УДК 004.048

ІНДУКТИВНА ГІБРИДНА МОДЕЛЬ КЛАСТЕРИЗАЦІЇ ДАНИХ НА ОСНОВІ ЩІЛЬНІСНИХ АЛГОРИТМІВ

Л. М. Ясінська-Дамрі

Українська академія друкарства,
вул. Під Голоском, 19, Львів, 79020, Україна

Подано результати дослідження щодо практичної реалізації гібридної індуктивної моделі кластеризації даних на основі щільнісних алгоритмів DBSCAN та OPTICS. Проведено порівняльний аналіз різних типів внутрішніх критеріїв якості кластеризації та відповідних їм зовнішніх критеріїв для різних типів синтетичних даних. Показано, що вибір внутрішніх критеріїв є суттєвим для оцінки якості групування об'єктів у кластерній структурі, а для кожного типу даних формування комбінації внутрішніх критеріїв якості необхідно проводити з урахуванням характеру розподілу об'єктів і кластерів у просторі. Сформовано функції розрахунку критерію балансу для кожного типу даних, який містить як компоненти внутрішні та зовнішні критерії якості кластеризації. Показано, що запропонована модель дозволяє оптимізувати визначення параметрів щільнісних алгоритмів кластеризації DBSCAN та OPTICS з точки зору характеру розподілу об'єктів у відповідних кластерах.

Ключові слова: *кластеризація даних, щільнісні алгоритми кластеризації, внутрішні та зовнішні критерії якості кластеризації, індуктивна технологія об'єктивної кластеризації.*

Постановка проблеми. Щільнісні алгоритми є одними із сучасних алгоритмів кластерного аналізу, перевага яких полягає в тому, що вони дозволяють виділити кластери складної форми, з одного боку, та об'єкти, які за щільністю їх розподілу у просторі можуть бути ідентифіковані як шум, оскільки вони за щільністю не потрапляють у жоден кластер, з іншого боку. До даних типів алгоритмів можна віднести DBSCAN (Density-Based Spatial Clustering and Application with Noise) [1] та його логічне продовження алгоритм OPTICS (Ordering Points To Identify the Clustering Structure) [2]. Як показує аналіз покрокової процедури формування кластерної структури даних алгоритмів, результат кластеризації при застосуванні щільнісних алгоритмів визначається двома параметрами: *Eps* (радіус сфери ε-околиці точок, що формують кластерну структуру) та *MinPts* (мінімальна кількість точок всередині ε-околиці). Комбінація цих параметрів визначає кластерну структуру.

Аналіз останніх досліджень та публікацій. У [1] авторами запропоновано метод оцінки значення *Eps* при заздалегідь заданому значенні *MinPts* на основі *k-dist* графа (*k* у даному випадку дорівнює *MinPts*), який являє собою розподіл

значень *Eps* при різній кількості точок всередині *Eps* околиці. На думку авторів, оптимальне значення *Eps* при даному значенні *MinPts* знаходиться на коліні даного графа. Але слід зазначити, що точне визначення значення *Eps* аналіз *k-dist* графа не дозволяє, можна тільки встановити інтервал, у якому знаходиться оптимальне значення *Eps*. Цей факт ускладнює коректне застосування алгоритму кластеризації DBSCAN, оскільки результат роботи алгоритму має високу чутливість до значення параметра *Eps*, яке зазвичай підлаштовується емпіричним шляхом в процесі моделювання.

Щільнісний алгоритм кластеризації OPTICS [2] є логічним продовженням алгоритму DBSCAN [1]. Автори даного алгоритму запропонували метод визначення оптимальних значень *Eps* і *MinPts* на основі діаграми досяжності, яка візуалізує розподіл значень *Eps* при різних значеннях щільності розподілу об'єктів у просторі ознак. До переваг цього методу порівняно з методом *k-dist* графа слід віднести суттєво меншу чутливість результату роботи алгоритму до значення *Eps*. Діапазон зміни значень *Eps* встановлюється шляхом аналізу діаграми досяжності так, щоб він вміщував необхідну кількість впадин, розділених вертикальними лініями. Кожна впадина відповідає кластеру. Але слід зазначити, що цьому алгоритму також властивий великий відсоток суб'єктивізму, оскільки оптимальне значення *Eps* встановлюється емпіричним шляхом у процесі моделювання.

У [3, 4] авторами запропоновано модель реалізації алгоритму кластеризації OPTICS та DBSCAN із застосуванням принципів індуктивної технології об'єктивної кластеризації. Апробація запропонованої моделі проводилася із використанням синтетичних даних малої розмірності (набір точок у двовимірному просторі). У межах цього дослідження вирішення проблеми об'єктивного визначення оптимальної комбінації значень *Eps* і *MinPts* на основі індуктивної технології об'єктивної кластеризації із застосуванням як внутрішніх, так і зовнішніх критеріїв якості кластеризації отримало подальший розвиток завдяки більш ретельному вибору внутрішніх критеріїв кластеризації та більш глибокому аналізу результатів роботи моделі із застосуванням кількісних критеріїв якості кластеризації даних.

Мета статті – практична реалізація індуктивної технології об'єктивної кластеризації на основі щільнісних алгоритмів OPTICS та DBSCAN, що передбачає оптимізацію внутрішніх критеріїв якості кластеризації з урахуванням форми кластерів та характеру розподілу об'єктів у відповідних кластерах.

