Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Кваліфікаційна наукова
праця на правах рукопису

ШАФРОНЕНКО АЛІНА ЮРІЇВНА

УДК 004.85:[004.62.048:004.275]

ДИСЕРТАЦІЯ

АДАПТИВНІ МЕТОДИ НЕЧІТКОЇ КЛАСТЕРИЗАЦІЇ ПОТОКІВ ДАНИХ З ВИКОРИСТАННЯМ ЕВОЛЮЦІЙНОГО САМОНАВЧАННЯ

05.13.23 – системи та засоби штучного інтелекту
технічні науки

Подається на здобуття наукового ступеня доктора технічних наук

Дисертація містить результати власних досліджень.
Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело А.Ю. Шафроненко

Цей примірник дисертації ідентичний за змістом
з іншими примірниками, що подані до спеціалізованої
вченої ради Д 64.052.11

Учений секретар спеціалізованої вченої ради Д 64.052.11 І.П. Плісс

Харків - 2024

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.

Метою дисертаційної роботи є 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.

Для досягнення мети дисертаційної роботи необхідно вирішити такі завдання:

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 для нечіткої кластеризації масивів даних.

5. Розробити гібридні еволюційні методи нечіткої кластеризації масивів даних.

6. Імплементація, тестування та експериментальна перевірка розроблених методів.

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.

В дисертаційній роботі отримані такі 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.

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.

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 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.

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).

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.

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

Список публікацій здобувачки

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).

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 thesis solves an important theoretical problem of developing new effective fuzzy methods of computational intelligence, namely, fuzzy data clustering, based on evolutionary self-learning and providing them with adaptive properties, which makes it possible to process non-stationary data streams distorted by outliers and omissions, which are sequentially processed online.

The purpose of the thesis is to conduct a set of research aimed at developing new approaches and methods of evolutionary self-learning for adaptive fuzzy clustering of distorted data streams online under conditions of a priori and current uncertainty using the latest achievements in this field: Computer Science, Computational Intelligence, Data Science, Data Streams, Big Data, Evolving Systems.

To achieve the goal of the thesis, it is necessary to solve the following tasks:

1. Analyze approaches and methods for processing data streams.

2. Develop adaptive methods for fuzzy clustering of data streams under conditions of intersecting classes and a priori uncertainty.

3. Develop methods for adaptive fuzzy clustering of data with different distribution densities.

4. Develop evolutionary optimization methods for fuzzy clustering of data sets.

5. Develop hybrid evolutionary methods for fuzzy clustering of data sets.

6. Implementation, testing, and experimental verification of the developed methods.

The object of research: online clustering of data streams using evolutionary self-learning.

The subject of research: adaptive fuzzy methods for processing distorted data streams online under conditions of a priori and current uncertainty using evolutionary self-learning.

The results of the thesis are based on the use of optimization theory, probability theory, random processes, theory of fuzzy systems, linear algebra, mathematical analysis, Data Science, machine self-learning, and hybrid computational intelligence systems.

The following scientific results were obtained in the thesis:

1. For the first time, adaptive probabilistic, possibilistic, and credibilistic methods for fuzzy clustering of distorted data streams are proposed, which are intended for solving Data Stream Mining and Big Data Mining problems, which allow processing a priori unknown amount of data sequentially, observation by observation as they arrive online.

2. For the first time, an online fuzzy clustering method was proposed, based on the ideas of analyzing data distribution densities, their peaks, and a credibilistic fuzzy approach, which allows for improvement in the quality of data clustering with arbitrary a priori unknown distribution densities.

3. For the first time, a method of fast fuzzy data clustering using analysis of data distribution density peaks based on a credibilistic approach was proposed, which allows the solving of a wide class of Data Stream Mining and Big Data Mining problems in situations where the data is contaminated with noise.

4. For the first time, fast methods of fuzzy clustering of data of arbitrary nature with a priori unknown distributions were proposed, which allowed to improve the quality of the results of dividing data arrays into classes under conditions of uncertainty.

5. For the first time, a method of sequential possibilistic fuzzy clustering of data was proposed, which is designed to work in online mode, which allows quickly finding the extrema (centroids) of clusters, regardless of the amount of data received for processing in vector or matrix forms.

6. For the first time, a method of fuzzy clustering of data arrays based on the improved evolutionary algorithm of the gray wolf, which made it possible to find global extrema of objective functions and reduce the time for their search was proposed.

7. For the first time, a method of fuzzy clustering of data arrays based on the combined optimization of density distribution functions and the evolutionary method of cat swarms, which made it possible to avoid getting stuck in local extrema was proposed.

8. For the first time, effective approaches to solving the multi-extreme problem of credibilistic fuzzy clustering based on modified optimization procedures of crazy cats and gray wolves, which made it possible to reduce the time for solving the problem were proposed.

9. For the first time, an effective approach to solving the problem of adaptive fuzzy clustering of data distorted by omissions and outliers based on the strategy of the nearest prototype centroid using evolutionary procedures, which allowed to increase the noise immunity of the optimization process, was proposed.

10. The evolutionary method based on fish schools has been improved, which has confirmed its effectiveness in solving the problems of fuzzy clustering of data received in both batch and online modes, which allows for reducing the time for searching for global extrema.

11. The Gustafson-Kessel clustering method has been improved, which is based on the likelihood approach to fuzzy clustering and forms intersection classes of a hyper ellipsoidal shape with arbitrary orientation of the axes in the feature space, which allows processing data flows as they arrive for processing in online mode.

12. The optimization method based on evolutionary cat swarms has been improved and a randomized modification of the basic procedure has been introduced

by introducing elements of global random search into the search and pursuit processes, which allows to increase in the accuracy of determining the direction of movement in the search mode and improve the global properties of the method in the pursuit mode.

The thesis was completed at the Department of Artificial Intelligence. The main scientific results of the research were obtained within the framework of state-funded fundamental research projects of NURE: "Dynamic intellectual analysis of sequences of fuzzy information under conditions of significant uncertainty based on hybrid computational intelligence systems" (DR No. 0116U002539), "Deep hybrid computational intelligence systems for data flow analysis and their fast learning" (DR No. 0119U001403) and "Adaptive bagging of hybrid computational intelligence systems based on optimal online learning in terms of speed" (DR No. 0124U000363), as well as applied research "Development of methods and algorithms for combined learning of deep neuro-neo-fuzzy systems under conditions of a short training sample" (DR No. 0122U001701), which were carried out on the basis of orders of the Ministry of Education and Science of Ukraine based on the results of a competitive selection of scientific projects. The applicant participated in the implementation of the specified research and is a co-author of the research reports.

The results of the thesis can be used to solve a wide class of applied problems and, above all, the problems of Data Mining, Data Stream Mining, Big Data Mining and Medical Data Mining, clustering, forecasting, diagnostics, decision-making, control, classification under conditions of a priori information deficiency.

The results of the dissertation work can be used to solve a wide class of applied problems and, above all, the problems of Data Mining, Data Stream Mining, Big Data Mining and Medical Data Mining, clustering, forecasting, diagnostics, decision-making, management, classification under conditions of a priori information deficiency.

The obtained results allow:

– to increase the accuracy of clustering of data flows received for processing in online mode according to estimates of the quality of data clustering by 8%;

– to increase the speed of fuzzy clustering methods of data flows under conditions of a priori and current uncertainty, due to the proposed optimization procedures by 10%;

– to increase the accuracy of data forecasting up to 7-8% due to the analysis of a large amount of information in online mode;

– to reduce the probability of error in dividing distorted data flows into classes under conditions of uncertainty up to 5%;

– to accelerate the analysis and adoption of justified decisions depending on the task at hand;

– increase the accuracy and objectivity of the medical diagnosis process, recovery of distorted and lost observations received for processing online;

– increase the reliability and objectivity of medical diagnosis of patients with a conditionally unknown diagnosis.

The results of the dissertation work were tested and implemented: in the KP "Sanitary and Ecological Center" of the Kharkiv City Council (implementation act dated June 29, 2023 and implementation act dated September 26, 2024); in LLC "Construction and Assembly Enterprise 168" (implementation act dated December 21, 2023); in LLC "Komunservice 2018" (implementation act dated April 12, 2023); in LLC Scientific and Production Firm "Help-Agro" (implementation act dated February 27, 2023); in the KNP "REGIONAL CENTER OF ONCOLOGY" (implementation act No. 1 dated November 14, 2023 and implementation act No. 2 dated April 22, 2024); in the educational process of the Kharkiv National University of Radio Electronics (implementation act dated 04/25/2024; implementation act dated 04/26/2024, implementation act dated 03/21/2024).

Based on the research results, 40 scientific papers have been published, including 2 monographs published abroad; 20 articles (19 articles in periodicals on technical sciences, 9 of which were published in Ukrainian professional publications of category "A", indexed in the scientometric international databases Scopus and/or

Web of Science, 1 article in a foreign English-language periodical on technical sciences of the European Union, Budapest, Hungary); 18 reports in the materials of international conferences, 12 of which are included in the scientometric international databases Scopus, Web of Science, DBLP.

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

List of the publications of the applicant

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).

ЗМІСТ

ПЕРЕЛІК ПРИЙНЯТИХ СКОРОЧЕНЬ

ALA – адаптивні лінійні асоціатори;

ADALINE – адаптивний лінійний елемент;

ШНМ – штучні нейронні мережі;

SM – міра подібності (similarity measure);

FCM – нечіткі С-середні (fuzzy C-means);

WTM – переможець отримує більше (winner takes more);

PSM – часткові міра подібності (partition similarity measure);

PD – часткова відстань (partition distance);

PCM – метод С-середніх;

GK – алгоритм Густафсона-Кесселя;

APrFC – адаптивний алгоритм імовірнісної нечіткої кластеризації;

APosFC – адаптивний алгоритм можливої нечіткої кластеризації;

ACrFC – адаптивна достовірна нечітка кластеризація;

RCM_GK – рекурентна модифікація алгоритму Густафсона-Кесселя правдоподібної нечіткої кластеризації;

SI – індекс силуету;

CHI – індекс Цалінскі-Харабаса;

DBI – індекс Девіса-Болдіна;

CFC – достовірна нечітка кластеризація;

PC – коефіцієнт розподілу спостережень;

CE – класифікаційна ентропія;

SC – індекс розподілу;

S – індекс поділу;

XB – індекс Ксі та Бені;

DI – індекс Данна;

DENCLUE – Density-based Clustering of Applications with Noise;

DBSCAN – Density-based spatial clustering of applications with noise

OPTICS – Ordering points to identify the clustering structure;

NMI – показник нормалізованої взаємної інформації;

FCDP – метод швидкої нечіткої правдоподібної кластеризації на основі аналізу піків щільності розподілу даних;

NCrCP – метод нечіткої довірчої кластеризації даних на основі аналізу щільності розподілу даних та їх піків;

CA - кластерна точність;

DENCLUE-SA – імітований відпал;

DENCLUE-GA – генетичний алгоритм;

PSO – «ройові» процедури (Particle Swarm Optimization);

SM* – режим пошуку (Seeking Mode);

TM - режим гонитви (Tracing Mode);

CS – зграя котів (Cat Swarm);

SMP – обсяг пам’яті пошуку (Seeking Memory Pool);

SRD – крок зміни по кожній координаті простору (Seeking Range of the selected Dimension);

CDC – координати, що змінюються (Counts of Dimension to Change);

SPC - параметр стану кота (self-position consideration);

GWO – алгоритм сірих вовків;

CSO – Cat Swarm optimization algorithm;

FSS – Fish School algorithm;

OMFS - модифікований метод оптимізації на основі Fish School;

PMGWO – ансамбль алгоритму можливісної нечіткої кластеризації та оптимізаційного алгоритму сірих вовків;

RI – коефіцієнт Ренда;

AFC_PCEP – метод адаптивної нечіткої кластеризації викривлених пропусками та викидами даних основі стратегії найближчого прототипу - центроїду з використанням еволюційних процедур;

FGWO – метод кластеризації масивів даних на основі модифікованого алгоритму сірого вовка;

FCSO – модифікований метод кластеризації на основі зграї котів;

CGWO – методу божевільних вовків;

SS – індекс суми квадратів помилок;

TI – індекс Trace;

ТОВ – товариство з обмеженою відповідальністю;

КП – комунальне підприємство;

НКП – некомерційне комунальне підприємство.

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ

X – масив багатовимірних спостережень;

$x(1), x(2), ..., x(N)$ – вектор-спостереження;

k – номер спостереження;

i - номер атрибуту;

N – кількість спостережень;

R – множина (масив);

X_G - масив викривлених спостережень;

X_F - повністю заповнений масив спостережень, без викривлених даних;

δ_{ki} - характеристична функція належності спостереження до кластера;

m – кількість кластерів;

q – номер кластера;

β – фаззіфікатор;

σ – розподіл Коші;

μ – рівень належності;

d – Евклідова відстань;

c – центроїд;

Cl – кластер;

d_p – часткова відстань;

$Goal$ – функція цілі;

$\eta(k)$ – параметр кроку навчання;

ω – параметр, що визначає відстань між спостереженнями та центроїдом, на який рівень належності μ набуває значення 0,5;

S – міра подібності;

σ – параметр ширини функції впливу;

L – функція Лагранжа;

$\lambda(k)$ – невизначені множники Лагранжа;

S_q – нечітка кореляційна матриця;

φ_q – функція сусідства робастного WTM-правила самонавчання;

$Credib$ – рівень правдоподібності;

$Credib_q(k)$ – рівень правдоподібності того, що спостереження $x(k)$ належить кластеру Cl_q;

μ^* – нормований рівень належності;

V_q – обернена нормована нечітка кореляційна матриця;

d_V – квадрат відстані Махаланобіса;

$f^{x(\bullet)}(x)$ – ядерна дзвонувата функція для будь-якого векторного спостереження $x(\bullet)$ з вихідного масиву X;

$Tr(\bullet)$ – символ сліду матриці;

ξ – поріг, що дозволяє формувати дійсно значущі кластери;

$\Gamma^x\left(x, x^{l-1}\right)$ – матриця перших похідних;

ρ_k – локальна щільність;

δ_k – відстань до точки з більш високою щільністю;

d_c – відстань зрізу, яка задається і варіюється користувачем для отримання необхідної точності рішення задачі;

D – матриця відстаней між спостереженнями;

δ_k^* – точки з максимальною щільністю;

$f_C^{\tilde{x}}(x)$ – функція Коші;

$f_E^{\tilde{x}}(x)$ – функція Єпанечнікова;

$f^x(x)$ – функція впливу;

η^l – параметр кроку пошуку, що визначає швидкість збіжності алгоритму;

x_q^P – піки-центроїди кластерів;

Q – кількість осіб в зграї котів;

cat_p – кіт;

f – фітнес-функція;

p – номер кота;

R_p – ймовірність вибору кожного змінного стану кота;

$\upsilon_{pi}(\tau)$ – швидкість руху p-го кота по i-й координаті на τ-й ітерації погоні;

τ – ітерація погоні;

η_{TM} – постійний крок погоні;

$x_{best,i}(\tau)$ – найкраще вирішення задачі оптимізації, отримане на τ-й ітерації;

$\hat{\nabla} f(x_p(\tau))$ – оцінки градієнта оптимізованої функції в точці $x_p(\tau)$;

η_{SM} – крок пошуку у просторі P_x^n;

e_i – координатні орти;

η_{SRD} – величина пробного кроку, яка визначається прийнятим значенням SRD;

Ξ – напрямок руху кота;

Dir_q^l – вектор, що задає напрямок руху q-го агента на l-й ітерації пошуку;

w_q – вага кожної риби;

$Rand\{0,1\}$ – рівномірно розподілене у інтервалі $(0,1)$ випадкове число;

x_q^l – інстинктивно - колективний рух косяка риб;

Bar^l – зважений центр ваги косяка риб;

x_{qbest}^0 – «найкраща» риба;

$\overline{x}^0$ – центр ваги n риб (без найгіршої);

x_q^{1*} – нова риба;

GW – вектор позиції сірого вовка;

r_1 та r_2 – випадкові числа;

t – номер ітерації;

T – максимальна кількість ітерацій, що задана;

w_1 – вага α-вовка;

w_2 – вага β-вовка;

w_3 – вага δ-вовка;

X_P – позиція здобичі,

X – позиція вовка;

Cl – початкова позиція вовка.

ВСТУП

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. Говорячи про обчислювальний інтелект, 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 згаданих методів. Зокрема, особливу увагу привертають еволюційні нейро-фаззі 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 проблеми, а також за фіксованих умов: кількості класів, їхньої форми та рівня збурень.

У сучасних умовах, особливо за 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 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, планами 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 і завдання 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.

Для досягнення мети дисертаційної роботи необхідно вирішити такі завдання:

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.

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 отриманих результатів.

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 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.

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 міжнародних конференціях та семінарах: 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.

РОЗДІЛ 1
АНАЛІЗ СТАНУ ПРОБЛЕМИ І ПОСТАНОВКА ЗАВДАННЯ ДОСЛІДЖЕННЯ

Проблема обробки потоків даних за умов апріорної невизначеності та їх викривленості є важливою і актуальною в умовах сучасних технологій обробки великих даних та інтелектуальних систем. Особливо гостра ця проблема постає в задачах кластеризації [1-4], оскільки ці задачі вирішуються на основі самонавчання, тобто рівень апріорної невизначеності є значно вищім ніж у задачах Data Mining, крім того відомі на сьогодні алгоритми кластеризації орієнтовані на роботу у пакетному режимі, тобто не здатні вирішувати задачі у онлайн режимі.

Апріорна невизначеність означає відсутність точних знань або попередніх припущень про структуру даних, характер їхнього розподілу або особливості поведінки окремих елементів в потоці створює значні труднощі для традиційних алгоритмів кластеризації та машинного навчання [5, 8], які часто передбачають певну початкову інформацію або статичні умови. У таких випадках необхідно використовувати адаптивні методи, які здатні змінювати свою поведінку в залежності від нових даних, що поступають.

Викривленість даних є ще однією серйозною проблемою при обробці потоків. Це явище вказує на наявність аномалій, відхилень від стандартних моделей або неправдивих вхідних даних, які можуть значно впливати на точність та ефективність обробки даних. Викривленість може бути зумовлена різними факторами, такими як зміна зовнішніх умов, помилки у вимірюваннях, випадкові чи систематичні збої в процесі збору даних або інші непередбачувані обставини. В умовах потоку даних важливо мати механізми для виявлення та корекції викривлених значень, а також для збереження стійкості до таких аномалій в режимі реального часу.

Однією з головних труднощів у розв’язанні проблеми обробки потоків даних є необхідність враховувати не тільки швидкість надходження нових даних, але й їх різноманітність та складність [8, 9, 12]. Стандартні методи обробки даних часто не підходять для таких задач, оскільки вони не здатні ефективно адаптуватися до змін у динамічних середовищах [13]. У зв’язку з цим розробка нових підходів, таких як онлайн-методи, машинне навчання з підкріпленням або методи, що використовують адаптивні моделі, стає надзвичайно важливою [16-18].

Незважаючи на численні труднощі, що виникають через апріорну невизначеність та викривленість, сучасні методи обробки потоків даних роблять значний прогрес [20, 24]. Вони включають в себе інтеграцію алгоритмів, що дозволяють постійно підвищувати точність класифікацій, оптимізувати параметри моделей у процесі навчання, а також інтегрувати елементи самонавчання. Такі підходи дозволяють знижувати вплив викривлених даних і використовувати накопичений досвід для покращення процесу обробки, що є важливим кроком на шляху до успішного вирішення проблеми в умовах реального часу [26, 29].

Таким чином, обробка потоків даних за умов апріорної невизначеності та викривленості є важливою проблемою, що потребує розвитку нових технологій і підходів для адаптації алгоритмів до змінюваних умов, забезпечення стійкості до аномалій та підвищення точності аналізу в реальному часі. Рішення цих завдань дозволить значно покращити ефективність обробки великих обсягів даних в різних галузях, від фінансових технологій до охорони здоров’я і науки.

