Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

ШАФРОНЕНКО АЛІНА ЮРІЇВНА

УДК 004.85:[004.62.048:004.275]

АДАПТИВНІ МЕТОДИ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ ПОТОКІВ ДАНИХ З ВИКОРИСТАННЯМ ЕВОЛЮЦІЙНОГО САМОНАВЧАННЯ

05.13.23 – системи та засоби штучного інтелекту
технічні науки

Реферат дисертації на здобуття наукового ступеня доктора технічних наук

Харків – 2024

Дисертацією є рукопис.

Робота виконана в Харківському національному університеті радіоелектроніки Міністерства освіти і науки України.

Офіційні опоненти: доктор технічних наук, професор
Антощук Світлана Григорівна
директор Інституту комп'ютерних систем
Національного університету «Одеська політехніка»,
МОН України

доктор технічних наук, професор
Субботін Сергій Олександрович,
завідувач кафедри програмних засобів
Національного університету
«Запорізька політехніка»,
МОН України

доктор технічних наук, професор
Удовенко Сергій Григорович,
завідувач кафедри інформатики та комп'ютерної техніки Харківського національного економічного університету імені Семена Кузнеця,
МОН України

Захист відбудеться «14» березня 2025 р. о 13.00 годині на засіданні спеціалізованої вченої ради Д 64.052.11 у Харківському національному університеті радіоелектроніки за адресою: Україна, 61166, м. Харків, пр. Науки, 14.

З дисертацією можна ознайомитись у бібліотеці Харківського національного університету радіоелектроніки за адресою: Україна, 61166, м. Харків, пр. Науки, 14 та на сайті вченої ради ХНУРЕ за адресою: https://nure.ua/branch/d-64-052-11/informacija-pro-zahist-disertacij.

Реферат розіслано «11» лютого 2025 р.

Учений секретар
спеціалізованої вченої ради І.П. Плісс

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність теми. На сьогодні методи штучного інтелекту, і перш за все, обчислювального інтелекту, отримали широке застосування для розв'язання різноманітних задач Data Mining, зокрема класифікації, розпізнавання образів, кластеризації, асоціації, оптимізації та екстраполяції.

Ці методи характеризуються здатністю вирішувати задачі будь-якої складності за умов перетинних класів, апріорної невизначеності щодо їхньої форми, кількості, тощо. В рамках обчислювального інтелекту, слід зазначити такі підходи, як штучні нейронні мережі (як мілкі, так і глибокі), нечіткі системи, еволюційні алгоритми та, так звані, гібридні системи обчислювального інтелекту. Ці системи обчислювального інтелекту поєднують переваги всіх розглянутих методів. Зокрема, особливу увагу привертають еволюційні нечіткі системи, що здатні ефективно вирішувати весь спектр зазначених задач.

Водночас слід зазначити, що всі ці методи призначені переважно для розв'язання задач класичного Data Mining, коли навчальна вибірка задана апріорі і залишається незмінною протягом усього процесу вирішення задач, а також за фіксованих умов: кількості класів, їхньої форми та рівня збурень.

У сучасних умовах, особливо в умовах воєнного стану, інформація що обробляється характеризуються нестабільністю та збуреністю даних навчальної вибірки, а також необхідністю обробки даних у форматі потоку. Це означає, що обсяг даних є апріорі невизначеним, вони можуть бути спотворені перешкодами або мати пропуски, змінювати свої властивості під час обробки, кількість класів може змінюватися, а самі класи можуть перетинатися довільним чином. За таких умов класичні методи обробки даних, зокрема кластеризації, стають неефективними.

Більше того, популярні на сьогодні глибокі нейронні мережі не пристосовані для вирішення такого класу задач. Це пояснюється тим, що вони потребують великих обсягів стабільних даних, характеристики яких не змінюються у часі. Крім того, ці задачі зазвичай вирішуються у режимі багатьох епох, що передбачає сталість властивостей інформації, які в реальних умовах може змінюватися. Тому глибокі нейронні мережі виявляються неефективними. На цей час інтенсивно розвивається перспективний напрямок Big Data Mining, де обсяг даних є принципово необмеженим. І в цьому випадку глибокі нейронні мережі залишаються неефективними, оскільки їх навчання по багатьом епохам є вкрай ускладненим за умов, коли вибірка апріорі невідома та постійно змінюється.

Аналіз публікацій провідних фахівців з цієї галузі, таких як I. Aizenberg (автор терміну «Deep Learning»), J. Kacprzyk, F. Klawonn, Yu. Tanaka, H. Takagi, P. Angelov, E. Lughofer, E. Rüstern, K. Moraga, Зайченко Ю., Субботін С., Пелешко Д., Філатов В., Бодянський Є., що працюють у напрямах інтелектуального аналізу даних, доводять, що ця проблематика є актуальною у всьому світі.

Усе перераховане вище, зумовлює потребу у розробленні нових ефективних нечітких методів кластеризації потоків даних, що здатні працювати за умов, коли дані надходять в онлайн режимі, можливо з високою частотою (є обмеження на продуктивність машини), можуть міняти свої властивості: сама структура даних також може змінюватись довільним чином (кількість класів, рівнів перетину, їх форми). На практиці досить часто виникають такі задачі, коли розмічена навчальна вибірка є відсутньою. Зрозуміло, що класичні методи тут непрацездатні, тому виникає задача аналізу потоків даних, які надходять на опрацювання в онлайн режимі (можливо з високою частотою), довільним чином можуть змінювати свої властивості, мати непередбачувані дрейфи, змінну кількість класів, їх рівнів перетину і, що саме головне і найбільш складне, немає розміченої вибірки.

Отже, розроблення нових методів та удосконалення існуючих методів нечіткої кластеризації даних за умов апріорної невизначеності на основі еволюційного самонавчання та надання їм адаптивних властивостей, що забезпечує можливість опрацьовування потоків нестаціонарних даних, викривлених завадами та пропусками, що послідовно надходять на обробку в онлайн режимі є актуальною теоретичною проблемою.

Зв'язок роботи з науковими програмами та темами. Дисертаційна робота виконана на кафедрі штучного інтелекту Харківського національного університету радіоелектроніки та відповідає науковому напряму кафедри «Гібридні системи обчислювального інтелекту для аналізу даних». Основні наукові результати досліджень отримано в рамках держбюджетних фундаментальних НДР ХНУРЕ: «Динамічний інтелектуальний аналіз послідовностей нечіткої інформації за умов суттєвої невизначеності на основі гібридних систем обчислювального інтелекту», (ДР №0116U002539), «Глибинні гібридні системи обчислювального інтелекту для аналізу потоків даних та їх швидке навчання» (ДР №0119U001403) та «Адаптивний бегінг гібридних систем обчислювального інтелекту на основі оптимального за швидкодією онлайн навчання» (ДР №0124U000363), а також прикладної НДР «Розробка методів та алгоритмів комбінованого навчання глибинних нейро-нео-фаззі систем за умов короткої навчальної вибірки» (ДР № 0122U001701), які виконувались на підставі наказів МОН України за результатами конкурсного відбору наукових проектів. Здобувачка брала участь у виконанні зазначених НДР і є співавтором звітів про НДР.

Мета дослідження - проведення комплексу досліджень, спрямованих на створення нових підходів та методів еволюційного самонавчання для адаптивної нечіткої кластеризації потоків викривлених даних в онлайн режимі за умов апріорної та поточної невизначеності з використанням найсучасніших досягнень у цій галузі: Computer Science, Computational Intelligence, Data Science, Data Streams, Big Data, Evolving Systems.

Задачі дослідження:

1. Провести аналіз підходів та методів для обробки потоків даних.

2. Розробити адаптивні методи нечіткої кластеризації потоків даних за умов перетинних класів та апріорної невизначеності.

3. Розробити методи адаптивної нечіткої кластеризації даних з різною щільністю розподілу.

4. Розробити еволюційні методи оптимізації в задачах нечіткої кластеризації масивів даних.

5. Розробити гібридні еволюційні методи нечіткої кластеризації масивів даних.

6. Експериментальна перевірка розроблених методів тестування та імплементація.

Об'єкт дослідження – онлайн кластеризація потоків даних з використанням еволюційного самонавчання.

Предмет дослідження - адаптивні нечіткі методи для обробки потоків викривлених даних в онлайн режимі за умов апріорної та поточної невизначеності з використанням еволюційного самонавчання.

Методи дослідження. Основними методами дослідження є методи обчислювального інтелекту: динамічний інтелектуальний аналіз даних - для знаходження прихованих залежностей в інформації; методи машинного навчання, за допомогою яких були синтезовані нові методи нечіткої кластеризації потоків даних, що дозволяють кластеризувати потоки даних в онлайн режимі; теорія нечіткої кластеризації – для розробки методів кластеризації викривлених потоків даних в умовах класів, що перетинаються та мають довільну форму; імітаційне моделювання - для визначення ефективності застосування розроблених методів.

Наукова новизна роботи. У дисертаційній роботі вирішено важливу теоретичну проблему створення нових ефективних нечітких методів обчислювального інтелекту, а саме, нечіткої кластеризації даних за умов апріорної невизначеності на основі еволюційного самонавчання та надання їм адаптивних властивостей, що забезпечує можливість опрацьовування потоків нестаціонарних даних, викривлених завадами та пропусками, що послідовно надходять на обробку в онлайн режимі.

Отримано такі нові наукові результати:

1. Уперше запропоновано адаптивні ймовірнісні, можливісні та правдоподібні методи нечіткої кластеризації потоків викривлених даних, які призначені для вирішення задач Data Stream Mining та Big Data Mining, що дозволяють опрацьовувати апріорі невідому кількість даних послідовно, спостереження за спостереженням в міру їх надходження у онлайн режимі.

2. Уперше запропоновано онлайн метод нечіткої кластеризації, який базується на ідеях аналізу щільностей розподілу даних, їх піків та правдоподібного нечіткого підходу, що дозволяє підвищити якість кластеризації даних з довільними апріорі невідомими щільностями розподілів.

3. Уперше запропоновано метод швидкої нечіткої кластеризації даних з використанням аналізу піків щільності розподілу даних на основі

правдоподібного підходу, що дозволяє вирішувати широкий клас задач Data Stream Mining та Big Data Mining у ситуаціях, коли дані забруднені завадами.

4. Уперше запропоновано швидкі методи нечіткої кластеризації даних довільної природи з апріорі невідомими розподілами, що дозволяє підвищити якість результатів розбиття масивів даних на класи за умов невизначеності.

5. Уперше запропоновано метод послідовної можливісної нечіткої кластеризації даних, який призначено для роботи в онлайн режимі, що дозволяє швидко знаходити екстремуми (центроїди) кластерів, незалежно від обсягів даних, що надходять на обробку у векторній або матричній формах.

6. Уперше запропоновано метод нечіткої кластеризації масивів даних на основі покращеного еволюційного алгоритму сірого вовка, що дозволяє відшукувати глобальні екстремуми цільових функцій та скоротити час їх пошуку.

7. Уперше запропоновано метод нечіткої кластеризації масивів даних на основі комбінованої оптимізації функцій щільності розподілу та еволюційного методу котячих зграй, що дозволяє уникнути застрягання в локальних екстремумах.

8. Уперше запропоновано підходи до вирішення багатоекстремальної задачі правдоподібної нечіткої кластеризації на основі модифікованих оптимізаційних процедур божевільної котячої зграї та зграї сірих вовків, що дозволяє скоротити час вирішення задачі.

9. Уперше запропоновано підхід до вирішення задачі адаптивної нечіткої кластеризації викривлених пропусками та викидами даних на основі стратегії найближчого прототипу-центроїду з використанням еволюційних процедур, що дозволяє підвищити завадостійкість процесу оптимізації.

10. Удосконалено еволюційний метод на основі косяків риб, що підвищив ефективність вирішення задач нечіткої кластеризації даних, які надходять як в пакетному, так і в онлайн режимах, що дозволяє скоротити час пошуку глобальних екстремумів.

11. Удосконалено метод кластеризації Густафсона-Кесселя, який базується на підході правдоподібності до нечіткої кластеризації та формує перетинні класи гіпереліпсоїдальної форми з довільною орієнтацією осей у просторі ознак, що дозволяє опрацьовувати потоки даних в міру їх надходження на обробку в онлайн режимі.

12. Удосконалено метод оптимізації на основі еволюційних котячих зграй шляхом введення в процеси пошуку та гонитви елементів глобального випадкового пошуку, що дозволяє підвищити точність визначення напрямку руху в режимі пошуку та покращити глобальні властивості методу у режимі гонитви.

Практичне значення одержаних результатів, полягає у підвищенні ефективності методів нечіткої кластеризації даних, коли дані надходять в онлайн режимі. В порівнянні з класичними методами кластеризації (*K*-means, FCM), розроблені адаптивні методи нечіткої кластеризації з використанням

еволюційного самонавчання забезпечують точність визначення кількості класів (кластерів) в умовах дефіциту апріорної інформації. Запропоновані методи нечіткої кластеризації на основі щільностей обробки потоків даних, в порівнянні з методами на основі щільностей (DBSCAN, OPTICS, DENCLUE) є більш точними та швидкими.

Розроблені адаптивні методи нечіткої кластеризації працездатні як в пакетному так і в онлайн режимах та здатні працювати на вибірках, що змінюють розмірність та форму кластерів; дозволяють обробляти великі обсяги даних, що можуть подаватись на обробку послідовно у формі потоків даних, ефективно працювати за умов суттєвої невизначеності, стохастичності, нелінійності, апріорної невизначеності, нестаціонарності та є найбільш пристосованими для вирішення задач Data Mining та Data Stream Mining, завдяки своїм універсальним апроксимуючим властивостям, здатності до самонавчання.

Результати дисертаційної роботи можуть бути використані для розв’язання широкого класу прикладних задач і, перш за все, задач Data Mining, Data Stream Mining, Big Data Mining та Medical Data Mining, кластеризації, прогнозування, діагностування, прийняття рішень, керування, класифікації за умов дефіциту апріорної інформації.

Отримані результати дають змогу:

- підвищити точність кластеризації потоків даних, що поступають на обробку в онлайн режимі за оцінками якості кластеризації даних на 8 %;
- підвищити швидкість роботи методів нечіткої кластеризації потоків даних за умов апріорної та поточної невизначеності, за рахунок запропонованих процедур оптимізації на 10%;
- підвищити точність прогнозування даних до 7-8% за рахунок аналізу великого об’єму інформації, що подається в онлайн режимі;
- зменшити ймовірність похибки розбиття потоків викривлених даних на класи, за умов невизначеності до 5%;
- прискорити аналіз та прийняття обґрунтованих рішень в залежності від поставленої задачі;
- підвищити точність і об’єктивність процесу медичного діагностування, відновлення викривлених та втрачених спостережень, що потрапляють на обробку в онлайн режимі;
- підвищити надійність та об’єктивність медичного діагностування пацієнтів з умовно невідомим діагнозом.

Результати дисертаційної роботи були апробовані і впроваджені: в КП «Санітарно-екологічний центр» Харківської міської ради (акт впровадження від 29 червня 2023р. та акт впровадження від 26 вересня 2024р.); в ТОВ «Будівельно-монтажне підприємство 168» (акт впровадження від 21 грудня 2023 р.); в ТОВ «Комунсервіс 2018» (акт впровадження від 12 квітня 2023р.); в ТОВ Науково-виробнича фірма «Хелп-Агро» (акт впровадження від 27

лютого 2023р.); в КНП «ОБЛАСНИЙ ЦЕНТР ОНКОЛОГІЇ», (акт впровадження №1 від 14 листопада 2023р. та акт впровадження №2 від 22 квітня 2024р.); в освітній процес Харківського національного університету радіоелектроніки (акт впровадження від 25.04.2024; акт впровадження від 26.04.2024, акт впровадження від 21.03.2024).

Особистий внесок здобувача. Дисертаційна робота виконана здобувачем особисто. Всі висновки, положення та рекомендації, подані в ній, сформульовано на основі особистих досліджень автора. В дисертації використано праці інших науковців, на які зроблено посилання. З колективних наукових праць у дисертації використано лише авторські ідеї та положення.