Виклад основного матеріалу дослідження. Практична реалізація індуктивної технології об'єктивної кластеризації на основі щільнісних алгоритмів передбачає наступні етапи.

Етап 1. Ініціалізація моделі.

1.1. Формування матриці експериментальних даних: $G = \{g_{ij}\}, i = \overline{1,n}, j = \overline{1,m}$, де g_{ij} – значення ознаки, що відповідає *i*-му рядку та *j*-му стовпчику; *n* та *m* – кількість рядків та стовпців відповідно.

1.2. Вибір метрики оцінки ступеня близькості об'єктів, кластерів, об'єктів та кластерів залежно від типу даних, що досліджуються.

1.3. Формування функцій розрахунку критеріїв якості кластеризації: внутрішніх, зовнішніх та балансу: QC_{int}, QC_{ext}, QC_{bal}.

1.4. Формування двох еквівалентних підмножин даних *A* і *B*.

Етап 2. Кластеризація даних. Розрахунок критеріїв якості кластеризації.

2.1. Формування діапазону зміни параметра *MinPts:* $MinPts_{min} = 3$, $MinPts_{max}$. Створення *k-dist* графа або діаграми досяжності для граничних значень даного параметра у разі застосування алгоритму DBSCAN або OPTICS відповідно. Оцінка діапазону та кроку зміни параметра *Eps*: Eps_{min}, Eps_{max}, шляхом аналізу *k-dist* графа або діаграми досяжності.

2.2. Ініціалізація вихідних значень параметрів *MinPts* та *Eps:* $e = Eps_{min}$; $k = MinPts_{min} = 3$.

2.3. Кластеризація даних на двох еквівалентних підмножинах *A* і *B*. Формування відповідних кластерних структур.

2.4. Якщо кількість кластерів більша або дорівнює двом та кількість кластерів у двох кластеризаціях однакова ($K_A = K_B \geq 2$), розрахунок внутрішніх та зовнішніх критеріїв якості кластеризації. У протилежному випадку – збільшення значення параметра *Eps* на один крок ($e = e + de$).

2.5. Якщо $e \leq Eps_{\max}$, перехід на крок 2.3 даної процедури. У протилежному випадку – розрахунок критерію балансу та збільшення значення *k* на одиницю ($k = k + 1$).

2.6. Якщо $k \leq k_{\max}$, перехід на крок 2.3 даної процедури. У протилежному випадку – остаточне формування матриці критеріїв якості кластеризації.

Етап 3. Аналіз отриманих результатів. Фіксація оптимальної кластеризації.

3.1. Створення діаграм залежності критерію балансу від значення *Eps* для кожного значення *MinPts.*

3.2. Фіксація найкращих кластеризацій за критерієм балансу у кожному випадку (відповідають максимумам критерію балансу).

3.3. Розрахунок критерію точності класифікації відповідних об’єктів за формулою (3.9) у разі наявності необхідної інформації в анотації даних.

3.4. Фіксація оптимальних параметрів відповідного алгоритму кластерного аналізу, що відповідають максимуму критерію точності класифікації. Формування оптимальної кластеризації.

Одним із важливих кроків ініціалізації моделі кластеризації даних є формування векторів внутрішніх та зовнішніх критеріїв для досліджуваних даних.

У статті подані дослідження щодо реалізації цього кроку для синтетичних даних *Compound*, що містять кластери різної форми об’єктів у двовимірному просторі (рис. 1). Як бачимо, частина об’єктів за густиною їх розподілу у просторі може бути ідентифікована як шум. До того ж, відповідно до анотації, дані *Compound* містять шість кластерів. Але детальній аналіз показує, що перший кластер може бути ідентифікований за густиною як шум, тобто п’ять кластерів у цьому випадку є коректною кластеризацією.

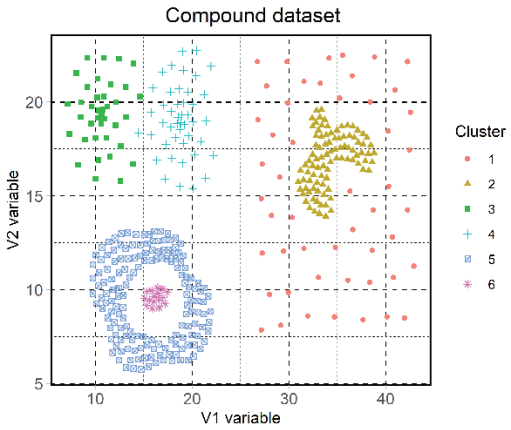

Рис. 1. Характер розподілу обєктів та кластерів у даних *Compound*

Таблиця

Внутрішні критерії оцінки якості кластерної структури, що використовувалися в процесі моделювання

№	Критерій	Правило	№	Критерій	Правило
1	WB index	*min*	9	Calinski Harabasz	*max*
2	Banfeld-Raftery	*min*	10	PBM	*max*
3	C index	*min*	11	GDI	*max*
4	Ray Turi	*min*	12	Ratkowsky Lance	*max*
5	Davies Bouldin	*min*	13	Dunn	*max*
6	McClain-Rao	*min*	14	Gamma	*max*
7	Scott-Symons	*min*	15	Silhouette	*max*
8	Xie-Beni	*min*	16	Wemmert-Gancarski	*max*

