поєднуються найбільш близькі об’єкти. Послідовність об’єднання легко піддається геометричній інтерпретації й може бути представлена у вигляді графа-дерева (дендрограми) [5]. Основною передумовою ієрархічних дивізивних методів є те, що спочатку всі об’єкти належать до одного кластеру. У процесі класифікації за певними правилами поступово від цього кластера відокремлюються групи схожих між собою об’єктів. Так, на кожному кроці кількість кластерів зростає, а міра відстані між кластерами зменшується. Складнощі ієрархічних методів кластеризації наступні: обмеження обсягу набору даних, вибір міри близькості, негнучкість отриманих класифікацій. Перевага цієї групи методів порівняно з неієрархічними методами полягає у їх наочності і можливості отримання детального уявлення про структуру даних. При використанні ієрархічних методів існує можливість досить легко ідентифікувати викиди в наборі даних і, в результаті, підвищити якість даних. Велика кількість методів ієрархічного кластерного аналізу різниться не тільки використаними мірами подібності (розходження), але й алгоритмами класифікації.

Неієрархічні методи виявляють більш високу стійкість по відношенню до викидів, невірного вибору метрики, включення незначущих змінних в базу для кластеризації та інше. Необхідно заздалегідь фіксувати результуючу кількість кластерів, правило зупинки і, якщо

на те є підстави, початковий центр кластеру, що суттєво впливає на ефективність роботи алгоритму. Якщо немає підстав штучно задавати ці умови, рекомендується використовувати ієрархічні методи.