1.1 Потоки даних та їх онлайн обробка

Потоки даних є важливим компонентом сучасних інформаційних систем, що дозволяють здійснювати обробку інформації в режимі реального

часу [5, 9, 12]. Вони утворюються як безперервні послідовності подій або повідомлень, що генеруються такими джерелами, як мобільні пристрої, датчики Інтернету речей, соціальні платформи чи фінансові системи. Основною особливістю потоків даних є їх безперервність, динамічність і висока швидкість надходження, що вимагає використання спеціалізованих підходів до обробки. Дані потоків обробляються на льоту, що відрізняє їх від традиційних статичних наборів, де обробка здійснюється пакетно.

Сучасні технології обробки потоків базуються на використанні спеціалізованих платформ. Наприклад, Apache Kafka забезпечує передачу та зберігання потоків у режимі реального часу, гарантує їх надійність та масштабованість. Apache Flink є потужним інструментом для виконання складного аналізу, зокрема трансформації даних, кластеризації або виявлення аномалій. Apache Spark Streaming, у свою чергу, дозволяє поєднувати потокову і пакетну обробку, надаючи можливість аналізу даних у вигляді мікро-батчів. Методи обробки потоків даних включають фільтрацію для відбору релевантної інформації, агрегацію для узагальнення даних та аналіз аномалій, що дозволяє виявляти нетипові події [11, 24, 31].

Потоки даних активно використовуються у багатьох галузях. У фінансовій сфері вони є основою для алгоритмічної торгівлі, виявлення шахрайства та оцінки ризиків. Інтернет речей генерує величезні обсяги потоків, що використовуються для оптимізації інфраструктури, прогнозування несправностей обладнання та управління інтелектуальними системами. В медицині потоки даних дозволяють у реальному часі аналізувати біометричні показники пацієнтів, забезпечуючи можливість оперативного реагування на зміни їхнього стану.

Однак, використання потоків даних супроводжується низкою викликів. Це, зокрема, необхідність забезпечення масштабованості обробки, мінімізації затримок та ефективного використання обчислювальних ресурсів. Зі зростанням обсягів потоків і розвитком технологій машинного навчання стає можливим впровадження нових підходів, що дозволяють інтегрувати аналіз

потокових даних із розумними алгоритмами. Таким чином, потоки даних є важливим інструментом для вирішення сучасних завдань обробки інформації, а технології, що забезпечують їх обробку, продовжують розвиватися.

Набір вхідних даних зручно представити у вигляді таблиці «об'єкт-властивість», де спостереження записані у вигляді вектора-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 в масиві даних X, або поточний дискретний час в задачах Data Stream Mining. Передбачається також, що дані, які надходять на обробку, нормовані в гіперкуб $[-1;1]$ так, що $-1 \le x_i(k) \le 1$, де $x_i(k), i = 1,2,...,n$ – i-та компонента вектора спостережень $x(k)$ [32].

Таблиця 1.1 – Приклад таблиці «об'єкт - властивість»

	Id	SepalLegth Cm	SepalWidth Cm	PentalLegth Cm	PentalWidth Cm	Species
0	1	5.1	3.5	1.4	0.2	Iris-Setosa
1	2	4.9	3.0	1.4	0.2	Iris-Setosa
2	3	4.7	3.2	1.3	0.2	Iris-Setosa
3	4	4.6	3.1	1.5	0.2	Iris-Setosa
4	5	5.0	3.6	1.4	0.2	Iris-Setosa

1.2 Викривлені дані

Сучасний стан питання обробки потоків даних, викривлених завадами та пропусками, характеризується постійним розвитком методів, які

дозволяють працювати з такими даними у реальному часі [34-36]. Оскільки збирання даних з різних джерел, таких як сенсори, соціальні мережі та інтернет речей, стає все більш актуальним, зростає потреба у ефективних підходах до очищення і корекції даних, щоб забезпечити точність і надійність результатів.

Важливою проблемою є збурення, яке може виникати через технічні неполадки, зовнішні впливи чи випадкові коливання, що спотворюють дані. Все це може істотно вплинути на точність обробки та прийняття рішень, якщо не застосовувати методи фільтрації чи усунення завад.

Пропуски даних також є серйозною проблемою, адже вони можуть виникати через недоступність джерела або технічні збої, що спричиняють втрату важливої інформації. Крім того, пропуски можуть ускладнювати аналіз і моделювання, оскільки втрата навіть невеликої частини даних може призвести до значних похибок.

Викривлення даних - ще одна проблема, з якою стикаються при обробці потоків інформації, що може бути результатом помилок при зборі даних, неправильно налаштованих сенсорів чи програмних помилок. Такі викривлення вимагають застосування складних алгоритмів [38-40] для корекції та нормалізації, щоб забезпечити достовірність та ефективність подальшої обробки.

Для вирішення цих проблем розробляються різні підходи, зокрема використання методів статистики, машинного навчання, фільтрації сигналів і методів заповнення пропусків, що дозволяє зменшити вплив збурення та пропусків і покращити якість оброблених даних [42].

У свою чергу, це сприяє підвищенню точності моделей прогнозування та прийняття рішень в умовах невизначеності та неповноти даних.

Набір вхідних даних можна представити у вигляді таблиці «об’єкт-властивість», де екземпляри записані вектором-спостережень $X = \{x(1), x(2), ..., x(k), ...x(N)\} \subset R^n$, при цьому допускається, що X_G рядків

можуть мати по одному викривленому значенню, а $X_F = X - X_G$ заповнені повністю.

Таблиця 1.2 – Приклад таблиці «об'єкт-властивість» з викривленими даними

	Id	SepalLegth Cm	SepalWidth Cm	PentalLegth Cm	PentalWidth Cm	Species
0	1	5.1	3.5	1.4	0.2	Iris-Setosa
1	2	4.9	3.0	1.4	0.2	Iris-Setosa
2	3	4.7	3.2	NaN	0.2	Iris-Setosa
3	4	4.6	3.1	1.5	0.2	Iris-Setosa
4	5	5.0	3.6	1.4	0.2	Iris-Setosa

Тоді, можна записати

$$\delta_{ki} = \begin{cases} 0 \mid x_{ki} \in X_G, \\ 1 \mid x_{ki} \in X_F, \end{cases}$$

$$\delta_{k\Sigma} = \sum_{i=1}^{n} \delta_{ki},$$

де $X_F = \{x_k \in X \mid x_k$ - вектор, що містить всі складові$\}$;

$X_P = \{x_{ki}, 1 \le i \le n, 1 \le k \le N \mid$ всі значення x_k, що містяться в $X\}$;

$X_G = \{x_{ki} = ?, 1 \le i \le n, 1 \le k \le N \mid$ всі значення x_k, що відсутні в X (викривлені дані, пропуски, аномальні викиди)$\}$.

1.3 Аналіз існуючих методів інтелектуального аналізу потоків даних

Інтелектуальний аналіз потоків даних є важливою галуззю сучасної науки, що досліджує методи обробки, аналізу та отримання знань із великих обсягів інформації, які надходять у режимі реального часу [90-93]. Ця сфера базується на досягненнях інформатики, статистики, математичного моделювання та машинного навчання. Вона має на меті створення ефективних алгоритмів і систем, здатних працювати зі швидко змінними та гетерогенними потоками даних.

У сучасній науці та технологіях аналіз потоків даних характеризується специфічними викликами. Однією з ключових проблем є висока швидкість надходження інформації, що вимагає негайної обробки, інакше дані втрачають свою актуальність. Крім того, обсяги потоків можуть перевищувати можливості збереження, тому аналіз нерідко здійснюється на льоту, без проміжного зберігання. Гетерогенність даних також створює труднощі, оскільки потоки можуть містити текст, відео, звукові сигнали або числові дані.

Методи інтелектуального аналізу даних у потоках спрямовані на вирішення цих завдань. Вони включають алгоритми класифікації, регресії, кластеризації, детектування аномалій, а також інструменти для виявлення шаблонів і трендів. Наприклад, застосування методів машинного навчання забезпечує адаптивність систем, яка необхідна для роботи в умовах змінних потоків даних. Динамічні моделі, такі як рекурентні нейронні мережі та методи оновлення параметрів у реальному часі, дозволяють ефективно працювати з часовими рядами [63, 64].

Ще одним важливим напрямом є розробка механізмів забезпечення масштабованості, оскільки в умовах зростання обсягів інформації обчислювальні ресурси мають використовуватися максимально ефективно. Для цього широко застосовуються розподілені обчислювальні платформи та хмарні середовища, які підтримують паралельну обробку потоків даних.

У контексті прикладних задач, аналіз потоків даних знаходить застосування у фінансовій сфері, де він використовується для виявлення підозрілих транзакцій, а також у моніторингу ринкових трендів. У сфері медицини та охорони здоров’я методи обробки потокових даних сприяють реальному моніторингу стану пацієнтів та ранньому виявленню критичних станів. У промисловості аналіз потоків застосовується для оптимізації процесів, прогнозування збоїв у обладнанні та зменшення простоїв.

Сучасні дослідження в цій галузі також зосереджені на розв’язанні питань забезпечення приватності та безпеки даних у потоках, адже великі обсяги інформації часто містять конфіденційні або персональні дані, що породжує необхідність розробки методів, які враховують не лише ефективність обробки, але й вимоги до етичності та захисту інформації.

Серед основних методів інтелектуального аналізу потоків даних виділяють кластеризацію, класифікацію, асоціативний аналіз, методи прогнозування та алгоритми виявлення аномалій. Незважаючи на їх ефективність, кожен із цих методів має свої недоліки, що ускладнюють застосування у певних умовах [65-71].

Одним із суттєвих викликів є обмеження обчислювальних ресурсів, адже потоки даних часто характеризуються високою швидкістю та великим обсягом інформації. Методи кластеризації та класифікації, наприклад, потребують значних обчислювальних потужностей, що може створювати затримки при роботі з великими даними. Іншим аспектом є складність підтримки актуальності моделей, оскільки в потоках даних часто спостерігаються зміни структури чи характеристик інформації. Це явище, відоме як *concept drift*, вимагає постійного оновлення моделей, що збільшує витрати часу та ресурсів.

Крім того, багато методів виявлення аномалій або асоціативного аналізу схильні до проблеми збурень в даних, що може призводити до значної кількості хибнопозитивних або хибнонегативних результатів. Інший аспект полягає в тому, що ці методи часто розроблені для роботи з певними типами

даних і можуть не враховувати специфіку потоків, які включають багатовимірні або нерівномірно структуровані дані.

Таким чином, існуючі методи аналізу потоків даних стикаються з низкою обмежень, які потребують подальшого наукового вдосконалення. Зокрема, перспективними напрямами є розвиток адаптивних методів кластеризації потоків даних, що враховують динамічність потоків, та створення гібридних методів, які поєднують переваги різних підходів.

1.4 Аналіз сучасного стану адаптивних методів кластеризації потоків даних

Сучасні адаптивні методи кластеризації потоків даних розробляються для аналізу динамічних та великих обсягів інформації в реальному часі. Вони спрямовані на виявлення структур даних, які змінюються у часі, враховуючи явище концептуального зсуву та особливості потоку, такі як висока швидкість і варіативність. До таких методів належать алгоритми, що адаптуються до змін у даних, наприклад, STREAM, CluStream, DenStream, а також їх модифікації, які враховують збуреність, багатовимірність і нерівномірність розподілу даних. Однак ці методи мають певні недоліки, які обмежують їх ефективність у різних умовах [65, 66].

Одним із ключових викликів є висока обчислювальна складність, оскільки адаптивність вимагає постійного оновлення моделей, що може перевантажувати обчислювальні ресурси. Це особливо проблематично для великих потоків даних з високою швидкістю надходження.

Іншим недоліком є складність роботи з багатовимірними даними, адже більшість алгоритмів неефективно справляються з високою розмірністю, що ускладнює процес кластеризації та може знижувати точність результатів.

Ще однією проблемою є чутливість до параметрів моделі, таких як кількість кластерів або радіус сусідства. У потоках даних ці параметри часто змінюються, і їх неправильний вибір може значно погіршити результати.

Крім того, адаптивні методи зазвичай погано справляються з високим рівнем збуреності або аномаліями у потоках даних, що може призводити до утворення хибних кластерів. Додатково слід зазначити, що багато з цих методів мають обмежену інтерпретованість, оскільки складні механізми адаптації та оновлення можуть бути важко зрозумілими для користувача.

Таким чином, хоча сучасні адаптивні методи кластеризації є потужним інструментом для аналізу потоків даних, їх недоліки вказують на необхідність подальшого вдосконалення [32, 42, 49]. Зокрема, перспективними напрямками розвитку є зниження обчислювальної складності, підвищення стійкості до збурень та покращення роботи з багатовимірними структурами даних. Подолати складнощі, щодо обробки потоків даних за умов апріорної невизначеності можна за допомогою гібридизації та вдосконаленню відомих методів з оптимізаційними процедурами, що базуються на еволюційному самонавчанні [44, 45].

1.5 Аналіз сучасного стану гібридних методів кластеризації потоків даних

Гібридні методи кластеризації потоків даних поєднують переваги кількох підходів для забезпечення точнішого та стійкішого аналізу великих динамічних обсягів інформації [51]. Вони спрямовані на подолання обмежень окремих традиційних методів, таких як складність роботи з шумом, недостатня адаптивність або обмежена здатність до обробки багатовимірних даних. Сучасні гібридні алгоритми активно використовуються в таких сферах, як фінансовий аналіз, кібербезпека, прогнозування потоків трафіку та моніторинг інтернету речей.

Одним із ключових підходів є інтеграція методів кластеризації на основі центрів, таких як *k*-середніх, із алгоритмами щільнісного аналізу, такими як DBSCAN. Це дозволяє створювати моделі, які одночасно добре працюють із даними, що мають явно виражену структуру, і з нерівномірно розподіленими потоками. Інший напрямок полягає у використанні гібридів із включенням елементів машинного навчання, таких як нейронні мережі, які допомагають адаптуватися до концептуального зсуву та забезпечувати більш високу точність кластеризації.

Сучасний стан гібридних методів також включає поєднання онлайн- і офлайн-стратегій [44, 45-50]. Перший етап обробки виконується в реальному часі для грубої кластеризації, тоді як більш точний аналіз виконується офлайн, коли обсяг даних дозволяє знизити затримки. Це забезпечує баланс між швидкістю обробки та якістю результатів.

Однак гібридні методи мають і свої проблеми. Вони зазвичай характеризуються високою обчислювальною складністю, оскільки комбінування підходів вимагає значних ресурсів. Крім того, вони потребують складної налаштованості параметрів, що може бути непростим завданням для нових користувачів. Складність їхньої реалізації та інтерпретації результатів також може обмежувати використання цих методів у практичних задачах.

Загалом, гібридні методи кластеризації потоків даних демонструють значний потенціал, проте їхній розвиток потребує подальших досліджень, спрямованих на підвищення ефективності, зменшення обчислювальних витрат і полегшення впровадження в реальні системи.

Покращити роботу гібридних методів можна за допомогою еволюційних алгоритмів завдяки своїй здатності працювати в умовах великого пошукового простору.

1.6 Еволюційний підхід в інтелектуальному аналізі потоків даних

Гібридні методи кластеризації потоків даних тісно пов'язані з еволюційними підходами, оскільки обидва підходи спрямовані на адаптацію до динамічних умов і пошук оптимальних рішень в умовах невизначеності.

Еволюційні алгоритми, такі як генетичні алгоритми, рої часток або методи колоній мурах, використовують біологічно натхненні принципи для пошуку глобально оптимальних рішень у складних задачах. У контексті гібридних методів ці алгоритми часто інтегруються як один із компонентів, що відповідає за оптимізацію параметрів або структури кластерів.

Еволюційні методи сприяють підвищенню ефективності гібридних підходів завдяки здатності працювати в умовах великого пошукового простору. Наприклад, вони можуть автоматично визначати оптимальну кількість кластерів, налаштовувати радіуси сусідства або вагові коефіцієнти в багатовимірних потоках даних. Це особливо корисно для адаптації моделей до концептуального зсуву в потоках, коли зміни у даних вимагають перегляду кластерної структури [51-55].

Крім того, гібридизація еволюційних алгоритмів із традиційними методами кластеризації, такими як k-середніх [60, 61] або DBSCAN, дозволяє покращити їхню продуктивність. Еволюційний компонент може виконувати глобальний пошук, знижуючи ймовірність потрапляння в локальні мінімуми, тоді як традиційні методи забезпечують швидку локальну оптимізацію. Це створює баланс між точністю кластеризації та обчислювальними витратами.

Важливою перевагою поєднання гібридних і еволюційних методів є їхня здатність до самоадаптації [184, 186]. Еволюційні алгоритми дозволяють створювати моделі, які поступово вдосконалюються під впливом змін у даних, що підвищує їхню стійкість і ефективність у динамічних середовищах. Це робить їх особливо корисними для роботи з потоками даних, які характеризуються високою швидкістю і варіативністю.

Таким чином, еволюційні алгоритми виступають фундаментом для вдосконалення гібридних методів кластеризації, забезпечуючи гнучкість, адаптивність та здатність до вирішення складних задач в умовах змінних потоків даних.

1.6.1 Генетичний алгоритм

Генетичні алгоритми є конкретною реалізацією еволюційних алгоритмів, що застосовують спільну основу, адже обидва базуються на біологічних принципах еволюції, таких як природний відбір, схрещування та мутація для розв'язання задач оптимізації та пошуку [185].

Еволюційні алгоритми, як узагальнений підхід, використовують ітеративний процес, у якому множина можливих рішень (популяція) поступово вдосконалюється через механізми натхненні природними процесами. Цей клас алгоритмів включає кілька підходів, серед яких генетичні алгоритми, еволюційне програмування, еволюційні стратегії, генетичне програмування та інші. Вони орієнтовані на розв'язання задач, які складно розв'язати традиційними методами через великий простір можливих рішень або складність математичної моделі.

Генетичні алгоритми, як найбільш популярний підклас еволюційних алгоритмів, розроблені для оптимізації та адаптації рішень шляхом моделювання генетичних процесів. У них рішення представляються у вигляді хромосом, які кодують можливі варіанти вирішення задачі. Хромосоми проходять через кілька основних етапів:

Особливістю генетичного алгоритму є акцент на використання оператора «схрещування», який виконує операцію рекомбінації рішень-кандидатів, роль якої аналогічна ролі схрещування в живій природі.

Основною особливістю генетичних алгоритмів є використання оператора рекомбінації (схрещення) як основного механізму пошуку. Це ґрунтується на припущенні, що частини оптимального розв'язку можуть бути знайдені незалежно та рекомбіновані для отримання кращого розв'язку.

Задача кодується таким чином, щоб її вирішення могло бути представлено у вигляді масиву подібного до інформації складу хромосоми. Цей масив часто називають саме так «хромосома». Випадковим чином в масиві створюється деяка кількість початкових елементів «осіб», або початкова популяція. Особи оцінюються з використанням функції допасованості, в результаті якої кожній особі присвоюється певне її значення, яке визначає можливість виживання особи. Після цього з використанням отриманих значень допасованості вибираються особи, допущені до схрещування (селекція). До осіб застосовується «генетичні оператори» (в більшості випадків це оператор схрещення (crossover) і оператор мутації (mutation)), створюючи таким чином наступне покоління осіб. Особи наступного покоління також оцінюються застосуванням генетичних операторів і виконується селекція і мутація. Так моделюється еволюційний процес, що продовжується декілька життєвих циклів (поколінь), поки не буде виконано критерій зупинки алгоритму. Таким критерієм може бути:

– знаходження глобального, або надоптимального рішення;
– вичерпання числа поколінь, що відпущені на еволюцію;
– вичерпання часу, відпущеного на еволюцію.