Процес моделювання щодо визначення оптимальної комбінації внутрішніх критеріїв, поданих у таблиці, та відповідних їм зовнішніх критеріїв передбачав штучне розділення вихідної множини даних на дві еквівалентні підмножини на першому кроці. На другому кроці відбувалося покрокове збільшення кількості штучних кластерів від 2 до 8. Розрахунок внутрішніх критеріїв здійснювався шляхом застосування функцій пакета *clusterCrit* [6] мови програмування *R* [7]. На рис. 1 та 2 зображені діаграми залежності значень внутрішніх критеріїв від кількості кластерів у разі застосування двох еквівалентних підмножин *A* і *B*. За анотацією критеріїв, перша група діаграм (рис. 2) відповідає правилу *min*, тобто мінімальне значення критерію відповідає оптимальній кластеризації. Діаграми, зображені на рис. 3, відповідають правилу *max*. На рис. 4 зображені діаграми розподілу значень зовнішніх критеріїв, розрахованих як нормалізована різниця відповідних внутрішніх критеріїв.

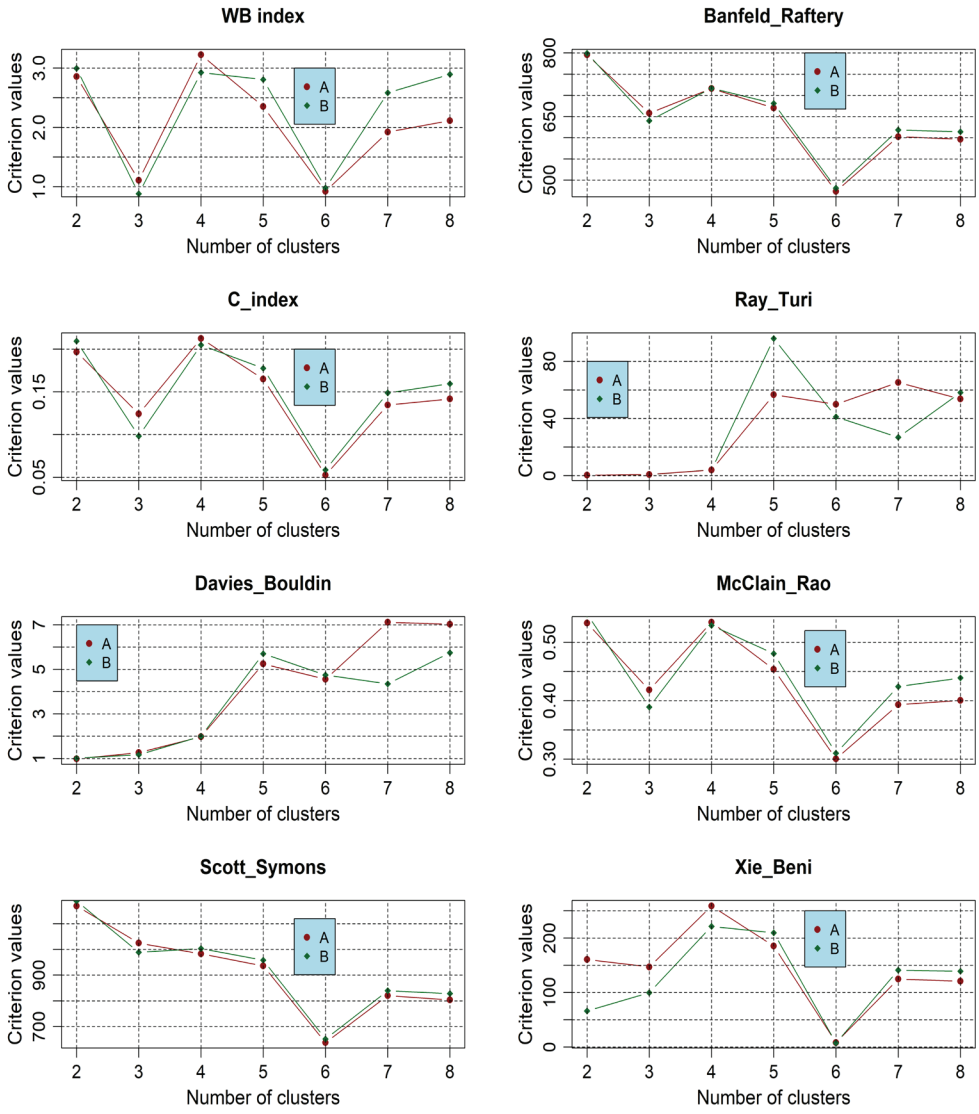

Рис. 2. Діаграми розподілу внутрішніх критеріїв якості кластеризації, що відповідають правилу *min*, розрахованих для двох еквівалентних підмножин *A* і *B* даних *Compound* для кластеризацій з кількістю кластерів від 2 до 8

Аналіз отриманих результатів дозволяє зробити висновок, що для даних *Compound* наступні внутрішні критерії дозволяють виділити оптимальну кластеризацію (шість кластерів) при застосуванні двох еквівалентних підмножин даних: за правилом *min* – Banfied_Raftery, C-index, McClain_Rao, Scott_Symons та Xie_Beni; за правилом *max* – Dunn і Gamma. Аналіз діаграм, представлених на рис. 3, показує, що здебільшого внутрішні та зовнішні критерії суперечать один одному. Цей

факт підтверджує висновок щодо наявності похибки відтворюваності і, як результат, необхідності у компромісному рішенні шляхом застосування критерію балансу. Детальний аналіз діаграм залежності зовнішніх критеріїв від кількості кластерів показує також, що із вищезазначеного переліку внутрішніх критеріїв якості кластеризації мінімуму похибки відтворюваності для кластерної структури, яка містить шість кластерів (мінімальне значення зовнішнього критерію), відповідають наступні критерії: Dunn, Gamma та Xie_Beni.