У друкованих працях, опублікованих у співавторстві, ідеї та принципи, що використані в дисертаційному дослідженні, є результатом індивідуальної праці автора, а саме: [1] – розроблено методи нечіткої кластеризації викривлених даних на основі еволюційного алгоритму котячої зграї; [2] – модифіковано методи нечіткої кластеризації викривлених даних, що базуються на адаптивному самонавчанні; [3] - запропоновано онлайн нечітку кластеризацію неповних даних із використанням правдоподібного підходу та міри схожості спеціального типу; [4] - розроблено онлайн метод до нечіткої можливісної кластеризації даних із використанням еволюційних алгоритмів; [5] - запропоновано модифікацію оптимізаційної процедури божевільних котів; [6] – запропоновано онлайн швидку нечітку правдоподібну кластеризацію на основі аналізу піків щільності розподілу даних; [7] – запропоновано модифікацію нечіткої правдоподібної кластеризації масивів даних, що базується на ідеях аналізу щільностей розподілу цих даних та їх піків; [8] - запропоновано кластеризацію потоків даних на основі піків щільності розподілу даних та еволюційного методу котячих зграй; [9] – запропоновано підхід до кластеризації масивів даних, що описано як у векторній, так і матричній формах на основі оптимізації функцій щільності розподілу даних у цих масивах; [10] – покращено алгоритм сірого вовка; [11] – розроблено метод адаптивної правдоподібної нечіткої кластеризації даних, призначений для вирішення проблем Data Stream Mining, коли дані надходять на обробку в онлайн режимі; [12] - введено рандомізовану модифікацію базової процедури котячих зграй; [13] - введено прискорену модифікацію методу котячих зграй; [14] - введено процедуру градієнтної оптимізації нечіткої правдоподібної кластеризації; [15] – запропоновано модифікацію онлайн нечіткої кластеризації потоків даних на основі еволюційної оптимізації котячої зграї; [16] – модифіковано метод правдоподібної нечіткої кластеризації потоків даних; [17]- розроблено метод адаптивної правдоподібної нечіткої кластеризації даних на основі еволюційного алгоритму; [18] - запропоновано стратегію найближчого прототипу - центроїда з використанням оптимізаційних процедур; [19] - введено оптимізаційну функцію модифікованого алгоритму косяків риб, випадкового пошуку та еволюційної оптимізації; [20] - введено еволюційну оптимізацію алгоритму косяків риб; [21] – розроблено рекурентну модифікацію алгоритму Густафсона – Кесселя;

[22] - модифіковано алгоритм сірого вовка; [23] – введено процедуру оптимального розширення з використанням часткових відстаней; [24] - запропоновано модифікацію нечіткої кластеризації спотворених наборів даних за допомогою еволюційної оптимізації; [25] - запропоновано стратегію завершення; [26] - запропоновано міру подібності; [27] - модифіковано алгоритм зграї котів; [28] – розроблено метод онлайн правдоподібної нечіткої кластеризації потоків даних; [29] - запропоновано еволюційну оптимізацію котячої зграї; [30] - введено модифікацію процедури нечіткої кластеризації, що заснована на алгоритмі Густафсона-Кесселя; [31] - розроблено процедуру адаптивного відновлення спотворених потоків даних; [32] – введено функцію розподілу щільності Коші; [33] – модифіковано еволюційний алгоритм сірих вовків; [34] – запропоновано міру подібності спеціального типу; [35] - введено модифікацію нечіткої кластеризації масивів даних; [36] - введено спеціальну процедуру вимірювання подібності; [37] - запропоновано онлайн рекурентний підхід до нечіткої правдоподібної кластеризації; [38] – модифіковано онлайн нечіткий правдоподібний метод кластеризації викривлених даних; [39] - запропоновано адаптивну нечітку кластеризацію даних на основі еволюційних процедур; [40] - запропоновано правдоподібну нечітку кластеризацію даних на основі еволюційних процедур.

Апробація результатів дисертації. Основні теоретичні та практичні результати дисертаційної роботи були представлені та обговорені на наукових міжнародних конференціях: IEEE Second International Conference “Data Stream Mining & Processing”, DSMP 2018, 20-25 August, Lviv 2018; International Workshop on Computer Modeling and Intelligent Systems (CMIS) 2019, 2020, 2023; The 8th International Conference on Advanced Optoelectronics and Lasers (CAOL*2019); International Conference on Advanced Computer Information Technologies, ACIT 2019, 2020, 2021; IEEE Ukraine Conference on Electrical and Computer Engineering (UKRCON), 2019; Міжнародному науковому симпозіумі «Інтелектуальні рішення» (IntSol-2019); Міжнародній науковій конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту», ISDMCI 2019, 2021; V International Scientific and Practical Conference Sofia, Bulgaria, 15-17 January 2020; I International Scientific and Practical Conference Graz, Austria 30-31 January 2020; 5th International Conference on Computational Linguistics and Intelligent Systems (COLINS-2021); V International Scientific and Practical Conference, Tokyo, Japan 8-10 December 2021; V International Scientific and Practical Conference, Kharkiv, Ukraine 28-30 November 2021.

Публікації. За результатами досліджень опубліковано 40 наукових праць серед яких: 2 монографії, що видано за кордоном; 20 статей (19 статей у періодичних фахових виданнях з технічних наук, 9 з яких опубліковано у фахових виданнях України категорії «А», що проіндексовано у наукометричних міжнародних базах Scopus та/або Web of Science, 1 стаття у періодичному закордонному англомовному виданні з технічних наук Європейського Союзу, Будапешт, Угорщина); 18 доповідей у матеріалах

міжнародних конференцій, 12 з яких включено до наукометричних міжнародних баз Scopus, Web of Science, DBLP.

Структура та обсяг роботи. Дисертаційна робота складається зі вступу, шести розділів, висновків, списку використаних джерел із 312 найменувань та додатків. Загальний обсяг дисертації становить 345 сторінок, у тому числі 254 сторінок основного тексту. Робота містить 65 рисунків та 56 таблиць.

Здобувачка висловлює подяку Бодянському Євгенію Володимировичу, доктору технічних наук, професору, професору кафедри штучного інтелекту Харківського національного університету радіоелектроніки за консультування та всебічну підтримку під час підготовки дисертації.

ОСНОВНИЙ ЗМІСТ РОБОТИ

У **вступі** обґрунтовано актуальність проблеми яка вирішується в дисертаційній роботі, розглянуто сучасний стан проблеми, визначені мета, об’єкт, предмет і методи дослідження, наведені задачі, що розв’язуються у дисертаційній роботі, зв’язок з науковими НДР, наведено наукову новизну та практичне значення отриманих результатів, перелік публікацій за темою роботи, надано інформацію про особистий внесок здобувачки.

У **першому розділі** проведено аналіз методів обробки потоків даних в умовах апріорної невизначеності та викривленості, що є актуальною науковою проблемою, яка потребує розвитку адаптивних методів, здатних працювати в динамічних і нестабільних середовищах; зроблено акцент на важливості вирішення задач кластеризації та аналізу даних за умов змінюваних характеристик потоку, що включає зміну кількості класів, їхньої структури та непередбачуваних дрейфів; встановлено, що традиційні алгоритми та методи машинного навчання є неефективними в задачах обробки потокових даних через їхню нездатність адаптуватися до швидких змін та викривлень вхідної інформації; обґрунтовано необхідність розробки онлайн-методів і механізмів самонавчання, які забезпечують стійкість до аномалій і знижують вплив викривлень, зокрема через адаптацію моделей у реальному часі; проведено аналіз сучасних підходів, що демонструє поступове вдосконалення алгоритмів обробки потоків даних шляхом інтеграції методів оптимізації, самонавчання та підвищення точності класифікації, що створює перспективи для успішного вирішення поставлених завдань у різних галузях. Проведений аналіз стану проблеми зі створення нових ефективних методів обчислювального інтелекту, а саме, кластеризації даних, на основі еволюційного самонавчання та надання їм адаптивних властивостей, що дає можливість опрацьовувати потоки нестаціонарних даних, збурених завадами та пропусками, які послідовно надходять на обробку в режимі реального часу, дає можливість зробити висновок про недостатню ефективність існуючих методів та підходів для вирішення задач інтелектуального аналізу потоків даних та необхідністю створення нових підходів та методів. Тому, розроблення адаптивних гібридних методів нечіткої кластеризації з використанням еволюційного

самонавчання, що здатні ефективно працювати за умов невизначеності, збурень, обмежених обчислювальних ресурсів та орієнтовані на онлайн-обробку даних, забезпечуючи високу точність навіть за відсутності повної апріорної інформації, є актуальною проблемою. Необхідно реалізувати такі задачі: розробити адаптивні методи нечіткої кластеризації потоків даних за умов перетинних класів та апріорної невизначеності; розробити методи адаптивної нечіткої кластеризації даних з різною щільністю розподілу; розробити еволюційні методи оптимізації в задачах нечіткої кластеризації масивів даних; розробити гібридні еволюційні методи нечіткої кластеризації масивів даних; тестування та експериментальна перевірка розроблених методів.

Другий розділ присвячено розробці та дослідженню методів адаптивної кластеризації потоків даних за умов перетинних класів та апріорної невизначеності.

При великій кількості спостережень ієрархічні методи кластерного аналізу непрацездатні. У таких випадках використовують неієрархічні методи, засновані на розділенні, що являє собою ітеративні методи дроблення вихідної сукупності. В процесі розподілу, нові кластери формуються до того часу, доки не буде виконано правило зупинки.

Така неієрархічна кластеризація полягає у розділенні набору даних на певну кількість окремих кластерів. Існує два підходи. Перший полягає у визначенні меж кластерів як найбільш щільних ділянок у багатовимірному просторі вихідних даних, тобто. визначення кластера там, де є велике «згущення точок». Другий підхід полягає в мінімізації міри відмінності об’єктів.

Вихідною інформацією для вирішення задачі кластеризації є масив багатовимірних векторів спостережень $X = \{x(1), x(2), ..., x(k), ...x(N)\} \subset R^n$, де $x(k) \in R^n$ - k-тий вектор – спостереження, k - або номер цього спостереження у масиві даних X, або поточний дискретний час у задачах Data Stream Mining. Якщо дані надходять на опрацювання послідовно у online режимі, ці дані повинні бути розбиті на m перетинних класів, при цьому для кожного $x(k)$ повинен бути також розрахований рівень нечіткої належності до кожного з кластерів $U_q(k), q = 1, 2, ...m$. Передбачається також, що дані, що надходять на опрацювання, передоброблені так, що $-1 \le x_i(k) \le 1$, де $x_i(k), i = 1, 2, ..., n$ - i-та компонента вектора спостережень $x(k)$.

Переважна більшість відомих алгоритмів нечіткої кластеризації передбачає, що вихідний масив даних X містить N спостережень і не змінюється в процесі аналізу. В той же час існує досить широкий клас задач Data Stream Mining, де дані надходять на обробку послідовно і їх обсяг апріорі є невідомим та Big Data Mining коли цей обсяг є настільки великим, що просто не дозволяє опрацьовувати ці дані у пакетному режимі. У таких ситуаціях на перший план виходять рекурентні алгоритми нечіткої кластеризації, за

допомогою яких ці дані аналізуються послідовно вектор за вектором в міру їх надходження в систему.

Запропоновано рекурентні ймовірнісні, можливісні та правдоподібні методи нечіткої кластеризації викривлених потоків даних, які призначені для вирішення задач Data Stream Mining, коли дані надходять на обробку послідовно і їх обсяг апріорі є невідомим та Big Data Mining, коли цей обсяг є настільки великим, що просто не дозволяє опрацьовувати ці дані у пакетному режимі, за допомогою яких ці дані аналізуються послідовно спостереження за спостереженням в міру їх надходження в систему.

Запропоновано рекурентний метод ймовірнісної нечіткої кластеризації, що пов'язаний з мінімізацією цільової функції

$$Goal\left(\mu_q(k), c_q\right) = \sum_{k=1}^{N}\sum_{q=1}^{m}\mu_q^{\beta}(k)\, d^2\left(x(k), c_q\right) \tag{1}$$

за наявності обмежень

$$\sum_{q=1}^{N}\mu_q(k) = 1, 0 < \sum_{q=1}^{N}\mu_q(k) < N, \tag{2}$$

де $\mu_q(k)$ – рівень нечіткої належності спостереження $x(k)$ до q-го кластера $Cl_q (1 \le q \le m)$;

c_q – прототип-центроїд q-го кластеру, що має бути уточнений в процесі послідовної рекурентної кластеризації;

$\beta > 1$ – параметр фаззіфікації, що задає «розмитість» границь кластерів;

$d\left(x(k), c_q\right)$ – відстань між $x(k)$ та c_q у прийнятій метриці.

Вирішення задачі нелінійного програмування (2.1), (2.2) за допомогою алгоритму Ерроу-Гурвіца-Удзави веде до процедури рекурентної кластеризації

$$\begin{cases} \mu_q(k+1) = \dfrac{\left(d^2\left(x(k+1), c_q(k)\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(d^2\left(x(k+1), c_l(k)\right)\right)^{\frac{1}{1-\beta}}}, \\ c_q(k+1) = c(k) + \eta(k+1)\,\mu_q^{\beta}(k+1)\left(x(k+1) - c_q(k)\right), \end{cases} \tag{3}$$

де $\eta(k)$ – параметр кроку навчання.

При значенні фаззіфікатора $\beta = 2$ приходимо до рекурентної версії нечітких *C*-середніх у вигляді

$$\begin{cases} \mu_q(k+1) = \dfrac{\left(d^2\left(x(k+1), c_q(k)\right)\right)^{-1}}{\sum\limits_{l=1}^{m}\left(d^2\left(x(k+1), c_l(k)\right)\right)^{-1}}, \\ c_q(k+1) = c(k) + \eta(k+1)\mu_q^{\beta}(k+1)\left(x(k+1) - c_q(k)\right). \end{cases} \quad (4)$$

Шляхом нескладних перетворень можна переписати співвідношення (4) у вигляді

$$\begin{cases} \mu_q(k+1) = \dfrac{1}{1 + \dfrac{d^2\left(x(k+1), c_q(k)\right)}{\sigma_q^2(k+1)}}, \\ \sigma_q^2(k+1) = \left(\sum\limits_{\substack{l=1 \\ l \neq q}}^{m}\left(d^2\left(x(k+1), c_l(k)\right)^{-1}\right)\right)^{-1}, \end{cases} \quad (5)$$

що по суті є функцією щільності розподілу Коші з параметром ширини $\sigma^2(k+1)$, тобто відповідає умовам, що висуваються до функцій сусідства у процедурах Т. Кохонена.

Запропоновано рекурентний можливісний метод нечіткої кластеризації. Можливісні методи нечіткої кластеризації пов'язані з мінімізацією цільової функції

$$Goal\left(\mu_q(k), c_q, \omega_q\right) = \sum_{k=1}^{N}\sum_{q=1}^{m}\mu_q^{\beta}(k)d^2\left(x(k), c_q\right) + \sum_{q=1}^{m}\omega_q\sum_{k=1}^{N}\left(1 - \mu_q(k)\right)^{\beta}, \quad (6)$$

де параметр ω_q визначає відстань між спостереженням та центроїдом c_q, на якій рівень належності $\mu_q(k)$ набуває значення 0,5.

Онлайн версія методу Крішнапурама-Келлера має вигляд:

$$\begin{cases} \mu_q(k+1) = \left(1 + \left(\dfrac{d^2\left(x(k+1), c_q(k)\right)}{\omega_q(k)}\right)^{\frac{1}{\beta-1}}\right)^{-1}, \\ c_q(k+1) = c_q(k) + \eta(k+1)\mu_q^{\beta}(k+1)\left(x(k+1) - c_q(k)\right), \\ \omega_q(k+1) = \dfrac{\sum\limits_{p=1}^{k+1}\mu_q^{\beta}(p)d^2\left(x(p), c_q(k+1)\right)}{\sum\limits_{p=1}^{k+1}\mu_q^{\beta}(p)} \end{cases} \quad (7)$$

або при $\beta = 2$

$$\begin{cases} \mu_q(k+1) = \left(1 + \dfrac{d^2\left(x(k+1), c_q(k)\right)}{\omega_q(k)}\right)^{-1}, \\ c_q(k+1) = c_q(k) + \eta(k+1)\mu_q^2(k+1)\left(x(k+1) - c_q(k)\right), \\ \omega_q(k+1) = \dfrac{\sum\limits_{p=1}^{k+1} \mu_q^2(p) d^2\left(x(p), c_q(k+1)\right)}{\sum\limits_{p=1}^{k+1} \mu_q^2(p)}. \end{cases} \quad (8)$$

І знов таки тут у першому співвідношенні (8) виникає функція Коші з параметром ширини ω_q, що визначається третім співвідношенням (8).

