

УДК 004.048

МЕТОДОЛОГІЧНІ ЗАСАДИ РОЗРОБКИ ГІБРИДНИХ ІНДУКТИВНИХ МОДЕЛЕЙ КЛАСТЕРИЗАЦІЇ ВЕЛИКИХ ДАНИХЛ. М. Ясінська-Дамрі¹, Б. В. Дурняк¹, С. А. Бабічев²¹Українська академія друкарства,
вул. Під Голоском, 19, Львів, 79020, Україна²Херсонський державний університет,
вул. Університетська, 27, Херсон, 73000, Україна

Подано результати дослідження щодо розробки методологічних засад створення гібридної індуктивної моделі кластеризації великих даних. Представлено архітектуру та концептуальний опис формування кластерної структури на основі принципів індуктивної технології об'єктивної кластеризації. Практична реалізація запропонованої концепції передбачає застосування двох еквівалентних підмножин даних, що містять однакову кількість попарно близьких об'єктів. Для оцінки якості розподілу об'єктів у кластерній структурі запропоновано три типи критеріїв: внутрішній, що визначає характер розподілу об'єктів у кластерах стосовно медіани відповідного кластера; зовнішній, що розраховується як нормалізована різниця відповідних внутрішніх критеріїв якості; критерій балансу, що містить як компоненти внутрішні та зовнішні критерії і розраховується на основі застосування функції бажаності Харрінгтона. Остаточне рішення щодо формування кластерної структури у рамках запропонованої концепції приймається на основі оцінки якості класифікації досліджуваних об'єктів.

Ключові слова: кластеризація, індуктивні методи, критерії якості кластеризації, функція бажаності Харрінгтона, точність класифікації.

Постановка проблеми. Сучасний стан наукового напрямку, орієнтованого на інтелектуальний аналіз великих даних, визначається збільшенням можливостей для підвищення ефективності вирішення відповідної проблеми завдяки удосконаленню чинних та розробці нових методів та алгоритмів інтелектуального аналізу даних та машинного навчання. У статті представлені теоретичні дослідження щодо застосування індуктивних моделей кластеризації профілів експресій генів з метою формування підмножин взаємно корельованих профілів, що дозволяють з високою точністю ідентифікувати об'єкти, які містять значення експресій виділених генів як атрибути. Під експресією гена у даному випадку розуміється рівень його активності, який пропорційний кількості гена даного типу, що визначають відповідні функції досліджуваного об'єкта. Профіль експресій гена є вектором значень експресій даного гена, що визначені при різних умовах проведення експерименту. Складність розв'язання даної задачі визначається високою розмірністю

профілів експресій генів. Кількість значень експресій генів у відповідних профілях дорівнює кількості досліджуваних об'єктів (у загальному випадку $\approx 100\text{--}300$). Більшість існуючих алгоритмів кластеризації орієнтовані на низько-розмірні дані, що містять невелику кількість атрибутів. До того ж більшості існуючих алгоритмів кластеризації даних властива похибка відтворюваності, тобто задовільні результати кластеризації, отримані на одній множині даних, не завжди повторюються при застосуванні іншої аналогічної підмножини даних.

Аналіз останніх досліджень та публікацій. Вирішенню поставленої проблеми на сьогодні присвячено велику кількість наукових праць [1–5]. Як показує аналіз літературних джерел, її вирішення може бути досягнуто шляхом ретельної валідації отриманої моделі із застосуванням «свіжих» даних (не використовуються при створенні моделі), яка містить об'єкти, що близькі за своїми властивостями до об'єктів вихідної підмножини даних. У рамках наших досліджень ця проблема вирішується шляхом реалізації сучасних алгоритмів кластеризації даних із застосуванням індуктивної технології об'єктивної кластеризації [6, 7]. Крім того, велика кількість профілів експресій генів у вихідних даних (≈ 10000) викликає необхідність розробки ієрархічних гібридних моделей покрокової кластеризації профілів експресії генів з оцінкою якості їх групування на кожному ієрархічному рівні шляхом оцінки якості класифікації об'єктів, які містять виділені профілі експресій генів як атрибути.

Метою статті є розробка методологічних засад гібридних моделей аналізу та обробки даних експресій генів на основі індуктивних методів аналізу складних систем.