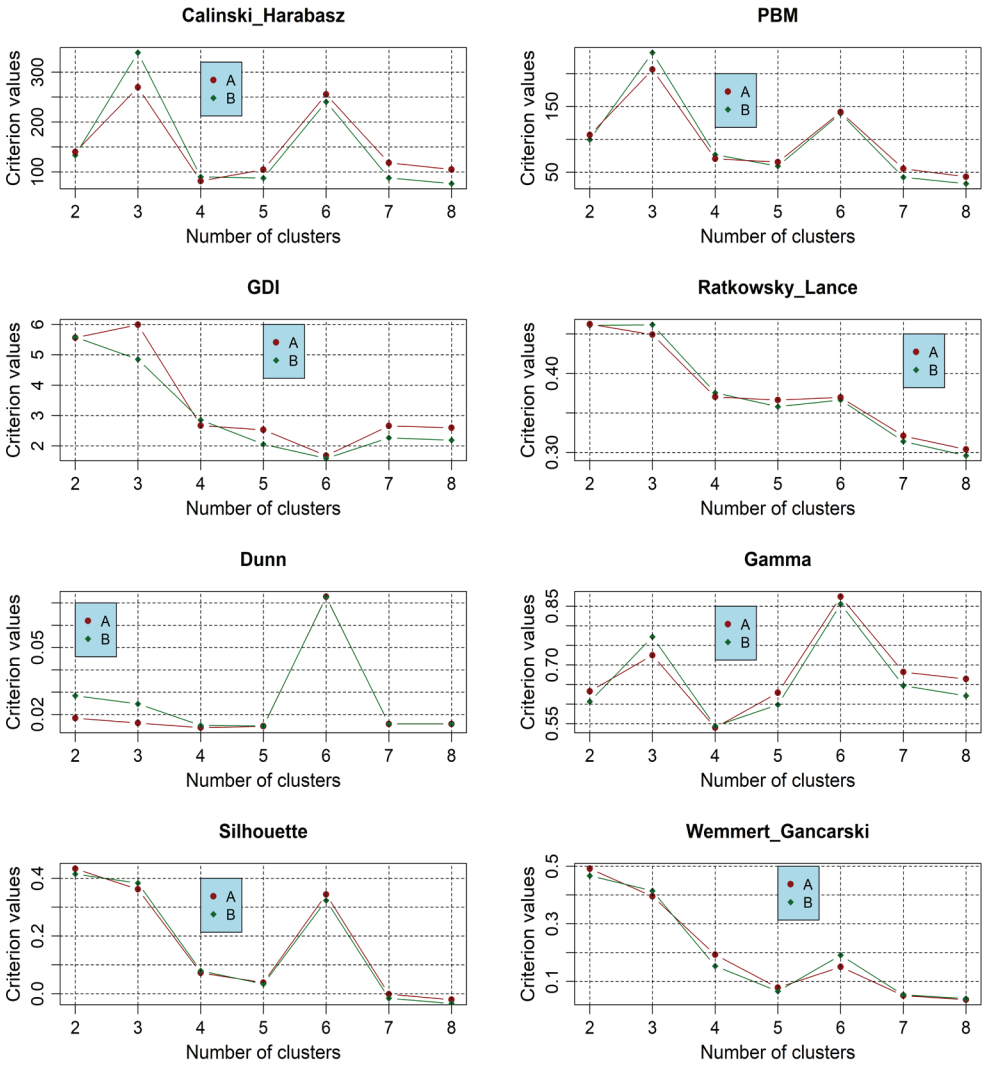

Рис. 3. Діаграми розподілу внутрішніх критеріїв якості кластеризації, що відповідають правилу *max*, розрахованих для двох еквівалентних підмножин *A* і *B* даних *Compound* для кластеризацій з кількістю кластерів від 2 до 8