Запропоновано адаптивну нечітку робастну кластеризацію даних на основі міри подібності. Вводячи у розгляд цільову функцію, засновану на мірі подібності

$$E_S(\mu_q(k), c_q) = \sum_{k=1}^{N}\sum_{q=1}^{m} \mu_q^{\beta}(k) S(x_k, c_q) = \sum_{k=1}^{N}\sum_{q=1}^{m} \frac{\mu_q^{\beta}(k)\sigma^2}{\sigma^2 + \left\|x_k - c_q\right\|^2},$$

ймовірнісні обмеження

$$\sum_{q=1}^{m} \mu_q(k) = 1,$$

функцію Лагранжа

$$L_S(\mu_q(k), c_q, \lambda(k)) = \sum_{k=1}^{N}\sum_{q=1}^{m} \frac{\mu_q^{\beta}(k)\sigma^2}{\sigma^2 + \left\|x_k - c_q\right\|^2} + \sum_{k=1}^{N} \lambda(k)\left(\sum_{q=1}^{m} \mu_q(k) - 1\right) \quad (9)$$

і вирішуючи систему рівнянь Каруша-Куна-Таккера, приходимо до вирішення

$$\begin{cases} \mu_q(k) = (S(x_k, c_q))^{\frac{1}{\beta-1}} \Big/ \sum\limits_{l=1}^{m} (S(x_k, c_l))^{\frac{1}{\beta-1}}, \\ \lambda(k) = -\left(\sum\limits_{l=1}^{m} (\beta S(x_k, c_l))^{\frac{1}{\beta-1}}\right)^{\beta-1}, \\ \nabla_{c_q} L_S(\mu_q(k), c_q, \lambda(k)) = \sum\limits_{k=1}^{N} \mu_q^{\beta}(k) \dfrac{x_k - c_q}{(\sigma^2 + \left\|x_k - c_q\right\|^2)^2} = \vec{0}. \end{cases} \quad (10)$$

Вважаючи значення фаззіфікатора $\beta = 2$ приходимо до робастного варіанту FCM:

$$\begin{cases} \mu_q(k+1) = \dfrac{S(x_{k+1}, c_q(k))}{\sum\limits_{l=1}^{m} S(x_{k+1}, c_l(k))}, \\ c_q(k+1) = c_q(k) + \eta(k+1)\dfrac{\mu_q^2(k+1)}{(\sigma^2 + \left\| x_{k+1} - c_q(k) \right\|^2)^2}. \end{cases}$$

Використовуючи далі концепцію прискореного часу, можна ввести робастну адаптивну ймовірнісну процедуру нечіткої кластеризації виду (11), при цьому рішення про належність кожного x_k до конкретного кластера приймається за максимальним значенням міри подібності

$$\begin{cases} \mu_q^{(\tau+1)}(k) = \dfrac{(S(x_k, c_q^{(\tau)}(k)))^{\frac{1}{\beta-1}}}{\sum\limits_{l=1}^{m} (S(x_k, c_l^{(\tau)}(k)))^{\frac{1}{\beta-1}}}, \\ c_q^{(Q)}(k) = c_q^{(0)}(k+1), \\ c_q^{(\tau+1)}(k+1) = c_q^{(\tau)}(k+1) + \eta(k+1)\dfrac{(\mu_q^{(Q)}(k))^{\beta}}{(\sigma^2 + \left\| x_{k+1} - c_q^{(\tau)}(k+1) \right\|^2)^2}(x_{k+1} - c_q^{(\tau)}(k+1)). \end{cases} \quad (11)$$

Аналогічним чином може бути синтезований метод робастної адаптивної можливісної нечіткої кластеризації. Вводячи цільову функцію

$$Goal_S(\mu_q(k), c_q, \omega_q) = \sum_{k=1}^{N}\sum_{q=1}^{m} \mu_q^{\beta}(k) S^{-1}(x_k, c_q) + \sum_{q=1}^{m} \omega_q \sum_{k=1}^{N} (1 - \mu_q(k))^{\beta}$$

і вирішуючи задачу її оптимізації, приходимо до процедури:

$$\begin{cases} \mu_q(k+1) = \left(1 + \left(\dfrac{S^{-1}(x_{k+1}, c_q(k))}{\omega_q(k)} \right) \right)^{-1}, \\ c_q(k+1) = c_q(k) + \eta(k+1)\mu_q^{\beta}(k+1)\dfrac{x_{k+1} - c_q(k)}{(\sigma^2 + \left\| x_{k+1} - c_q(k) \right\|^2)^2}, \\ \omega_q(k+1) = \dfrac{\sum\limits_{p=1}^{k+1} \mu_q^{\beta}(p) S^{-1}(x_p, c_q(k+1))}{\sum\limits_{p=1}^{k+1} \mu_q^{\beta}(p)}, \end{cases} \quad (12)$$

що приймає при $\beta = 2$ вигляд

$$\begin{cases} \mu_q(k+1) = \dfrac{1}{1+\dfrac{S^{-1}(x_{k+1}, c_q(k))}{\omega_q(k)}}, \\ c_q(k+1) = c_q(k) + \eta(k+1)\dfrac{\mu_q^2(k+1)}{(\sigma^2 + \left\| x_{k+1} - c_q(k) \right\|^2)^2}(\tilde{x}_{k+1} - w_q(k)), \\ \omega_q(k+1) = \dfrac{\sum\limits_{p=1}^{k+1} \mu_q^2(p) S^{-1}(x_p, c_q(k+1))}{\sum\limits_{p=1}^{k+1} \mu_q^2(p)}. \end{cases}$$

І, нарешті, вводячи прискорений час, отримуємо процедуру

$$\begin{cases} \mu_q^{(\tau+1)}(k) = \dfrac{1}{1+\left(\dfrac{S^{-1}(x_k, c_q^{(\tau)}(k))}{\omega_q^{(\tau)}(k)}\right)^{\frac{1}{\beta-1}}}, \\ c_q^{(Q)}(k) = c_q^{(0)}(k+1), \\ c_q^{(\tau+1)}(k+1) = c_q^{(\tau)}(k+1) + \eta(k+1)\dfrac{(\mu_q^{(Q)}(k))^\beta}{(\sigma^2 + \left\| x_{k+1} - c_q^{(\tau)}(k+1) \right\|^2)^2}(x_{k+1} - c_q^{(\tau)}(k+1)), \\ \omega_q^{(\tau+1)}(k) = \dfrac{\sum\limits_{p=1}^{k} (\mu_q^{(\tau+1)}(p))^\beta S^{-1}(x_p, c_q^{(\tau+1)}(k))}{\sum\limits_{p=1}^{k} (\mu_q^{(\tau+1)}(p))^\beta}. \end{cases}$$

Запропоновано метод адаптивної нечіткої робастної кластеризації даних з пропусками на основі міри подібності. Для вирішення задачі робастної кластеризації даних з пропусками, введемо до розгляду часткову міру подібності (PSM), що є гібридом часткової відстані (PD) та міри подібності (SM). Така PSM має вигляд:

$$S_P(x_k, c_q) = \frac{\sigma^2}{\sigma^2 + d_P^2(x_k, c_q)}, \tag{13}$$

що дозволяє отримати алгоритми з бажаними властивостями на основі процедур, описаних вище. Так, на основі процедури (11) можна ввести адаптивний робастний ймовірнісний алгоритм нечіткої кластеризації даних з пропусками:

$$\begin{cases} \mu_q^{(\tau+1)}(k) = \dfrac{(S_P(x_k^{(\tau)}, c_q^{(\tau)}(k)))^{\frac{1}{\beta-1}}}{\sum\limits_{l=1}^{m}(S_P(x_k^{(\tau)}, c_l^{(\tau)}))^{\frac{1}{\beta-1}}}, \\ x_{ki}^{(\tau)} = c_{qi}^{(\tau)},\ c_q^{(\tau)}(k) = \arg\max\limits_q \{S_P(x_k^{(\tau)}, c_1^{(\tau)}(k)), ..., S_P(x_k^{(\tau)}, c_m^{(\tau)}(k))\}, \\ c_q^{(Q)}(k) = c_q^{(0)}(k+1), \\ c_q^{(\tau+1)}(k+1) = c_q^{(\tau)}(k+1) + \eta(k+1)\dfrac{(\mu_q^{(Q)}(k))^{\beta}}{(\sigma^2 + \left\| x_{k+1}^{(\tau)} - c_q^{(\tau)}(k+1) \right\|^2)^2}(x_{k+1}^{(\tau)} - c_q^{(\tau)}(k+1)), \end{cases} \tag{14}$$

а також можна записати адаптивний робастний можливісний метод нечіткої кластеризації даних з пропусками (14). Таким чином, використання часткової міри подібності, заснованої на частковій відстані, дозволяє вирішувати задачі нечіткої кластеризації даних, що містять як пропуски, так і аномальні спостереження:

$$\begin{cases} \mu_q^{(\tau+1)}(k) = \dfrac{1}{1 + \left(\dfrac{S^{-1}(x_k, c_q^{(\tau)}(k))}{\omega_q^{(\tau)}(k)} \right)^{\frac{1}{\beta-1}}}, \\ x_{ki}^{(\tau)} = c_{qi}^{(\tau)},\ c_q^{(\tau)}(k) = \arg\max\limits_q \{S_P(a_k^{(\tau)}, c_1^{(\tau)}(k)), ..., S_P(x_k^{(\tau)}, c_m^{(\tau)}(k))\} \\ c_q^{(Q)}(k) = c_q^{(0)}(k+1), \\ c_q^{(\tau+1)}(k+1) = c_q^{(\tau)}(k+1) + \eta(k+1)\dfrac{(\mu_q^{(Q)}(k))^{\beta}}{(\sigma^2 + \left\| x_{k+1} - c_q^{(\tau)}(k+1) \right\|^2)^2}(x_{k+1}^{(\tau)} - c_q^{(\tau)}(k+1)), \\ \omega_q^{(\tau+1)}(k) = \dfrac{\sum\limits_{p=1}^{k}(\mu_q^{(\tau+1)}(p))^{\beta} S_P^{-1}(x_p, c_q^{(\tau+1)}(k))}{\sum\limits_{p=1}^{k}(\mu_q^{(\tau)}(p))^{\beta}}. \end{cases}$$

Запропоновано рекурентний правдоподібний метод нечіткої кластеризації. Правдоподібні методи нечіткої кластеризації пов'язані з мінімізацією цільової функції:

$$Goal\left(Credib_q(k), c_q\right) = \sum_{k=1}^{N}\sum_{q=1}^{m} Credib_q^{\beta}(k) d^2\left(x(k), c_q\right) \tag{15}$$

за наявності обмежень

$$\begin{cases} 0 \le Credib_q(k) \le 1 \forall q, k, \\ \sup Credib_q(k) \ge 0,5 \forall k, \\ Credib_q(k) + \sup Credib_l(k) = 1 \end{cases} \tag{16}$$

для всіх q,k, для яких $Credib_q(k) \geq 0$, де $Credib_q(k)$ – рівень правдоподібності того, що спостереження $x(k)$ належить кластеру Cl_q. Таким чином, якщо пакетний метод правдоподібної нечіткої кластеризації має вигляд:

$$\begin{cases} \mu_q(k) = \dfrac{1}{1 + d^2\left(x(k), c_q\right)}, \\ \mu_q^*(k) = \dfrac{\mu_q(k)}{\sup \mu_l(k)}, \\ Credib_q(k) = \dfrac{1}{2}\left(\mu_q^*(k) + 1 - \sup \mu_l^*(k)\right), \\ c_q = \dfrac{\sum\limits_{k=1}^{N} Credib_q^{\beta}(k) x(k)}{\sum\limits_{k=1}^{N} Credib_q^{\beta}(k)}, \end{cases} \tag{17}$$

то його рекурентна версія описується наступним виразом:

$$\begin{cases} \sigma_q^2(k+1) = \sum\limits_{\substack{l=1 \\ l \neq q}}^{m} \left(d^2\left(x(k+1), c_l(k)\right)^{\frac{1}{1-\beta}} \right)^{-1}, \\ \mu_q(k+1) = \dfrac{1}{1 + \dfrac{\left(d^2\left(x(k+1), c_q(k)\right)\right)^{\beta-1}}{\sigma_q^2(k+1)}}, \\ \mu_q^*(k+1) = \dfrac{\mu_q(k+1)}{\sup \mu_l(k+1)}, \\ Credib_q(k+1) = \dfrac{1}{2}\left(\mu_q^*(k+1) + 1 - \sup \mu_l^*(k+1)\right), \\ c_q(k+1) = c_q(k) + \eta(k+1) Credib_q^{\beta}(k+1)\left(x(k+1) - c_q(k)\right). \end{cases} \tag{18}$$

З обчислювальної точки зору рекурентний метод правдоподібної нечіткої кластеризації не є складнішим у порівнянні з онлайн версіями ймовірнісних, можливісних та робастних процедур.

Онлайн версія методу правдоподібної нечіткої кластеризації набуває вигляду

$$\begin{cases} \sigma_q^2(k+1) = \left(\sum\limits_{\substack{l=1 \\ l \neq q}}^{m} \left\| x(k+1) - c_q(k) \right\|^{-2} \right)^{-1}, \\ \mu_q(k+1) = \left(1 + \dfrac{\left\| x(k+1) - c_q(k) \right\|^2}{\sigma_q^2(k+1)} \right)^{-1} \\ \mu_q^*(k+1) = \mu_q(k+1) \left(\sup \mu_l(k+1) \right)^{-1}, \\ Credib_q(k+1) = \dfrac{1}{2} \left(\mu_q^*(k+1) + 1 - \sup\limits_{l \neq q} \mu_l^*(k+1) \right), \\ c_q(k+1) = c_q(k) + \eta(k+1) Credib_q^{\beta}(k+1) \left(x(k+1) - c_q(k) \right). \end{cases} \quad (19)$$

Нескладно також модифікувати метод Густафсона-Кесселя на випадок можливісної нечіткої кластеризації. При цьому цільова функція (6) набуває вигляду

$$Goal\left(\mu_q(k), c_q, \omega_q \right) = \sum_{k=1}^{N} \sum_{q=1}^{m} \mu_q^{\beta}(k) d_{V_q}^2 \left(x(k), c_q \right) + \sum_{q=1}^{m} \omega_q \sum_{k=1}^{N} \left(1 - \mu_q(k) \right)^{\beta}, \quad (20)$$

а пакетна форма методу:

$$\begin{cases} \mu_q(k) = \left(1 + \dfrac{d_{V_q}^2 \left(x(k), c_q \right)^{\frac{1}{\beta-1}}}{\omega_q} \right)^{-1}, \\ c_q = \dfrac{\sum\limits_{k=1}^{N} \mu_q^{\beta}(k) x(k)}{\sum\limits_{k=1}^{N} \mu_q^{\beta}(k)}, \\ S_q = \sum\limits_{k=1}^{N} \mu_q^{\beta}(k) \left(x(k) - c_q \right) \left(x(k) - c_q \right)^T, \\ V_q = \left(\det S_q \right)^{\frac{1}{n}} S_q^{-1}, \\ \omega_q(k) = \dfrac{\sum\limits_{k=1}^{N} \mu_q^{\beta}(k) \left(x(k) - c_q \right)^T V_q \left(x(k) - c_q \right)}{\sum\limits_{k=1}^{N} \mu_q^{\beta}(k)}. \end{cases} \quad (21)$$

Перепишемо метод (21) у рекурентній формі:

$$
\begin{cases}
\mu_q(k)=\left(1+\left(\dfrac{d^2_{V_q}(k)\left(a(k+1),c_q(k)\right)}{\omega_q(k)}\right)^{\frac{1}{\beta-1}}\right)^{-1},\\
S_q(k+1)=S_q(k)+\mu_q^{\beta}(k+1)\left(x(k+1)-c_q(k)\right)\left(x(k+1)-c_q(k)\right)^T,\\
S_q^{-1}(k+1)=S_q^{-1}(k)-\dfrac{\mu_q^{\beta}(k+1)S_q^{-1}(k)\left(x(k+1)-c_q(k)\right)\left(x(k+1)-c_q(k)\right)^T S_q^{-1}(k)}{1+\mu_q^{\beta}(k+1)\left(x(k+1)-c_q(k)\right)^T S_q^{-1}(k)\left(x(k+1)-c_q(k)\right)},\\
\det S_q(k+1)=\left(\det S_q(k)\right)\left(1+\mu_q^{\beta}(k+1)\left(x(k+1)-c_q(k)\right)^T\left(x(k+1)-c_q(k)\right)\right),\\
V_q(k+1)=\left(\det S_q(k+1)\right)^{\frac{1}{n}}S_q^{-1}(k+1),\\
c_q(k+1)=c_q(k)+\eta(k+1)\mu_q^{\beta}(k+1)V_q(k+1)\left(x(k+1)-c_q(k)\right),\\
\omega_q(k+1)=\dfrac{\sum\limits_{p=1}^{k+1}\mu_q^{\beta}(p)\left(x(p)-c_q(k+1)\right)^T V_q(k+1)\left(x(p)-c_q(k+1)\right)}{\sum\limits_{p=1}^{k+1}\mu_q^{\beta}(p)}.
\end{cases}
\tag{22}
$$

Модифікація правдоподібного варіанту методу Густафсона-Кесселя має вигляд:

$$
Goal\left(Credib_q(k),c_q\right)=\sum_{k=1}^{N}\sum_{q=1}^{m}Credib_q^{\beta}(k)\,d^2_{V_q}\left(x(k),c_q\right)
\tag{23}
$$

$$
\begin{cases}
S_q=\sum\limits_{k=1}^{N}\mu_q^{\beta}(k)\left(x(k)-c_q\right)\left(x(k)-c_q\right)^T,\\
V_q=\left(\det S_q\right)^{\frac{1}{n}}S_q^{-1},\\
\mu_q(k)=\dfrac{1}{1+d^2_{V_q}\left(x(k),c_q\right)},\\
\mu_q^{*}(k)=\dfrac{\mu_q(k)}{\sup\mu_l(k)},\\
Credib_q(k)=\dfrac{1}{2}\left(\mu_q^{*}(k)+1-\sup\mu_l^{*}(k)\right),\\
c_q=\dfrac{\sum\limits_{k=1}^{N}Credib_q^{\beta}(k)x(k)}{\sum\limits_{k=1}^{N}Credib_q^{\beta}(k)}.
\end{cases}
\tag{24}
$$

Введено рекурентну модифікацію методу Густафсона–Кесселя правдоподібної нечіткої кластеризації:

$$
\begin{cases}
\mu_q(k+1) = \left(1 + d^2_{V_q}(k)\left(x(k+1), c_q(k)\right)\right)^{-1}, \\
\mu^*_q(k+1) = \dfrac{\mu_q(k+1)}{\sup \mu_l(k+1)}, \\
Credib_q(k+1) = \dfrac{1}{2}\left(\mu^*_q(k+1) - 1 - \sup \mu^*_l(k+1)\right), \\
S_q(k+1) = S_q(k) + \mu^\beta_q(k+1)\left(x(k+1) - c_q(k)\right)\left(x(k+1) - c_q(k)\right)^T, \\
S^{-1}_q(k+1) = S^{-1}_q(k) - \dfrac{\mu^\beta_q(k+1) S^{-1}_q(k)\left(x(k+1) - c_q(k)\right)\left(x(k+1) - c_q(k)\right)^T S^{-1}_q(k)}{1 + \mu^\beta_q(k+1)\left(x(k+1) - c_q(k)\right)^T S^{-1}_q(k)\left(x(k+1) - c_q(k)\right)}, \\
\det S_q(k+1) = \left(\det S_q(k)\right)\left(1 + \mu^\beta_q(k+1)\left(x(k+1) - c_q(k)\right)^T \left(x(k+1) - c_q(k)\right)\right), \\
V(k+1) = \left(\det S_q(k+1)\right)^{\frac{1}{n}} S^{-1}_q(k+1), \\
c_q(k+1) = c_q(k) + \eta(k+1)\mu^\beta_q(k+1) V_q(k+1)\left(x(k+1) - c_q(k)\right).
\end{cases}
$$

Проведені експериментальні дослідження розроблених та модифікованих методів підтвердили, що в порівнянні з класичними методами кластеризації (K-means, FCM), розроблені адаптивні методи нечіткої кластеризації дозволяють підвищити точність визначення кількості класів (кластерів) в умовах дефіциту апріорної інформації, працездатні як в пакетному так і в онлайн режимах та здатні працювати на вибірках, що змінюють розмірність та форму кластерів; дозволяють обробляти великі обсяги даних, що можуть подаватись на обробку послідовно у формі потоків даних, ефективно працювати за умов суттєвої невизначеності, стохастичності, нелінійності, апріорної невизначеності, нестаціонарності та є найбільш пристосованими для вирішення задач Data Mining та Data Stream Mining, завдяки своїм універсальним апроксимуючим властивостям, здатності до самонавчання.