Виклад основного матеріалу дослідження. В основу індуктивної технології об'єктивної кластеризації покладені три фундаментальних принципи, що запозичені з методології індуктивного моделювання складних систем [1–5] і дозволяють створити цілісну теорію формування кластерної структури із застосуванням внутрішніх та зовнішніх критеріїв, аналіз яких створює умови для мінімізації похибки відтворюваності при отриманні бажаної кластерної структури:

1. *Принцип послідовного перебору.* Реалізація даного принципу передбачає послідовний перебір кластеризацій на двох еквівалентних підмножинах даних (містять однакову кількість попарно близьких об'єктів) з розрахунком як внутрішніх критеріїв якості групування об'єктів у кожній кластеризації (за умовою рівної кількості кластерів), так і зовнішніх критеріїв, які визначають різницю між результатами кластеризацій об'єктів, що містяться в еквівалентних підмножинах даних.

2. *Принцип зовнішнього доповнення.* У рамках методології індуктивного моделювання складних систем реалізація даного принципу передбачає застосування «свіжої інформації» для верифікації отриманої моделі, тобто підмножини об'єктів, які за своїми властивостями є близькими до об'єктів, що містяться в основній множині даних. У випадку коректної моделі результати кластеризації на основній множині даних і додатковій підмножині повинні збігатися у межах допустимої помилки. При реалізації індуктивної технології об'єктивної кластеризації роль «свіжої інформації» відіграє еквівалентна підмножина даних, тобто на першому кроці вихідні дані, кластеризація яких передбачається, розділяються на дві еквівалентні підмножини, що містять однакову кількість попарно близьких об'єктів.

3. *Принцип неостаточності рішень.* Реалізація даного принципу у рамках індуктивної технології об'єктивної кластеризації передбачає фіксацію проміжних кластеризацій, які відповідають екстремумам внутрішніх та зовнішніх критеріїв якості кластеризації. Але слід зазначити, що внутрішні і зовнішні критерії можуть суперечити між собою, тому остаточний вибір оптимальної кластеризації здійснюється на основі комплексного критерію балансу, який повинен вміщувати як компоненти внутрішні і зовнішні критерії якості кластеризації. Для розрахунку комплексного балансу у [6, 8] авторами запропоновано використовувати функцію бажаності Харрінгтона [9], рівняння якої представлено формулою (1), а графік зображений на рис. 1:

$$d = \exp(-\exp(-Y)). \quad (1)$$

де d – значення бажаності, що варіюється від 0 (найгірша кластеризація) до 1 (найкраща кластеризація), а значення безрозмірного показника Y змінюється за лінійним законом від -2 ($d = 0$) до 5 ($d = 1$).

Особливості даної функції полягають у тому, що вона має дві точки перегину: $0,37 = 1/e$ та $0,63 = 1 - 1/e$. Ці точки визначають межу допустимих значень за шкалою бажаності. У таблиці представлені діапазони допустимих значень за шкалою бажаності Харрінгтона та відповідні їм терми. Метод бажаності Харрінгтона являє собою один із методів розв'язання багатокритеріальних задач, при цьому нормалізація значень різних критеріїв, що застосовуються у рамках моделі, відбувається на першому кроці шляхом трансформації шкал відповідних критеріїв у шкалу показника Y за лінійною (у більшості випадків) або нелінійною (в окремих випадках, враховуючи характер розподілу даних) шкалою.

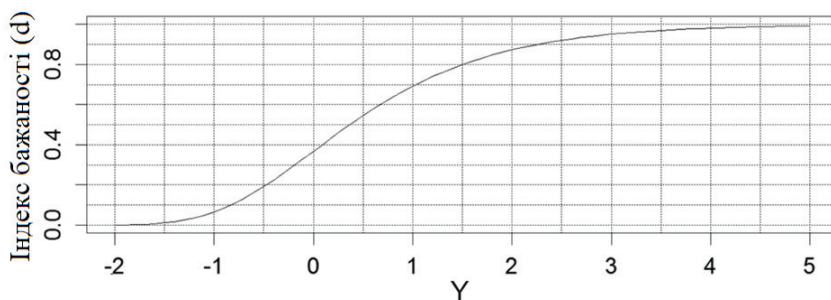


Рис. 1. Функція бажаності Харрінгтона

Таблиця

Діапазони допустимих значень за шкалою бажаності Харрінгтона та відповідні їм терми