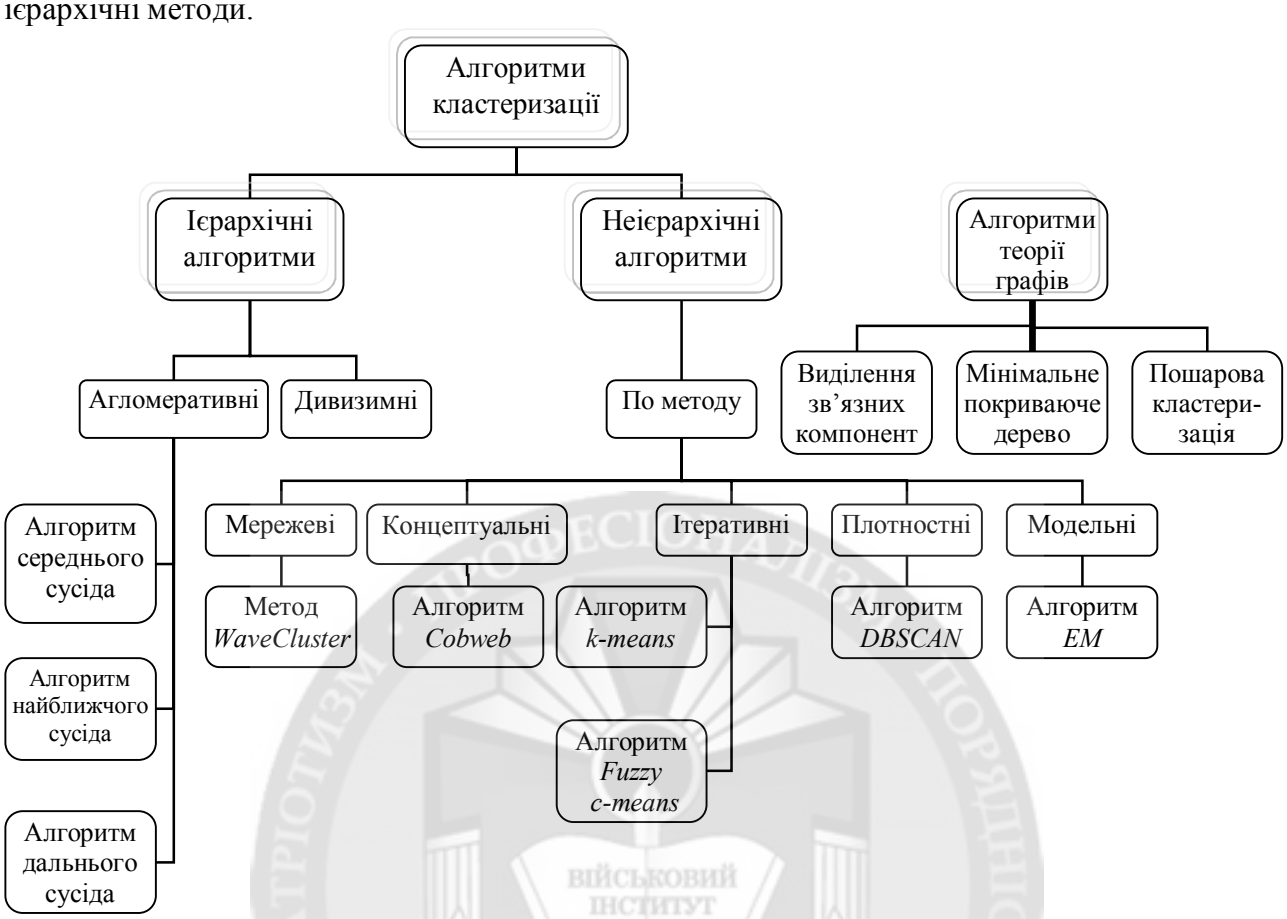

Рис. 1. Класифікація алгоритмів та методів кластерного аналізу

Огляд алгоритмів. В процесі розгляду особливостей найбільш поширених у практичному використанні алгоритмів з'ясовано наступне:

Алгоритми ієрархічної кластеризації. Серед алгоритмів ієрархічної кластеризації виділяються два основні типи: агломеративні і дивизимні алгоритми. Дивизимні алгоритми працюють за принципом «зверху-униз»: на початку всі об'єкти поміщаються в один кластер, який потім розбивається на більш дрібні кластери. Більш поширені агломеративні алгоритми, які на початку роботи поміщають кожний об'єкт в окремий кластер, а потім поєднують кластери в більш великі, поки всі об'єкти вибірки не будуть утримуватися в одному кластері. У такий спосіб будується система вкладених розбивок. Результати таких алгоритмів звичайно представляють у вигляді дендрограми. Класичним прикладом такого дерева є класифікація тварин і рослин.

До недоліку ієрархічних алгоритмів можна віднести систему повних розбивок, яка може бути зайвою в контексті розв'язуваної задачі.

Алгоритми квадратичної помилки. Задачу кластеризації можна розглядати як побудову оптимальної розбивки об'єктів на групи. При цьому оптимальність може бути визначена як вимога мінімізації середньоквадратичної помилки розбивки:

$$e^2(X,L) = \sum_{j=1}^{K} \sum_{i=1}^{n_j} \left\| x_i^{(j)} - c_j \right\|,$$

де c_j – «центр мас» кластеру j (крапка із середніми значеннями характеристик для даного кластеру).

Найпоширенішим алгоритмом цієї категорії є метод *k*-середніх (*k-means*). Цей алгоритм будує задане число кластерів, розташованих максимально віддалено один від одного. Робота алгоритму ділиться на кілька етапів:

1. Випадкове обрання *k* крапок, які є початковими «центрами мас» кластерів.
2. Віднесення кожного об'єкту до кластеру з найближчим «центром мас».
3. Перерахунок «центрів мас» кластерів згідно з їхнім поточним складом.
4. Повернення до п. 2 за умови, що критерій зупинки алгоритму не вдоволений.

У якості критерію зупинки роботи алгоритму звичайно вибирають мінімальну змінну середньоквадратичної помилки. Також можливо припиняти роботу алгоритму за умови, якщо на кроці 2 не було об'єктів, які перемістилися із кластера в кластер. Слід зазначити, що основним недоліком алгоритмів на базі методу *k*-середніх є вимога початкового визначення кількості та положення центрів кластерів. Інформація про ці параметри на початковому етапі дослідження інформаційного простору, як правило, відсутня.

Нечіткі алгоритми. Найбільш популярним алгоритмом нечіткої кластеризації є алгоритм *c*-середніх (*c-means*). Він представляє собою модифікацію методу *k*-середніх. Кроки роботи алгоритму:

1. Обрання початкової нечіткої розбивки *n* об'єктів на *k* кластерів шляхом вибору матриці приналежності *U* розміру *n* x *k*.
2. Визначення значення критерію нечіткої помилки із застосуванням матриці *U*:

$$E^2(X,U) = \sum_{i=1}^{N}\sum_{k=1}^{K} U_{ik} \left\| x_i^{(k)} - c_k \right\|^2,$$

де c_k – «центр мас» нечіткого кластера *k*:

3. Перегрупування об'єктів з метою зменшення цього значення критерію нечіткої помилки.
4. Повернення до п. 2 до тих пір, поки зміни матриці U не стануть незначними.