Генетичні алгоритми можуть використовуватися для пошуку рішень в дуже великих і важких просторах пошуку.

Можна виділити такі етапи генетичного алгоритму:

1. Створення початкової популяції.
2. Обчислення функції пристосованості для осіб популяції (оцінювання).
3. Повторювання до виконання критерію зупинки алгоритму:

– вибір індивідів із поточної популяції (селекція);
– схрещення або/та мутація;
– обчислення функції пристосованості для всіх осіб;
– формування нового покоління.

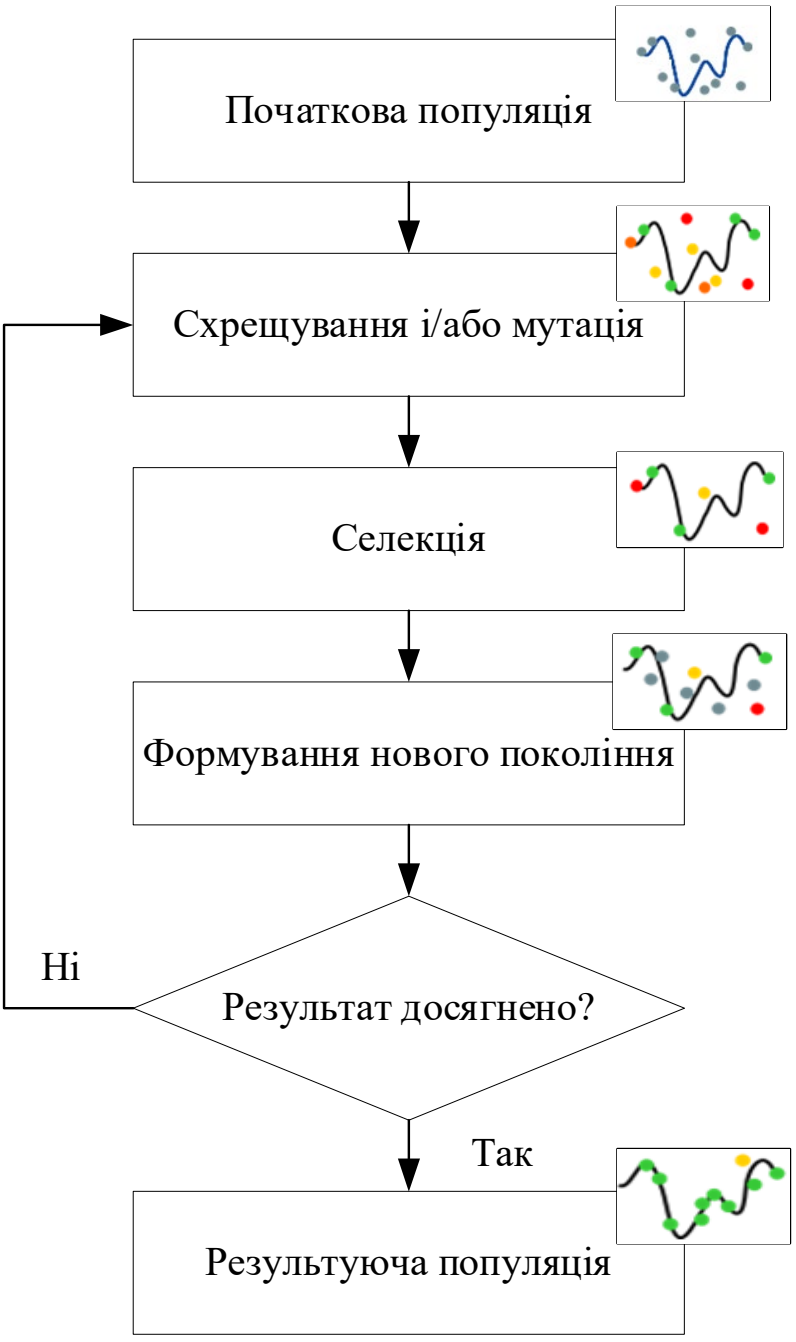


Рисунок 1.1 – Схема роботи генетичного алгоритму

Перед першим кроком необхідно випадковим чином створити деяку початкову популяцію. Навіть якщо популяція виявиться абсолютно неконкурентоздатною, генетичний алгоритм все одно достатньо швидко переведе її в придатну для життя популяцію.

Основна перевага генетичних алгоритмів у тому, що вони не потребують аналітичного визначення градієнтів або похідних, як це роблять деякі традиційні оптимізаційні методи. Вони ефективно працюють у задачах із багатьма локальними екстремумами, складними функціями або великою кількістю змінних.

Однак генетичні алгоритми мають і недоліки. Вони можуть бути обчислювально затратними, оскільки для досягнення збіжності потрібне значне число ітерацій. Крім того, вони не гарантують глобальної оптимальності, а результат може залежати від налаштувань параметрів, таких як розмір популяції, ймовірність мутації або метод відбору. Ці виклики призводять до необхідності комбінування генетичних алгоритмів із іншими методами, такими як градієнтні підходи чи локальна оптимізація, що породжує гібридні еволюційні методи.

1.6.2 Аналіз сучасного стану ройових еволюційних алгоритмів при обробці потоків даних

Генетичні алгоритми є важливим інструментом у галузі обробки потоків даних, що дозволяє вирішувати складні оптимізаційні завдання завдяки застосуванню принципів природного відбору та еволюції [181-186]. Їхнє використання у цьому контексті має значний потенціал через здатність адаптуватися до швидкозмінних умов, характерних для потоків даних, а також ефективно працювати з великою кількістю параметрів і розв'язувати нелінійні завдання.

Однією з основних причин зацікавленості в генетичних алгоритмах при роботі з потоками даних є їхня здатність до пошуку глобально оптимальних розв'язків у складних просторах, які часто виникають у задачах реального часу. Ця властивість особливо корисна для задач класифікації, прогнозування, кластеризації та виявлення аномалій у потоках даних. Генетичні алгоритми забезпечують можливість ефективного пошуку, використовуючи еволюційні

операції, такі як селекція, кросовер та мутація, що дозволяє їм долати локальні оптимуми, які можуть обмежувати продуктивність традиційних методів.

У сучасному стані розвитку генетичних алгоритмів спостерігається активне застосування гібридних методів, які комбінують еволюційний підхід з іншими технологіями, такими як машинне навчання, нейронні мережі та методи статистичного аналізу. Це дозволяє підвищити їхню ефективність, зокрема шляхом адаптивного налаштування параметрів алгоритмів залежно від характеристик потоку даних. Наприклад, використання нейронних мереж у поєднанні з генетичними алгоритмами дає змогу навчати моделі у реальному часі, постійно вдосконалюючи їхні прогнози на основі нових даних.

Ще однією важливою рисою сучасних досліджень є застосування розподілених і паралельних обчислювальних платформ для реалізації генетичних алгоритмів. Враховуючи великі обсяги потоків даних та необхідність їхньої обробки у реальному часі, розподілені обчислення дозволяють значно підвищити швидкість роботи алгоритмів і забезпечити їхню масштабованість. Наприклад, використання хмарних технологій дає змогу інтегрувати генетичні алгоритми у складні інфраструктури обробки потоків даних, що охоплюють численні джерела інформації.

Також слід зазначити, що генетичні алгоритми активно використовуються в задачах обробки потоків даних для підвищення якості результатів аналізу. В реальних застосунках, таких як фінансова аналітика, медичний моніторинг, розумне управління транспортними системами або енергетичними мережами, ці алгоритми дозволяють оптимізувати розподіл ресурсів, прогнозувати ризики та виявляти потенційно критичні події.

Проте використання генетичних алгоритмів при обробці потоків даних також стикається з певними викликами. Зокрема, необхідність швидкої адаптації до зміни умов потоку потребує вдосконалення механізмів селекції та мутації, а також розробки нових стратегій управління популяцією. Крім того, важливо забезпечувати баланс між точністю аналізу та швидкістю роботи алгоритмів, щоб відповідати вимогам реального часу. Однією з головних

проблем залишається висока обчислювальна складність, яка може бути критичною в умовах роботи з великими потоками.

Таким чином, сучасний стан генетичних алгоритмів у контексті обробки потоків даних характеризується їхньою еволюцією від класичних методів до адаптивних, розподілених і гібридних моделей, що забезпечує широкий спектр можливостей для аналізу даних у реальному часі, що є важливим аспектом у багатьох сферах діяльності.

На сьогодні, розвиток генетичних алгоритмів спрямований на підвищення їхньої швидкості, адаптивності та здатності до самооновлення робить їх одним із провідних інструментів для вирішення задач обробки потоків даних.

1.7 Оцінки якості кластеризації методів кластерного аналізу потоків даних

Оцінка якості кластеризації потоків даних є важливим аспектом у аналізі та обробці великих обсягів інформації, оскільки кластеризація зазвичай застосовується для виявлення структури в даних без попереднього розподілу на категорії. Потоки даних, як правило, мають особливості, зокрема, динамічний характер і великий обсяг, що робить оцінку кластеризації складнішою.

Для оцінки якості кластеризації зазвичай застосовуються дві основні категорії методів: внутрішні та зовнішні метрики.

Внутрішні метрики зосереджуються на характеристиках самих кластерів. Вони аналізують, наскільки добре дані згруповані в кластери та наскільки ці кластери розрізняються один від одного. Один із популярних способів оцінки внутрішньої якості кластеризації - це використання показника компактності, який вимірює, наскільки об’єкти в одному кластері схожі один на одного. Також використовується розрізненість, що характеризує, наскільки

відмінні один від одного різні кластери. Важливим інструментом для оцінки таких характеристик є силует, який вимірює, наскільки добре кожен елемент класифіковано в свій кластер порівняно з іншими кластерами.

Окрім цього, застосовуються більш складні методи, що дозволяють оцінити баланс між згортанням і розрізненістю кластерів, такі як індекс Девіса-Болдіна. Ці метрики зручні, оскільки вони не вимагають наявності зовнішніх еталонів або розмічених даних і можуть бути використані для оцінки кластеризації в реальному часі, коли потоки даних постійно змінюються.

Зовнішні методи оцінки якості кластеризації порівнюють отримані результати з еталонними даними або вручну розміченими класами. Наприклад, можна використовувати індекс Нормана-Моргана або індекс коригування Рандомізованого індексу (ARI), щоб оцінити ступінь відповідності між отриманими кластерами та відомими категоріями. Ці методи особливо корисні, коли існують попередні знання або еталонні розмітки, які можна використовувати для порівняння.

Одним з основних викликів при оцінці якості кластеризації потоків даних є те, що ці потоки є змінними і можуть змінювати свою структуру з часом. Тому важливо застосовувати методи, які можуть адаптуватися до нових даних, а також враховувати часову складову при оцінці кластеризації. Наприклад, методи, що використовують вимірювання стабільності кластерів на різних етапах потоку даних, можуть бути корисними для того, щоб виявити, чи зберігаються кластери в межах потоку або чи змінюються з часом.

Крім того, потоки даних можуть бути дуже великими, тому традиційні методи кластеризації, які потребують зберігання всіх даних у пам’яті, можуть бути непідходящими. Тому важливо використовувати метрики, які здатні працювати в різних умовах, використовуючи оновлення кількість кластерів у реальному часі та забезпечуючи ефективність обробки потокових даних.

Оцінка якості кластеризації потоків даних потребує обліку багатьох аспектів, включаючи характеристики самих даних, змінність їх структури і здатність алгоритмів кластеризації адаптуватися до цих змін.

1.7.1 Основні оцінки якості результатів кластеризації потоків даних

Метрика нормованої взаємної інформації (NMI, Normalized Mutual Information) є популярною метрикою для оцінки якості кластеризації. Вона використовується для вимірювання того, наскільки добре результати кластеризації узгоджуються з попередньо заданими еталонними категоріями або класами, при цьому NMI нормалізує значення, щоб забезпечити значення в діапазоні від 0 до 1, де 1 означає ідеальну відповідність, а 0 - повну відсутність взаємозв'язку.

Нормована взаємна інформація заснована на понятті взаємної інформації (MI), яка вимірює кількість інформації, що передається між двома змінними або наборами категорій. У контексті кластеризації, ці дві змінні - це набір еталонних класифікацій і результати кластеризації. Взаємна інформація показує, скільки інформації про класи можна отримати за допомогою кластерів і навпаки.

Взаємна інформація визначається за формулою:

$$NMI(U,V) = \frac{2 \cdot I(U,V)}{H(U) + H(V)}. \quad (1.1)$$

Таким чином, NMI є нормованим відношенням між взаємною інформацією і геометричним середнім ентропій двох змінних. Це дозволяє метриці залишатися незалежною від кількості класів або кластерів, а також забезпечує значення в діапазоні від 0 до 1.

Значення NMI = 1 вказує на повну відповідність між кластеризацією та еталонними класами, що означає, що класи і кластери ідеально збігаються.

Значення NMI = 0 вказує на відсутність будь-якої взаємної інформації між класифікацією та кластеризацією, що означає, що класи і кластери зовсім не пов’язані між собою.

Ця метрика є дуже корисною для порівняння результатів кластеризації, особливо коли класи не є чітко розділеними або коли дані мають високу варіативність. Її можна застосовувати як для оцінки кластеризації в статичних наборах даних, так і в поточних потоках даних, порівнюючи нові результати з наявними еталонами.

Індекс Девіса-Болдіна (DBI) є метрикою, яка використовується для оцінки якості кластеризації, зокрема, для вимірювання компактності і розрізненості кластерів. Цей індекс враховує як згортання кластерів (компактність), так і відстань між ними (розрізненість), що робить його корисним для оцінки структури кластеризації.

DBI базується на середніх відстанях між кластерами та їхньому внутрішньому згортанні. Чим менше значення DBI, тим краща якість кластеризації, оскільки це вказує на те, що кластери є компактними і добре відокремленими один від одного.

Формула для обчислення індексу Девіса-Болдіна виглядає так:

$$DBI = \frac{1}{k}\sum_{i=1}^{k}\max_{j\neq i}\left(\frac{\sigma_i + \sigma_j}{d_{ij}}\right). \tag{1.2}$$

Основні компоненти DBI:

Середня відстань всередині кластеру (compactness) міра того, наскільки компактним є кожен кластер. Якщо елементи в кластері близькі до одного центроїду, то ця величина буде малою.

Відстань між центроїдами кластерів (separation) міра того, наскільки далеко один кластер від іншого. Якщо кластери добре відокремлені, відстань між їхніми центроїдами буде великою.

Таким чином, DBI поєднує ці два аспекти в одну метрику:

– низьке значення DBI вказує на хорошу кластеризацію, тобто кластери є компактними і добре розрізняються;

– високе значення DBI свідчить про погану кластеризацію, коли кластери або занадто великі (погано згруповані), або надто близькі один до одного (погано відокремлені).

Індекс Девіса-Болдіна є корисним інструментом для порівняння різних алгоритмів кластеризації або для вибору оптимальної кількості кластерів, оскільки дозволяє врахувати і внутрішню згуртованість кластерів, і їхнє відокремлення.

Кластерна точність (Cluster Accuracy, CA) — це метрика, яка використовується для оцінки якості кластеризації, і зокрема, вона вимірює, наскільки добре кластери відповідають заданим класам або еталонним категоріям. Кластерна точність зазвичай застосовується в контексті зовнішніх метрик, коли є доступ до еталонної класифікації або правильно розмічених даних, з якими можна порівнювати результат кластеризації.

Основна ідея кластерної точності полягає в тому, щоб перевірити, наскільки класи в результаті кластеризації відповідають класам у розмічених даних. Вона оцінює, наскільки добре елементи одного кластеру відповідають одному класу в еталонних даних.

Формула для кластерної точності виглядає наступним чином:

$$CA = \frac{1}{n}\sum_{i=1}^{n} 1(l_i = map(c_i)). \tag{1.3}$$

Кластерна точність має кілька ключових аспектів:

– вимірює точність класифікації, ґрунтуючись на еталонних даних, тому для її обчислення потрібно мати попередньо розмічені дані.

– може бути корисною для порівняння результатів різних алгоритмів кластеризації, оскільки дає уявлення про те, наскільки добре ці алгоритми відображають істинні категорії у даних.

– може бути обмеженою, якщо класи не є чітко визначеними або якщо кластеризація вимагає більшої гнучкості у розпізнаванні складних структур даних.

Однак слід зауважити, що кластерна точність може бути не завжди найкращою метрикою для оцінки кластеризації, особливо в випадках, коли класи не є чітко окресленими або в даних присутні перекриття між класами. У таких випадках можуть бути використані інші метрики, такі як нормоване взаємне інформаційне вимірювання або індекс Девіса-Болдіна.

Індекс Дана (Dana Index, DI) є метрикою, яка використовується для оцінки якості кластеризації, зокрема для вимірювання того, наскільки добре результати кластеризації відповідають істинній структурі даних. Індекс Дана зосереджується на оцінці якості кластеризації в контексті її здатності правильно групувати схожі об’єкти разом і правильно відокремлювати різні об’єкти.

Індекс Данна - це одна з зовнішніх метрик кластеризації, що враховує не лише якість самих кластерів, але й їхнє співвідношення з еталонними класами (якщо такі є). Це дозволяє порівняти результати кластеризації з вже відомими категоріями або класами в розмічених даних.

Формула для індексу Данна:

$$DI = 1 - \sum_{i=1}^{S} p_i^2 \ . \tag{1.4}$$

Індекс Данна є корисним для оцінки якості кластеризації, особливо коли важливо не тільки оцінити згуртованість елементів всередині кластерів, але й відстань між різними кластерами, що робить його важливим інструментом при виборі найбільш ефективних методів кластеризації.

Індекс Силуету (Silhouette Index, SI) є популярною метрикою для оцінки якості кластеризації, яка вимірює, наскільки добре елементи одного кластеру згруповані разом і наскільки добре вони відокремлені від елементів інших кластерів. Індекс Силуету комбінує два основні аспекти кластеризації: компактність і відокремлення, дозволяючи оцінити, наскільки правильними є сформовані кластери.

Формула індексу Силуету:

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} . \quad (1.5)$$

Індекс Цалінскі-Харабаса (Calinski-Harabasz Index, CHI) -це одна з метрик для оцінки якості кластеризації, яка використовується для вимірювання того, наскільки добре кластеризація поділяє дані на окремі, добре відокремлені групи.

Індекс Цалінскі-Харабаса обчислюється за співвідношенням між варіацією між кластерами (міжкластерна дисперсія) і варіацією всередині кластерів (внутрішньокластерна дисперсія). Ідеальний випадок кластеризації - це коли варіація між кластерами висока (класи чітко відокремлені), а варіація всередині кластерів низька (елементи в кожному кластері дуже схожі один на одного):

$$CHI = \frac{Tr(Bk)}{Tr(Wk)} * \frac{n-k}{k-1} . \quad (1.6)$$

Коефіцієнт розподілу (Partition Coefficient, PC) - це метрика для оцінки якості кластеризації, яка вимірює ступінь «розпорошення» елементів по кластерах. Вона визначає, наскільки чітко елементи належать до конкретних кластерів, і наскільки добре кожен елемент ідентифікується з одним

кластером. Ідея полягає в тому, що чим вище значення коефіцієнта розподілу, тим краще елементи розподілені по кластерах.

Формула коефіцієнта розподілу:

$$PC = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij}^{2}. \tag{1.7}$$

Коефіцієнт розподілу є корисною метрикою для оцінки якості кластеризації, що дозволяє вимірювати, наскільки чітко елементи належать до своїх кластерів. Однак, для комплексної оцінки кластеризації, особливо для тих випадків, коли важливим є також відокремлення між кластерами, необхідно використовувати його разом з іншими метриками, такими як індекс Силуету або індекс Калінскі-Харабаса.