№	Діапазон варіювання індексу бажаності	Терми
1	0.8 – 1	Відмінно
2	0.63 – 0.8	Добре
3	0.37 – 0.63	Задовільно
4	0.2 – 0.37	Погано
5	0 – 0.2	Несприйнятно

Кількість внутрішніх критеріїв якості кластеризації даних визначаються типом даних, об'єкти яких необхідно розділити на кластери. Застосування індуктивної технології об'єктивної кластеризації передбачає формування векторів внутрішніх критеріїв на еквівалентних підмножинах A та B в процесі послідовного перебору кластеризацій:

$$QC_{int}^A = \{QC_{int_i}^A\}; QC_{int}^B = \{QC_{int_i}^B\}, i = \overline{1, m}, \quad (2)$$

де m – кількість внутрішніх критеріїв, що розраховуються для оцінки якості групування об'єктів в окремих кластеризаціях.

Як було зазначено вище, найкращі кластеризації за внутрішніми критеріями якості можуть не бути об'єктивними за похибкою відтворюваності, тобто екстремуми внутрішніх критеріїв (2), отриманих для еквівалентних підмножин даних, можуть суперечити між собою. Для мінімізації похибки відтворюваності розраховується зовнішній критерій якості кластеризації, який враховує розбіжності між результатами групування об'єктів у еквівалентних підмножинах даних. Застосування стандартних зовнішніх критеріїв якості розподілу об'єктів у різних кластеризаціях (таких, як індекси Жаккара, Кульчинського тощо) не є ефективним у даному випадку, оскільки їх використання передбачає кластеризації на однакових об'єктах. Дана умова у випадку індуктивної технології об'єктивної кластеризації не виконується, оскільки еквівалентні підмножини даних містять однакову кількість близьких за своїми властивостями об'єктів. Але слід зазначити, що мінімум похибки відтворюваності відповідає найбільшій збіжності характеру розподілу об'єктів у кластери в еквівалентних підмножинах даних. Оскільки внутрішні критерії відображають характер розподілу об'єктів у відповідних кластерах, то природньо припустити, що мінімальне значення похибки відтворюваності може бути досягнуто при мінімальній різниці між значеннями внутрішніх критеріїв, розрахованих на еквівалентних підмножинах об'єктів A та B . У [6, 7] авторами запропоновано метод розрахунку зовнішнього критерію якості кластеризації як нормалізованої різниці відповідних внутрішніх критеріїв, отриманих на еквівалентних підмножинах даних:

$$QC_{ext} = \left\{ \frac{|QC_{int}^A - QC_{int}^B|}{C_{int}^A - QC_{int}^B} \right\}, i = \overline{1, m}. \quad (3)$$

Алгоритм розрахунку комплексного критерію балансу на основі методу бажаності Харрінгтона передбачає наступні кроки:

1. Трансформування шкал отриманих заздалегідь внутрішніх та зовнішніх критеріїв якості кластеризації у шкалу показника Y відповідно до системи лінійних рівнянь:

$$Y_i = \begin{cases} a \pm b \cdot QC_{int_i} \\ a - b \cdot QC_{ext_i} \end{cases}, i = \overline{1, m}, \quad (4)$$

де знак у верхньому рівнянні визначається характером зміни значень відповідного внутрішнього критерію в процесі формування кластерної структури. Знак «+»

обирається у випадку, якщо оптимальній заданим критерієм кластеризації відповідає максимальне значення даного критерію (позитивна кореляція між векторами Y і QC), у протилежному випадку ставиться знак «-». Значення параметрів a і b визначаються емпіричним шляхом, враховуючи граничні значення відповідних критеріїв, характер їх зміни та граничні значення шкали показника Y :

$$\begin{cases} Y_{\max} = a + b \cdot QC_{\max} \\ Y_{\min} = a + b \cdot QC_{\min} \end{cases} \quad \text{або} \quad \begin{cases} Y_{\max} = a - b \cdot QC_{\min} \\ Y_{\min} = a - b \cdot QC_{\max} \end{cases}. \quad (5)$$