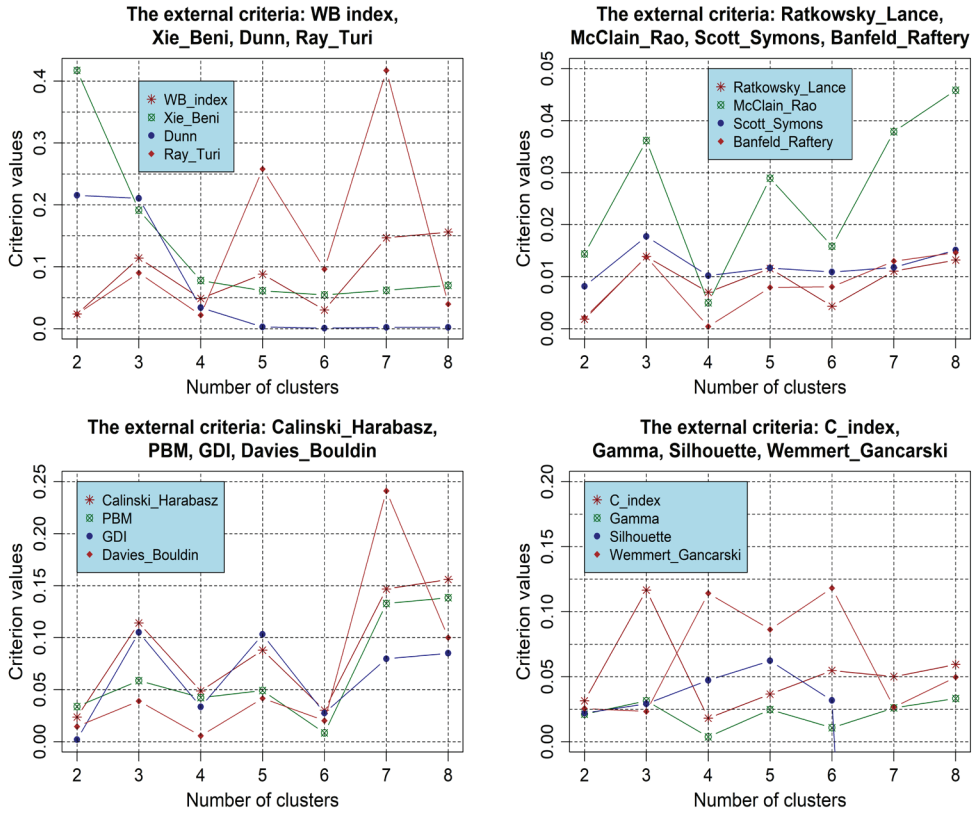

Рис. 4. Діаграми розподілу зовнішніх критеріїв для кластеризацій з кількістю кластерів від 2 до 8

Розрахунок критерію балансу для даних *Compound* здійснюється у відповідності з наступним алгоритмом:

1. Розрахунок коефіцієнтів лінійного рівняння перетворення значень відповідних внутрішніх та зовнішніх критеріїв у значення показника Y для даних еквівалентних підмножин A і B, враховуючи граничні значення відповідних величин на відносний характер їх зміни.

Внутрішні критерії:

$$\begin{cases} Y_{\max} = a_{XB}^{A,B} - b_{XB}^{A,B} \cdot QC_{XB}^{\min(A,B)} \\ Y_{\min} = a_{XB}^{A,B} - b_{XB}^{A,B} \cdot QC_{XB}^{\max(A,B)} \end{cases}$$

$$\begin{cases} Y_{\max} = a_{Dn}^{A,B} + b_{Dn}^{A,B} \cdot QC_{Dn}^{\max(A,B)} \\ Y_{\min} = a_{Dm}^{A,B} + b_{Dn}^{A,B} \cdot QC_{Dn}^{\min(A,B)} \end{cases} \quad (1)$$

$$\begin{cases} Y_{\max} = a_{Gm}^{A,B} + b_{Gm}^{A,B} \cdot QC_{Gm}^{\max(A,B)} \\ Y_{\min} = a_{Gm}^{A,B} + b_{Gm}^{A,B} \cdot QC_{Gm}^{\min(A,B)} \end{cases}$$

Зовнішні критерії:

$$\begin{cases} Y_{\max} = a - b \cdot QC_{ext}^{\min} \\ Y_{\min} = a - b \cdot QC_{ext}^{\max} \end{cases}, \quad (2)$$

де $QC_{XB}^{\min(A,B)}$, $QC_{XB}^{\max(A,B)}$, $QC_{Dn}^{\max(A,B)}$, $QC_{Dn}^{\min(A,B)}$, $QC_{Gm}^{\max(A,B)}$, $QC_{Gm}^{\min(A,B)}$ – мінімальне та максимальне значення критеріїв Xie_Beni, Dunn і Gamma, що розраховані для еквівалентних підмножин A і B відповідно; $QC_{ext}^{\min}$, $QC_{ext}^{\max}$ – мінімальне та максимальне значення відповідного зовнішнього критерію; $Y_{max} = 5$, $Y_{min} = -2$ – максимальне та мінімальне значення безрозмірного показника Y за методом бажаності Харрінгтона, a і b – коефіцієнти відповідних лінійних рівнянь.

2. Трансформування значень відповідних критеріїв якості кластеризації у значення показника Y відповідно до системи рівнянь:

$$\begin{cases} Y_{XB}^{A,B} = a_{XB}^{A,B} - b_{XB}^{A,B} \cdot QC_{XB}^{A,B} \\ Y_{Dn}^{A,B} = a_{Dn}^{A,B} + b_{Dn}^{A,B} \cdot QC_{Dn}^{A,B} \\ Y_{Gm}^{A,B} = a_{Gm}^{A,B} + b_{Gm}^{A,B} \cdot QC_{Gm}^{A,B} \\ Y_{ext}^{A,B} = a_{ext}^{A,B} - b_{ext}^{A,B} \cdot QC_{ext}^{A,B} \end{cases} \quad (3)$$