Застосовування вказаного алгоритму може бути недоцільним за умови, якщо заздалегідь невідоме число кластерів, або є необхідність однозначного віднесення кожного об'єкту до одного кластеру.

Алгоритми, засновані на теорії графів. Сутність таких алгоритмів полягає в тому, що вибірка об'єктів представляється у вигляді графа $G = (V, E)$, вершинам якого відповідають об'єкти, а ребра мають вагу, рівну «відстані» між об'єктами. Перевагою графових алгоритмів кластеризації є наочність, відносна простота реалізації й можливість внесення різних удосконалень, заснованих на геометричних міркуваннях. Основними алгоритмами є алгоритм відокремлення зв'язних компонентів, алгоритм побудови мінімального покриваючого дерева й алгоритм пошарової кластеризації.

Алгоритм виділення зв'язних компонентів. В алгоритмі виділення зв'язних компонентів задається вхідний параметр R і в графі видаляються всі ребра, для яких «відстані» менше R. Сполученими залишаються тільки найбільш близькі пари об’єктів. Сенс алгоритму полягає в тому, щоб підібрати таке значення R, що лежить у діапазоні усіх «відстаней», при якому граф буде розбитий на кілька зв’язних компонентів. Отримані компоненти і є кластерами.

Для підбору параметра R зазвичай будується гістограма розподілів попарних відстаней. У завданнях з добре вираженою кластерною структурою даних на гістограмі буде два піки: один відповідає внутрикластерним відстаням, другий – міжкластерним відстаням. Параметр R обирається із зони мінімуму між цими піками. При цьому керувати кількістю кластерів за допомогою порога відстані досить важко.

Алгоритм мінімального покриваючого дерева. Сутність алгоритму полягає в представленні всього набору даних у вигляді графа, вершини якого – це елементи даних, а вага кожного ребра дорівнює відстані між відповідними елементами. Зазначений алгоритм будує на графі мінімальне покриваюче дерево, а потім послідовно видаляє ребра з

найбільшою вагою. Кластером вважається множина елементів, з'єднаних «залишком» дерева. З кожним прибраним ребром кількість кластерів збільшується.

Алгоритм пошарової кластеризації заснований на виділенні зв'язних компонент графа на деякому рівні відстаней між об'єктами (вершинами). Рівень відстані задається порогом відстані *c*. Так, якщо відстань між об'єктами $0 \le \rho(x, x') \le 1$, то $0 \le c \le 1$. Алгоритм пошарової кластеризації формує послідовність підграфів графа G, які відображають ієрархічні зв'язки між кластерами:

$$G^0 \subseteq G^1 \subseteq ... \subseteq G^m,$$

де $G^t = (V, E^t)$ – граф на рівні c^t, c^t – *t*-й поріг відстані, m – кількість рівнів ієрархії. За допомогою зміни порогів відстані $\{c^o, ..., c^m\}$, де $0 = c^o < c^t < ... < c^m = 1$ можливо контролювати глибину ієрархії кластерів, що одержуються. Відповідно, алгоритм пошарової кластеризації здатний створювати як плоску розбивку даних, так і ієрархічну.

Так, необхідність розроблення методів і алгоритмів неієрархічної кластеризації обумовлена тим, що вони дозволяють досягти великої гнучкості у багатоваріантному розрахунку кластерів. До недоліків цих методів доцільно віднести необхідність визначення початкової кількості та положення центрів кластерів і зазначення умови зупинки роботи алгоритмів.

У табл. 1 наведено порівняльний аналіз найбільш поширених алгоритмів кластеризації.