Результати розділу 2 відображено у публікаціях [2, 3, 14, 16, 21, 23, 25, 26, 28, 30-33, 36-38].

Третій розділ присвячено розробці методів адаптивної нечіткої кластеризації даних з різною щільністю розподілу. Слід відзначити, що в загальному випадку вирішення задачі кластеризації суттєво ускладняється, якщо вихідні вектори (тут у загальному випадку матриці) спостереження мають велику різноманітність, викривлені збуреннями та завадами, містять пропуски, самі вихідні масиви або занадто великі (Big Data) або занадто короткі, кластери можуть мати досить складну форму, а їх кількість апріорі невідома. У цьому випадку найбільш ефективними (але й найбільш складними) є алгоритми, що базуються на аналізі щільностей розподілу даних, серед яких в якості одного з найбільш «популярних» є DENCLUE (Density-based Clustering of Applications with Noise) та його модифікації, що були

запропоновані для вирішення задач кластеризації великих масивів векторних даних високої розмірності, при цьому класи, що формуються у процесі кластеризації, можуть мати будь яку складну форму. В основі цих алгоритмів полягає пошук екстремумів максимумів функції щільності розподілу даних у масиві, що аналізується (багатоекстремальна оптимізація), при цьому ця функція формується, як суперпозиція ядерних (дзвонуватих) функцій, пов'язаних з кожним спостереженням. Фактично ця функція будується на основі вікон Парзена та оцінок Надарая - Ватсона.

З обчислювальної точки зору задача кластеризації перетворюється у проблему пошуку локальних екстремумів багатоекстремальної функції векторного аргументу щільності зо допомогою градієнтних процедур, які багаторазово запускаються з різних точок вихідного масиву даних. Зрозуміло, що це займає досить багато часу, оскільки апріорі навіть невідомо скільки ж екстремумів має сформована функція щільності.

Пришвидшити процес пошуку цих екстремумів можна, скориставшись ідеями еволюційної оптимізації, що включає в себе алгоритми, інспіровані природою, ройові алгоритми, популяційні алгоритми, тощо. При цьому пошук ведеться одночасно групою агентів, що діють або незалежно, або у взаємодії, що дозволяє суттєво пришвидшити процес пошуку екстремумів, кожен з яких «відповідає» тому або іншому кластеру, що формується.

Вихідною інформацією для вирішення задачі кластеризації традиційно є масив векторів-спостережень $X = \{x(1), x(2), ..., x(k), ..., x(N)\}$, $x(k) = \{x_i(k)\} \in R^n$, при цьому дані попередньо відцентровано на гіперкуб так, що $x(k) = \{x_{i,i_2}(k)\} \in R^{n_1 \times n_2}$. Така ситуація може виникати у випадку обробки масивів зображень.

Основними поняттями, на яких базується DENCLUE є функція впливу, функція щільності та атрактори щільності, що за суттю є локальними екстремумами функції щільності.

У загальному випадку функція впливу для будь-якого векторного спостереження $x(\bullet)$ з вихідного масиву X є ядерною дзвонуватою функцією $f^{x(\bullet)}(x)$, при цьому найбільш популярною є традиційна гаусівська функція:

$$f_G^{x(\bullet)}(x) = \exp\left(-\frac{d^2(x, x(\bullet))}{2\sigma^2}\right) = \exp\left(-\frac{\|x - x(\bullet)\|^2}{2\sigma^2}\right) \quad (25)$$

(тут $d^2(x, x(\bullet))$ - евклідова відстань, σ^2 - параметр ширини функції впливу), завдяки простоті обчислення її похідних.

У матричному випадку замість евклідової можна використати метрику Фробеніуса, при цьому функція впливу набуває вигляду:

$$f_G^{x(\bullet)}(x) = \exp\left(-\frac{d^2(x, x(\bullet))}{2\sigma^2}\right) = \exp\left(-\frac{Tr\left(x - x(\bullet)\right)\left(x - x(\bullet)\right)^T}{2\sigma^2}\right), \quad (26)$$

де $Tr(\bullet)$ - символ сліду матриці.

На основі функцій впливу формується функція щільності розподілу даних у масиві X у вигляді

$$f^x(x)=\sum_{k=1}^{N} f\left(x,x(k)\right), \tag{27}$$

що по суті є оцінкою Надарая-Ватсона. Нескладно бачити, що функція $f^x(x)$ може приймати значення в інтервалі $1\le f^x(x)\le N,$ при цьому крайні значення з цього інтервалу приймаються, коли вибірка містить лише одне спостереження або усі N спостережень співпадають, тобто існує лише один кластер - вироджена ситуація. Для знаходження $m>1$ кластерів необхідно ввести у розгляд деякий поріг $\xi>1,$ що дозволяє формувати дійсно значущі кластери, відстежуючи аномальні спостереження та класи, що містять занадто мало даних. Власне процес формування кластерів пов'язаний з відшуканням усіх екстремумів функції щільності (27) за допомогою градієнтної процедури

$$x^l=x^{l-1}+\eta^l\frac{\nabla f^x\left(x^l,x^{l-1}\right)}{\left\|\nabla f^x\left(x^l,x^{l-1}\right)\right\|},\ \ x_0=x\left(k\right),l=0,1,2,...;\forall k=1,2,...,N, \tag{28}$$

тобто кількість запусків алгоритму (28) визначається обсягом навчальної вибірки N. Зрозуміло, що при великих N процес кластеризації - пошуку локальних екстремумів може потребувати дуже багато часу. Тому запропоновані модифікації DENCLUE пов'язані з пришвидшенням процесу пошуку локальних екстремумів (27) шляхом модифікації градієнтної процедури (28). У випадку, коли спостереження $x(k)$ у вибірці $X\in\left(n_1\times n_2\right)$ - матриця, нескладно ввести у розгляд матричний варіант процедури (28):

$$x^l=x^{l-1}+\eta^l\Gamma^x\left(x,x^{l-1}\right)\left(Tr\Gamma^x\left(x,x^{l-1}\right)\Gamma^{xT}\left(x,x^{l-1}\right)\right)^{-\frac{1}{2}},$$

де $\Gamma^x\left(x,x^{l-1}\right)=\left\{\frac{\partial f^x\left(x,x^{l-1}\right)}{\partial x_{i_1i_2}}\right\}\in R^{n_1\times n_2}.$

Процес градієнтної оптимізації закінчується відшуканням m локальних екстремумів функції (27), при цьому чим менше значення ξ, тим більше кластерів може бути сформовано. Пришвидшити процес відшукання локальних екстремумів можна, використовуючи замість градієнтного пошуку методи еволюційної оптимізації, серед яких в якості достатньо ефективного, чисельно простого і швидкого можна відзначити, так званий, пошук на основі котячих ігрaй, що повинен бути модифікований для вирішення задачі кластеризації.

Запропоновано модифікацію нечіткого методу кластеризації на основі піків щільності розподілу даних. Процес нечіткої кластеризації на основі аналізу піків щільності розподілу даних зручно представити у вигляді послідовності кроків, при цьому вихідною інформацією як і в інших методах, заснованих на парадигмі самонавчання, є нерозмічена вибірка векторних

спостережень $X=\{x(1),x(2),...,x(k),...x(N)\}$, $x(k)\in R^n$, при цьому для зручності розрахунків всі компоненти цих векторів попередньо закодовані в деякому обмеженому інтервалі, наприклад, $-1\le x_i(k)\le 1$, $x_i(k), i=1,2,...,n$.

Роботу методу сформульовано наступною послідовністю елементарних кроків:

Крок 1. На першому кроці на основі вихідної $(n\times N)$ матриці «об'єкт – властивість» вводиться $(N\times N)$ – матриця відстаней між спостереженнями:

$$D=\{d_{kl}\},\ d_{kl}=\|x(k)-x(l)\|\ \forall k,l,$$

при цьому може бути використана будь-яка метрика, яка використовується в інтелектуальному аналізі даних і, зокрема, в кластерному аналізі.

Крок 2. На другому кроці розраховується $(N\times 1)$ - вектор локальних щільностей $\rho=\{\rho_k\}\in R^N$:

$$\rho_k=\sum_{k=1}^{N}\chi(d_{kl}-d_c),\ \text{де}\ \chi(d)=\begin{cases}1,\ \text{якщо}\ d<0,\\ 0,\ \text{в іншому випадку.}\end{cases}$$

Крок 3. Розрахунок вектора мінімальних відстаней $\delta=\{\delta_k\}\in P^N$ до точок з більш високою щільністю $\delta_k=\min\limits_{\forall l,\rho_l>\rho_k}\{d_{kl}\}$, а для точки з максимальною щільністю δ_k^* розраховується: $\delta_k^*=\max\limits_{l}\{d_{kl}\}$.

Крок 4. Формування центроїдів кластерів $c_q, q=1,2,...,m,$ при цьому в якості центроїдів $c_q=x(k)$ обираються точки з найвищою щільністю, тобто обираються деякі спостереження з вихідної вибірки X. До кожного з центроїдів c_j приписуються точки, найближчі до нього в сенсі $\min(d_{kl})\equiv d_{ql}$.

Далі всі центроїди впорядковуються за зменшенням цього добутку $c_1,...,c_q,...c_m$, а якість одержуваного рішення оцінюється за допомогою будь-якого з критеріїв, прийнятих в чіткій кластеризації. Якщо з точки зору використаного критерію якість кластеризації виявляється незадовільною, можна або зменшити значення d_c, або збільшити число можливих кластерів, тобто $q=1,...,m,m+1,m+2,....$ Далі процедура нечіткої кластеризації повторюється, починаючи з першого кроку.

Крок 5. Починаючи з п'ятого кроку реалізується процедура нечіткої кластеризації. При цьому для кожної точки $x(k)\ne c_q$ розраховуються рівні нечіткої належності в стандартній формі

$$\mu_q(k)=\frac{d_{qk}^{-2}}{\sum\limits_{l=1}^{m}d_{lk}^{-2}},\qquad(29)$$

або на основі функції щільності розподілу Коші

$$\mu_q(k)=\left(1+\frac{d_{qk}^{-2}}{\sigma_q^2}\right). \tag{30}$$

Крок 6. На основі оцінок імовірнісної нечіткої належності (29), (30) розраховується рівень довіри до отриманих результатів на основі стандартного правдоподібного підходу

$$Cred_q(k)=\frac{1}{2}\left(\mu_q^*(k)+1-\sup\mu_q^*(k)\right). \tag{31}$$

Крок 7. Завершення процедури нечіткої кластеризації шляхом оцінки якості результатів за допомогою будь-якого з критеріїв, що застосовуються в нечіткій кластеризації, хоча оцінка (31) вже сама по собі надає наскільки можна довіряти правдоподібності отриманих результатів.

Оцінки (29), (30) пов'язані з, так званою, ймовірнісною нечіткою кластеризацією. На основі оцінок можуть бути розраховані рівні довіри отриманих результатів за допомогою співвідношень

$$\begin{cases} Cred_q\left(k\right)=\frac{1}{2}\left(\mu_q^*\left(k\right)+1-\sup\mu_r^*\left(k\right)\right), \\ \mu_q^*\left(k\right)=\frac{\mu_q\left(k\right)}{\sup\mu_r\left(k\right)}. \end{cases}$$

Таким чином, введена процедура нечіткої кластеризації, що базується на аналізі щільностей розподілу даних та їх піків, дозволяє скоротити час вирішення задачі за рахунок зменшення кількості звернень до блоку оптимізації, що відшукує екстремуми-атрактори прийнятої функції щільності.

Проведені експериментальні дослідження дозволяють рекомендувати запропоновані методи для використання на практиці для вирішення проблем автоматичної кластеризації великих даних. Особливістю запропонованих методів є обчислювальна простота і висока швидкість, пов'язана з тим, що весь масив обробляється тільки один раз, тобто виключається необхідність в багатоепоховому самонавчанні, що реалізується в традиційних алгоритмах нечіткої кластеризації.

Результати розділу 3 відображено у публікаціях [3, 6-9, 13, 14, 32].

Четвертий розділ присвячено розробці еволюційних методів оптимізації в задачах кластеризації масивів даних різної природи. Однією з основних складових частин обчислювального інтелекту є еволюційні алгоритми, які, за суттю, є певними математичними моделями відтворення або розвитку біологічних організмів, навіяні природою та призначені, у найзагальнішому випадку, для пошуку глобального оптимуму мультиекстремальних функцій в умовах невизначеності. Історично першими еволюційними алгоритмами були, так звані, генетичні алгоритми, в основі яких лежать селекційно-генетичні механізми, які реалізують виживання найсильніших особин у процесі еволюції. Деякі з еволюційних алгоритмів включають у себе так звані «ройові» процедури (Particle Swarm Optimization - PSO). Ці алгоритми надихаються поведінкою рою птахів чи

інших соціальних організмів, які спільно функціонують для досягнення спільної мети. В основі PSO лежить ідея руху потенційних рішень (часток) в просторі пошуку, де кожна частка користується інформацією про найкраще рішення, яке вона і її сусіди знаходили до цього часу. Ці алгоритми підтвердили свою ефективність у вирішенні ряду досить складних завдань і вже «встигли» зазнати ряд модифікацій, серед яких процедури на основі гармонійного пошуку, дробових похідних, адаптації параметрів пошуку, тощо. Водночас ці процедури не позбавлені деяких недоліків, які погіршують властивості процесу пошуку глобального екстремуму.

Для пошуку глобального екстремуму скалярної функції векторного аргументу $x = (x_1, x_2, ..., x_n)^T \in R^n$ запропоновано використовувати модель поведінки зграй котів (Cat Swarm - CS) при цьому передбачається, що кожен кіт cat_p зграї, що складається з Q осіб $(p = 1,2,...,Q)$, може перебувати в одному з двох станів: режим пошуку (Seeking Mode - SM) і режим гонитви (Tracing Mode - TM). При цьому режим пошуку пов'язаний з повільними рухами з незначною амплітудою біля вихідної позиції (сканування простору в поточній позиції), а режим гонитви визначається швидкими стрибками з великою амплітудою і дозволяє вивести кота cat_p з локального екстремуму, якщо він потрапив туди. Поєднання локального сканування та різких змін поточного стану дозволяє з більшою ймовірністю відшукати глобальний екстремум у порівнянні з традиційними методами багатоекстремальної оптимізації.

Процес відшукання екстремуму за допомогою котячої зграї реалізовано у вигляді такої послідовності кроків:

Крок CS 1. Створити зграю з Q котів у вигляді набору *n*-вимірних векторів $x_p^{(0)}$, випадковим чином розподілених на безлічі допустимих значень аргументів P_x^n, тобто $x_p^{(0)} \in P_x^n \subset P^n$; оцінити значення оптимізованої функції (фітнес-функції) $f(a_p(0))$ у всіх Q точках, при цьому передбачається, що метою оптимізації є відшукання глобального мінімуму $f(a)$.

Крок CS 2. Ввести параметр стану SPC (self-position consideration), який приймає два значення 1 або 0; випадково розділити зграю на дві групи: коти в пошуку (SPC=1) і коти в режимі гонитви (SPC=0).

Крок CS 3. Якщо SPC=1, запустити відповідну групу котів у пошук, коти що залишилися з SPC=0 запустити в режим гонитви.

Крок CS 4. Оцінити значення фітнес-функції та зберегти нові стани $x_p(1),$ відповідні найменшим значенням $f(x_p(1))$.

Крок CS 5. Провернутись до кроку CS 1 з оновленою зграєю $x_p(1), p = 1,2,...,Q.$

Режими пошуку та переслідування можуть бути реалізовані паралельно і також складатися з послідовності кроків. При цьому режим пошуку котячої зграї відповідає процесу локального пошуку завдання оптимізації. Режим

пошуку визначається трьома основними факторами: обсягом пам'яті пошуку (Seeking Memory Pool - SMP), який визначає кількість копій кожного кота, що створюються. cat_p, кроком зміни по кожній координаті простору (Seeking Range of the selected Dimension - SRD) та змінюваних координат (Counts of Dimension to Change - CDC). Власне, режим пошуку може бути реалізований у вигляді такої послідовності кроків:

Крок SM 1. Якщо SPC = 1, створити C (C=SMP) копій cat_p.