Ентропія класифікації (Classification Entropy, CE) - це метрика, що вимірює ступінь невизначеності або хаосу в кластеризації або класифікації. Вона базується на понятті ентропії з теорії інформації і використовується для оцінки того, наскільки «чистими» є кластери або класи в розділених даних.

Ентропія класифікації зазвичай оцінює, наскільки добре класи або кластери містять елементи одного класу (у випадку класифікації) або схожі елементи (у випадку кластеризації). Чим менша ентропія, тим чіткіше класи або кластери, оскільки елементи одного класу або кластера менш різноманітні. Висока ентропія вказує на те, що класи або кластери мають значне змішування елементів різних типів або класів.

Формула ентропії класифікації:

$$CE = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{c} u_{ij} log(u_{ij}). \tag{1.8}$$

Ентропія класифікації є корисним інструментом для оцінки якості кластеризації або класифікації, оскільки дозволяє виміряти, наскільки добре

елементи згруповані в кластери або класи. Чим нижча ентропія, тим чіткіше класи або кластери, що свідчить про високу якість класифікації або кластеризації.

Індекс розподілу (Separation Criterion, SC) є метрикою, яка вимірює наскільки добре кластери або класи відокремлені один від одного. Це показник для оцінки того, чи є кластери чітко видимими та відокремленими, або ж вони перекриваються, що ускладнює класифікацію або кластеризацію.

Індекс розподілу зазвичай враховує різні аспекти розподілу елементів у кластерах або класах. Він оцінює рівень відокремлення між кластерами, тобто наскільки великі відстані між центроїдами різних кластерів. Зазвичай чим більша відстань між кластерами, тим вищий індекс розподілу, що вказує на більш чітке відокремлення.

Формула індексу розподілу (SC):

$$SC = \frac{1}{1-n}\sum_{i=1}^{n} max_{j}\left(u_{ij}\right). \tag{1.9}$$

Індекс розподілу є корисною метрикою для оцінки того, наскільки чітко кластери відокремлені один від одного. Це дозволяє зрозуміти, наскільки добре алгоритм кластеризації або класифікації працює, зокрема в контексті відокремлення класів чи кластерів.

Індекс розділення (Separation Index, S) є метрикою, яка використовується для оцінки чіткості відокремлення між кластерами або класами в кластеризації або класифікації. Це показник того, наскільки добре різні кластери або класи розмежовані і наскільки сильно елементи одного кластеру чи класу відрізняються від елементів інших кластерів чи класів.

Індекс розділення може допомогти в аналізі того, чи є кластери добре відокремленими, і чи є класифікація ефективною. Чим вищий індекс розділення, тим більша відстань між кластерами, і тим чіткіше вони відокремлені. З іншого боку, низький індекс розділення свідчить про те, що

кластери або класи можуть бути нечітко відокремленими, що призводить до поганої кластеризації або класифікації.

Формула індексу розділення (S):

$$S = \frac{\text{Between-class variance}}{\text{Total variance}}. \tag{1.10}$$

Індекс розділення є корисною метрикою для оцінки якості кластеризації або класифікації, оскільки він допомагає визначити, наскільки чітко відокремлені класи або кластери. Чим вище значення індексу, тим краще відокремлені класи або кластери, що свідчить про високу якість класифікації або кластеризації.

Індекс Ксі та Бені (Xie-Beni index, XB) — це метрика, яка використовується для оцінки якості кластеризації, зокрема для вимірювання компактності кластерів і їхнього відокремлення. Цей індекс дозволяє порівнювати різні результати кластеризації і вибирати найкраще рішення для визначеної задачі.

Формула індексу Ксі та Бені (XB):

$$XB = \frac{\sum_{k=1}^{K} \sum_{x_i \in C_k} d\left(x_i, c_k\right)^2}{\min_{j \neq k} \sum_{x_i \in C_k} \sum_{x_j \in C_j} d\left(x_i, x_j\right)}. \tag{1.11}$$

Індекс Ксі та Бені є корисним інструментом для оцінки якості кластеризації, оскільки він допомагає вимірювати, наскільки кластери є компактними та відокремленими один від одного. Менше значення індексу вказує на кращу кластеризацію, в той час як більше значення свідчить про погану кластеризацію.

1.8 Формулювання проблеми дослідження

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. Говорячи про обчислювальний інтелект, 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 згаданих методів. Зокрема, особливу увагу привертають еволюційні нейро-фаззі 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 проблеми, а також за фіксованих умов: кількості класів, їхньої форми та рівня збурень.

У сучасних умовах, особливо за 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 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.

1.9 Висновок до розділу 1

1. 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.

2. 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.

3. 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.

4. Виділено необхідність розробки онлайн-методів і механізмів самонавчання, які забезпечують стійкість до аномалій і знижують вплив викривлень, зокрема через адаптацію моделей у реальному часі.

5. 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.

6. 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.

7. 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.

Результати розділу 1 відображено у публікаціях [1, 2] (Додаток А).

РОЗДІЛ 2
АДАПТИВНА КЛАСТЕРИЗАЦІЯ ПОТОКІВ ДАНИХ ЗА УМОВ ПЕРЕТИННИХ КЛАСІВ ТА АПРІОРНОЇ НЕВИЗНАЧЕННОСТІ

В першому розділі було обґрунтовано необхідність 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.

Другий розділ містить розробку методів адаптивної кластеризації потоків даних за умов перетинних класів, апріорної та поточної невизначеності, збурень, обмежених обчислювальних ресурсів та орієнтовані на онлайн-обробку даних, забезпечуючи високу точність навіть за відсутності повної апріорної інформації.

Однією з головних труднощів у розв’язанні проблеми обробки потоків даних є необхідність враховувати не тільки швидкість надходження нових даних, але й їх різноманітність та складність [8, 9, 12]. Стандартні методи обробки даних часто не підходять для таких задач, оскільки вони не здатні ефективно адаптуватися до змін у динамічних середовищах [13].

Відомі методи кластеризації, а саме, ієрархічні методи – це процедури прямого перебору. Вони не здатні працювати на вибірках, що змінюють розмірність. Методи, засновані на центроїдах – найбільш популярні методи і яскравим представником цих методів є K-means. Методи нечіткої кластеризації (FCM, алгоритм Густафсона-Кесселя, метод Ягера-Филева) – працюють тільки в пакетному офлайн режимі, тобто якщо вибірка фіксована і задана апріорі, або якщо ця вибірка іде у онлайн режимі, то ці методи не працюють. Більшість відомих методів формують кластери у формі опуклих тіл (гіперкуль або гіпереліпсоїдів з довільною орієнтацією осей), хоча на практиці

це далеко не так. Необхідно також підкреслити, що методи нечіткої кластеризації, засновані на центроїдах – найбільш популярні методи, які працюють тільки за умов апріорі заданої кількості кластерів опуклої форми.

В роботі поставлено задачу розробки адаптивних методів нечіткої кластеризації потоків даних за умов перетинних класів та апріорної невизначеності. В рамках вирішення даної задачі необхідно розробити:

– адаптивні методи нечіткої кластеризації потоків даних за умов перетинних класів, апріорної та поточної невизначеності;
– провести експериментальні дослідження розроблених та модифікованих методів.

2.1 Формальна постановка задачі адаптивної кластеризації потоків даних за умов перетинних кластерів

Задача кластеризації масивів даних, які описуються наборами векторів-образів (спостережень), досить часто зустрічається в багатьох задачах, пов’язаних з інтелектуальним аналізом даних, при цьому останнім часом особлива увага приділяється так званій нечіткій кластеризації [3, 6, 7, 13-28], коли оброблюваний вектор-образ (вектор-спостереження) ознак з різними рівнями ймовірностей, можливостей, достовірності чи належності може відноситися одночасно до кількох класів (кластерів), які в свою чергу мають різні форми в просторі ознак, можуть перетинатись, перекриватись одне з одним, бути досить сильно розмитими, що, в свою чергу, достатньо сильно ускладнює роботу методів кластеризації [18-44, 76-93]. Слід також звернути увагу на природу даних, які по своїй суті, мають низку особливостей, різну якість, форму, розмірність, тип, тощо.

Разом з тим, у багатьох задачах інтелектуального аналізу даних, вихідні масиви даних можуть містити викривлені дані, інформація в яких з тих чи

інших причин відсутня, спотворена або зовсім містить аномальні спостереження.

Крім зазначених, існує ще цілий ряд підходів [11, 13, 24, 29, 32, 39, 42, 43, 84, 90-115] до обробки даних, що містять пропуски, проте всі вони працездатні тільки у випадках, коли масив вихідних спостережень заданий заздалегідь в повному обсязі і не змінюється в процесі обробки. У той же час існує широкий клас задач, коли дані надходять на обробку послідовно в режимі реального часу так, як це відбувається при навчанні самоорганізовних мап Кохонена [27] або їх модифікацій [29, 32].

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 в масиві даних X, або поточний дискретний час в задачах Data Stream Mining.

Якщо дані надходять на обробку послідовно у режимі реального часу, ці дані повинні бути розбиті на m перетинних класів (кластерів), при цьому для кожного спостереження $x(k)$ повинен бути також розрахований рівень нечіткої належності до кожного з кластерів $\mu_q(k), q = 1,2,...m$. Передбачається

також, що дані, які надходять на обробку, нормовані в гіперкуб $[-1;1]$ 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.

Адаптивна кластеризація є важливим напрямом у сфері аналізу даних, що забезпечує автоматичне групування об’єктів у динамічно змінних середовищах. Її актуальність зумовлена необхідністю аналізу великих обсягів інформації, що постійно оновлюється, зокрема в контексті потокової обробки даних, кібербезпеки, управління транспортними системами та прогнозування поведінки користувачів.

На відміну від традиційних алгоритмів, адаптивні методи здатні коригувати кількість кластерів, оновлювати параметри групування в реальному часі та враховувати змінність простору ознак. Це забезпечує їхню стійкість до викидів, динамічність і застосовність у задачах, де структура даних не є статичною.

Попри виклики, що стоять перед цими методами, подальший розвиток алгоритмів, зокрема шляхом інтеграції з технологіями машинного навчання, відкриває нові можливості для підвищення їхньої ефективності та застосовності у складних інформаційних середовищах.

2.2 Адаптивний метод кластеризації

Найбільш поширений серед неієрархічних методів алгоритм k-середніх, також називається швидким кластерним аналізом. Повний опис алгоритму можна знайти у роботі Хартігана та Вонга [1, 31]. На відміну від ієрархічних методів, які не вимагають попередніх припущень щодо числа кластерів, для

можливості використання цього методу необхідно мати гіпотезу про найбільш ймовірну кількість кластерів. Загальна ідея алгоритму: задане фіксоване число k кластерів спостереження зіставляються кластерам отже середні в кластері (для всіх змінних) максимально можливо відрізняються друг від друга.

Для вирішення задачі кластеризації в онлайн режимі доцільно скористатися самоорганізовною картою Т. Кохонена [27], що має просту архітектуру з прямою передачею інформації і крім нульового рецепторного шару містить єдиний шар нейронів, найчастіше тих же адаптивних лінійних асоціаторів (ALA).

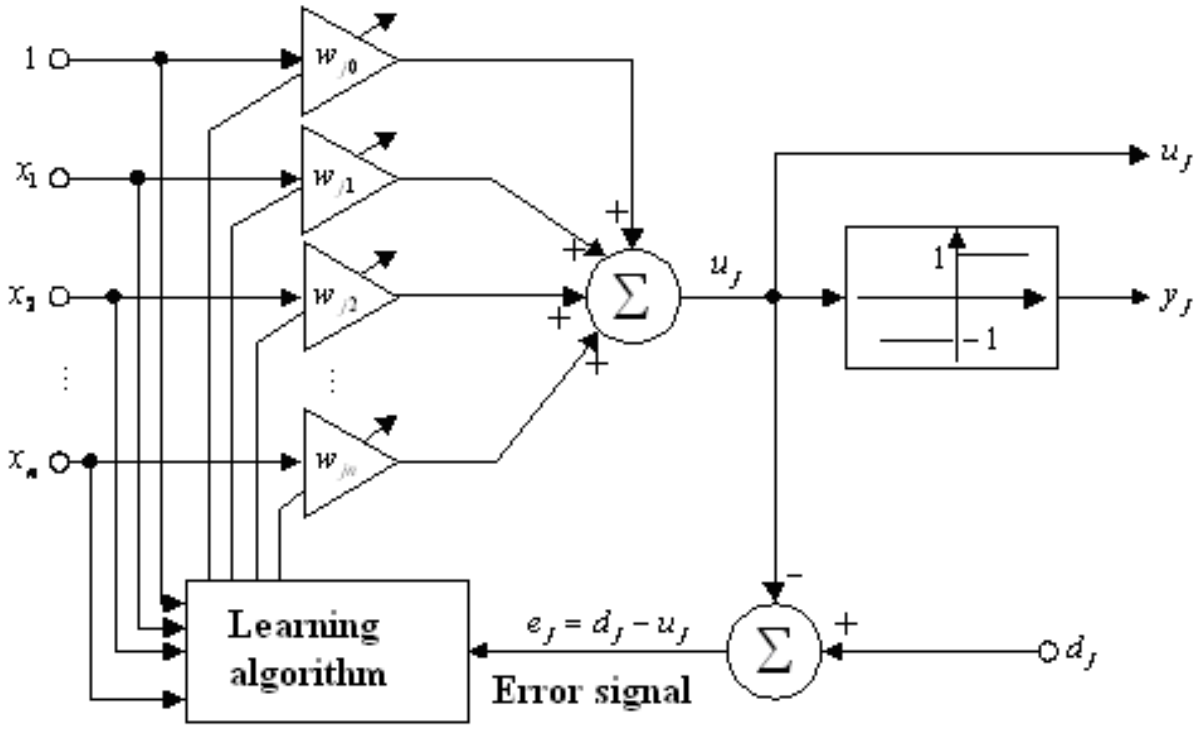


Рисунок 2.1 – Адаптивний лінійний асоціатор (ALA)

Одним з найпростіших нейронів, що навчаються, є адаптивний лінійний елемент (ADALINE), запропонований Б. Уідроу і наведений на рисунку 2.1. Адаліна може використовуватися як елементарний нейрон у складі штучних нейронних мережах (ШНМ), так і самостійно в задачах розпізнавання образів, обробки сигналів, реалізації логічних функцій. У цьому випадку Адаліни

підходять для роботи з даними, які надходять на обробку в послідовному режимі.

Кожен нейрон пов’язаний з кожним рецептором нульового шару прямими зв’язками та з рештою нейронів поперечними внутрішньошаровими (латеральними) зв’язками. Саме латеральні зв’язки забезпечують збудження одних нейронів та гальмування інших.

Завдяки такій організації мережі, кожен нейрон-ALA отримує всю інформацію про аналізований вектор-образ і генерує на своєму виході відповідний відгук, після чого між нейронами виникає конкуренція, в результаті якої визначається єдиний нейрон-переможець з максимальним вихідним сигналом. Цей сигнал по латеральним зв’язкам забезпечує збудження найближчих «сусідів» переможця та придушення реакції далеко віддалених нейронів.

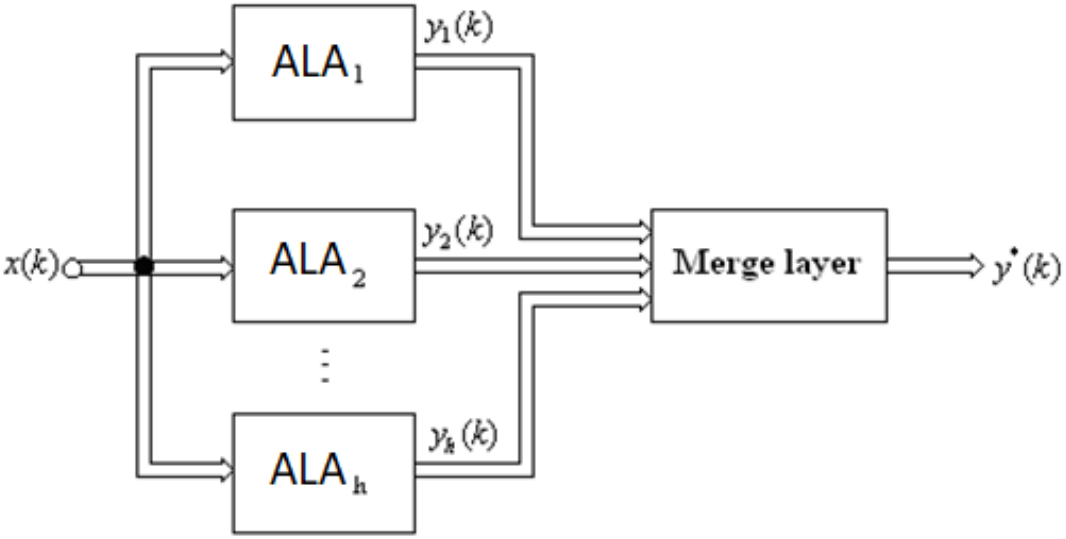


Рисунок 2.2 - Ансамбль h - паралельно працюючих штучних нейронних мереж

Таким чином, в процесі конкурентного самонавчання формуються групи нейронів, кожен з яких максимально реагує на образи «своїх» кластерів-підобластей вхідного простору сигналів.

Якість розв'язання різноманітних задач прогнозування, розпізнавання образів, зворотного моделювання, управління, відновлення даних тощо, може бути підвищена за допомогою використання ансамблів нейронних мереж, у яких одні й ті ж дані обробляються відразу декількома h - паралельно працюючими штучними нейронними мережами.

Вихідні сигнали деяким чином комбінуються в загальну оцінку, яка дає уявлення про якість отриманих результатів за допомогою локальних мереж, що входять в ансамблі, як це показано на рисунку 2.2.

2.3 Рекурентний ймовірнісний метод нечіткої кластеризації

Задача кластеризації (класифікації в режимі самонавчання) багатовимірних даних є важливою частиною інтелектуального аналізу даних (Data Mining), в рамках якої склався ряд напрямків і підходів [1, 2]. Один з таких напрямків утворюють методи нечіткої (фаззі-) кластеризації, в основі яких лежить припущення що класи-кластери, які формуються, взаємно перетинаються так, що кожен вектор-спостереження з різними рівнями належності-ймовірності-можливості може належати одночасно до декількох класів одночасно.

Тут найбільш широкого поширення набули алгоритми ймовірнісної нечіткої кластеризації і, перш за все, метод нечітких с-середніх (FCM) [3,4]. Цей підхід обмежується ймовірнісними обмеженнями на рівні належності так, що «забруднені» збуреннями і викидами спостереження можуть бути віднесені до різних класів з практично однаковими рівнями належності.

Метод ймовірнісної нечіткої кластеризації пов'язаний з мінімізацією цільової функції

$$Goal\left(\mu_q(k), c_q\right) = \sum_{k=1}^{N}\sum_{q=1}^{m}\mu_q^{\beta}(k)\, d^2\left(x(k), c_q\right) \qquad (2.1)$$

за наявності обмежень

$$\sum_{q=1}^{N}\mu_q(k)=1,0<\sum_{q=1}^{N}\mu_q(k)<N, \tag{2.2}$$

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, найчастіше метриці Ітакури-Сайто [53], окремим випадком якої є, наприклад, відстань Махаланобіса [76].