2. Формування вектора значень показника $Y = \{Y_i\}$ відповідно до лінійного рівняння:

$$Y_i = a \pm b \cdot QC_i. \quad (6)$$

3. Розрахунок приватних бажаностей для кожного значення показника Y , отриманого на попередньому кроці:

$$d_i = \exp(-\exp(-Y_i)). \quad (7)$$

4. Розрахунок комплексного критерію балансу як середнє геометричне всіх приватних бажаностей, розрахованих за формулою (7):

$$QC_{bal} = \sqrt[n]{\prod_{i=1}^n d_i}. \quad (8)$$

Відповідно до даних таблиці, максимальне значення критерію балансу відповідає оптимальній кластеризації як з точки зору розподілу об'єктів у кластерах, так і з точки зору мінімального значення похибки відтворюваності. Слід зазначити, що у процесі моделювання можлива наявність декількох локальних максимумів критерію балансу, які мало розрізняються між собою. У даному випадку остаточне рішення приймається з огляду на мету поставленої задачі та з урахуванням бажаної кількості кластерів, на які необхідно розділити експериментальні дані.

Наведені вище теоретичні дослідження та висновки створюють умови для розробки покрокової процедури формування кластерної структури на основі принципів індуктивної технології об'єктивної кластеризації.

Архітектура та концептуальний опис покрокової процедури формування кластерної структури на основі принципів індуктивної технології об'єктивної кластеризації. Структурна блок-схема покрокової процедури формування кластерної структури на основі принципів індуктивної технології індуктивної кластеризації зображена на рис. 2.

Реалізація даної процедури передбачає наступні етапи:

Етап I. Постановка проблеми. Аналіз експериментальних даних. Формування мети. Ініціалізація моделі.

1.1. Постановка проблеми. Аналіз експериментальних даних. Формування мети.

1.2. Ініціалізація моделі. Реалізація даного кроку передбачає вибір метрики оцінки ступеня близькості досліджуваних даних, формування вектора внутрішніх критеріїв кластеризації залежно від типу даних, вибір алгоритму кластеризації та ініціалізація діапазону та кроку зміни даних параметрів ($p_{\min}$, $p_{\max}$, dp).