Крок SM 2. Відповідно до прийнятого CDC змінити стан cat_p.

Крок SM 3. Оцінити значення оптимізованої фітнес-функції для кожного зміненого стану cat_p.

Крок SM 4. Ввести ймовірність вибору кожного змінного стану

$$R_p = \frac{f(x_p(\tau)) - f_{\min}(x_p(\tau))}{f_{\max}(x_p(\tau)) - f_{\min}(x_p(\tau))}, \tau = 1,2,...,T \qquad (32)$$

та кота з максимальним значенням R_p виключити з подальшого розгляду. Кіт з $R_p = 0$ є «найкращою» копією cat_p, оскільки їй відповідає найменше значення оптимізованої функції $f_{\min}(x_p(\tau))$. Режим гонитви відповідає процесу глобального пошуку, що дозволяє «проскакувати» локальні екстремуми оптимізованої функції, і може бути реалізований у вигляді послідовності кроків:

Крок TM 1. Якщо SPC = 0, для групи котів у гонитві розрахувати для кожної швидкості руху по кожній координаті за допомогою рекурентного виразу

$$\upsilon_{pi}(\tau+1) = \upsilon_{pi}(\tau) + r(\tau)\eta_{TM}(x_{best,i}(\tau) - x_{pi}(\tau)), \qquad (33)$$

де $\upsilon_{pi}(\tau)$ - швидкість руху p-го кота по i-й координаті на τ-й ітерації гонитви; $0 < r(\tau) < 1$ - випадковий параметр гонитви; η_{TM} - постійний крок гонитви; $x_{best,i}(\tau)$ - найкраще вирішення задачі оптимізації, отримане на τ-й ітерації.

Крок TM 2. Ввести гранично можливі значення швидкостей $\upsilon_{\min}$ і $\upsilon_{\max}$, для кожного кота перевірити умову $\upsilon_{\min} < \upsilon_{pi}(\tau+1) < \upsilon_{\max}$ і якщо воно порушується, покласти $\upsilon_{pi}(\tau+1)$ рівним відповідному значенню $\upsilon_{\min}$ або $\upsilon_{\max}$.

Крок TM 3. Змінити становище кожного кота в гонитві відповідно до співвідношення

$$x_{pi}(\tau+1) = x_{pi}(\tau) + x_{pi}(\tau). \qquad (34)$$

Крок TM 4. Перевірити, чи належить $x_p(\tau+1)$ P_{ai}^n.

У режимі гонитви реалізується градієнтний пошук із великим кроком, що у загальному випадку гарантує відшукання глобального екстремуму. У зв'язку з цим доцільно модернізувати процедуру оптимізації на основі котячих

зграй шляхом її рандомізації на основі випадкового пошуку, що володіє цілою низкою переваг перед детермінованими процедурами пошуку екстремуму.

Оскільки режим пошуку SM є за суттю процесом локальної оптимізації, рух кожного з котів cat_p з SPC=1 доцільно організувати в антиградієнтному напрямку відповідно до стандартної рекурентної градієнтної процедури

$$x_p(\tau+1) = x_p(\tau) - \eta_{SM}\hat{\nabla} f(x_p(\tau)), \qquad (35)$$

де $\hat{\nabla} f(x_p(\tau))$ - оцінки градієнта оптимізованої функції в точці $x_p(\tau)$; η_{SM} - крок пошуку у просторі P_x^n.

Складові градієнта $\nabla f(x_p(\tau))$, що є частковими похідними $\frac{\partial f(x_p(\tau))}{\partial x_p}$, можуть бути оцінені шляхом вимірювання оптимізованої функції в пробних станах в околі точки $x_p(\tau)$. Найбільш простим з обчислювальної точки зору є пошук з центральною пробою, при цьому проводиться оцінка оптимізованої функції в $(n+1)$-й точці $(CDC = n)$: $x_p(\tau)$, $x_p(\tau)+\eta_{SRD}e_1$ $x_p(\tau)+\eta_{SRD}e_1$, $x_p(\tau)+\eta_{SRD}e_2,...,x_p(\tau)+\eta_{SRD}e_n$, де e_i - координатні орти; η_{SRD} - величина пробного кроку, яка визначається прийнятим значенням SRD.

Реалізувавши далі крок у просторі P_x^n відповідно до (35), приходимо до нового стану cat_p у режимі пошуку з координатами

$$\begin{cases} x_{p1}(\tau+1) = x_{p1}(\tau) - \dfrac{\eta_{SM}}{\eta_{SRD}}(f(x_p(\tau)) + \eta_{SRD}e_1) - f(x_p(\tau)), \\ x_{p2}(\tau+1) = x_{p2}(\tau) - \dfrac{\eta_{SM}}{\eta_{SRD}}(f(x_p(\tau)) + \eta_{SRD}e_2) - f(x_p(\tau)), \\ x_{pn}(\tau+1) = x_{pn}(\tau) - \dfrac{\eta_{SM}}{\eta_{SRD}}(f(x_p(\tau)) + \eta_{SRD}e_n) - f(x_p(\tau)). \end{cases} \qquad (36)$$

Відзначено, що у випадку $f(x_p(\tau+1)) < f(x_p(\tau))$, cat_p наближається до локального мінімуму, тобто. покращує свій стан і може залишатися в режимі пошуку. Якщо ж $f(x_p(\tau+1)) \geq f(x_p(\tau))$, cat_p знаходиться в околиці локального мінімуму, вивести з якого її можна, при переході в режим гонитви. Як недолік цієї процедури оптимізації можна відзначити фіксоване значення $CDC = n$, що вимагає послідовної зміни всіх координат у просторі P_x^n. Розширити можливості процесу пошуку можна, звернувшись до рандомізованих процедур, найпростішою з яких є суто випадкова оцінка напрямку спуску, сенс якого полягає в тому, що зі стану $x_p(\tau)$ робиться випадкова проба $x_p(\tau)+\eta_{SRD}\Xi$, де $\Xi = (\xi_1, \xi_2, ..., \xi_n)^T$ - одиничний випадковий вектор, рівномірно розподілений у просторі P_x^n. У разі якщо $x_p(\tau)+\eta_{SRD}\Xi < f(x_p(\tau))$, робиться робочий крок пошуку:

$$x_p(\tau+1) = x_p(\tau) - \eta_{SM}\Xi \tag{37}$$

(при цьому можна прийняти $\eta_{SRD} = \eta_{SM}$), в іншому випадку проба визнається невдалою та реалізується спроба з новим вектором Ξ. Узагальненням цієї процедури є оцінка напряму пошуку за найкращою з кількох випадкових спроб. За напрямок спуску вибирається той напрямок Ξ^*, яке забезпечило найменше значення функції $f(x_p)$, тобто cat_p переводиться в новий стан згідно з виразом

$$x_p(\tau+1) = x_p(\tau) + \eta_{SRD}\Xi^*. \tag{38}$$

Об'єднуючи процедури пошуку (35), (36), (38), вводиться на розгляд пошук на основі градієнта. В цьому випадку за оцінку градієнта приймається середньозважене L випадкових напрямків, кожен з яких береться з вагою, що відповідає варіації $f(x_p)$ вздовж цього напряму:

$$\hat{\nabla} f(x_p(\tau)) = -\frac{\sum_{l=1}^{L}\Xi_l\left(f\left(x_p(\tau)+\eta_{SRD}\Xi_l\right)-\nabla f(x_p(\tau))\right)}{\left\|\sum_{l=1}^{L}\Xi_l\left(f\left(x_p(\tau)+\eta_{SRD}\Xi_l\right)-\nabla f(x_p(\tau))\right)\right\|}. \tag{39}$$

Підставляючи далі (39) в (38), отримуємо процедуру градієнтного спуску в напрямку мінімуму функції, що оптимізується. Таким чином, всі кішки з SPC=1 зміщуються в напрямку локальних мінімумів функції, що оптимізується. Режим гонитви ТМ на відміну від локального режиму пошуку SM забезпечує загальну процедуру оптимізації на основі CS глобальні властивості, що дозволяють не застрягати їй у локальних екстремумах. Зрозуміло, що крім процедури (33), (34) існують інші алгоритми, що володіють необхідними властивостями.

Одним із таких найбільш ефективних чисельно простих методів є метод важкої кульки, що спирається на аналогію руху важкого тіла по викривленій поверхні з урахуванням сил тяжіння та тертя. При цьому через інерцію кулька-кішка «проскакує» локальні екстремуми, а через тертя рух має зупинитися в глобальному екстремумі. Даний метод для кішок у режимі гонитви (SPC=0) може бути записаний у вигляді

$$x_p(\tau+1) = x_p(\tau) - \alpha(x_p(\tau) - x_p(\tau-1)) - \eta_{TM}\hat{\nabla} f(x_p(\tau)), \tag{40}$$

де α - параметр, що визначає інерційні властивості процесу гонитви. При (40) повністю збігається з (35), відрізняючись лише кроком η_{SM}. При $\alpha = 1$ процес гонитви стає незагасаючим, тому цей параметр вибирається в інтервалі $0 < \alpha < 1$, при цьому чим ближче α до одиниці, тим сильніше виявляються інерційні властивості, проте процес слабо згасає в околиці екстремуму. У зв'язку з цим доцільно кожній кішці з SPC=0 призначити різні значення параметра α. Зауважимо також, що в процедуру (40) введено випадкову компоненту, що вводить додаткове «гойдання» в процес гонитви, яка покращує глобальні властивості методу. При цьому (40) модифікується до

вигляду $x_p(\tau+1)=a_p(\tau)-\alpha(x_p(\tau)-x_p(\tau-1))-\eta_{TM}\hat{\nabla}f(x_p(\tau))+\eta_{SRD}\Xi$, тобто cat_p одночасно знаходиться і в режимі гонитви, і в режимі пошуку-сканування простору P_x^n.

Запропоновано модифікацію методу оптимізації на основі косяків риб. При використанні методів еволюційної оптимізації, що за суттю є методами оптимізації нульового порядку, припускається, що при відшуканні екстремумів деякої функції $f^x(x)$ застосовується популяція агентів, кожен з яких діє або самостійно, або взаємодіючі з іншими, при цьому рух кожного q-го агента $(q=1,2,...,Q)$ на l-й ітерації пошуку може бути записаний за допомогою співвідношення $x_q^l=x_q^{l-1}+\eta_q^l Dir_q^l,\ q=1,2,...,Q,$ де $x_q^l=\left(x_{q1}^l,x_{q2}^l,...,x_{qn}^l\right)^T$, Dir_q^l - вектор, що задає напрямок руху q-го агента на l-й ітерації пошуку.

У великій родині таких методів слід відзначити метод на основі косяків риб, де кожен агент популяції імітує рух окремої риби. Основною перевагою цього методу є достатня ефективність відшукання глобального екстремуму досить складних функцій, до яких можна віднести і функцію щільності розподілу даних в задачах кластеризації. Автори методу вводять у розгляд ітерації, пов'язані з рухом косяка: годування та плавання.

Оператор годування відповідає за вагу кожної риби як елемента косяка - агента. Чим важче риба, тим ближче вона до екстремума - максимума. Вага кожної риби w_q налаштовується згідно із виразом

$$w_q^l=w_q^{l-1}+\frac{f^x\left(x_q^l\right)-f^x\left(x_q^{l-1}\right)}{\max\limits_p\left\{f^x\left(x_q^l\right)-f^x\left(x_q^{l-1}\right)\right\}}\quad \forall q=1,2,...,Q, \tag{41}$$

при цьому $0<w_q^l<w_{\max},\ w_l^0=0,5w_{\max}$.

Оператор плавання описує як індивідуальний рух кожної риби, так і колективний рух косяка в цілому. Тут розглядається три типи руху: індивідуальний, інстинктивно-колективний та колективно-рольовий. Індивідуальний рух описується співвідношенням

$$x_{qi}^l=\begin{cases}x_{qi}^l+\eta_q^l Rand\{0,1\},\ \text{if } f^x\left(x_q^l\right)>f^a\left(x_q^{l-1}\right),\\ x_q^{l-1}\ \text{інакше,}\end{cases} \tag{42}$$

де $Rand\{0,1\}$ - рівномірно розподілене у інтервалі $(0,1)$ випадкове число.

Фактично це процедура «зондування» функції $f^x\left(x\right)$ в околі точки x_q^{l-1} при цьому крім (4.10) тут можна бути заснований будь-який інший алгоритм випадкового пошуку.

На базі зондування функції щільності за допомогою індивідуального руху (4.10) реалізується інстинктивно - колективний рух у напряму зростання цієї функції

$$x_q^l = x_q^{l-1} + \frac{\left(\sum_{p=1}^{Q}\left(x_p^l - x_p^{l-1}\right)\right)\left(f^x\left(x_q^l\right) - f^x\left(x_q^{l-1}\right)\right)}{\sum_{p=1}^{Q}\left(f^x\left(x_p^l\right) - f^x\left(x_p^{l-1}\right)\right)}. \quad (43)$$

Вводячи у розгляд зважений центр ваги косяка риб

$$Bar^l = \frac{\sum_{p=1}^{Q} x_p^l w_p^l}{\sum_{p=1}^{Q} w_p^l}, \quad (44)$$

можна записати цей рух у вигляді

$$x_q^l = \begin{cases} x_q^l - \eta_q^l Rand\{0,1\}\dfrac{x_q^{l-1} - Bar^{l-1}}{\left\|x_q^{l-1} - Bar^{l-1}\right\|}, \ if \ \ \sum_{p=1}^{Q} w_p^l > \sum_{p=1}^{Q} w_p^{l-1}, \\ x_q^l + \eta_q^l Rand\{0,1\}\dfrac{x_q^{l-1} - Bar^{l-1}}{\left\|x_q^{l-1} - Bar^{l-1}\right\|}, \ if \ \ \sum_{p=1}^{Q} w_p^l < \sum_{p=1}^{Q} w_p^{l-1}. \end{cases} \quad (45)$$

Для підвищення ефективності FSS у розгляд водиться додатковий оператор розведення, що дозволяє створювати нових риб - агентів, що мають покращені характеристики у порівнянні з вже існуючими членами косяка. Для цього можна скористатися ідеями еволюційної оптимізації, серед яких з обчислювальної точки зору та ефективності - надійності відшукання екстремуму можна відзначити послідовний симплекс-метод та його модифікації.

Сформуємо косяк, що містить $Q = n + 1$ риб-агентів, при цьому ця кількість залишається незмінною у процесі пошуку, тобто популяція $x_1^0, x_2^0, ..., x_Q^0$ генерується випадковим чином. В цій популяції знайдемо «найгіршу» рибу x_{qworst}^0, що моє найменшу вагу $w_{q\min}^0$ та «найкращу» рибу x_{qbest}^0 з найбільшою вагою $w_{q\max}^0$. Основна операція руху симплекса полягає у відображенні x_{qworst}^0 через центр ваги n риб (без найгіршої), який може бути записаний у вигляді $\overline{x}^0 = \frac{1}{n}\sum_{q=1}^{Q}\left(x_q^0 - x_{qworst}^0\right)$.

В результаті цієї операції створюється нова риба $x_q^{1*} = \overline{x}^0 + \alpha\left(\overline{x}^0 - x_{qworst}^0\right)$, яка заміняє у косяку найгіршу особину x_{qworst}^0. Таким чином формується нова популяція $x_1^1, x_2^1, ..., x_Q^1$. Отже, рух косяка-симплекса описується за допомогою співвідношень

$$\begin{cases} \overline{x}^{l-1} = \dfrac{1}{n}\sum\limits_{q=1}^{Q}\left(x_q^{l-1} - x_{qwost}^{l-1}\right), \\ x_q^{l} = \overline{x}^{l-1} + \alpha\left(\overline{x}_q^{l-1} - x_{qwost}^{l-1}\right), \end{cases} \tag{46}$$

що у загальному випадку є за своєю суттю алгоритмом оптимізації Нелдера-Міда. Таким чином, з косяка у процесі пошуку екстремуму вилучаються найгірші риби з найнижчою вагою та створюються нові агенти з більшою вагою. Оскільки задача, що розглядається, є за своєю суттю проблемою багатоекстремальної оптимізації, необхідно відшукати множину екстремумів, кожен з яких є центроїдом деякого кластера. При знаходженні якогось з екстремумів з вихідної вибірки X виключаються спостереження, що розташовані безпосередньо в його околі. Після цього вилучення запропонована процедура комбінованої еволюційної оптимізації повторюється до відшукання всіх екстремумів-центроїдів.

Запропоновано модифікацію методу сірих вовків. За правилами цього методу, сірі вовки живуть разом і полюють групами. Процес пошуку та полювання описується формулою (47), якщо жертву знайдено, вони спочатку вистежують, переслідують і наближаються до неї; якщо здобич біжить, тоді сірі вовки переслідують, оточують і спостерігають за здобиччю, поки вона не перестане рухатися; далі нарешті починається атака.

Стандартний метод сірих вовків (GWO). Метод імітує поведінку пошуку і полювання на здобич сірих вовків в зграї. В математичній моделі найкращій результат вовка в зграї називається альфа (α), а другий найкращий є бета (β), і, отже, третій найкращій називається дельта (δ). Інші рішення кандидатів зграї - омегами (ω). Всі омеги будуть керуватися цими трьома сірими вовками під час пошуку (оптимізації) та полювання.