3. Розрахунок приватних бажаностей для кожного критерію:

$$\mathrm{d_i} = \exp(-\exp(-Y_i)). \quad (4)$$

4. Розрахунок критерію балансу як середнє геометричне усіх приватних бажаностей (2 + 2 + 2 (внутрішні критерії для підмножин A і B) + 3 (зовнішні критерії для кожної пари внутрішніх критеріїв)):

$$QC_{bal} = \sqrt[9]{\prod_{i=1}^{9} d_i}. \quad (5)$$

Максимальне значення критерію (5) відповідає оптимальній кластеризації за даною групою критеріїв. На рис. 5 зображено діаграми розподілу критерію балансу при застосуванні алгоритму кластеризації DBSCAN. Значення *MinPts* при цьому змінювалося від 3 до 8. Аналіз отриманих результатів дозволяє зробити висновок, що здебільшого оптимальна кластеризація за максимальним значенням критерію балансу відповідає чотирьохкластерній структурі. За характером розподілу об’єктів у кластерах усі кластеризації є коректними, але один кластер, що знаходиться всередині більшого кластера, не ідентифікується. Кластеризації у даних випадках розрізняються тільки кількість об’єктів, що ідентифікуються як шум. При значенні *minPts* = 3 об’єкти у просторі розділені на три кластера, що також може бути сприйнятливим, оскільки кластери не перетинаються між собою. Але оптимальною кластеризацією у цьому випадку є кластеризація, що відповідає параметрам алгоритму *minPts* = 5, *Ep*s = 1.444. У цьому випадку отримуються 5 кластерів. Характер розподілу об’єктів у просторі при оптимальних параметрах алгоритму зображений на рис. 6.

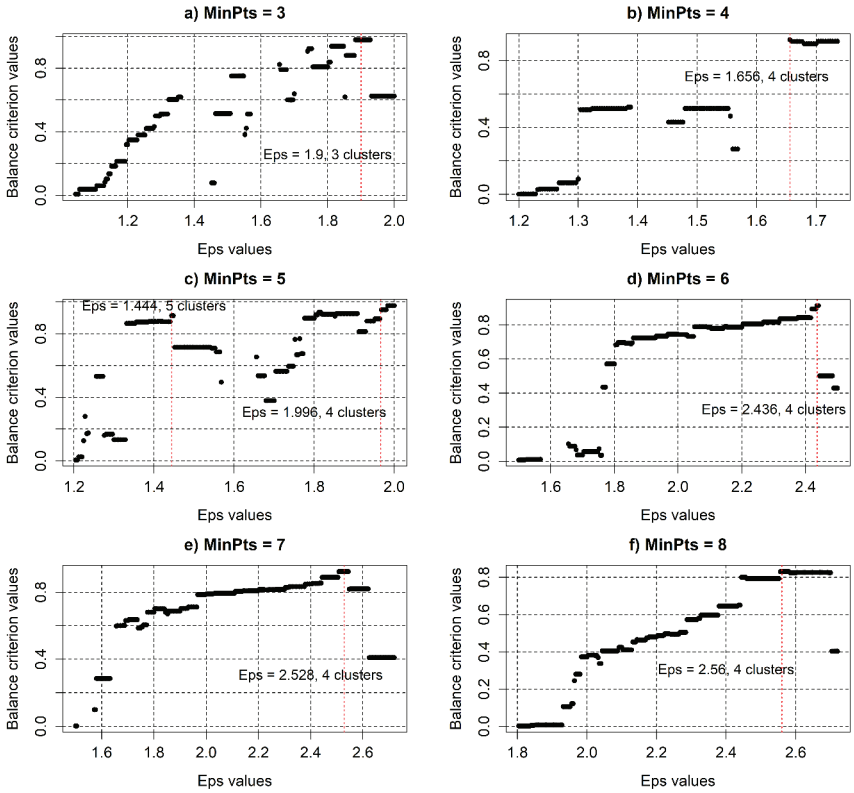

Рис. 5. Результати моделювання щодо визначення оптимальних параметрів алгоритму DBSCAN за критерієм балансу для даних *Compound*

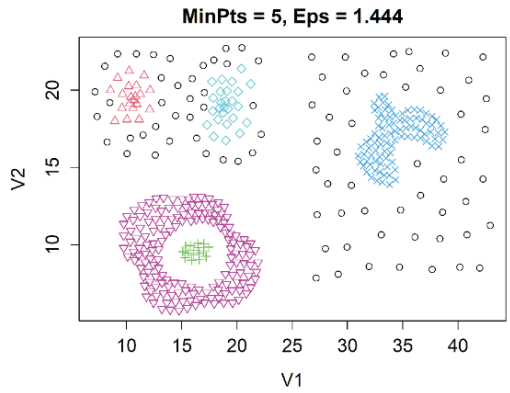

Рис. 6. Результати моделювання щодо кластеризації даних *Compound* при застосуванні гібридної індуктивної моделі алгоритму DBSCAN

На рис. 7 зображені аналогічні результати у разі застосування щільнісного алгоритму кластеризації OPTICS. Аналіз отриманих діаграм дозволив визначити оптимальне значення *Eps* для кожного *minPts*. Результати моделювання для значень *minPts* від 3 до 6 зображені на рис. 8. При *minPts* = 7 значення критерію балансу показує незадовільні результати як за характером розподілу об'єктів у кластери, так і за похибкою відтворюваності. При *minPts* = 8, як показали результати моделювання, результати кластеризації також не відповідали характеру розподілу об'єктів у кластери.