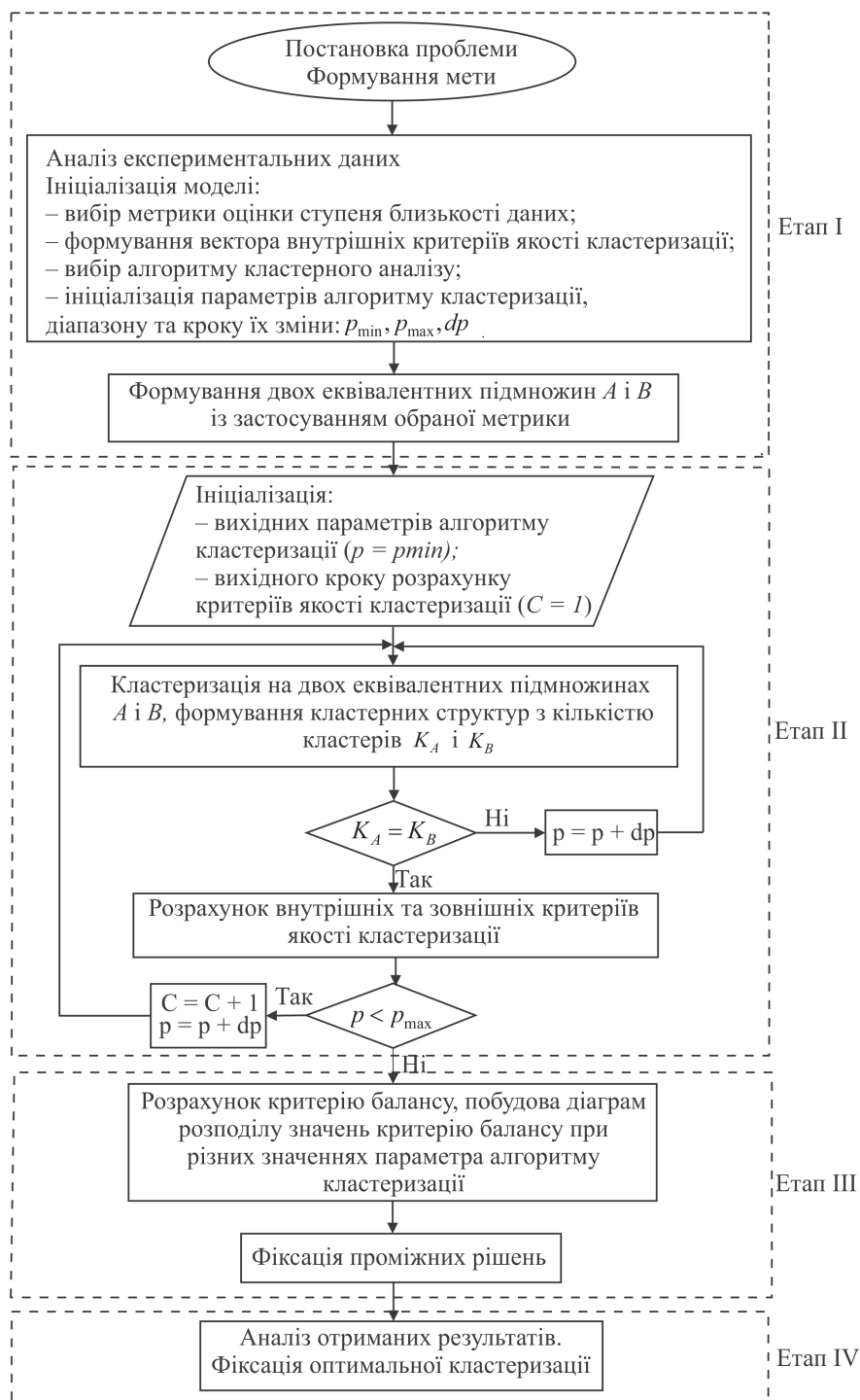


Рис. 2. Структурна блок-схема покрокової процедури формування кластерної структури на основі принципів індуктивної технології об'єктивної кластеризації

1.3. Формування двох еквівалентних підмножин даних A і B за метрикою, визначеною на кроці 1.2 даної процедури.

Етап II. Кластеризація даних. Формування векторів значень внутрішніх та зовнішніх критеріїв якості кластеризації.

2.1. Ініціалізація вихідних параметрів алгоритму кластеризації ($p = p_{\min}$) та вихідного кроку розрахунку внутрішніх критеріїв якості кластеризації ($C = 1$).

2.2. Кластеризація даних на двох еквівалентних підмножинах A і B . Формування кластерної структури з кількістю кластерів K_A і K_B .

2.3. Якщо кількість кластерів у кластеризаціях на підмножинах A і B є однаковою ($K_A = K_B$), розрахунок внутрішніх критеріїв QC_{int}^A і QC_{int}^B та зовнішніх критеріїв QC_{ext} для кожної пари внутрішніх критеріїв за формулою (3).

Етап III. Розрахунок критерію балансу. Формування проміжних рішень щодо оптимальної кластерної структури.

3.1. Розрахунок критерію балансу за формулами (4) – (8).

3.2. Побудова діаграм розподілу значень критерію балансу при різних значеннях параметра алгоритму кластеризації.

3.3. Фіксація проміжних рішень, тобто параметрів алгоритму, які відповідають максимальним значенням критерію балансу.

Етап IV. Аналіз отриманих результатів. Фіксація оптимальної кластеризації.

4.1. Аналіз отриманих результатів. Реалізація даного кроку передбачає вибір критеріїв, за якими обирається оптимальна кластеризація. У випадку застосування даної технології для кластеризації профілів експресії генів кластеризуються атрибути відповідних об'єктів (профілі експресії генів). Здебільшого заздалегідь відомі класи, на які можуть бути розділені дані об'єкти (анотація даних). Таким чином, оцінка кластерної структури може бути виконана шляхом реалізації процедури класифікації об'єктів, які містять як атрибути профілі експресії генів, що сконцентровані у відповідних кластерах із розрахунком критеріїв якості класифікації з урахуванням похибок першого та другого роду. У рамках запропонованої моделі індуктивної технології об'єктивної кластеризації якість групування профілів експресії генів оцінювалася шляхом розрахунку точності класифікації об'єктів, яка у загальному випадку визначає ймовірність коректного розподілу об'єктів у відповідні класи:

$$ACC = \frac{\sum_{i=1}^{N_{cl}} TP_i}{N_{obj}}, \quad (9)$$

де TP_i – кількість коректно розпізнаних об'єктів у i -му класі; N_{cl} – кількість класів; N_{obj} – загальна кількість об'єктів, що містять тестові дані.