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

$$\begin{cases}\mu_q(k+1)=\dfrac{\left(d^2\left(x(k+1),c_q(k)\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(d^2\left(x(k+1),c_l(k)\right)\right)^{\frac{1}{1-\beta}}},\\ c_q(k+1)=c(k)+\eta(k+1)\mu_q^{\beta}(k+1)\left(x(k+1)-c_q(k)\right),\end{cases} \tag{2.3}$$

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

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

$$\begin{cases} \mu_q\left(k+1\right)=\dfrac{\left(d^2\left(x\left(k+1\right),c_q\left(k\right)\right)\right)^{-1}}{\sum\limits_{l=1}^{m}\left(d^2\left(x\left(k+1\right),c_l\left(k\right)\right)\right)^{-1}}, \\ c_q\left(k+1\right)=c\left(k\right)+\eta\left(k+1\right)\mu_q^{\beta}\left(k+1\right)\left(x\left(k+1\right)-c_q\left(k\right)\right), \end{cases} \tag{2.4}$$

при цьому цікаво помітити, що другі співвідношення (2.3), (2.4) є по суті правилом самонавчання Т. Кохонена [27] «Winner Takes More» з функцією сусідства $\mu_q^{\beta}\left(k+1\right)$ або $\mu_q^{2}\left(k+1\right)$ на кожному кроці налаштування.

Шляхом нескладних перетворень можна переписати перше співвідношення (2.4) у вигляді

$$\mu_q\left(k+1\right)=\frac{1}{1+\dfrac{d^2\left(x\left(k+1\right),c_q(k)\right)^{\beta-1}}{\sum\limits_{\substack{l=1\\ l\neq q}}^{m}d^2\left(x\left(k+1\right),c_l(k)\right)^{\frac{1}{1-\beta}}}} \tag{2.5}$$

або для $\beta=2$

$$\begin{cases} \mu_q\left(k+1\right)=\dfrac{1}{1+\dfrac{d^2\left(x\left(k+1\right),c_q(k)\right)}{\sigma_q^2\left(k+1\right)}}, \\ \sigma_q^2\left(k+1\right)=\left(\sum\limits_{\substack{l=1\\ l\neq q}}^{m}\left(d^2\left(x\left(k+1\right),c_l\left(k\right)\right)^{-1}\right)\right)^{-1}, \end{cases} \tag{2.6}$$

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

У зв'язку з цим в [13] був запропонований можливісний підхід до нечіткої кластеризації (PCM) більш стійкий до шумів і збурень.

2.4 Рекурентний можливісний метод нечіткої кластеризації

Можливісний метод нечіткої кластеризації є узагальненням традиційних підходів до нечіткої кластеризації, таких як алгоритм нечітких с-середніх. Основна відмінність цього методу полягає в тому, що він використовує можливісну міру належності (possibility membership), яка дозволяє кожному об’єкту незалежно належати до певного кластера без обмеження, що сума всіх належностей для даної точки має дорівнювати одиниці.

Це особливо важливо у випадках, коли дані містять шум або аномальні точки, оскільки традиційні нечіткі методи, такі як FCM, можуть змушувати такі точки штучно належати до одного з кластерів. Можливісна кластеризація вирішує цю проблему, дозволяючи окремим об'єктам мати низький ступінь приналежності до всіх кластерів, що робить її більш стійкою до викидів.

Можливісні методи нечіткої кластеризації пов’язані з мінімізацією цільової функції [27, 42, 59, 112]

$$Goal\left(\mu_q(k), c_q, \omega_q\right) = \sum_{k=1}^{N}\sum_{q=1}^{m}\mu_q^{\beta}(k)d^2\left(x(k), c_q\right) + \sum_{q=1}^{m}\omega_q\sum_{k=1}^{N}\left(1-\mu_q(k)\right)^{\beta}, \quad (2.7)$$

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 алгоритму Крішнапурама-Келлера [22] має вигляд:

$$\begin{cases} \mu_q(k+1) = \left(1 + \left(\dfrac{d^2\left(x(k+1), c_q(k)\right)}{\omega_q(k)}\right)^{\frac{1}{\beta-1}}\right)^{-1}, \\ c_q(k+1) = c_q(k) + \eta(k+1)\mu_q^{\beta}(k+1)\left(x(k+1) - c_q(k)\right), \\ \omega_q(k+1) = \dfrac{\sum\limits_{p=1}^{k+1}\mu_q^{\beta}(p)d^2\left(x(p), c_q(k+1)\right)}{\sum\limits_{p=1}^{k+1}\mu_q^{\beta}(p)} \end{cases} \quad (2.8)$$

або при $\beta = 2$

$$\begin{cases} \mu_q(k+1) = \left(1 + \dfrac{d^2\left(x(k+1), c_q(k)\right)}{\omega_q(k)}\right)^{-1}, \\ c_q(k+1) = c_q(k) + \eta(k+1)\mu_q^{2}(k+1)\left(x(k+1) - c_q(k)\right), \\ \omega_q(k+1) = \dfrac{\sum\limits_{p=1}^{k+1}\mu_q^{2}(p)d^2\left(x(p), c_q(k+1)\right)}{\sum\limits_{p=1}^{k+1}\mu_q^{2}(p)}. \end{cases} \quad (2.9)$$

У першому співвідношенні (2.9) виникає функція Коші з параметром ширини ω_q, що визначається третім співвідношенням (2.9).

2.5 Адаптивна нечітка робастна кластеризація даних на основі міри подібності

Як вже зазначалося, для вирішення завдання нечіткої кластеризації даних, що містять викиди, можна використовувати цільові функції

спеціального виду [6, 13, 20, 32, 86]. З практичної точки зору зручнішим є використання замість цільових функцій, заснованих на метриках, так званих, мір подібності (SM) [116, 117], до яких пред'являються більш м'які ніж для метрик умови:

$$\begin{cases} S_q(x_k, x_p) \geq 0, \\ S_q(x_k, x_p) = S_q(x_p, x_k) \\ S_q(x_k, x_k) = 1 \geq S_q(x_k, x_p) \end{cases}$$

(відсутня нерівність трикутника), а задача кластеризації може бути «прив'язана» до максимізації цих мір.

Якщо дані перетворені так, що $-1 \leq x_{ki} \leq 1$, то міра подібності може бути сконструйована так, щоб придушити небажані дані, що лежать на краях інтервалу $[-1,1]$.

Рисунок 2.3 ілюструє використання подібності функції Коші з різними параметрами ширини $\sigma^2 < 1$.

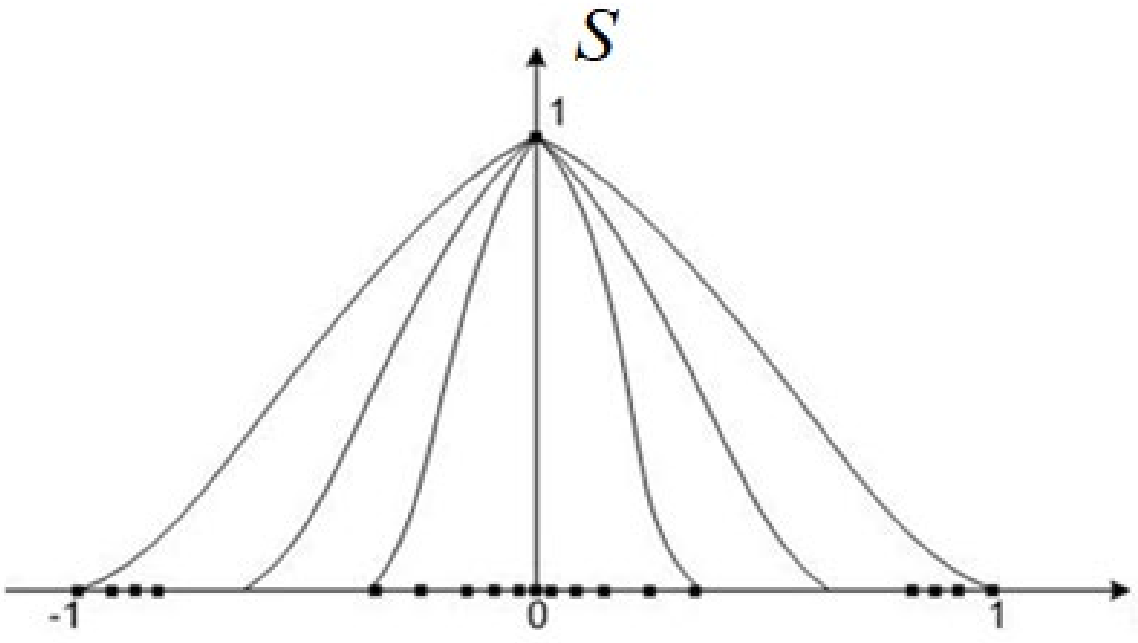


Рисунок 2.3 – Міра подібності на основі функції Коші

Вибираючи параметр ширини σ^2 функції

$$S(x_k, c_q) = \frac{1}{1 + \frac{\left\| x_k - c_q \right\|^2}{\sigma^2}} = \frac{\sigma^2}{\sigma^2 + \left\| x_k - c_q \right\|^2} = \frac{\sigma^2}{\sigma^2 + d^2(x_k, c_q)}, \qquad (2.10)$$

можна виключити вплив аномальних спостережень, що в принципі неможливо зробити за допомогою евклідової метрики

$$d^2(x_k, c_q) = \left\| x_k - c_q \right\|^2, \qquad (2.11)$$

при цьому можна помітити, що $S(x_k, c_q)$ є величина зворотна $d^2(x_k, c_q)$.

Далі, вводячи у розгляд цільову функцію, засновану на мірі подібності (2.10),

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

ймовірнісні обмеження

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

функцію Лагранжа

$$L_S(\mu_q(k), c_q, \lambda(k)) = \sum_{k=1}^{N} \sum_{q=1}^{m} \frac{\mu_q^{\beta}(k)\sigma^2}{\sigma^2 + \left\| x_k - c_q \right\|^2} + \sum_{k=1}^{N} \lambda(k) \left(\sum_{q=1}^{m} \mu_q(k) - 1 \right) \qquad (2.12)$$

(тут $\lambda(k)$ - невизначені множники Лагранжа) і вирішуючи систему рівнянь Каруша-Куна-Таккера, приходимо до вирішення

$$\begin{cases} \mu_q(k) = (S(x_k, c_q))^{\frac{1}{\beta-1}} \Big/ \sum_{l=1}^{m} (S(x_k, c_l))^{\frac{1}{\beta-1}}, \\ \lambda(k) = -(\sum_{l=1}^{m} (\beta S(x_k, c_l))^{\frac{1}{\beta-1}})^{\beta-1}, \\ \nabla_{c_q} L_S(\mu_q(k), c_q, \lambda(k)) = \sum_{k=1}^{N} \mu_q^{\beta}(k) \dfrac{x_k - c_q}{(\sigma^2 + \left\| x_k - c_q \right\|^2)^2} = \vec{0}. \end{cases} \quad (2.13)$$

Останнє рівняння системи (2.13) не має аналітичного рішення, тому для знаходження сідлової точки лагранжиана (2.12) можна скористатися процедурою Ерроу-Гурвіца-Удзави, в результаті застосування якої приходимо до алгоритму

$$\begin{cases} \mu_q(k+1) = \dfrac{(S(x_{k+1}, c_q(k)))^{\frac{1}{\beta-1}}}{\sum_{l=1}^{m} (S(x_{k+1}, c_l(k)))^{\frac{1}{\beta-1}}}, \\ c_q(k+1) = c_q(k) + \eta(k+1)\mu_q^{\beta}(k+1) \dfrac{x_{k+1} - c_q(k)}{(\sigma^2 + \left\| x_{k+1} - c_q(k) \right\|^2)^2} = \\ = c_q(k) + \eta(k+1)\varphi_q(k+1)(x_{k+1} - c_q(k)), \end{cases} \quad (2.14)$$

де $\varphi_q(k+1) = \dfrac{x_{k+1} - c_q(k)}{(\sigma^2 + \left\| x_{k+1} - c_q(k) \right\|^2)^2}$ - функція сусідства робастного WTM-правила самонавчання.

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 (2.15) при цьому рішення про належність кожного x_k до конкретного кластера приймається за максимальним значенням міри подібності

$$\begin{cases} \mu_q^{(\tau+1)}(k) = \dfrac{(S(x_k, c_q^{(\tau)}(k)))^{\frac{1}{\beta-1}}}{\sum\limits_{l=1}^{m} (S(x_k, c_l^{(\tau)}(k)))^{\frac{1}{\beta-1}}}, \\ c_q^{(Q)}(k) = c_q^{(0)}(k+1), \\ c_q^{(\tau+1)}(k+1) = c_q^{(\tau)}(k+1) + \eta(k+1)\dfrac{(\mu_q^{(Q)}(k))^{\beta}}{(\sigma^2 + \left\| x_{k+1} - c_q^{(\tau)}(k+1) \right\|^2)^2}(x_{k+1} - c_q^{(\tau)}(k+1)). \end{cases} \quad (2.15)$$

Аналогічним чином може бути синтезований алгоритм 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 завдання її оптимізації, приходимо до вирішення:

$$\begin{cases} \mu_q(k+1) = \left(1 + \left(\dfrac{S^{-1}(x_{k+1}, c_q(k))}{\omega_q(k)}\right)\right)^{-1}, \\ c_q(k+1) = c_q(k) + \eta(k+1)\mu_q^{\beta}(k+1)\dfrac{x_{k+1} - c_q(k)}{(\sigma^2 + \left\|x_{k+1} - c_q(k)\right\|^2)^2}, \\ \omega_q(k+1) = \dfrac{\sum\limits_{p=1}^{k+1}\mu_q^{\beta}(p)S^{-1}(x_p, c_q(k+1))}{\sum\limits_{p=1}^{k+1}\mu_q^{\beta}(p)}, \end{cases} \quad (2.16)$$

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

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

(2.17)

2.6 Адаптивна нечітка робастна кластеризація 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). Нескладно бачити, що така PSM має вигляд

$$S_P(x_k, c_q) = \frac{\sigma^2}{\sigma^2 + d_P^2(x_k, c_q)}, \tag{2.18}$$

що дозволяє отримати алгоритми з бажаними властивостями на основі процедур, описаних вище.

Так, на основі процедури (2.15) можна ввести адаптивний робастний ймовірнісний алгоритм нечіткої кластеризації даних з пропусками:

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

(2.19)

а також можна записати адаптивний робастний можливісний алгоритм нечіткої кластеризації даних з пропусками (2.20). Таким чином, використання часткової міри подібності, заснованої на частковій відстані, дозволяє вирішувати задачі нечіткої кластеризації даних, що містять як пропуски, так і аномальні спостереження

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

(2.20)

2.7 Рекурентний правдоподібний метод нечіткої кластеризації

РСМ-алгоритми страждають від, так званої, проблеми співпадіння, коли в процесі обробки інформації деякі кластери починають зливатися один з одним, що в результаті веде до невірної оцінки кількості сформованих кластерів [22, 39].

Цих недоліків позбавлені алгоритми правдоподібної нечіткої кластеризації [97-100], засновані на апараті теорії достовірності [98].

В рамках цього підходу в процесі розрахунків оцінюються не тільки рівні нечіткої належності, але і рівні довіри, що засновані на мірі належності спеціального виду.

Правдоподібні методи нечіткої кластеризації пов'язані з мінімізацією цільової функції [97-107, 118-125]

$$Goal\left(Credib_q(k), c_q\right) = \sum_{k=1}^{N}\sum_{q=1}^{m} Credib_q^{\beta}(k) d^2\left(x(k), c_q\right) \tag{2.21}$$

за наявності обмежень

$$\begin{cases} 0 \le Credib_q(k) \le 1 \forall q,k, \\ \sup Credib_q(k) \ge 0,5 \forall k, \\ Credib_q(k) + \sup Credib_l(k) = 1 \end{cases} \tag{2.22}$$

для всіх q,k, для яких $Credib_q(k) \ge 0$,

де $Credib_q(k)$ – рівень правдоподібності того, що спостереження $x(k)$ належить кластеру Cl_q.

В процедурах довірчої кластеризації рівень нечіткої належності визначається функцією належності [97-101]

$$\mu_q(k) = \varphi_q\left(d\left(x(k), c_q\right)\right), \tag{2.23}$$

де $\varphi_q\left(d\left(x(k), c_q\right)\right)$ – монотонно зменшується на інтервалі $[0, \infty]$ та $\varphi_q(0) = 1, \varphi_q(\infty) \to 0$.

Процедура (2.23) є не що інше, як міра подібності заснована на відстані [85, 86]. В якості такої міри у [118, 119] було запропоновано використовувати функцію

$$\mu_q(k) = \frac{1}{1 + d^2\left(x(k), c_q\right)}, \tag{2.24}$$

що X знов таки є функцією Коші з одиничним параметром ширини при цьому ніяк не враховується характер розподілу даних у вхідному масиві X.

Тому, більш прийнятним є вибір замість (2.24) співвідношень (2.5), (2.6), що прив'язані саме до характеру даних як в цілому у масивів X, так і у кластерах $Cl_q, q = 1, 2...m$.

Тут цікаво помітити, що функція Коші постійно виникає в задачах нечіткої кластеризації, як вже відзначених ймовірнісної, можливісної, правдоподібної, так, робастної кластеризації стійкої до аномальних викидів у вихідних даних [24, 36, 85, 86, 116, 126].

Таким чином, якщо пакетний метод правдоподібної нечіткої кластеризації має вигляд [118, 119-122]

$$\begin{cases} \mu_q(k) = \dfrac{1}{1 + d^2\left(x(k), c_q\right)}, \\ \mu_q^*(k) = \dfrac{\mu_q(k)}{\sup \mu_l(k)}, \\ Credib_q(k) = \dfrac{1}{2}\left(\mu_q^*(k) + 1 - \sup \mu_l^*(k)\right), \\ c_q = \dfrac{\sum\limits_{k=1}^{N} Credib_q^{\beta}(k)\, x(k)}{\sum\limits_{k=1}^{N} Credib_q^{\beta}(k)}, \end{cases} \tag{2.25}$$

то його рекурентна версія може бути записана у формі

$$\begin{cases} \sigma_q^2(k+1) = \sum\limits_{\substack{l=1 \\ l \neq q}}^{m} \left(d^2\left(x(k+1), c_l(k)\right)^{\frac{1}{1-\beta}} \right)^{-1}, \\ \mu_q(k+1) = \dfrac{1}{1 + \dfrac{\left(d^2\left(x(k+1), c_q(k)\right)\right)^{\beta-1}}{\sigma_q^2(k+1)}}, \\ \mu_q^*(k+1) = \dfrac{\mu_q(k+1)}{\sup \mu_l(k+1)}, \\ Credib_q(k+1) = \dfrac{1}{2}\left(\mu_q^*(k+1) + 1 - \sup \mu_l^*(k+1)\right), \\ c_q(k+1) = c_q(k) + \eta(k+1)\, Credib_q^{\beta}(k+1)\left(x(k+1) - c_q(k)\right) \end{cases} \tag{2.26}$$

або для найбільш розповсюдженого фаззіфікатора $\beta = 2$

$$\begin{cases} \sigma_q^2(k+1) = \left(\sum\limits_{\substack{l=1 \\ l \neq q}}^{m} \left\| x(k+1) - c_l(k) \right\|^{-2} \right)^{-1}, \\ \mu_q(k+1) = \left(1 + \dfrac{\left\| x(k+1) - c_q(k) \right\|^2}{\sigma_q^2(k+1)} \right)^{-1}, \\ \mu_q^*(k+1) = \dfrac{\mu_q(k+1)}{\sup \mu_l(k+1)}, \\ Credib_q(k+1) = \dfrac{1}{2}\left(\mu_q^*(k+1) + 1 - \sup \mu_l^*(k+1) \right), \\ c_q(k+1) = c_q(k) + \eta(k+1) Credib_q^2(k+1)\left(x(k+1) - c_q(k) \right). \end{cases} \tag{2.27}$$

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

2.8 Онлайн нечітка правдоподібна кластеризація викривлених даних на основі міри подібності спеціального типу

Правдоподібна нечітка кластеризація пов’язана з мінімізацією цільової функції (2.21) за обмежень (2.22). При цьому рівень правдоподібності розраховується на основі функції належності [118-132]

$$\mu(k) = \varphi\left(d(x,c) \right) \tag{2.28}$$

що задовольняє умовам:

$\varphi_q(\bullet)$ монотонно зменшується в інтервалі $[0,\infty]$,

$$\varphi_q(0)=1,$$

$$\varphi_q(\infty)\to 0.$$

Помітимо, що функція (2.28) є за суттю мірою подібності, заснованій на відстані [38].