Коли жертва знайдена, починається ітерація (t=1). Згодом α- , β- та δ-вовки керуватимуть ω, щоб переслідувати здобич і, зрештою, оточити її. Три коефіцієнти A, B і C пропонуються для опису поведінки оточення:

$$\begin{aligned} C_\alpha &= \left|B_1 * GW_\alpha - X(t)\right|, \\ C_\beta &= \left|B_2 * GW_\beta - X(t)\right|, \\ C_\delta &= \left|B_3 * GW_\delta - X(t)\right|, \end{aligned} \tag{47}$$

де t вказує на поточну ітерацію;

GW вектор позиції сірого вовка,

GW_1, GW_2 і GW_3 - є векторами положення α- , β- та δ- вовків, що обчислюється наступним чином:

$$\begin{aligned} GW_1 &= GW_\alpha - A_1 * C_\alpha, \\ GW_2 &= GW_\beta - A_2 * C_\beta, \\ GW_3 &= GW_\delta - A_3 * C_\delta, \end{aligned} \tag{48}$$

$$GW(t) = \frac{GW_1 + GW_2 + GW_3}{3}. \tag{49}$$

Параметри A та B є комбінаціями керуючого параметра α та випадкових чисел r_1 та r_2:

$$A = 2\alpha r_1 - \alpha,$$
$$B = 2r_2. \tag{50}$$

Контрольний параметр α замінюється значенням параметра A і, нарешті, змушує омега-вовків наближатися або тікати від домінуючих вовків, таких як альфа, бета та дельта. Якщо $|A| > 1$, сірі вовки втікають від домінантів, а це означає, що омега-вовки втечуть від здобичі та досліджуватимуть більше простору, що в оптимізації називається глобальним пошуком. Та якщо $|A| < 1$, вони наближаються до домінант, а значить δ-вовки будуть слідувати за домінантами, які наближаються до здобичі, і це називається локальним пошуком в оптимізації. Контрольний параметр α визначається як лінійне зниження від максимального значення 2 до 0 під час ітерацій:

$$\alpha = 2\left(1 - \frac{t}{T}\right),$$

де t - номер ітерації, T - максимальна кількість ітерацій, що задана.

Багато алгоритмів ройового інтелекту імітують поведінку полювання та пошуку деяких тварин. Однак GWO моделює внутрішню ієрархію керівництва вовків, таким чином, в процесі пошуку позиція найкращого рішення може бути комплексно оцінена трьома рішеннями. Але для інших алгоритмів ройового інтелекту, найкраще рішення шукається лише на основі одного рішення – локального оптимуму. Отже, GWO може значно зменшити ймовірність передчасного потрапляння в локальний оптимум. Щоб досягти належного компромісу між розвідкою та полюванням, пропонується покращений GWO.

Розглядаючи рівняння (49) видно, що в процесі пошуку, однакову роль відіграють домінанти. Кожен із сірих вовків зграї наближається або тікає в пошуку здобичі. Однак, слід зауважити, що найближче до здобичі домінанти із середньою вагою альфа, ніж бета і дельта. Таким чином, на початку процедури пошуку в рівнянні (49) слід враховувати лише положення альфа, або його вага має бути набагато більшою, ніж ваги інших домінант. Рівняння (49) можна переписати у вигляді:

$$GW(t+1) = \frac{w_1 GW_1 + w_2 GW_2 + w_3 GW_3}{3}, \tag{51}$$

де $w_1 + w_2 + w_3 = 1$, при w_1 - вага α-вовка, w_2 - вага β-вовка, w_3 - вага δ-вовка, при цьому $w_1 \geq w_2 \geq w_3$.

Зміна позицій вовків описана наступними виразами:

$$C = \left|B * X_P(t) - X(t)\right|, \tag{52}$$

$$X\left(t+1\right) = X_P(t) - A * C, \tag{53}$$

де X_P - позиція здобичі,

X - позиція вовка.

На етапі параметрів налаштування, задаються початкові позиції вовків – домінант

$$Cl_1 = C_\alpha;$$
$$Cl_2 = C_\beta; \text{ при } t = 0,$$
$$Cl_3 = C_\delta;$$

беручи за початкові позиції центри кластерів, знайдених за допомогою методу можливісної нечіткої кластеризації.

Проведені експериментальні дослідження підтверджують, що запропоновані методи підвищили швидкість роботи методів нечіткої кластеризації потоків даних за умов апріорної та поточної невизначеності, за рахунок запропонованих процедур оптимізації на 10%.

Результати розділу 4 відображено у публікаціях [1, 4, 5, 9, 10, 12, 13, 15, 17-20, 22, 24, 27, 29, 35, 39, 40].

П'ятий розділ присвячено розробці гібридних еволюційних методів кластеризації масивів даних. Гібридні еволюційні методи кластеризації представляють собою поєднання технік еволюційного обчислення та методів кластеризації. Ці методи спроєктовані для покращення результатів кластеризації шляхом використання переваг обидвох підходів.

Запропоновано метод нечіткої кластеризації масивів даних на основі еволюційного методу оптимізації котячих зграй. Завдання нечіткої кластеризації може бути зведена до пошуку глобального екстремуму цільових функцій (1), (6). Для вирішення завдання використані еволюційні біоінспіровані «ройові» процедури оптимізації серед яких, в якості одного з найбільш швидкодіючих, можна відзначити, метод котячих зграй. Зауважимо, що саме котячі зграї з успіхом були використані для вирішення завдань чіткої кластеризації в рамках процедури c-середніх, що породжується цільовими функціями (1), (6) при $\beta \to 1, \mu_j(k) = \{0,1\}$. В рамках цього підходу передбачається, що кожен центроїд c_j представлений одним з котів зграї, а кінцеве рішення визначається котами, що забезпечують мінімум цільової функції $Goal(c_j)$ (1) або (6). В рамках класичного «котячого» методу передбачається, що кожен кіт cat_p зграї, що складається з Q особин $(p = 1,2,...,Q)$ може перебувати в одному з двох станів: режимі пошуку (Seeking Mode - SM) і режимі погоні (Tracing Mode - TM). При цьому режим пошуку пов'язаний з повільними рухами з незначною амплітудою біля вихідної позиції (сканування простору в околиці поточної позиції), а режим гонитви визначається швидкими стрибками з великою амплітудою і дозволяє вивести кота cat_p з локального екстремуму, якщо він потрапив туди. Поєднання локального сканування та різких змін поточного стану дозволяє з більшою ймовірністю відшукати глобальний екстремум у порівнянні з традиційними методами багатоекстремальної оптимізації.

Запропоновано онлайн метод для правдоподібної нечіткої кластеризації на основі еволюційної оптимізації котячої зграї. Цільова функція правдоподібної кластеризації має вигляд (15) за наявності обмежень (16).

Для знаходження глобального екстремуму (15) доцільно використовувати так звані еволюційні алгоритми оптимізації роїв частинок, серед яких є методи котячої зграї, які виявилися ефективними при вирішенні широкого кола завдань інтелектуального аналізу даних. В рамках цього підходу оптимізація котячої зграї передбачає, що кожен кіт зграї може перебувати в одному з двох станів: режимі пошуку та режимі відстеження.

В загальному випадку обидва режими для кожної із зграї котів можна описати процедурою рекурентної оптимізації:

$$c_p(\tau+1) = c_p(\tau) - \alpha(c_p(\tau) - c_p(\tau-1)) - \eta\hat{\nabla}Goal_M(c_p(\tau)) + \eta_\xi \Xi(\tau), \quad (54)$$

де $c_p(\tau+1)$ - стан p-того кота в зграї на τ-ій ітерації пошуку, α - параметр, що визначає інерційні властивості режиму трасування, η- крок пошуку, $\hat{\nabla}Goal(c_p(\tau))$ - градієнтна оцінка цільової функції (15) в околі точки $c_p(\tau)$ $\Xi(\tau)$ - випадкова складова, яка вносить додаткові стохастичні рухи в процес трасування, η_ξ - параметр, що задає амплітуду цих рухів.

Таким чином, кожен кіт може одночасно перебувати в режимах пошуку і відстеження і при достатній кількості котів в зграї забезпечується пошук глобального екстремуму.

Запропоновано модифікацію методу глобальної оптимізації божевільної котячої зграї для задачі нечіткої кластеризації. Для пошуку глобального екстремуму цільової функції нечіткої кластеризації пропонується використовувати модифікований метод оптимізації божевільної котячої зграї, синтезованого на основі оптимізаційного підходу котячої зграї і методів глобального випадкового пошуку.

Ідея оптимізації на основі еволюційного алгоритму котячої зграї полягає в тому, що формується група-зграя «котів», кожен з яких рухається у напрямку або локального, або глобального екстремуму прийнятої цільової функції $Goal\left(c_q\right)$. При цьому ця зграя складається з Q осіб $cat_p, p = 1,2,...,Q$, кожна з яких може перебувати в одному з двох можливих станів: режим пошуку (SM) локальних екстремумів і режим гонитви (TM), що ставить собі за мету відшукання глобального екстремуму. У загальному випадку обидва режими SM і TM реалізуються паралельно, при цьому SM фактично базується на основі покоординатного спуску, тобто в кожен конкретний момент може змінюватися тільки одна координата n-вимірного простору пошуку, що природно знижує швидкодію процедури. У режимі гонитви швидкості руху по кожній координаті також оцінюються незалежно одне від одної, що знову-таки знижує швидкодію.

Для подолання цих недоліків в був запропонований рандомізований метод оптимізації на основі котячих зграй, що забезпечує підвищену швидкодію в порівнянні з відомою процедурою - прототипом.

Рух кота в режимі погоні описується методом, що є «гібридом» популярного методу оптимізації «важкої кульки» і випадкового пошуку

$$c_p(\tau+1)=c_p(\tau)-\alpha\left(c_p(\tau)-c_p(\tau-1)\right)-\eta_{TM}\hat{\nabla}Goal\left(c_p(\tau)\right)+\Xi(\tau), \quad (55)$$

де $0<\alpha<1$ - параметр інерції режиму погоні, $\Xi(\tau)$ - випадкове збурення, що вводить додаткове сканування простору пошуку.

Для поліпшення процесу пошуку глобального екстремуму в режимі погоні в алгоритм руху кожної кішки додатково було введено «фактор божевільності», який описується набором випадкових параметрів і дозволяє здійснювати раптові стрибки, що змінюють траєкторію руху, шляхом варіювання характеристик сигналу збурення $\Xi(\tau)$. Для керування сигналом $\Xi(\tau)$ доцільно скористатися ідеєю «блукаючого» глобального пошуку Л. Растригіна, який довів свою ефективність при вирішенні багатоекстремальних задач. При цьому характеристики випадкового збурення змінюються відповідно до виразу

$$\Xi(\tau)=\gamma\Xi(\tau-1)-\delta\left(E\left(w_p(\tau)\right)-E\left(w_p(\tau-1)\right)\right)+\sigma^2H(k), \quad (56)$$

де γ - параметр корекції характеристик збурення,

$0\le\delta\le1$ - параметр швидкості самонавчання типу параметра інерції α у (60), σ^2 - дисперсія білого шуму $H(\tau)$.

Запропоновано метод кластеризації масивів даних на основі комбінованої оптимізації функцій щільності розподілу та еволюційного методу котячих зграй. Основними поняттями, на яких базується метод є функція впливу, функція щільності та атрактори щільності, що за суттю є локальними екстремумами функції щільності. У загальному випадку функція впливу для будь-якого векторного спостереження $x(\bullet)$ з вихідного масиву X є ядерною дзвонуватою функцією $f^{x(\bullet)}(x)$, при цьому найбільш популярною є традиційна гаусівська функція

$$f_G^{x(\bullet)}(x)=\exp\left(-\frac{d^2(x,x(\bullet))}{2\sigma^2}\right)=\exp\left(-\frac{\left\|x-x(\bullet)\right\|^2}{2\sigma^2}\right), \quad (57)$$

де $d^2(x,x(\bullet))$ - евклідова відстань, σ^2 - параметр ширини функції впливу, завдяки простоті обчислення її похідних.

У матричному випадку замість евклідової можна використати метрику Фробеніуса, при цьому функція впливу набуває вигляду

$$f_G^{x(\bullet)}(x)=\exp\left(-\frac{d^2(x,x(\bullet))}{2\sigma^2}\right)=\exp\left(-\frac{Tr\left(x-x(\bullet)\right)\left(x-x(\bullet)\right)^T}{2\sigma^2}\right), \quad (58)$$

де $Tr(\bullet)$ - символ сліду матриці. На основі функцій впливу формується функція щільності розподілу даних у масиві X у вигляді

$$f^x(x)=\sum_{k=1}^{N} f\left(x,x(k)\right), \tag{59}$$

що за суттю є оцінкою Надарая - Ватсона. Функція $f^x(x)$ може приймати значення в інтервалі $1\le f^x(x)\le N,$ при цьому крайні значення з цього інтервалу приймаються, коли вибірка містить лише одне спостереження або усі N спостережень співпадають, тобто існує лише один кластер - вироджена ситуація. Для знаходження $m>1$ кластерів необхідно ввести у розгляд деякий поріг $\xi>1$, що дозволяє формувати дійсно значущі кластери, відстежуючи аномальні спостереження та класи, що містять занадто мало даних. Власне процес формування кластерів пов'язаний з відшуканням усіх екстремумів функції щільності (59) за допомогою градієнтної процедури

$$x^l=x^{l-1}+\eta^l\frac{\nabla f^x\left(x^l,x^{l-1}\right)}{\left\|\nabla f^x\left(x^l,x^{l-1}\right)\right\|},\ \ x_0=x\left(k\right),l=0,1,2,...;\forall k=1,2,...,N, \tag{60}$$

тобто кількість запусків методу (60) визначається обсягом навчальної вибірки N. Зрозуміло, що при великих N процес кластеризації - пошуку локальних екстремумів потребує дуже багато часу. Тому запропоновані модифікації пов'язані з пришвидшенням процесу пошуку локальних екстремумів (59) шляхом модифікації градієнтної процедури (60). У випадку коли спостереження $x(k)$ у вибірці $X\in\left(n_1\times n_2\right)$ - є матрицями, введемо у розгляд матричний варіант процедури (60):

$$x^l=x^{l-1}+\eta^l\Gamma^x\left(x,x^{l-1}\right)\left(Tr\Gamma^x\left(x,x^{l-1}\right)\Gamma^{xT}\left(x,x^{l-1}\right)\right)^{-\frac{1}{2}},$$

де $\Gamma^x\left(x,x^{l-1}\right)=\left\{\frac{\partial f^x\left(x,x^{l-1}\right)}{\partial x_{i_1i_2}}\right\}\in X^{n_1\times n_2}.$

Процес градієнтної оптимізації закінчується відшуканням m локальних екстремумів функції (59), при цьому чим менше значення ξ, тим більше кластерів може бути сформовано.

Пришвидшити процес відшукання локальних екстремумів можна, використовуючи замість градієнтного пошуку методи еволюційної оптимізації, серед яких в якості достатньо ефективного, чисельно простого і швидкого можна відзначити, так званий, пошук на основі котячих зграй, що модифікований для вирішення задачі кластеризації.

Запропоновано метод адаптивної нечіткої кластеризації викривлених даних на основі стратегії найближчого прототипу-центроїда з використанням еволюційних процедур. Стратегія найближчого прототипу-центроїда розглянута в якості гібрида стратегії оптимального розширення та часткових відстаней і складається з послідовності кроків:

1. Завдання початкових умов для роботи методу: $\beta > 0$, m, необхідної точності $\varepsilon > 0$, прототипів (центроїдів) кластерів c_q, кількості епох $\tau = 1,2,...,Q$.

2. Розрахунок рівнів належності:

$$\mu_q^{(\tau+1)}(k) = \left(\sum_{l=1}^{m} \left(\left\| x^{(\tau)}(k) - c_l^{(\tau)} \right\|^2 \right)^{\frac{1}{1-\beta}} \right)^{-1} \left(\left\| x^{(\tau)}(k) - c_q^{(\tau)} \right\|^2 \right)^{\frac{1}{1-\beta}}.$$

3. Розрахунок центроїдів кластерів:

$$c_q^{(\tau+1)} = \left(\sum_{k=1}^{N} \left(\mu_q^{(\tau+1)}(k) \right)^{\beta} \right)^{-1} \sum_{k=1}^{N} \left(\mu_q^{(\tau+1)}(k) \right)^{\beta} x^{(\tau)}(k).$$

4. Перевірка умов зупину: якщо $\left\| c_q^{(\tau+1)} - c_q^{(\tau)} \right\| < \varepsilon \ \forall \ q$ або $\tau = Q$, останов; інакше йти до кроку 5.