Вочевидь, що оптимальна кластеризація у даному випадку відповідає максимальному значенню точності класифікації об'єктів, що містять виділені профілі експресії генів з урахуванням мети поставленої задачі.

Висновки. У статті наведені результати дослідження щодо розробки методологічних засад створення індуктивної гібридної моделі кластеризації складних даних на основі комплексного застосування існуючих алгоритмів кластеризації,

методів інтелектуального аналізу даних та машинного навчання. Розроблено архітектуру та наведено концептуальний опис покрокової процедури формування кластерної структури на основі індуктивних методів об'єктивної кластеризації. Детально описана покрокова процедура реалізації цієї процедури, яка передбачає комплексне застосування внутрішніх критеріїв якості кластеризації, що визначають характер розподілу об'єктів у кластерах стосовно медіани відповідного кластера, зовнішніх критеріїв, які розраховуються як нормалізована різниця внутрішніх критеріїв, розрахованих на еквівалентних підмножинах даних (містять однакову кількість попарно близьких об'єктів) та критерію балансу, який містить як компоненти внутрішні та зовнішні критерії і розраховується на основі функції бажаності Харрінгтона. Максимальне значення критерію балансу відповідає оптимальній кластеризації за обраними критеріями якості. Остаточне рішення щодо формування кластерної структури у рамках запропонованої моделі приймається на основі оцінки якості класифікації об'єктів, що містять виділені у кластери об'єкти як атрибуту.

Подальшою перспективою авторських досліджень є практична реалізація запропонованої моделі із застосуванням різних сучасних алгоритмів кластеризації даних.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Ивахненко А. Г. Индуктивный метод самоорганизации моделей сложных систем. Киев : Наукова думка, 1982. 296 с.
2. Ивахненко А. Г. Объективная кластеризация на основе теории самоорганизации моделей. *Автоматика*. 1987. № 5. С. 6–15.
3. Madala H. R., Ivakhnenko A. G. Inductive Learning Algorithms for Complex Systems Modeling. CRC Press, 1994. 365 p.
4. Степашко В. С. Елементи теорії індуктивного моделювання. Стан та перспективи розвитку інформатики в Україні : монографія / Колектив авторів. Київ : Наукова думка, 2010. С. 471–486.
5. Степашко В. С. Самоорганизация прогнозирующих моделей сложных процессов и систем. XV Всероссийская научно-техническая конференция «Нейроинформатика-2013»: Лекции по нейроинформатике. Москва : НИЯУ МИФИ, 2013. С. 150–170.
6. Babichev S., Taif M. A., Lytvynenko V., Korobchinskyi M. Objective clustering inductive technology of gene expression sequences features. Communications in Computer and Information Science. In the book «Beyond Databases, Architectures and Structures», edited by S. Kozelski and D. Mrozek, 2017. Pp. 359–372.
7. Babichev S., Gozhyj A., Kornelyuk A., Lytvynenko V. Objective clustering inductive technology of gene expression profiles based on SOTA clustering algorithm. Biopolymers and Cell. Kiev : National Academy of Science Ukraine, 2017. Vol. 33 (5). Pp. 379–392.
8. Бабічев С. А. Концептуальные основы и методология создания индуктивной технологии объективной кластеризации. Індуктивне моделювання складних систем. Київ : Міжнародний науково-навчальний центр інформаційних технологій та систем НАН України та МОН України, 2016. Вип. 8. С. 12–32.