В якості такої функції у [121] було запропоновано використовувати вираз

$$\mu_q(k)=\left(1+d^2\left(x(k),c_q\right)\right), \tag{2.29}$$

що є звичайною дзвонуватою функцією належності, яка використовується в системах нечіткого висновування.

Цікаво зауважити, що вираз (2.29) може бути переписаний у формі

$$\mu_q(k)=\left(d^2\left(x(k),c_q(k)\right)\right)^{\frac{1}{1-\beta}}\left(\sum_{l=1}^{m}\left(d^2\left(x(k),c_q(k)\right)\right)^{\frac{1}{1-\beta}}\right)^{-1}=$$

$$=\left(\left(d^2\left(x(k),c_q(k)\right)\right)^{\frac{1}{1-\beta}}\left(d^2\left(x(k),c_q(k)\right)\right)^{\frac{1}{1-\beta}}\right)+\left(\sum_{l=1}^{m}\left(d^2\left(x(k),c_q(k)\right)\right)^{\frac{1}{1-\beta}}\right)^{-1}=$$

$$=\left(1+\left(d^2\left(x(k),c_q(k)\right)\right)^{\frac{1}{1-\beta}}\sum_{l=1}^{m}\left(d^2\left(x(k),c_q(k)\right)\right)^{\frac{1}{1-\beta}}\right)^{-1},$$

а для метрики Евкліда і $\beta=2$ приймає вид функції щільності розподілу Коші з параметром ширини σ_q^2

$$\mu_q(k)=\left(1+\frac{\left\|x(k)-c_q(k)\right\|^2}{\sigma_q^2}\right)^{-1}, \tag{2.30}$$

$$\sigma_q^2 = \left(\sum_{\substack{l=1 \\ l \neq q}}^{m} \left\| x(k) - c_l(k) \right\|^{-2} \right)^{-1}. \tag{2.31}$$

Остаточно пакетний метод правдоподібної нечіткої кластеризації може бути записаний у формі [121]:

$$\begin{cases} \mu_q^{(\tau+1)}(k) = \left(1 + d^2\left(x(k), c_q^{(\tau)}\right)\right)^{-1} \\ \mu_q^{*(\tau+1)}(k) = \mu_q^{(\tau+1)}(k)\left(\sup \mu_l^{(\tau+1)}(k)\right)^{-1}, \\ Credib_q^{(\tau+1)}(k) = \dfrac{1}{2}\left(\mu_q^{*(\tau+1)}(k) + 1 - \sup\limits_{l \neq q} \mu_l^{*(\tau+1)}(k) \right), \\ c_q^{(\tau+1)} = \sum\limits_{k=1}^{N} \left(Credib_q^{(\tau+1)}(k) \right)^{\beta} x(k) \left(\sum\limits_{k=1}^{N} \left(Credib_q^{(\tau+1)}(k) \right)^{\beta} \right)^{-1}. \end{cases} \tag{2.32}$$

Нарешті, можна записати онлайн версію методу правдоподібної нечіткої кластеризації у вигляді

$$\begin{cases} \sigma_q^2(k+1) = \left(\sum\limits_{\substack{l=1 \\ l \neq q}}^{m} \left\| x(k+1) - c_q(k) \right\|^{-2} \right)^{-1}, \\ \mu_q(k+1) = \left(1 + \dfrac{\left\| x(k+1) - c_q(k) \right\|^2}{\sigma_q^2(k+1)} \right)^{-1} \\ \mu_q^*(k+1) = \mu_q(k+1)\left(\sup \mu_l(k+1)\right)^{-1}, \\ Credib_q(k+1) = \dfrac{1}{2}\left(\mu_q^*(k+1) + 1 - \sup\limits_{l \neq q} \mu_l^*(k+1) \right), \\ c_q(k+1) = c_q(k) + \eta(k+1) Credib_q^{\beta}(k+1)\left(x(k+1) - c_q(k) \right). \end{cases} \tag{2.33}$$

Як видно, з обчислювальної точки зору online алгоритм правдоподібної нечіткої кластеризації не складніше рекурентних версій FCM і PCM, зберігаючи при цьому переваги правдоподібного підходу.

2.9 Рекурентна модифікація методу Густафсона-Кесселя

В процесі нечіткої кластеризації за допомогою розглянутих алгоритмів та методів класи, що формуються, мають форму гіперсфер, що не завжди відповідає реальним умовам, коли ці кластери можуть мати довільну форму. Більш адекватними та зручними є кластери гіпереліпсоїальної форми з довільною орієнтацією осей у просторі ознак.

Такі кластери можуть бути сформовані за допомогою методу Густафсона-Кесселя [113] та його модифікацій, що засновані на мінімізації цільової функції (2.1) за наявності обмежень (2.2) (ймовірнісний підхід), але в якості відстані використовується метрика вигляду

$$d_{V_q}^2\left(x(k),c_q\right)=\left\|x(k)-c_q\right\|_{V_q}^2=\left(x(k)-c_q\right)^T V_q\left(x(k)-c_q\right), \tag{2.34}$$

де

$$\begin{cases} V_q=\left(\det S_q\right)^{\frac{1}{n}} S_q^{-1}, \\ S_q=\sum\limits_{k=1}^{N}\mu_q^{\beta}(k)\left(x(k)-c_q\right)\left(x(k)-c_q\right)^T. \end{cases} \tag{2.35}$$

Мінімізація (2.34) за обмежень (2.2) веде до відомого результату

$$\begin{cases} \mu_q(k) = \dfrac{\left(d^2_{V_q}\left(x(k), c_q\right)\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(d^2_{V_l}\left(x(k), c_l\right)\right)^{\frac{1}{1-\beta}}}, \\ c_q = \dfrac{\sum\limits_{k=1}^{N}\mu_q^{\beta}(k)x(k)}{\sum\limits_{k=1}^{N}\mu_q^{\beta}(k)} \end{cases} \tag{2.36}$$

або для $\beta = 2$

$$\begin{cases} \mu_q(k) = \dfrac{1}{1 + \dfrac{d^2_{V_q}\left(x(k), c_q\right)}{\sigma_q^2(k)}}, \\ \sigma_q^2(k) = \left(\sum\limits_{\substack{l=1 \\ l \neq q}}^{m} d^{-2}_{V_l}\left(x(k), c_l\right)\right)^{-1}, \\ c_q(k) = \dfrac{\sum\limits_{k=1}^{N}\mu_q^2(k)x(k)}{\sum\limits_{k=1}^{N}\mu_q^2(k)}. \end{cases} \tag{2.37}$$

Таким чином, співвідношення (2.35) – (2.37) є за суттю процедурою ймовірнісної нечіткої кластеризації, але класи, що формуються, мають форму гіпереліпсоїдів з довільною орієнтацією осей.

Для того, щоб ввести рекурентну модифікацію методу Густафсона-Кесселя типу (2.3), (2.4) можна скористатися формулою Шермана-Моррісона обернення матриць [113, 114] та лемою матричного детермінанта [115], що веде до онлайн процедури

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

(2.38)

Процедура (2.38) є узагальненням методу (2.3) на випадок гіпереліпсоїдальних кластерів.

2.10 Рекурентна модифікація методу Густафсона-Кесселя можливісної нечіткої кластеризації

Нескладно також модифікувати метод Густафсона-Кесселя на випадок можливісної нечіткої кластеризації. При цьому цільова функція (2.7) набуває вигляду

$$Goal\left(\mu_q(k), c_q, \omega_q\right) = \sum_{k=1}^{N}\sum_{q=1}^{m}\mu_q^{\beta}(k) d^2_{V_q}\left(x(k), c_q\right) + \sum_{q=1}^{m}\omega_q\sum_{k=1}^{N}\left(1 - \mu_q(k)\right)^{\beta}, \quad (2.39)$$

а пакетна форма алгоритму:

$$\begin{cases} \mu_q(k) = \left(1 + \dfrac{d_{V_q}^2\left(x(k), c_q\right)^{\frac{1}{\beta-1}}}{\omega_q} \right)^{-1}, \\ c_q = \dfrac{\sum\limits_{k=1}^{N} \mu_q^{\beta}(k) x(k)}{\sum\limits_{k=1}^{N} \mu_q^{\beta}(k)}, \\ S_q = \sum\limits_{k=1}^{N} \mu_q^{\beta}(k)\left(x(k) - c_q\right)\left(x(k) - c_q\right)^T, \\ V_q = \left(\det S_q\right)^{\frac{1}{n}} S_q^{-1}, \\ \omega_q(k) = \dfrac{\sum\limits_{k=1}^{N} \mu_q^{\beta}(k)\left(x(k) - c_q\right)^T V_q \left(x(k) - c_q\right)}{\sum\limits_{k=1}^{N} \mu_q^{\beta}(k)}. \end{cases} \tag{2.40}$$

Метод (2.40) можна також записати у рекурентній формі:

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

(2.41)

Не дивлячись на деяку громіздкість алгоритму (2.41), його реалізація не скільки не складніша у порівнянні з рекурентною процедурою можливісної кластеризації (2.8) або (2.9).

2.11 Рекурентна модифікація методу Густафсона-Кесселя правдоподібної нечіткої кластеризації

Що стосується правдоподібного варіанта методу Густафсона-Кесселя, то замість цільової функції (2.21) має бути використана її модифікація

$$Goal\left(Credib_q\left(k\right),c_q\right)=\sum_{k=1}^{N}\sum_{q=1}^{m}Credib_q^{\beta}\left(k\right)d_{V_q}^2\left(x\left(k\right),c_q\right) \tag{2.42}$$

з урахуванням умов (2.22) – (2.24). Тоді можна записати

$$\begin{cases} S_q=\sum_{k=1}^{N}\mu_q^{\beta}\left(k\right)\left(x(k)-c_q\right)\left(x\left(k\right)-c_q\right)^T, \\ V_q=\left(\det S_q\right)^{\frac{1}{n}}S_q^{-1}, \\ \mu_q\left(k\right)=\dfrac{1}{1+d_{V_q}^2\left(x\left(k\right),c_q\right)}, \\ \mu_q^*\left(k\right)=\dfrac{\mu_q\left(k\right)}{\sup\mu_l\left(k\right)}, \\ Credib_q\left(k\right)=\dfrac{1}{2}\left(\mu_q^*\left(k\right)+1-\sup\mu_l^*\left(k\right)\right), \\ c_q=\dfrac{\sum_{k=1}^{N}Credib_q^{\beta}\left(k\right)x\left(k\right)}{\sum_{k=1}^{N}Credib_q^{\beta}\left(k\right)}. \end{cases} \tag{2.43}$$

Співвідношення (2.43) є узагальненням методу (2.25) на випадок метрики (2.34).

І нарешті, можна ввести рекурентну модифікацію методу Густафсона – Кесселя правдоподібної нечіткої кластеризації:

Процедура (2.43) є узагальненим процедури правдоподібної кластеризації (2.26) та рекурентної модифікації методу Густафсона-Кесселя (2.38):

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

2.12 Апробація методів адаптивної кластеризації потоків даних за умов перетинних кластерів на тренувальних вибірках

Для оцінки ефективності методи адаптивної кластеризації потоків даних за умов перетинних кластерів апробовані на тренувальних вибірках з архіву UCI repository.

Для порівняльного аналізу використано шість відомих сучасних алгоритмів кластеризації: метод нечітких с-середніх (FCM), метод с-середніх

(PCM) [12-22], метод Густафсона-Кесселя (GK) [113], метод правдоподібної кластеризації (CCM) [97-129], адаптивний алгоритм імовірнісної нечіткої кластеризації (APrFC), адаптивний алгоритм можливісної нечіткої кластеризації (APosFC), адаптивна правдоподібна нечітка кластеризація (ACrFC) та рекурентна модифікація методу Густафсона-Кесселя правдоподібної нечіткої кластеризації (RCM_GK).

Апробація проведена на зразках трьох наборів даних: Abalone, Wine і Gas. Опис цих наборів даних наведено в Таблиці 2.1.

Середню похибку центроїдів кластерів запропонованого RCM_GK порівнювали з іншими добре відомими методами, отриманий результат продемонстровано в Таблиці 2.2 і Таблиці 2.3.

Таблиця 2.1 – Опис тренувальних наборів даних

Назва вибірки	**Кількість спростережень**	**Кількість атрибутів**	**Кількість кластерів**
Abalone	4177	8	3
Wine	178	13	3
Gas	296	2	6

Таблиця 2.2 – Порівняльний аналіз середньої похибки центроїдів кластерів різних методів кластеризації

Назва вибірки	**Методи кластеризації**				
	FCM	**PCM**	**CCM**	**GK**	**RCM_GK**
Abalone	2,61	1,54	0,13	0,10	0,05
Gas	2,69	1,73	0,21	0,13	0,049
Wine	2,71	2,86	0,33	0,183	0,037

Таблиця 2.3 – Порівняльний аналіз середньої похибки центроїдів кластерів різних методів кластеризації

Назва вибірки	**Методи кластеризації**			
	APrFC	**APosFC**	**ACrFC**	**RCM_GK**
Abalone	0,12	0,11	0,07	0,05
Gas	0,17	0,13	0,06	0,049
Wine	0,20	0,18	0,04	0,037

Порівняльний аналіз, наведений в Таблиці 2.3 демонструє точність роботи методів кластеризації, з використанням адаптивного підходу до кластеризації потоків даних. З таблиці видно, що ймовірністний та можливісний методи схожі за результатами. Адаптивний метод правдоподібної кластеризації потоків даних, демонструє кращі варіант. Тобто різниця між отриманими результатами десь 7%.

В рекурентній модифікації методу Густафсона-Кесселя правдоподібної нечіткої кластеризації, відповідно до важливого призначення нечітких і достовірних рівнів належності до всіх вибірок, вплив очевидний в отриманих результатах кластеризації та точному визначенні центроїдів кластерів.

Щоб оцінити практичність цих методів, порівняймо час роботи кластеризації на різних наборах даних (Таблиця 2.4).

Таблиця 2.4 - Порівняння часу виконання (у секундах) восьми методів на тестових наборах даних

Вибірка	**Методи кластеризації**							
	FCM	**PCM**	**CCM**	**GK**	**APrFC**	**APosFC**	**ACrFC**	**RCM_GK**
Abalone	1,62	0,28	0,25	0,27	0,12	0,11	0,17	0,15
Gas	0,43	0,28	0,25	0,22	0,17	0,18	0,16	0,15
Wine	0,22	0,21	0,22	0,24	0,13	0,18	0,14	0,16

Значення адаптивних алгоритмів, таких як: адаптивний алгоритм для імовірнісної нечіткої кластеризації, адаптивний алгоритм для можливісної нечіткої кластеризації, адаптивна правдоподібна нечітка кластеризація в деяких випадках краща, ніж правдоподібна модифікація методу Густафсона – Кесселя завдяки адаптивним функціям.

На рисунку 2.4 представлено порівняння часу роботи цих методів. Як видно з діаграми, швидкість роботи при розв'язанні задачі запропонованих методів є вищою, ніж відомих класичних алгоритмів майже на 10-15%.

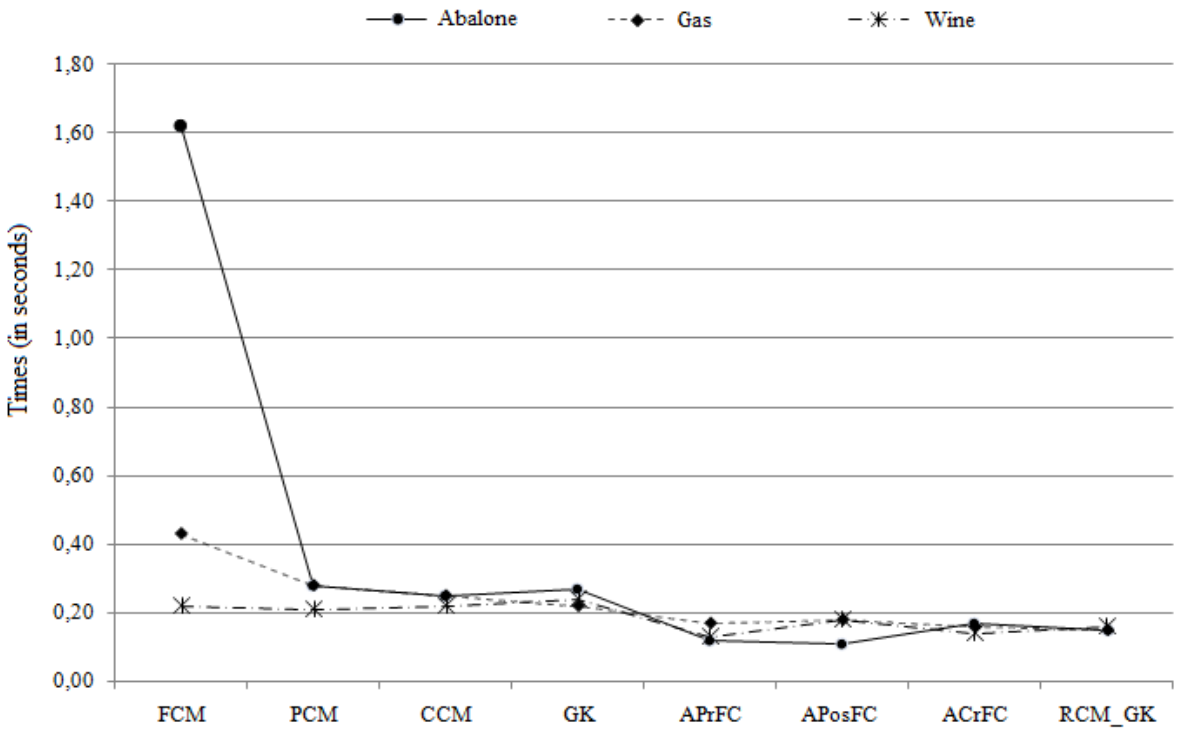


Рисунок 2.4 - Порівняння часу роботи восьми методів кластеризації на тестових наборах даних

Запропонована модифікація методу Густафсона-Кесселя базується на підході правдоподібності до нечіткої кластеризації та дозволяє формувати перетинні класи гіпереліпсоїдальної форми з довільною орієнтацією осей у просторі ознак.