5. Оцінка викривлених спостережень шляхом знаходження прототипу $c_q^{(\tau+1)}$ найближчого до $a(k)$ в сенсі часткової відстані

$$d_P^2\left(x(k), c_q\right) = \frac{n}{\delta_{k\sum}} \sum_{i=1}^{n} \left(x_i(k) - c_{qi} \right)^2 \delta_{ki},$$

тобто знаходження $c_q^{(\tau+1)} = \arg\min_q \left\{ d_P^2\left(x(k), c_1^{(\tau+1)}\right), ..., d_P^2\left(x(k), c_m^{(\tau+1)}\right) \right\}$ і заміна відсутніх спостережень $x_i(k)$ координатами $x_i^{(\tau+1)}(k) = c_{qi}^{(\tau+1)}$. Далі йти до кроку 2. Далі запишемо стратегію найближчого прототипу у рекурентній формі

$$\begin{cases} \mu_q^{(\tau+1)}(k) = \left(\sum_{l=1}^{m} \left(\left\| x_k^{(\tau)} - c_l(k) \right\|^2 \right)^{\frac{1}{1-\beta}} \right)^{-1} \left(\left\| x_k^{(\tau)} - c_q(k) \right\|^2 \right)^{\frac{1}{1-\beta}}, \\ \text{де } x_i^{(\tau)}(k) = c_{qi}(k), \\ c_q(k) = \arg\min_q \left\{ d_P^2\left(x(k), c_1(k)\right), ..., d_P^2\left(x(k), c_m(k)\right) \right\}, \\ c_q(k+1) = c_q(k) + \eta(k+1) \left(\mu_q^{(Q)}(k) \right)^{\beta} \left(x^{(Q)}(k) - c_q(k) \right) \ \ \forall q = 1,2,...,m. \end{cases}$$

Можливісна стратегія найближчого прототипу-центроїда у загублених спостереженнях записується у вигляді таких кроків:

1. Завдання початкових умов для роботи методу: $\beta > 0$, m, необхідної точності $\varepsilon > 0$, прототипів (центроїдів) кластерів c_q, кількість епох $\tau = 1,2,...,Q$.

2. Розрахунок рівнів належності:

$$\mu_q^{(\tau+1)}(k) = \frac{1}{1 + \left(\dfrac{\left\| x^{(\tau)}(k) - c_q^{(\tau)} \right\|^2}{\omega_q^{(\tau)}} \right)^{\frac{1}{\beta-1}}}.$$

3. Розрахунок центроїдів кластерів:

$$c_q^{(\tau+1)}(k)=\left(\frac{\sum\limits_{k=1}^{N}\left(\mu_q^{(\tau+1)}(k)\right)^{\beta}x^{(\tau)}(k)}{\sum\limits_{k=1}^{N}\left(\mu_q^{(\tau+1)}(k)\right)^{\beta}}\right).$$

4. Перевірка умов зупину: якщо $\left\|c_q^{(\tau+1)}-c_q^{(\tau)}\right\|<\varepsilon\ \forall\ q$ або $\tau=Q$, зупинитися; інакше перейти до кроку 5.

5. Оцінка викривлених спостережень шляхом знаходження прототипу $c_q^{(\tau+1)}$ найближчого до $c(k)$ в сенсі часткової відстані

$$d_P^2\left(x(k),c_q\right)=\frac{n}{\delta_{k\sum}}\sum_{i=1}^{n}\left(x_i(k)-c_{qi}\right)^2\delta_{ki},$$

тобто знаходження $c_q^{(\tau+1)}=\arg\min\limits_{q}\left\{d_P^{\ 2}\left(x(k),c_1^{(\tau+1)}\right),...,d_P^{\ 2}\left(x(k),c_m^{(\tau+1)}\right)\right\}$ і заміна відсутніх спостережень $x_i(k)$ координатами $x_i^{(\tau+1)}(k)=c_{qi}^{(\tau+1)}$.

6. Розрахунок скалярного параметра відстані

$$\omega_q^{(\tau+1)}=\frac{\sum\limits_{k=1}^{N}\left(\mu_q^{(\tau+1)}(k)\right)^{\beta}\left\|x^{(\tau+1)}(k)-c_q^{(\tau+1)}\right\|^2}{\sum\limits_{k=1}^{N}\left(\mu_q^{(\tau+1)}(k)\right)^{\beta}}.$$

7. Далі йти до кроку 2.

Аналогічно ймовірнісній адаптивній кластеризації на основі стратегії найближчого центроїда пропонується організувати процес можливісної кластеризації у вигляді

$$\begin{cases}\mu_q^{(\tau+1)}(k)=\dfrac{1}{1+\left(\dfrac{\left\|x_k^{(\tau)}-c_q(k)\right\|^2}{\omega_q^{(\tau)}}\right)^{\frac{1}{\beta-1}}},\\ \partial e\ a_i^{(\tau)}(k)=c_{qi}(k),\ \ c_q(k)=\arg\min\limits_{q}\left\{d_P^2\left(x(k),c_1(k)\right),...,d_P^2\left(x(k),c_m(k)\right)\right\},\\ c_q(k+1)=c_q(k)+\eta(k+1)\left(\mu_q^{(Q)}(k)\right)^{\beta}\left(x^{(Q)}(k)-c_q(k)\right)\ \ \forall q=1,2,...,m,\\ \omega_q^{(\tau+1)}=\dfrac{\sum\limits_{p=1}^{k}\left(\mu_q^{(\tau+1)}(p)\right)^{\beta}\left\|x^{(\tau)}(k)-c_q(k)\right\|^2}{\sum\limits_{p=1}^{k}\left(\mu_q^{(\tau+1)}(p)\right)^{\beta}}.\end{cases}$$

Для знаходження локальних екстремумів у вихідних даних, що надходять на обробку методами адаптивної нечіткої кластеризації даних на основі стратегії найближчого прототипу-центроїду доцільно використовувати еволюційні методи рою частинок. Одним з найшвидших методів рою частинок

є, так званий, метод котячої зграї, який підтвердив свою ефективність у вирішенні широкого кола задач від елементарних завдань Data Mining до більш складних задач Dynamic Data Mining, Data Stream Mining, Big Data Mining, Web Mining, Text Mining тощо.

Запропоновано кластеризацію масивів даних на основі модифікованого методу сірого вовка. В основі поширеного алгоритму імовірної нечіткої кластеризації лежить процедура мінімізації цільової функції (15), при обмеженнях (16). Задача умовної оптимізації була переформульована в задачу безумовної оптимізації цільової функції виду

$$Goal\left(c_j\right)=\sum_{k=1}^{N}\left(\sum_{j=1}^{m}\left\|x\left(k\right)-c_j\right\|^{2(1-\varphi)}\right)^{1-\varphi}, \tag{61}$$

при $\varphi=2$

$$Goal\left(c_j\right)=\sum_{k=1}^{N}\left(\sum_{j=1}^{m}\left\|x\left(k\right)-c_j\right\|^{-2}\right)^{-1}. \tag{62}$$

Таким чином, задача нечіткої кластеризації зведена до пошуку глобального екстремуму цільових функцій (61), (62). Для вирішення задачі використані еволюційні процедури оптимізації, а саме модифікований метод сірого вовка.

Для підвищення надійності знаходження саме глобального екстремуму цільової функції можна використати оптимізацію, модифіковану введенням випадкового блукання, яка довела свою ефективність при розв'язанні мультиекстремальні проблеми. Вводячи, додаткове пошукове збурення, можна записати рух вовка у вигляді:

$$\Xi\left(\tau\right)=\gamma\Xi\left(\tau\text{-}1\right)\text{-}\delta\left(X\left(t+1\right)\text{-}X\left(t\right)\right)+\sigma^2H\left(k\right),$$

де γ - параметр корекції характеристик збурення, $0\le\delta\le1$ - параметр швидкості самонавчання типу параметра інерції α, σ^2 - дисперсія білого шуму $H\left(\tau\right)$. Таким чином підвищується імовірність знаходження глобального екстремуму прийнятої цільової функції, що в кінцевому рахунку підвищує ефективність процесу нечіткої кластеризації.

Проведені експериментальні дослідження підтверджують, що запропоновані методи підвищили швидкість роботи методів нечіткої кластеризації потоків даних різної природи за умов викривленої інформації.

Результати розділу 5 відображено у публікаціях [1, 4, 5, 8-10, 12, 13, 15, 17-20, 22, 24, 27, 29,3 35, 39, 40].

Шостий розділ присвячено впровадженню низки розроблених методів для вирішення практичних задач.

Впроваджено метод нечіткої правдоподібної кластеризація даних на основі аналізу щільності розподілу даних та їх піків для підвищення врожайності озимої пшениці на ТОВ НАУКОВО-ВИРОБНИЧІЙ ФІРМІ «ХЕЛП-АГРО». Запропонований підхід дає можливість приймати ефективні управлінські рішення щодо підвищення врожайності сільськогосподарських

культур в умовах невизначеності зовнішнього середовища (акт впровадження від 27.02.2023р.).

Впроваджено метод адаптивної нечіткої кластеризації даних для оцінки стану будинків та визначення готовності до експлуатації в зимових умовах у підприємстві ТОВ «КОМУНСЕРВІС 2018». Впроваджено метод адаптивної кластеризації викривлених даних на основі правдоподібного підходу для аналізу та оцінки стану житлових будинків з урахуванням пошкоджень та зношеності, який дозволяє прискорити аналіз та прийняття обґрунтованих рішень щодо першочерговості відновлення будинків, в залежності від категорії пошкоджень та зношеності (акт впровадження від 12.04.2023р.).

Впроваджено метод адаптивної нечіткої кластеризації даних для вирішення практичних задач класифікації технологічних процесів та монтажних робіт загального призначення для отримання класифікації будівельних та монтажних робіт на будівництві з метою підвищення їх ефективності на ТОВ «Будівельно-монтажне підприємство - 168» (акт впровадження від 21.12.2023р.).

Впроваджено метод відновлення та фільтрації потоків даних за умов перетинних кластерів для задач покращення якості води на КП «Санітарно-екологічний центр». Надані рекомендації щодо поліпшення характеристик та усунення негативного впливу від використання води в залежності від класу згідно з нормативами для питної води (акт впровадження від 29.06.2023р.).

Впроваджено нечіткий метод кластеризації викривлених даних для класифікації пацієнтів з ознаками онкологічних захворювань на КНП «Обласний центр онкології». Підвищено точність та об’єктивність процесу медичного діагностування онкологічних захворювань на ранніх стадіях (акт впровадження від 14.11.2023р.).

Результати розділу 6 розвивають результати, що були відображені у публікаціях [1-6, 9, 10, 19, 24, 26-28, 31, 32, 36, 40].

ВИСНОВКИ

За результатами проведеного дослідження вирішено важливу теоретичну проблему зі створення нових ефективних нечітких методів обчислювального інтелекту, а саме, нечіткої кластеризації даних за умов апріорної невизначеності на основі еволюційного самонавчання та надання їм адаптивних властивостей, що забезпечує можливість опрацьовування потоків нестаціонарних даних, викривлених завадами та пропусками, що послідовно надходять на обробку в онлайн режимі. У результаті виконання цієї роботи одержані такі результати:

1. Проведено аналіз методів обробки потоків даних в умовах апріорної невизначеності та викривленості, обґрунтовано необхідність вирішення задач кластеризації та аналізу даних за умов змінних характеристик

потоку, що включає зміну кількості класів, їхньої структури та непередбачуваних дрейфів.

2. Обґрунтовано необхідність розроблення нових методів та удосконалення існуючих методів нечіткої кластеризації даних за умов апріорної невизначеності на основі еволюційного самонавчання та надання їм адаптивних властивостей, що забезпечує можливість опрацьовування потоків нестаціонарних даних, викривлених завадами та пропусками, що послідовно надходять на обробку в онлайн режимі.

3. Уперше запропоновано адаптивні ймовірнісні, можливісні та правдоподібні методи нечіткої кластеризації потоків викривлених даних, які призначені для вирішення задач Data Stream Mining та Big Data Mining, що дозволяють опрацьовувати апріорі невідому кількість даних послідовно, спостереження за спостереженням в міру їх надходження у онлайн режимі.

4. Уперше запропоновано онлайн метод нечіткої кластеризації, який базується на ідеях аналізу щільностей розподілу даних, їх піків та правдоподібного нечіткого підходу, що дозволяє підвищити якість кластеризації даних з довільними апріорі невідомими щільностями розподілів.

5. Уперше запропоновано метод швидкої нечіткої кластеризації даних з використанням аналізу піків щільності розподілу даних на основі правдоподібного підходу, що дозволяє вирішувати широкий клас задач Data Stream Mining та Big Data Mining у ситуаціях, коли дані забруднені завадами.

6. Уперше запропоновано швидкі методи нечіткої кластеризації даних довільної природи з апріорі невідомими розподілами, що дозволяє підвищити якість результатів розбиття масивів даних на класи за умов невизначеності.

7. Уперше запропоновано метод послідовної можливісної нечіткої кластеризації даних, який призначено для роботи в онлайн режимі, що дозволяє швидко знаходити екстремуми (центроїди) кластерів, незалежно від обсягів даних, що надходять на обробку у векторній або матричній формах.

8. Уперше запропоновано метод нечіткої кластеризації масивів даних на основі покращеного еволюційного алгоритму сірого вовка, що дозволяє відшукувати глобальні екстремуми цільових функцій та скоротити час їх пошуку.

9. Уперше запропоновано метод нечіткої кластеризації масивів даних на основі комбінованої оптимізації функцій щільності розподілу та еволюційного методу котячих зграй, що дозволяє уникнути застрягання в локальних екстремумах.

10. Уперше запропоновано підходи до вирішення багатоекстремальної задачі правдоподібної нечіткої кластеризації на основі модифікованих оптимізаційних процедур божевільної котячої зграї та зграї сірих вовків, що дозволяє скоротити час вирішення задачі.

11. Уперше запропоновано підхід до вирішення задачі адаптивної нечіткої кластеризації викривлених пропусками та викидами даних на основі

стратегії найближчого прототипу-центроїду з використанням еволюційних процедур, що дозволяє підвищити завадостійкість процесу оптимізації.

12. Удосконалено еволюційний метод на основі косяків риб, що підвищив ефективність вирішення задач нечіткої кластеризації даних, які надходять як в пакетному, так і в онлайн режимах, що дозволяє скоротити час пошуку глобальних екстремумів.

13. Удосконалено метод кластеризації Густафсона-Кесселя, який базується на підході правдоподібності до нечіткої кластеризації та формує перетинні класи гіпереліпсоїдальної форми з довільною орієнтацією осей у просторі ознак, що дозволяє опрацьовувати потоки даних в міру їх надходження на обробку в онлайн режимі.

14. Удосконалено метод оптимізації на основі еволюційних котячих зграй шляхом введення в процеси пошуку та гонитви елементів глобального випадкового пошуку, що дозволяє підвищити точність визначення напрямку руху в режимі пошуку та покращити глобальні властивості методу у режимі гонитви.

15. Проведено експериментальні дослідження розроблених та модифікованих методів. Розроблені методи нечіткої кластеризації забезпечують точність визначення кількості класів (кластерів) в умовах дефіциту апріорної інформації, працездатні як в пакетному так і в онлайн режимах та здатні працювати на вибірках, що змінюють розмірність та форму кластерів; дозволяють обробляти великі обсяги даних, що можуть подаватись на обробку послідовно у формі потоків даних, ефективно працювати за умов суттєвої невизначеності, стохастичності, нелінійності, апріорної невизначеності, нестаціонарності та є найбільш пристосованими для вирішення задач Data Mining та Data Stream Mining, завдяки своїм універсальним апроксимуючим властивостям, здатності до самонавчання.

Отримані результати дають змогу:

- підвищити точність кластеризації потоків даних, що поступають на обробку в онлайн режимі за оцінками якості кластеризації даних на 8 %;
- підвищити швидкість роботи методів нечіткої кластеризації потоків даних за умов апріорної та поточної невизначеності, за рахунок запропонованих процедур оптимізації на 10%;
- підвищити точність прогнозування даних до 7-8% за рахунок аналізу великого об’єму інформації, що подається в онлайн режимі;
- зменшити ймовірність похибки розбиття потоків викривлених даних на класи, за умов невизначеності до 5%;
- прискорити аналіз та прийняття обґрунтованих рішень в залежності від поставленої задачі;
- підвищити точність і об’єктивність процесу медичного діагностування, відновлення викривлених та втрачених спостережень, що потрапляють на обробку в онлайн режимі;

- підвищити надійність та об’єктивність медичного діагностування пацієнтів з умовно невідомим діагнозом.

16. Розроблені методи можуть бути використані для розв’язання широкого класу прикладних задач і, перш за все, задач Data Mining, Data Stream Mining, Big Data Mining та Medical Data Mining, кластеризації, прогнозування, діагностування, прийняття рішень, керування, класифікації за умов дефіциту апріорної інформації.

17. Достовірність наукових та практичних результатів підтверджується відповідними актами про впровадження дисертаційних досліджень, а також за рахунок порівняння одержаних результатів з результатами застосування існуючих класичних методів та підходів кластеризації потоків даних.

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА

1. Shafronenko, A., Bodyanskiy, Y., & Rudenko, D. (2020). Neuro-fuzzy clustering of distorted data using cat swarm optimization. United Kingdom, London. LAP LAMBERT Academic Publishing, 60.

2. Шафроненко, А., Бодянський, Є., & Плісс, І. (2022). Нечіткі методи інтелектуального аналізу даних. United Kingdom, London. GlobeEdit, 104.

3. Bodyanskiy, Y. V., Shafronenko, A. Y., & Klymova, I. N. (2021). Online fuzzy clustering of incomplete data using credibilistic approach and similarity measure of special type. *Radio Electronics, Computer Science, Control*, (1), 97-104. DOI: 10.15588/1607-3274-2021-1-10 (Web of Science, категорія «А»).

4. Бодянський, Є. В., Шафроненко, А. Ю., & Климова, І. М. (2021). Онлайн метод можливісної кластеризації даних на основі еволюційної оптимізації котячих зграй. *Радіоелектроніка, інформатика, управління*, (2), 65-70. DOI: 10.15588/1607-3274-2021-2-7 (Web of Science, категорія «А»).