9. Harrington J. The desirability function. *Industrial Quality Control*, 1965, Vol. 21 (10). Pp. 494–498.

REFERENCES

1. Ivahnenko, A. G. (1982). *Induktivnyj metod samoorganizacii modelej slozhnyh sistem*. Kiev : Naukova dumka (in Russian).
2. Ivahnenko, A. G. (1987). Ob#ektivnaja klasterizacija na osnove teorii samoorganizacii modelej: *Avtomatika*, 5, 6–15 (in Russian).
3. Madala, H. R., & Ivakhnenko, A. G. (1994). *Inductive Learning Algorithms for Complex Systems Modeling*. CRC Press (in English).
4. Stepashko, V. S. (2010). Elementy teorii induktivnoho modeliuvannia. Stan ta perspektyvy rozvytku informatyky v Ukraini / Kolektyv avtoriv. Kyiv : Naukova dumka, 471–486 (in Ukrainian).
5. Stepashko, V. S. (2013). Samoorganizacija prognozirujushhih modelej slozhnyh processov i sistem. HV Vserossijskaja nauchno-tehnicheskaja konferencija «Nejroinformatika-2013»: Lekcii po nejroinformatike. Moskva : NIJaU MIFI, 150–170 (in Russian).
6. Babichev, S., Taif, M. A., Lytvynenko, V., & Korobchinskyi, M. (2017). Objective clustering inductive technology of gene expression sequences features. *Communications in Computer and Information Science*. In the book «Beyond Databases, Architectures and Structures», edited by S. Kozelski and D. Mrozek, 359–372 (in English).
7. Babichev, S., Gozhij, A., Kornelyuk, A., Lytvynenko, V. (2017). Objective clustering inductive technology of gene expression profiles based on SOTA clustering algorithm. *Biopolymers and Cell*. Kiev : National Academy of Science Ukraine, 33 (5), 379–392 (in English).
8. Babichev, S. A. (2016). Konceptual'nye osnovy i metodologija sozdanija induktivnoj tehnologii ob#ektivnoj klasterizacii. *Induktivne modeliuvannia skladnykh system*. Kyiv : Mizhnarodnyi naukovo-navchalnyi tsentr informatsiinykh tekhnolohii ta system NAN Ukrainy ta MON Ukrainy, 8, 12–32 (in Russian).
9. Harrington, J. (1965). The desirability function. *Industrial Quality Control*, 21 (10), 494–498 (in English).

doi: 10.32403/0554-4866-2021-2-82-141-150

METHODOLOGICAL BASIS OF HYBRID INDUCTIVE MODELS DEVELOPMENT FOR BIG DATA CLUSTERING

L. Yasinska-Damri¹, B. Durnyak¹, S. Babichev²

¹*Ukrainian Academy of Printing,
19, Pid Holoskom St., Lviv, 79020, Ukraine
Lm.yasinska@gmail.com, durnyak@uad.lviv.ua*

²*Kherson State University,
27, University St., Kherson, 73000, Ukraine
sbabichev@ksu.ksu.ua*

The paper presents the results of the research regarding the development of a methodological basis for creating a big data clustering hybrid inductive model. As a result of the research, the architecture and stepwise procedure conceptual description of the cluster structure formation based on the principles of the objective clustering inductive technology has been presented. The practical implementation of the proposed concept involves the use of two equivalent data subsets containing the same number of pairwise similar objects. To assess the quality of the objects distribution within the appropriate cluster structure, three types of criteria have been proposed: the internal, which determines the particularities of the objects distribution within the clusters relative to the median of the appropriate cluster; the external that is calculated as the normalized difference of the relevant internal quality criteria; the balance criterion, which contains as the components both the internal and the external criteria and it is calculated based on the application of the Harrington desirability function. The final decision concerning the cluster structure formation within the framework of the proposed concept is done on the basis of assessing the accuracy of the studied objects classification. The implementation of the proposed technique assumes possibilities of the use of various types of existing clustering algorithms. The main advantage of the proposed concept is minimizing the reproducibility error with the formation of the optimal cluster structure taking into account the goal of the solved task.

A further perspective of the author's research is the practical implementation of the proposed technique using various current data clustering algorithms.

Keywords: *clustering, inductive methods, clustering quality criteria, Harrington desirability function, classification accuracy.*

Стаття надійшла до редакції 06.07.2021.

Received 06.07.2021.