Таблиця 2.5 - Порівняння кількості ітерацій і часу виконання (у секундах) восьми методів на наборі даних Abalone

	Методи кластеризації							
	FCM	**PCM**	**CCM**	**GK**	**APrFC**	**APosFC**	**ACrFC**	**RCM_GK**
Кількість ітерацій	40	99	75	100	45	78	55	76
Час	1,63	4,51	3,43	1,22	1,41	1,58	1,49	1,15

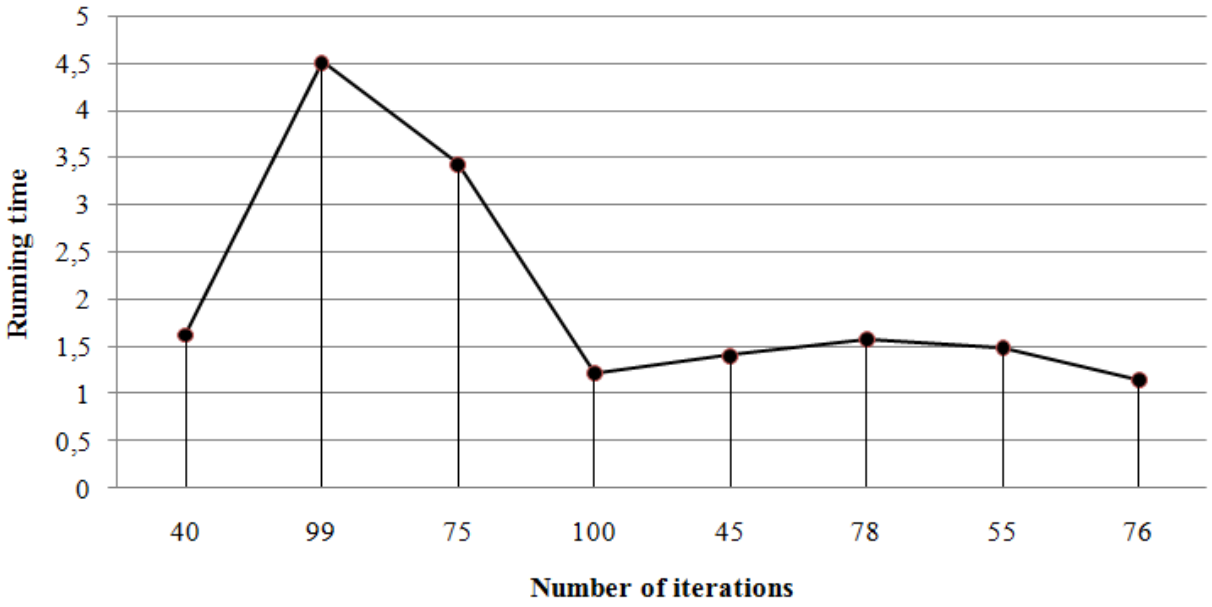


Рисунок 2.5 - Порівняння кількості ітерацій восьми методів кластеризації на тестовому наборі даних Abalone

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.

Проведені експериментальні дослідження підтвердили ефективність запропонованих методів адаптивної кластеризації потоків даних за умов перетинних кластерів, що дозволяє рекомендувати їх для використання на практиці для вирішення задач автоматичної кластеризації.

Запропоновані методи призначені для використання в гібридних системах обчислювального інтелекту і, перш за все, в задачах навчання штучних нейронних мереж, нейронечітких систем, а також в задачах кластеризації та класифікації.

2.12.1 Апробація адаптивних методів кластеризації за умов перетинних кластерів на пошкоджених викидами та пропусками тренувальних вибірках

У більшості завдань кластеризації, пов'язаних з опрацюванням реальних даних, вихідна інформація, як правило, викривлена аномальними викидами і пропусками, причому кількість цих викидів і «дир» може бути співрозмірним з об'ємом «чистих» даних, при цьому можлива ситуація, коли всі дані є «брудними». Зрозуміло, що «класичні» методи (як пакетні, так і онлайнові) в цій ситуації неефективні. Для боротьби з аномальними викидами в задачах нечіткої кластеризації були запропоновані рекурентні методи, засновані на використанні як робастних цільових функцій спеціального виду, так і мір подібності, нечутливі до викидів і призначені для роботи як в пакеті, так і послідовно. Що ж стосується відсутніх спостережень-пропусків, то тут також був розроблений ряд методів (в рамках ймовірнісного та можливісного підходів) як пакетних, так і онлайн. І, нарешті, в [116, 120] була введена робастна ймовірнісна процедура нечіткої кластеризації даних, виявлених як викиди, так і пропуски на основі міри подібності спеціального типу.

Щоб перевірити розроблені методи, а також аналіз якості кластеризації даних порівняно з іншими більш відомими підходами, дослідження було

проведено з використанням добре відомих тестових наборів даних UCI репозиторію, таких як Wine, Gas, Glass та Iris. Опис цих наборів даних наведено в таблиці 2.6.

Таблиця 2.6 – Опис тестових вибірок UCI репозиторію

Вибірка	**Кількість спостережень**	**Кількість атрибутів**	**Кількість кластерів**	**Ресурс**
Вина	178	13	3	Forina et al.(1988)
Газ	296	2	6	Box and Jenkins (1970)
Скло	214	9	6	Maskey and Glass (1977)
Іриси	150	4	3	Fisher (1936)

Результати кластеризації наборів даних показані в таблиці 2.7. Як показано в таблиці, метод нечіткої правдоподібної кластеризації високі результати.

Порівняльний аналіз даних проводився з даними, запропонованими методами кластеризації, які містять відсутні значення, такими як дані адаптивної імовірнісної нечіткої кластеризації з відсутніми значеннями, адаптивної можливої нечіткої кластеризації з відсутніми значеннями та класичні алгоритми FCM і *K*-means.

Для оцінки якості кластеризації даних ми використовували індекс Силуету (SI), індекс Цалінскі-Харабаса (CHI) та індекс Девіса-Болдіна (DBI).

Індекс силуету показує, наскільки середня відстань до об’єктів кластера відрізняється від середньої відстані до об’єктів інших кластерів. Це значення знаходиться в діапазоні [-1, 1]. Значення, близькі до -1, відповідають «поганим» (розрізненим) типам кластеризації. Значення, близькі до нуля, вказують на те, що кластери перетинаються і перекриваються. Значення, близькі до 1, відповідають «щільним» чітко виділеним кластерам. Таким чином, чим більший силует, тим чіткіші скупчення, і вони являють собою компактні, щільно згруповані хмари точок.

Таблиця 2.7 - Результат кластеризації набору даних «Іриси» за допомогою різних методів

Метод кластеризації	**SI**	**CHI**	**DBI**
Адаптивна імовірнісна нечітка кластеризація даних з пропусками	0,2326	921,58	1,28
Адаптивна можливісна нечітка кластеризація даних з пропусками	0,2325	922,01	1,25
Адаптивна правдоподібна нечітка кластеризація даних з пропусками	0,3335	965,42	1,05
FCM	0,2354	986,39	1,23
K-means	0,3676	1419,28	1,09

Згідно з таблицею 2.7, можна зробити такі висновки щодо результатів кластеризації набору даних "Іриси" за допомогою різних методів.

Адаптивна правдоподібна нечітка кластеризація даних з пропусками показує найкращий показник SI (0,3335), що вказує на кращу відокремленість кластерів і вищу якість кластеризації порівняно з іншими методами.

K-means має найвищий показник SI (0,3676), що свідчить про ще кращу відокремленість кластерів, але цей метод показує більшу варіативність в кластеризації порівняно з адаптивними методами.

K-means також демонструє найкращий результат за CHI (1419,28), що вказує на найкращу згуртованість кластерів та їх розмежованість в порівнянні з іншими методами.

Адаптивні методи також мають високе значення CHI, але вони трохи поступаються K-means, що може свідчити про дещо гіршу згуртованість кластерів.

Адаптивна правдоподібна нечітка кластеризація даних з пропусками має найкращий результат за DBI (1,05), що свідчить про найменшу середню дисперсію між кластерами і вищу якість кластеризації.

FCM має близький до найкращого результат DBI (1,23), що також свідчить про хорошу відокремленість і згуртованість кластерів.

K-means має найкращий результат за CHI, але результат за DBI є нижчим (1,09), що свідчить про дещо більшу дисперсію між кластерами.

Адаптивна правдоподібна нечітка кластеризація показує найкращі результати за SI та DBI, що вказує на високу якість кластеризації з відокремленістю та згуртованістю кластерів.

K-means є сильним конкурентом за CHI, але має трохи гірший показник за DBI та SI, що може свідчити про меншу якість кластеризації в порівнянні з адаптивними методами.

FCM та інші адаптивні методи забезпечують хорошу якість кластеризації, однак не досягають рівня адаптивного правдоподібного методу за DBI.

Таким чином, для набору даних "Іриси" адаптивна правдоподібна нечітка кластеризація даних з пропусками є найкращим методом серед представлених, з найбільш високими значеннями індексів, що вказує на хорошу відокремленість та компактність кластерів.

Таблиця 2.8 - Порівняння 100 експериментів для різних наборів даних

Назва вибірки	**Метод кластеризації**	**Загальна точність**		
		Найвища	**Середня**	**Дисперсія**
Вина	FCM	68.54	68.54	0
	CFC	67.98	67.98	0

Продовження таблиці 2.8

Скло	FCM	49.53	49.08	0.01
	CFC	44.86	44.86	0
Газ	FCM	79.05	77.33	11.33
	CFC	68.58	68.55	0.01
Іриси	FCM	89.33	89.33	0
	CFC	91.33	90.06	0.04

Згідно з таблицею 2.8, можна зробити наступні висновки щодо порівняння методів кластеризації FCM і CFC на різних наборах даних, зокрема для Вина, Скло, Газ, і Іриси.

Для вибірки Іриси метод CFC показує кращі результати як за найвищою, так і за середньою точністю порівняно з FCM, а також забезпечує невелику дисперсію, що свідчить про стабільність результатів.

Для вибірки Газ метод FCM показує найвищу точність, але з високою дисперсією, що може вказувати на варіативність результатів в залежності від конкретних характеристик даних. CFC тут також має стабільність, але з нижчою точністю.

Для вибірки Скло та Вина обидва методи мають схожі результати, але FCM виявляється трохи кращим для вибірки Скло через стабільнішу середню точність.

Метод CFC показує більш стабільні результати на більшості наборів даних, зокрема для вибірки Iris, проте FCM демонструє перевагу в точності для вибірки Gas. Це вказує на те, що для деяких задач може бути корисним використовувати FCM, особливо коли важлива висока точність, навіть за наявності варіативності. CFC ж буде кращим варіантом, коли необхідна стабільність і консистентність результатів.

Звичайно, якість запропонованого методу слід оцінити. З цієї причини ми використали загальне порівняння точності 100 експериментів для різних наборів даних і двох алгоритмів кластеризації: метод нечітких середніх (FCM) і достовірну нечітку кластеризацію (CFC).

Метод правдоподібної нечіткої кластеризації працює не тільки з повними даними, але й з даними, які містять відсутні значення.

Для проведення експериментальних досліджень штучно введено 10 відсутніх значень у набір даних Iris.

Рисунок 2.6 демонструє набір даних нечіткої правдоподібної кластеризації (CFC) Iris із 10 відсутніми значеннями.

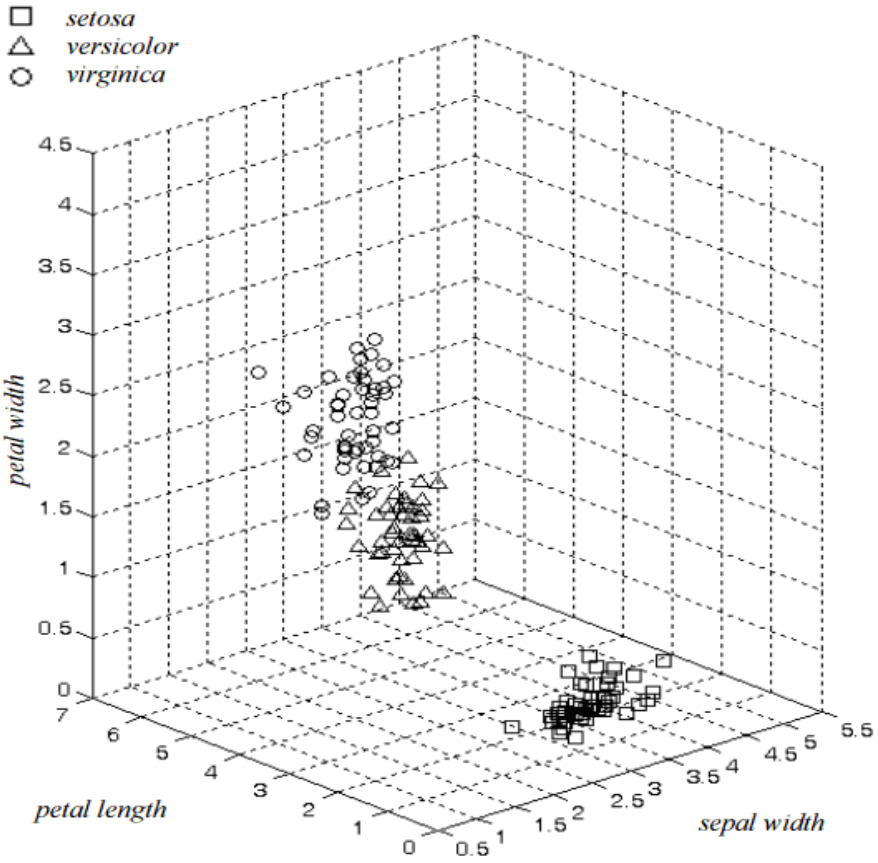


Рисунок 2.6 – Набір даних Iris для правдоподібної нечіткої кластеризації з 10 відсутніми значеннями

Проведені експерименти підтвердили дієвість запропонованих методів правдоподібної нечіткої кластеризації спотворених даних і дозволяють

рекомендувати його для використання на практиці для вирішення задач автоматичної кластеризації спотворених даних. Загальна точність методу правдоподібної нечіткої кластеризації покращена на 5-9% і не залежить від розміру та типу вибірки.

Методи призначені для використання в гібридних систем обчислювального інтелекту і, а саме, в проблемах навчання штучних нейронних мереж, нейронечітких систем, а також в проблемах кластеризації та класифікації.

2.12.2 Апробація адаптивних методів кластеризації на основі модифікованої міри подібності спеціального типу за умов перетинних кластерів на пошкоджених викидами та пропусками тренувальних вибірках

Експериментальні дослідження проводились на вибірці UCI репозиторію Iris.

Для оцінки якості роботи робастного методу кластеризації використовували критерії якості кластеризації, такі як: коефіцієнт розподілу (PC), ентропія класифікації (CE), індекс розподілу (SC), індекс розділення (S), індекс Ксі та Бені (XB), індекс Данна (DI).

Отже, результат кластеризації потрібно перерахувати, оскільки це був жорсткий алгоритм розділення. Результати експериментальних досліджень наведено в таблиці 2.9 на основі вибірки Iris і таблиці 2.10 на основі вибірки Wine.

Із таблиці 2.9 можна зробити аналіз результатів кластеризації для двох методів - Адаптивного робастного методу кластеризації на основі модифікованої міри подібності та FCM (Fuzzy C-Means) - на вибірці Iris із UCI Repository дозволяє зробити кілька важливих висновків про їх ефективність та якість кластеризації.

Таблиця 2.9 – Порівняльний аналіз адаптивного ймовірнісного робастного методу кластеризації на основі міри подібності та FCM на основі вибірки Iris

Методи кластеризації	**Iris UCI repository**					
	PC	**CE**	**SC**	**S**	**XB**	**DI**
Адаптивний робастний метод кластеризації на основі модифікованої міри подібності	0,8199	0,2122	0,2567	0,0022	0,0015	1
FCM	0,8011	0,3410	0,1345	0,0030	7,1965	0,0080

Таблиця 2.10 – Порівняльний аналіз адаптивного можливісного робастного методу кластеризації на основі міри подібності та FCM на основі вибірки Iris

Методи кластеризації	**Iris UCI repository**					
	PC	**CE**	**SC**	**S**	**XB**	**DI**
Адаптивний робастний метод кластеризації на основі модифікованої міри подібності	0,0230	0,0219	-0,3007	0,0032	0,0065	0,9999
FCM	0,7411	0,2389	0,3112	0,0040	6,9945	0,0078

Метод FCM показує гіршу точність кластеризації (0,8011) порівняно з адаптивним робастним методом (0,8199). Це свідчить про те, що FCM має гірші результати в плані чіткої класифікації об’єктів в конкретні кластери. Вища точність у адаптивного робастного методу кластеризації на основі модифікованої міри подібності вказує на більш точну і стабільну

ідентифікацію кластерів.

Помилка класифікації (CE). Результати показують, що адаптивний метод має меншу помилку класифікації (0,2122), що є позитивним аспектом, оскільки це свідчить про менше число помилок при класифікації об'єктів. У свою чергу, FCM має більшу помилку класифікації (0,3410), що означає, що в деяких випадках класифікація була менш точною. Це може свідчити про більшу чутливість методу до шуми або неоднозначних ситуацій, що виникають у даних.

Адаптивний робастний метод має кращий коефіцієнт силуета (0,2567), що свідчить про більшу відокремленість кластерів. Вищий коефіцієнт силуета означає, що кластери, сформовані методом FCM, є більш чіткими й добре визначеними. У свою чергу, FCM має значення SC рівне 0,1345, що вказує на меншу відокремленість кластерів, і це може бути ознакою того, що метод не так добре визначає чіткі кордони між кластерами.

Метод FCM показує менше значення S (0,0030), що свідчить про кращу здатність методу до визначення чітких меж між різними групами. У той час як адаптивний метод має значення S на рівні 0,0022, що є ще меншим показником. Це може свідчити про те, що адаптивний метод є менш здатним до чіткого розділення різних кластерів, при цьому може бути більше перетину або скупчення точок між кластерами.

Індекс XB для FCM є досить високим (7,1965), що може вказувати на погану внутрішню згуртованість кластерів та більшу варіативність всередині самих кластерів. Для адаптивного методу цей індекс є набагато нижчим (0,0015), що свідчить про високу згуртованість і компактність кластерів, тобто вони більш однорідні. Тому адаптивний метод кластеризації здатний створювати більш компактні та стійкі до варіативності кластери, хоча й може мати деякі труднощі з їх чітким визначенням.

DI для адаптивного методу становить 1, що вказує на високу щільність кластерів, що є гарною ознакою для методу, оскільки кластери утворюються щільними й стійкими до можливих змін. Натомість FCM має значення DI рівне

0,0080, що вказує на менш щільні кластери і може бути ознакою того, що вони більш розріджені та менш стійкі.

Адаптивний робастний метод має кращі результати за точністю кластеризації та відокремленістю кластерів, однак він показує значні проблеми з внутрішньою згуртованістю та щільністю кластерів, що може призводити до високої варіативності та нестабільності результатів. У той час як FCM виявляється менш точним, але забезпечує кращу компактність та стабільність кластерів, що може бути корисно в умовах, коли необхідно досягти більшого узгодження в класифікації, навіть якщо відокремленість кластерів не є ідеальною.

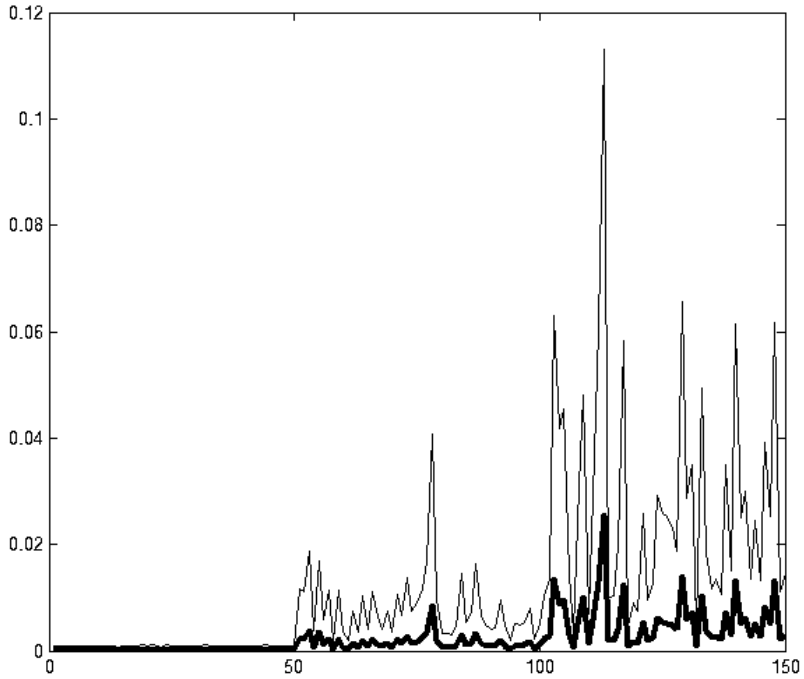


Рисунок 2.7 - Кластеризація даних, що не пошкоджені аномальними викидами, де суцільна лінія – рівень належності; жирна лінія є функцією модифікованої міри подібності.