Таблиця 1

Порівняльна таблиця алгоритмів

Алгоритм кластеризації	Форма кластерів	Вхідні дані	Результати
Ієрархічний	довільна	число кластерів чи поріг відстані для усікання ієрархії	бінарне дерево кластерів
k-середніх	гіперсфера	число кластерів	центри кластерів
c-середніх	гіперсфера	число кластерів, ступінь нечіткості	центри кластерів, матриця приналежності
Виділення зв'язних компонент	довільна	поріг відстані R	деревовидна структура кластерів
Мінімальне покриваюче дерево	довільна	число кластерів чи поріг відстані для видалення ребер	деревовидна структура кластерів
Пошарова кластеризація	довільна	послідовність порогів відстані	деревовидна структура кластерів з різними рівнями

Також широкого поширення в методах кластеризації даних набули так звані *Карти Кохонена*. Вони виконують проекцію багатомірних даних в простір меншої розмірності (зазвичай двомірної) і використовуються на практиці, як правило, при візуалізації даних з метою дослідження наявності чи відсутності кластерної структури в даних, числа кластерів, законів сумісного розподілу ознак, залежностей.

При використанні тільки карт Кохонена задача кластерного аналізу й ідентифікації залежностей не вирішується, вони тільки дозволяють по "розфарбуваннях" карти висунути гіпотези щодо наявності кластерної структури й кількості кластерів, залежностях між значеннями окремих змінних. Висунуті гіпотези повинні перевірятися й підтверджуватися іншими способами. Більш того, показано [10], що карти Кохонена можуть приводити як до

формування неправильних гіпотез, так і до неможливості побачити окремі реально наявні й статистично достовірні залежності в даних.

Так, карти Кохонена є методом головних компонентів, тільки з нейрофізіологічним ухилом. Відповідно, навіть після фільтрації однаково існує необхідність вирішувати задачу кластеризації, однак, вже в просторі меншої розмірності (наприклад, на площині). Крім того, карти Кохонена можна використовувати не тільки в якості фільтра, що знижує розмірність, але й у якості самостійного механізму кластеризації. Взагалі, якщо вузлів ("нейронів") у картах Кохонена дуже й дуже багато, то функціонування такої мережі стає схожим на метод головних компонентів, а якщо зробити кількість вузлів дуже маленькою (наприклад, що збігаються з передбачуваною кількістю кластерів), то надалі робота мережі не відрізняється від методу *k*-середніх; інакше кажучи, в останньому випадку вага кожного виходу карти може розглядатися як центр одного з кластерів.

Існує ще клас *інкрементних алгоритмів кластеризації*, спеціально адаптованих для роботи в мінливих умовах, тобто при додаванні/видаленні крапок даних. В основному ці алгоритми використовують методики класифікації: коли з'являється нова крапка даних, такий алгоритм відносить її до одного з наявних кластерів або створює новий, можливо склеюючи деякі з існуючих. Крім того, все спроектовано так, щоб існувала можливість додавати/видаляти крапки з мінімальними витратами часу/пам'яті. Це дуже ефективні й економічні алгоритми кластеризації, які, однак, не позбавлені недоліків. Так, розбивка суттєво залежить від порядку вступу даних, на що практично неможливо впливати.

Висновки з даного дослідження і перспективи подальшого розвитку у даному напрямку.

Таким чином, незважаючи на досить велику кількість розроблених моделей і алгоритмів кластерного аналізу, при розв'язанні прикладних задач дослідники часто зустрічаються із низкою проблем, до яких належать:

– складності в обґрунтуванні якості результатів аналізу, що враховує специфіку конкретної задачі;

– формулювання імовірнісних моделей досліджуваних об'єктів, особливо у випадку малого об'єму вибірки;

– необхідність обробки великої кількості різнотипних (кількісних або якісних) факторів;

– нелінійність взаємозв'язків; наявність пропусків, погрішностей виміру змінних;

– необхідність представлення результатів аналізу у формі, зручній й зрозумілій фахівцям прикладної області;

– проблема пошуку глобального екстремуму в критерії якості угрупування;

– нестійкість групуючих розв'язків при невеликих змінах вибірки або параметрів роботи алгоритму.

Так, доцільно зробити висновок щодо необхідності подальшого розроблення таких алгоритмів, які потребують мінімальної кількості початкових параметрів, прості в реалізації, дозволяють проводити багатоваріантний аналіз і дають задовільні результати.

ЛІТЕРАТУРА:

1. Voroncov, K. V. (2007), Algoritmy klasterizacii i mnogomernogo shkalirovanija : kurs lekcij [Algorithms for clustering and multidimensional scaling : course of lectures], Moskovskij gosudarstvennyj universitet, Moscow, Russia.