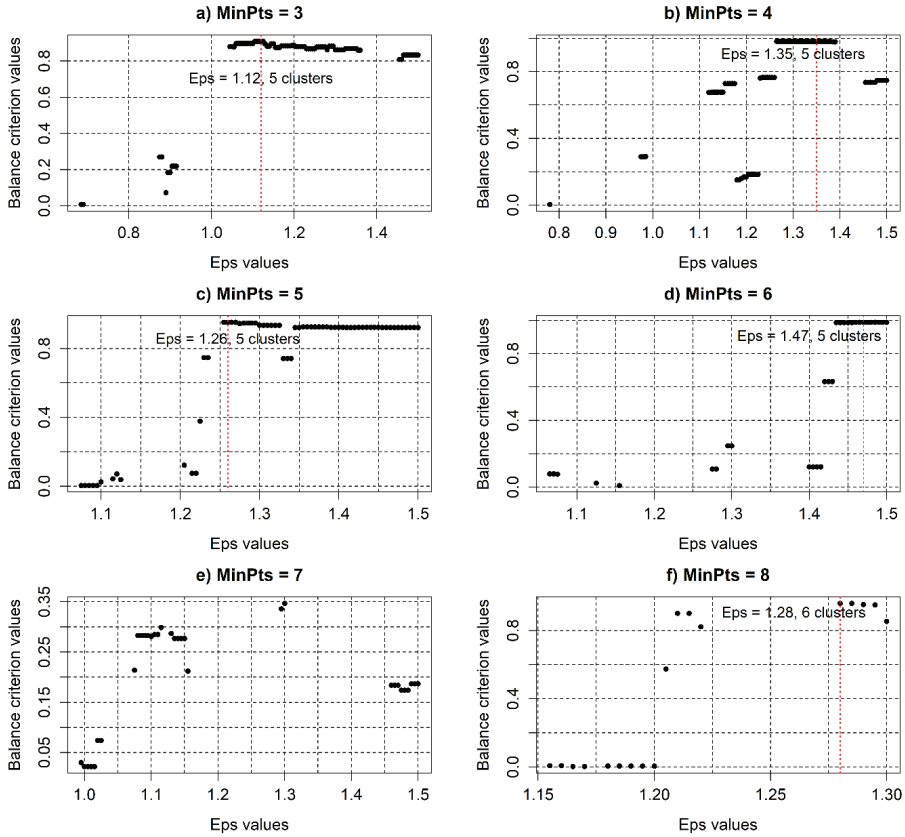

Рис. 7. Результати моделювання щодо визначення оптимальних параметрів алгоритму OPTICS за критерієм балансу для даних *Compound*

Аналіз отриманих результатів дозволяє зробити висновок, що в усіх чотирьох випадках характер розподілу об'єктів у кластери є адекватним. Є невелика різниця у кількості об'єктів, яка ідентифікована як шум, але порівняно з роботою алгоритму DBSCAN алгоритм OPTICS є більш стійкий до зміни параметрів, що спрощує його адекватне застосування.

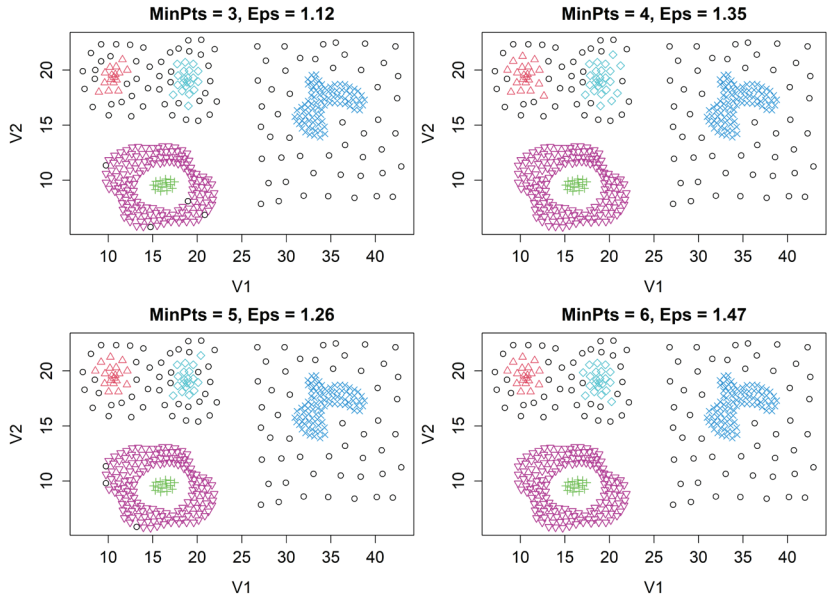

Рис. 8. Результати моделювання щодо кластеризації об'єктів даних *Compound* при застосуванні алгоритму кластеризації OPTICS

Висновки. У статті наведені результати дослідження щодо практичної реалізації індуктивної моделі об'єктивної кластеризації на основі застосування щільнісних алгоритмів DBSCAN і OPTICS. Модель представлена як покрокова процедура формування кластерної структури в інтервалу зміни параметрів відповідного алгоритму із розрахунком на кожному кроці реалізації цієї процедури внутрішніх і зовнішніх критеріїв якості кластеризації та критерію балансу, що містить як компоненти як внутрішні, так і зовнішні критерії. Аналіз отриманих результатів свідчить про високу ефективність застосування цієї технології для визначення оптимальних параметрів певного алгоритму, які відповідають максимальним значенням критерію балансу. Аналіз результатів моделювання показав, що сформовані кластерні структури адекватно відображають характер розподілу об'єктів у просторі, враховуючи анотацію даних, при цьому формується окрема група об'єктів, яка за більш низькою щільністю їх розподілу у просторі ідентифікується як шум. Цей факт може бути використаний у системах фільтрації зашумлених даних на основі щільнісних алгоритмів кластеризації. У процесі моделювання проведено також порівняльний аналіз гібридних індуктивних моделей кластеризації даних на основі щільнісних алгоритмів OPTICS і DBSCAN. Аналіз отриманих результатів показав перевагу алгоритму OPTICS завдяки вищій стабільності цього алгоритму при формуванні кластерної структури та меншій чутливості до варіації параметрів алгоритму.