5. Бодянський, Є. В., Шафроненко, А. Ю.,& Плісс, І. П. (2021). Правдоподібна нечітка кластеризація даних на основі еволюційного методу божевільних котів. *Системні дослідження та інформаційні технології*, (3), 110-119. DOI: 10.15588/1607-3274-2021-2-7 (Scopus, категорія «А»).

6. Бодянський, Є. В., Плісс, І. П., & Шафроненко, А. Ю. (2022). Швидка нечітка правдоподібна кластеризація на основі аналізу піків щільності розподілу даних. *Радіоелектроніка, інформатика, управління*, (1), 76-81. DOI: 10.15588/1607-3274-2022-1-9 (Web of Science, категорія «А»).

7. Бодянський, Є. В., Шафроненко, А. Ю., & Калиниченко, О. В. (2022). Нечітка довірча кластеризація даних на основі аналізу щільності розподілу даних та їх піків. *Радіоелектроніка, інформатика, управління*, (3), 58-68. DOI: 10.15588/1607-3274-2022-3-6 (Web of Science, категорія «А»).

8. Бодянський, Є. В., Плісс, І. П., & Шафроненко, А. Ю. (2022). Кластеризація масивів даних на основі комбінованої оптимізації функцій

щільності розподілу та еволюційного методу котячих зграй. *Радіоелектроніка, інформатика, управління*, (4), 61-70. DOI: 10.15588/1607-3274-2022-4-5 (Web of Science, категорія «А»).

9. Bodyanskiy, Y., Shafronenko, A., & Pliss, I. (2022). Clusterization of vector and matrix data arrays using the combined evolutionary method of fish schools. *System Research and Information Technologies, №4*. DOI: 10.20535/SRIT.2308-8893.2022.4.07 (Scopus, категорія «А»).

10. Шафроненко, А. Ю., Бодянський, Є. В., & Головін, О. О. (2023). Кластеризація масивів даних на основі модифікованого алгоритму сірого вовка. *Радіоелектроніка, інформатика, управління* (1), 73-79. DOI: 10.15588/1607-3274-2023-1-7 (Web of Science, категорія «А»).

11. Shafronenko, A. Y., Kasatkina, N. V., Bodyanskiy, Y. V., & Shafronenko, Y. O. (2023). Credibilistic robust online fuzzy clustering in data stream mining tasks. *Radio Electronics, Computer Science, Control*, (3), 97-103. DOI: 10.15588/1607-3274-2021-1-10 (Web of Science, категорія «А»).

12. Бодянський, Є. В., & Шафроненко, А. Ю. (2018). Рандомізована модифікація методу оптимізації на основі котячих зграй. *Системи обробки інформації*, (1), 142-147. DOI: 10.30748/soi.2018.152.20 (категорія «Б»).

13. Бодянський, Є. В., Шафроненко, А. Ю., & Патлань, К. В. (2018). Нечітка кластеризація масивів даних на основі еволюційного методу оптимізації котячих зграй. *Біоніка інтелекту*, *2*(91), 3-8. DOI: 10.30837/bi.2018.2(91).01 (категорія «Б»).

14. Бодянський, Є. В., Шафроненко, А. Ю., & Климова, І. М. (2019). Онлайн достовірна нечітка кластерізація даних з використанням функції належності спеціального типу. *Біоніка інтелекту*, *2*(93), 3-6. DOI: 10.30837/bi.2019.2(93).01 (категорія «Б»).

15. Shafronenko, A., & Bodyanskiy, Y. (2019). Online algorithm for possibilitic fuzzy clustering based on evolutionary cat swarm optimization. *Science and Education a New Dimension. Natural and Technical Sciences*, *193*, 86-88. DOI: 10.31174/SEND-NT2019-193VII23-22 (Будапешт, Угорщина, країна ЄС).

16. Бодянський, Є. В., Шафроненко, А. Ю., & Климова, І. М. (2020). Рекурентна достовірна нечітка кластеризація великих даних з використанням функції належності спеціального типу. *Біоніка інтелекту*, *2*(95), 77-81. DOI: 10.30837/bi.2020.2(95).10 (категорія «Б»).

17. Бодянський, Є. В., Шафроненко, А. Ю., & Климова, І. М. (2021). Метод адаптивної достовірної нечіткої кластеризації даних на основі еволюційного алгоритму. *Збірник наукових праць Харківського національного університету Повітряних Сил*, (2 (68)), 80-83. DOI: 10.30748/zhups.2021.68.10. (категорія «Б»).

18. Бодянський, Є. В., Плісс, І. П., & Шафроненко, А. Ю. (2022). Адаптивна нечітка кластеризація викривлених даних на основі стратегії найближчого прототипа-центроїда з використанням еволюційних процедур. *Artificial intelligence*, (1), 239-244, DOI: 10.15407/jai2022.01.239 (категорія «Б»).

19. Шафроненко А. Ю., Бодянський Є. В. (2022). Адаптивна кластеризація багатоекстремальних масивів даних з використанням модифікованого алгоритму риб'ячої зграї. *АСУ і прилади автоматики*. №178. 33-37. DOI: 10.30837/0135-1710.2022.178.033 (категорія «Б»).

20. Шафроненко, А. Ю., Бодянський, Є. В., & Руденко, Д. О. (2023). Модифікований рекурентний метод достовірної нечіткої кластеризації з використанням оптимізаційної процедури на основі косяків риб. *Системи обробки інформації*, (1 (172)), 92-96. DOI: 10.30748/soi.2023.172.11. (категорія «Б»).

21. Шафроненко, А. Ю., & Бодянський, Є. В. (2023). Нечітка достовірна кластеризація великих масивів даних з гіпереліпсоїдальними класами з довільною орієнтацією осей. *Наука і техніка Повітряних Сил Збройних Сил України*, (1 (50)), 93-99. DOI: 10.30748/nitps.2023.50.11. (категорія «Б»).

22. Шафроненко, А. Ю., & Бодянський, Є. В. (2023). Адаптивний підхід до нечіткої кластеризації на основі еволюційної оптимізації алгоритму сірих вовків. *Збірник наукових праць Харківського національного університету Повітряних Сил*, (1 (75)), 77-81. DOI: 10.30748/zhups.2023.75.11 (категорія «Б»).

23. Shafronenko, A., Dolotov, A., Bodyanskiy, Y., & Setlak, G. (2018, August). Fuzzy clustering of distorted observations based on optimal expansion using partial distances. In *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)* (pp. 327-330). IEEE. DOI: 10.1109/DSMP.2018.8478489 (Scopus, DBLP).

24. Shafronenko, A., Bodyanskiy, Y., Pliss, I., & Patlan, K. (2019, June). Fuzzy clusterization of distorted by missing observations data sets using evolutionary optimization. In *2019 9th International Conference on Advanced Computer Information Technologies (ACIT)* (pp. 217-220). IEEE. DOI: 10.1109/ACITT.2019.8779888 (Web of Science, Scopus, DBLP).

25. Bodyanskiy, Y. V., Shafronenko, A., & Rudenko, D. (2019). Online neuro fuzzy clustering of data with omissions and outliers based on completion strategy. *CEUR-WS,* (pp. 18-27) (Scopus, DBLP).

26. Hu, Z., Bodyanskiy, Y. V., Tyshchenko, O. K., & Shafronenko, A. (2019, July). Fuzzy clustering of incomplete data by means of similarity measures. In *2019 IEEE 2nd Ukraine Conference on Electrical and Computer Engineering (UKRCON)* (pp. 957-960). IEEE. DOI: 10.1109/UKRCON.2019.8879844 (Scopus).

27. Shafronenko, A. Y., Bodyanskiy, Y. V., & Pliss, I. P. (2019, September). The fast modification of evolutionary bioinspired cat swarm optimization method. In *2019 IEEE 8th International Conference on Advanced Optoelectronics and Lasers (CAOL)*. (pp. 548-552). IEEE. DOI: 10.1109/CAOL46282.2019.9019583 (Scopus, DBLP).

28. Shafronenko, A., Bodyanskiy, Y. V., Klymova, I., & Holovin, O. (2020, May). Online credibilistic fuzzy clustering of data using membership functions of special type. *CEUR-WS* (pp. 744-753). (Scopus, DBLP).

29. Shafronenko, A., & Bodyanskiy, Y. V. (2020). Adaptive fuzzy clustering approach based on evolutionary cat swarm optimization. *CEUR-WS* (pp. 832-842) (Scopus, DBLP).

30. Bodyanskiy, Y., Shafronenko, A., & Mashtalir, S. (2020). Online robust fuzzy clustering of data with omissions using similarity measure of special type. In Lecture Notes in Computational Intelligence and Decision Making: Proceedings of the XV International Scientific Conference “Intellectual Systems of Decision Making and Problems of Computational Intelligence” (ISDMCI'2019), Ukraine, May 21–25, 2019 15 (pp. 637-646). Springer International Publishing. DOI: 10.1007/978-3-030-26474-1_44 (Scopus, DBLP).

31. Bodyanskiy, Y. V., Shafronenko, A., & Klymova, I. (2021, April). Adaptive Recovery of Distorted Data Based on Credibilistic Fuzzy Clustering Approach. *CEUR-WS* (pp. 6-15) (Scopus, DBLP).

32. Shafronenko, A., Bodyanskiy, Y., Pliss, I., & Klymova, I. (2021, September). Online Credibilistic Fuzzy Clustering Method Based on Cauchy Density Distribution Function. In *2021 11th International Conference on Advanced Computer Information Technologies (ACIT)* (pp. 704-707). IEEE. DOI: 10.1109/ACIT52158.2021.9548572 (Web of Science, Scopus, DBLP).

33. Bodyanskiy, Y., Shafronenko, A., Klymova, I., & Polyvoda, V. (2022). Robust recurrent credibilistic modification of the Gustafson-Kessel algorithm. In Lecture Notes in Computational Intelligence and Decision Making: 2021 International Scientific Conference" Intellectual Systems of Decision-making and Problems of Computational Intelligence”, Proceedings (pp. 613-623). Springer International Publishing. DOI: 10.1007/978-3-030-82014-5_42 (Scopus, DBLP).

34. Shafronenko, A., Bodyanskiy, Y. V., & Pliss, I. (2023). Credibilistic fuzzy clustering method based on evolutionary approach of crazy wolfs in online mode. *CEUR-WS* (pp. 141-150) (Scopus, DBLP).

35. Бодянський Є., Шафроненко А., Плісс І., Патлань К. (2019). Нечітка кластеризація масивів даних за допомогою еволюційних ройових алгоритмів. In *Міжнародний науковий симпозіум «Інтелектуальні рішення». Обчислювальний інтелект (результати, проблеми, перспективи): праці міжнар.наук. - практ. конф., 15-20 квітня 2019р.*, 74-75.

36. Bodyanskiy Ye., Shafronenko A., Mashtalir S. (2019) Corrupted data online robust fuzzy clustering by special type similarity measure. In *Інтелектуальні системи прийняття рішень і проблеми обчислювального інтелекту: матеріали міжнар. наук. конф., с. Залізний Порт, 21-25 травня 2019 р.– Херсон: Видавництво ФОП Вишемирський В. С.,* 17-18.

37. Shafronenko, A. Y., & Rudenko, D. A. (2020). Online recurrent method of credibilistic fuzzy clustering. In: *5th International scientific and practical conference “Topical of the development of modern science” (January 15-17, 2020), Sofia, Bulgaria*, 37-40.

38. Bodyanskiy, Y. V., & Shafronenko, A. Y. (2020). Online credibilistic fuzzy clustering of data with gaps. *Problems and perspectives of modern science and practice*, *43*.

39. Шафроненко А.Ю., Свистунов І.О., Танянський О.С. (2021). Адаптивна нечітка кластеризація даних на основі еволюційних процедур. *Topical issues of modern science, society and education. Proceedings of the 5th International scientific and practical conference. SPC – Sci-conf.com.ua. Kharkiv, Ukraine. 2021,* 644-647.

40. Шафроненко, А. Ю., & Москаленко, В. В. (2021, December). Правдоподібна нечітка кластеризація даних на основі еволюційних процедур. In *The 5th International scientific and practical conference "Science, innovations and education: problems and prospects" (December 8-10, 2021) CPN Publishing Group, Tokyo, Japan. 2021. 1068 p.* (p. 383).

АНОТАЦІЯ

Шафроненко А.Ю. Адаптивні методи нечіткої кластеризації потоків даних з використанням еволюційного самонавчання. – Кваліфікаційна наукова робота на правах рукопису.

Дисертаційна робота на здобуття наукового ступеня доктора технічних наук за спеціальністю 05.13.23 – системи та засоби штучного інтелекту. – Харківський національний університет радіоелектроніки Міністерства освіти і науки України, Харків, 2024.

У дисертаційній роботі вирішено важливу теоретичну проблему створення нових ефективних нечітких методів обчислювального інтелекту, а саме, нечіткої кластеризації даних за умов апріорної невизначеності на основі еволюційного самонавчання та надання їм адаптивних властивостей, що забезпечує можливість опрацьовування потоків нестаціонарних даних, викривлених завадами та пропусками, що послідовно надходять на обробку в онлайн режимі. Проведено аналіз існуючих методів обробки потоків даних в умовах апріорної невизначеності та викривленості, обґрунтовано необхідність вирішення задач кластеризації та аналізу даних за умов змінних характеристик потоку, що включає зміну кількості класів, їхньої структури та непередбачуваних дрейфів. Розроблено адаптивні методи нечіткої кластеризації, які здатні працювати як в пакетному, так і в онлайн режимах, а також на вибірках, що змінюють розмірність та форму кластерів; дозволяють обробляти великі обсяги даних, що можуть надходити на обробку послідовно у формі потоків даних, ефективно працювати за умов поточної та апріорної невизначеності, стохастичності, нелінійності, нестаціонарності та є найбільш пристосованими для вирішення задач Data Mining та Data Stream Mining, завдяки своїм універсальним апроксимуючим властивостям, здатності до самонавчання.

Отримані результати дають змогу підвищити точність кластеризації потоків даних, що надходять на обробку в онлайн режимі за оцінками якості кластеризації даних на 8%; підвищити швидкість роботи методів нечіткої кластеризації потоків даних за умов апріорної та поточної невизначеності, за рахунок запропонованих процедур оптимізації на 10%; підвищити точність

прогнозування даних до 7-8% за рахунок аналізу великого обсягу інформації в онлайн режимі; зменшити ймовірність похибки розбиття потоків викривлених даних на класи за умов невизначеності до 5%; прискорити аналіз та прийняття обґрунтованих рішень в залежності від поставленої задачі; підвищити точність та об'єктивність процесу медичного діагностування, відновлення викривлених та втрачених спостережень, що надходять на обробку в онлайн режимі; підвищити надійність та об'єктивність медичного діагностування пацієнтів з умовно невідомим діагнозом.

Достовірність наукових та практичних результатів підтверджується відповідними актами про впровадження дисертаційних досліджень, а також за рахунок порівняння одержаних результатів з результатами застосування існуючих класичних методів та підходів кластеризації потоків даних.

Ключові слова: інтелектуальний аналіз даних, нечітка (фаззі) кластеризація, еволюційні методи та алгоритми, адаптація, потоки даних, машинне навчання, самонавчання, гібридні системи, методи оптимізації, онлайн.

ABSTRACT

Shafronenko A. Yu. Adaptive methods of fuzzy clustering of data streams using evolutionary self-learning. – Qualification of scientific work in the form of a manuscript.

Thesis for the degree of Doctor degree of Technical Sciences in the specialty 05.13.23 – systems and tools of artificial intelligence. – Kharkiv National University of Radio Electronics of the Ministry of Education and Science of Ukraine, Kharkiv, 2024.

The dissertation solves an important theoretical problem of creating new effective fuzzy methods of computational intelligence, namely, fuzzy data clustering under conditions of a priori uncertainty based on evolutionary self-learning and providing them with adaptive properties, which provides the possibility of processing non-stationary data flows distorted by noise and gaps, which are sequentially received for processing online. An analysis of methods for processing data flows under conditions of a priori uncertainty and distortion is carried out. The need to solve clustering problems and data analysis under conditions of changing flow characteristics, which includes changing the number of classes, their structure, and unpredictable drifts, is substantiated. Adaptive fuzzy clustering methods have been developed that are efficient in both batch and online modes, are able to work on samples that change the dimension and shape of clusters, allow processing large amounts of data that can be submitted for processing sequentially in the form of data streams, work effectively under conditions of significant and a priori uncertainty, stochasticity, nonlinearity, non-stationarity and are most adapted for solving Data Mining and Data Stream Mining problems, due to their universal approximating properties, ability to self-learning.

The results obtained make it possible to increase the accuracy of clustering data streams submitted for processing in online mode by assessing the quality of data clustering by 8%; increasing the speed of fuzzy clustering methods for data streams under conditions of a priori and current uncertainty, due to the proposed optimization procedures by 10%; increase the accuracy of data forecasting to 7-8% by analyzing a large amount of information submitted online; reduce the probability of error in dividing distorted data flows into classes, under conditions of uncertainty up to 5%; accelerate analysis and making informed decisions depending on the task; increase the accuracy and objectivity of the medical diagnostic process, and the recovery of distorted and lost observations that are processed online; increase the reliability and objectivity of medical diagnostics of patients with a conditionally unknown diagnosis.

The reliability of scientific and practical results is confirmed by relevant materials on the implementation of dissertation research, as well as by comparing the obtained practical results with the results of applying existing classical methods and approaches to data flow clustering.

Keywords: data mining, fuzzy clustering, evolutionary methods and algorithms, adaptation, data streams, machine learning, self-learning, hybrid systems, optimization methods, online.