2. Steh, Ju. V., Fajsal, M. E. Sardih, Lobur, M. V., Dombrova, M.S. and Arcibasov, V. E. (2010), "Development and study of clustering algorithms for large collections of documents" Zbirnik naukovih prats IPPME im.G.E. Puhova NAN Ukrayini, Kiev, Ukraine, no. 58, pp. 283–290.

3. A Tutorial on Clustering Algorithms, available at: http://home.dei.polimi.it/ /Clustering/tutorial_html/kmeans.htm (access date December 05, 2014).

4. Bradley P. Scaling Clustering Algorithms to Large Databases / P. Bradley,U. Fayyad, C. Reina // Proc. 4th Int'l Conf. Knowledge Discovery and Data Mining,AAAI Press, Menlo Park, Calif., 1998. – 180 p.

5. Hajkin Sajmon (2006), Nejronnye seti [Neural network], textbook, Izdatel'skij dom «Vil'jams», Moscow, Russia, 1104 p.

6. Jain A., Murty M., Flynn P. Data Clustering: A Review. // ACM Computing Surveys. 1999. Vol. 31, no. 3.

7. Ajvazjan, S. A., Buhshtaber, V. M., Enjukov, I. S. and Meshalkin, L. D. (1989), Prikladnaja statistika: klassifikacija i snizhenie razmernosti [Applied statistics: classification and dimensionality reduction], Finansy i statistika, Moscow, Russia.

8. Information-analytical resource that is devoted to machine learning, pattern recognition and data mining, available at: www.machinelearning.ru/ (access date December 05, 2014).

9. Chubukova, I.A. Course of lectures «Data Mining, available at: www.intuit.ru/department/database/datamining/ (access date December 05, 2014).

10. Yin H. Learning Nonlinear Principal Manifolds by Self-Organising Maps, In: Gorban A. N. et al (Eds.), LNCSE 58, Springer, 2007 ISBN 978-3-540-73749-0

Рецензент: д.т.н., проф. Дубовенко К.В., завідувач кафедри Електротехнологій і електропостачання Миколаївського національного аграрного університету

к.т.н. Волосюк Ю.В.

АНАЛИЗ АЛГОРИТМОВ КЛАСТЕРИЗАЦИИ ДЛЯ ЗАДАЧ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

В статье дана общая характеристика процедуры кластерного анализа. Приведен обзор существующих подходов к решению задачи кластеризации и математических методов кластерного анализа данных.

Описаны этапы процесса кластеризации, рассмотрены вопросы выбора степени расстояния и весов для класифицирующих свойств объектов. Проведена классификация и анализ существующих алгоритмов кластерного анализа, рассмотрены преимущества и недостатки этих алгоритмов. Обоснована целесообразность использования карт Кохонена в методиках кластеризации с целью исследования наличия или отсутствия кластерной структуры в данных, числа кластеров, законов совместного распределения признаков, зависимостей. Предоставлена сравнительная таблица алгоритмов.

Сделан вывод о необходимости дальнейшей разработки простых в реализации алгоритмов, требующих минимального количества начальных параметров, позволяющих проводить многовариантный анализ и дающих удовлетворительные результаты.

Ключевые слова: кластерный анализ, функция расстояния, поиск центров, алгоритм пороговой величины, алгоритм k-средних.

Ph.D. Volosyuk Y.

ANALYSIS OF CLUSTERING ALGORITHMS FOR DATA MINING TASKS

The article gives a general description of the procedure of cluster analysis. Provides an overview of existing approaches to solving the problem of clustering and mathematical methods of cluster analysis.

Stages of the clustering process, the issues of choice and the power of the distance weights for classification object properties. The classification and analysis of the existing algorithms for cluster analysis, discusses the advantages and disadvantages of these algorithms. The expediency of using Kohonen maps in clustering techniques to investigate the presence or absence of the cluster structure of the data, the number of clusters, the laws of the joint distribution of characteristics, dependencies. Provided a comparative table of algorithms.

The conclusion about the need for further development of easy-to-implement algorithms that require a minimum amount of initial parameters that allow for multivariate analysis and giving satisfactory results.

Keywords: cluster analysis, search centers, threshold algorithm, k-means algorithm.