Подальшою перспективою досліджень авторки є реалізація запропонованої технології із застосуванням інших сучасних алгоритмів кластерного аналізу та інших типів експериментальних даних.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Ester M., Kriegel H. P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial datasets with noise. In: Proceedings of the second international conference on knowledge discovery and data mining, Portland, Oregon, 1996. Pp. 226–231.
2. Ankerst M., Breunig M. M., Kriegel H. P., Sander J. OPTICS: Ordering Points To Identify the Clustering Structure. In: ACM special interest group on management of data record SIGMOD, 1999. Vol 28 (2). Pp. 49–60. doi: https://doi.org/10.1145/304181. 304187 30.
3. Babichev S., Durnyak B., Pikh I., Senkivskyy V. An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms (2020). Advances in Intelligent Systems and Computing, Vol. 1020. Pp. 532–553.
4. Babichev S., Durnyak B., Zhydetskyy V., Pikh I., Senkivskyy V. Application of Optics Density-Based Clustering Algorithm Using Inductive Methods of Complex System Analysis (2019). International Scientific and Technical Conference on Computer Sciences and Information Technologies, 1, art. no. 8929869. Pp. 169–172.
5. Zahn C. T. Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on Computers, 1971. Vol. 100 (1). Pp. 68–86.
6. El. URL: https://cran.r-project.org/web/packages/clusterCrit.
7. Ihaka R., Gentleman R. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*. 1996. Vol. 5 (3). Pp. 299–314.

REFERENCES

1. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial datasets with noise. In: Proceedings of the second international conference on knowledge discovery and data mining, Portland, Oregon, 226–231 (in English).
2. Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). OPTICS: Ordering Points To Identify the Clustering Structure. In: ACM special interest group on management of data record SIGMOD, 28 (2), 49–60. doi: https://doi.org/10.1145/304181. 304187 30 (in English).
3. Babichev, S., Durnyak, B., Pikh, I., & Senkivskyy, V. An Evaluation of the Objective Clustering Inductive Technology Effectiveness Implemented Using Density-Based and Agglomerative Hierarchical Clustering Algorithms (2020). Advances in Intelligent Systems and Computing, 1020, 532–553 (in English).
4. Babichev, S., Durnyak, B., Zhydetskyy, V., Pikh, I., & Senkivskyy, V. Application of Optics Density-Based Clustering Algorithm Using Inductive Methods of Complex System Analysis (2019). International Scientific and Technical Conference on Computer Sciences and Information Technologies, 1, art. no. 8929869, 169–172 (in English).
5. Zahn, C. T. (1971). Graph-theoretical methods for detecting and describing gestalt clusters. IEEE Transactions on Computers, 100 (1), 68–86 (in English).
6. El. Retrieved from https://cran.r-project.org/web/packages/clusterCrit (in English).
7. Ihaka, R., & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5 (3), 299–314 (in English).

doi: 10.32403/1998-6912-2021-2-63-64-76

INDUCTIVE HYBRID MODEL OF DATA CLUSTERING USING DENSITY-BASED ALGORITHMS

L. Yasinska-Damri

Ukrainian Academy of Printing,
19, Pid Holoskom St., Lviv, 79020, Ukraine
Lm.yasinska@gmail.com

The paper presents the results of the research concerning the practical implementation of a hybrid inductive model of data clustering using DBSCAN and OPTICS density-based algorithms. A comparative analysis of different types of internal clustering quality criteria and the corresponding external quality criteria for various types of synthetic data is performed. It is shown that the choice of internal clustering quality criteria is essential for assessing the quality of the objects grouping in a cluster structure and, for each type of the studied dataset the formation of a combination of internal quality criteria should be done considering the nature of both the objects and clusters distribution in the feature space. The simulation procedure is carried out using the synthetic dataset Compound, which contains according to the data annotation six various shapes clusters. The simulation results regarding comparison analysis of various types of the internal clustering quality criteria have shown that for the dataset Compound the optimal criteria in terms of both minimal reproducibility error and the optimal cluster structure are the following ones: DUNN, Gamma and Xie Beni. The functions of calculating the balance criterion, which contains as components of the selected internal clustering quality criteria and respective external clustering quality criteria are formed. As the simulation results, the charts of balance criterion versus the Eps value for each MinPts parameter are created. It is shown that the proposed model allows optimizing the definition of parameters of density clustering algorithms DBSCAN and OPTICS in terms of the nature of the distribution of objects in the respective clusters. Moreover, the simulation results allow one to conclude about the advantage of the OPTICS algorithm due to the higher stability of this algorithm operation during the cluster structure formation on the one hand and, less sensitivity to variation of the algorithm parameters on the other hand.

Keywords*: data clustering, density-based clustering algorithms, internal and external clustering quality criteria, objective clustering inductive technology.*

Стаття надійшла до редакції 12.07.2021.
Received 12.07.2021.