На рисунку 2.7 продемонстрована робота методів кластеризації даних, що не пошкоджені аномальними викидами. Далі викиди були додані штучно. Рисунок 2.8 демонструє чутливість до викидів у наборі даних Іриси.

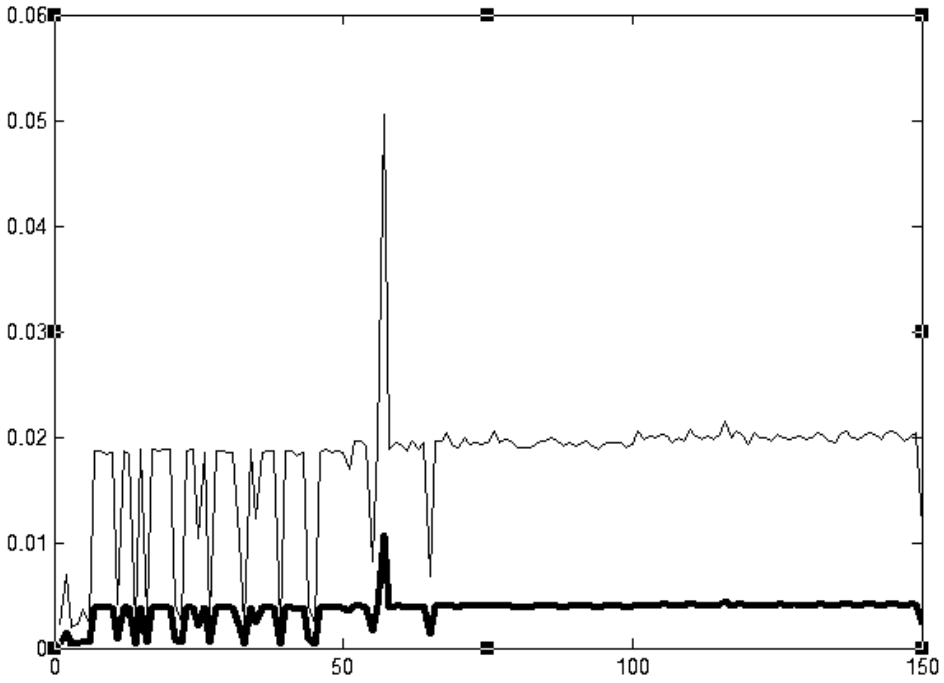


Рисунок 2.8 - Чутливість до викидів у наборі даних Iris, де суцільна лінія – рівень належності; жирна лінія є функцією модифікованої міри подібності.

На рисунку 2.9 показано роботу методу кластеризації даних, який не пошкоджено аномальними викидами та відсутніми спостереженнями, на рисунку 2.10 показано роботу методу кластеризації даних, який пошкоджено відсутніми спостереженнями, а на рисунку 2.11 показано роботу методу кластеризації даних що спотворено аномальними викидами та відсутніми спостереженнями.

Запропонований підхід базується на гібридизації імовірнісних, можливістних та правдоподібних процедур нечіткої кластеризації, робастних критеріїв оцінювання, мір подібності спеціального типу та самонавчання Т. Кохонена за принципом «Переможець отримує більше» для оптимізації цільової функції спеціального типу.

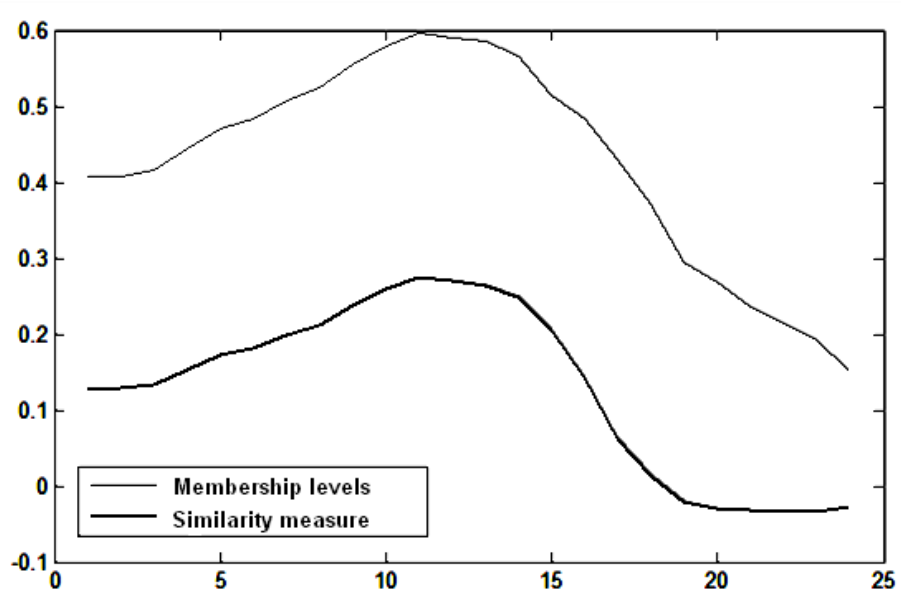


Рисунок 2.9 - Кластеризація даних, яка не пошкоджена аномальними викидами, де суцільна лінія – рівень членства; жирна лінія є функцією міри подібності

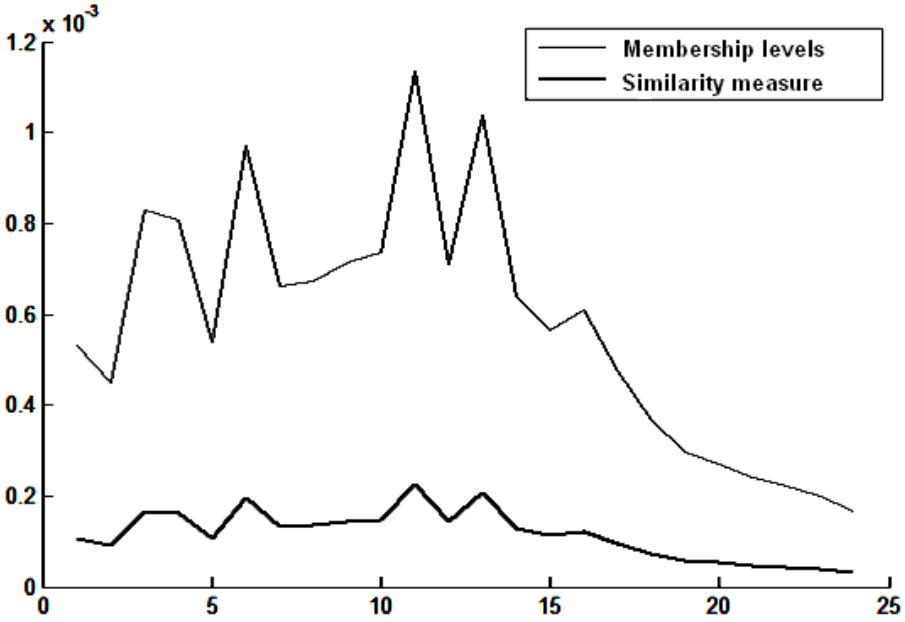


Рисунок 2.10 - Робота методу кластеризації даних, які спотворено відсутніми спостереженнями

Оптимізація цільової функції при обмеженнях за допомогою методу множників Лагранжа дозволила отримати аналітичні вирази для розрахунку

рівнів належності та рекурентних співвідношень для налаштування центроїдів сформованих кластерів.

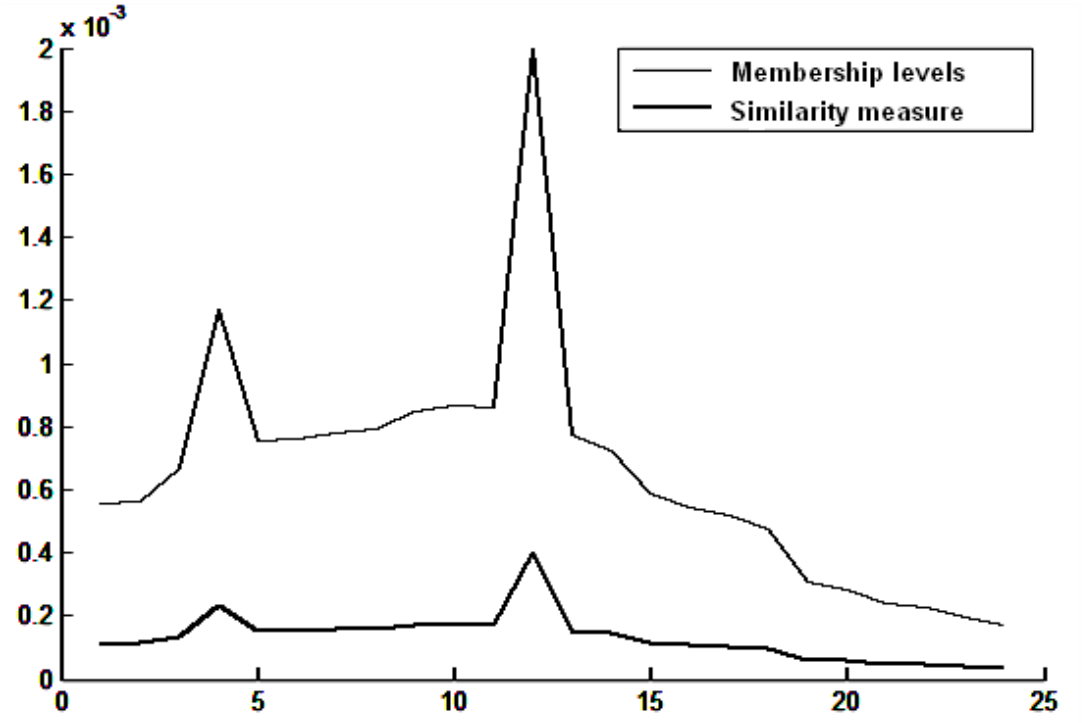


Рисунок 2.11 - Робота методу кластеризації даних, який спотворено аномальними викидами та відсутніми спостереженнями

Запропоновані методи є, по суті, процедурами градієнтної оптимізації. Показано, що запропоновані процедури є узагальненням відомих алгоритмів нечіткої кластеризації та збігаються з ними, якщо вихідні дані є «чистими» (без викидів та пропущених спостережень).

Проведені імітаційні експерименти підтверджують ефективність розробленого підходу.

2.13 Висновок до 2 розділу

1. Виділено основні характеристики потоків даних, які надходять в онлайн режимі та запропоновано постановку задачі розробки адаптивних методів нечіткої кластеризації потоків даних за умов перетинних класів та апріорної невизначеності.

2. 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.

3. 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.

4. Проведено експериментальні дослідження розроблених та модифікованих методів. В порівнянні з класичними методами кластеризації (*K*-means, FCM), розроблені адаптивні 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.

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

РОЗДІЛ 3
АДАПТИВНА НЕЧІТКА КЛАСТЕРИЗАЦІЯ ПОТОКІВ ДАНИХ З РІЗНОЮ ЩІЛЬНІСТЮ РОЗПОДІЛУ

Головну увагу у даному розділі приділено розробці методів адаптивної нечіткої кластеризації потоків даних з різною щільністю розподілу.

Задача кластеризації масивів спостережень довільної природи є невід'ємною частиною Data Mining, а у більш загальному випадку Data Science, а для її вирішення запропонована дуже велика кількість підходів , що відрізняються між собою як апріорними припущеннями що до фізичної природи даних та задачі, що вирішуються на їх основі, так і математичним апаратом, що використовується [1-29].

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 другому розділі було вирішено проблему розробки адаптивних методів нечіткої кластеризації потоків даних за умов перетинних класів та апріорної невизначеності. В порівнянні з класичними методами кластеризації (*K*-means, FCM), розроблені адаптивні 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.

Однак ефективність кластеризації значною мірою залежить від характеристик даних, зокрема їхньої щільності та варіативності розподілу. В задачах, де об’єкти розподілені нерівномірно, класичні алгоритми можуть давати некоректні результати через надмірне згладжування кластерних структур або їхню надмірну деталізацію.

Адаптивні методи кластеризації покликані вирішувати цю проблему, автоматично підлаштовуючись до локальних особливостей розподілу даних. Вони використовують механізми, що дозволяють динамічно змінювати параметри кластеризації залежно від щільності, розмірів та просторового розташування кластерів. Такий підхід забезпечує більшу точність групування та підвищує стійкість алгоритмів до шуму й аномальних значень.

Один із підходів до врахування щільності полягає у використанні методів кластеризації на основі щільності (Density-Based Clustering) [133] та його модифікації [134-171], 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 [141] та оцінок Надарая - Ватсона [142, 143]. До них належить DBSCAN (Density-Based Spatial Clustering of Applications with Noise), який визначає кластери як області з високою густиною точок, відокремлені зонами меншої щільності. Основною перевагою цього методу є здатність виявляти кластери довільної форми та відокремлювати шумові точки.

З обчислювальної точки зору задача кластеризації перетворюється у проблему пошуку локальних екстремумів багатоекстремальної функції

векторного аргументу щільності зо допомогою градієнтних процедур, які багатократно 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.

Для вирішення такої задачі потрібно:

– розробити методи адаптивної нечіткої кластеризація потоків даних з різною щільністю розподілу;

– провести імітаційне моделювання та експериментальні дослідження розроблених методів.

3.1 Передобробка потоків даних різної щільності для кластеризації

Передобробка даних є ключовим етапом у процесі кластеризації, особливо коли йдеться про потокові дані з неоднорідною щільністю. Потокові дані характеризуються високою швидкістю надходження, змінною структурою та наявністю шуму, що створює значні виклики для алгоритмів кластеризації. Окрім цього, нерівномірна щільність даних може суттєво впливати на якість групування, ускладнюючи виділення меж між кластерами та правильне визначення їхньої кількості.

Розробка ефективних методів передобробки потокових даних спрямована на нормалізацію щільності, усунення шумових об'єктів, масштабування та виявлення локальних структур у даних. Така передобробка дозволяє покращити результати кластеризації та забезпечити стабільність алгоритмів у режимі реального часу.

Проблеми передобробки потокових даних різної щільності:

1. Динамічна зміна щільності. Потокові дані можуть змінювати свою щільність у часі: певні регіони можуть бути густонаселеними, тоді як інші залишаються малозаповненими. Це створює труднощі для алгоритмів

кластеризації, які працюють із глобальними параметрами та можуть погано адаптуватися до локальних змін. У зв’язку з динамічною природою щільності потоків даних необхідно застосовувати спеціалізовані алгоритми, здатні адаптуватися до таких змін. Такі алгоритми повинні не лише ефективно обробляти великий обсяг даних, але й виявляти і аналізувати зміни в їхній щільності. Класичні методи кластеризації, такі як DBSCAN або DENCLUE, не завжди ефективно працюють з потоками даних, оскільки вони припускають статичний характер щільності. Для роботи з динамічними потоками використовуються адаптивні варіанти цих алгоритмів, здатні змінювати параметри кластеризації в реальному часі.

2. Наявність шуму та викидів. У потокових даних часто зустрічаються аномальні значення, які можуть зміщувати центри кластерів або призводити до неправильного розподілу об'єктів. Традиційні методи кластеризації, такі як k-means, чутливі до таких точок і можуть давати некоректні результати.

3. Різний масштаб і розмірність даних. У багатовимірних потоках даних часто спостерігається значна різниця в масштабах та розмірностях ознак, що ускладнює ефективне застосування алгоритмів кластеризації або аналізу. Щільність даних може варіюватися залежно від конкретних характеристик або ознак, і деякі з них можуть мати більшу варіативність, ніж інші. Наприклад, ознаки, що вимірюються у різних одиницях вимірювання (наприклад, вага в кілограмах і температура в градусах Цельсія), можуть мати суттєво різні масштаби, що призводить до домінування одних ознак над іншими під час аналізу. Тому важливо враховувати різницю в масштабах та впливати на алгоритм за допомогою нормалізації, щоб привести всі ознаки до однакової шкали і уникнути спотворення результатів кластеризації або інших методів.

4. Обмеженість ресурсів у потокових системах. Передобробка потоків даних повинна виконуватися в режимі реального часу, що обмежує можливості використання обчислювально складних методів. Це вимагає

застосування ефективних алгоритмів, які здатні швидко оновлювати модель без необхідності повторного аналізу всього масиву даних.

Один із підходів до передобробки потокових даних різної щільності полягає у використанні функції щільності розподілу даних. Вона дозволяє коригувати вагу об'єктів залежно від їхньої близькості до інших точок у багатовимірному просторі.

3.2 Формування функції щільності розподілу даних у масиві, що підлягає кластеризації

Формування функції щільності розподілу даних є важливим етапом у процесі кластеризації, оскільки дозволяє виявити внутрішню структуру даних, визначити області з високою концентрацією об’єктів і на їхній основі класифікувати дані в окремі кластери. Для того щоб сформувати таку функцію, необхідно оцінити щільність точок у різних частинах простору даних. Оцінка щільності є критично важливою для визначення кластерів, оскільки дозволяє виділяти області з високою щільністю точок, що, як правило, відповідають природним кластерам.

Зазвичай для цього використовуються різні методи оцінки щільності, серед яких одним із найбільш поширених є метод ядрового згладжування. Розглянемо покроковий процес формування функції щільності розподілу даних у масиві, що підлягає кластеризації.

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 (поліном) 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 по суті є локальними екстремумами функції щільності.

Функція впливу описує, як кожен елемент (точка) даних впливає на інші елементи в наборі, залежно від їхньої відстані або близькості. Це може бути, наприклад, функція, що визначає вагу впливу однієї точки на іншу в просторі даних. У методах, що базуються на щільності, таких як ядрове згладжування (KDE), функція впливу визначає, як розподіляється «вплив» однієї точки на навколишні точки, тобто як кожен об'єкт впливає на загальний розподіл щільності.

Функція впливу може бути побудована за допомогою ядра (наприклад, гаусівського ядра), яке присвоює більш високі ваги точкам, що знаходяться ближче до центральної точки, і зменшує їх із збільшенням відстані.

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

$$f_G^{x(\bullet)}(x) = \exp\left(-\frac{d^2(x, x(\bullet))}{2\sigma^2}\right) = \exp\left(-\frac{\|x - x(\bullet)\|^2}{2\sigma^2}\right) \qquad (3.1)$$

(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

$$f_G^{x(\bullet)}(x) = \exp\left(-\frac{d^2(x, x(\bullet))}{2\sigma^2}\right) = \exp\left(-\frac{Tr\left(x - x(\bullet)\right)\left(x - x(\bullet)\right)^T}{2\sigma^2}\right), \qquad (3.2)$$

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

Нескладно бачити, що (3.2) є узагальненням (3.1).

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

$$f^x(x) = \sum_{k=1}^{N} f\left(x, x(k)\right), \tag{3.3}$$

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 (3.3) за допомогою градієнтної процедури

$$x^l = x^{l-1} + \eta^l \frac{\nabla f^x\left(x^l, x^{l-1}\right)}{\left\|\nabla f^x\left(x^l, x^{l-1}\right)\right\|}, \;\; x_0 = x\left(k\right), l = 0,1,2,...; \forall k = 1,2,...,N, \tag{3.4}$$

тобто кількість запусків алгоритму (3.4) 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 (3.3) шляхом модифікації градієнтної процедури (3.4) [138-149].

Коли спостереження $x(k)$ у вибірці $X \in (n_1 \times n_2)$ - матриця, нескладно ввести у розгляд матричний варіант процедури (3.4):

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

де $\Gamma^x \left(x, x^{l-1}\right) = \left\{ \dfrac{\partial f^x \left(x, x^{l-1}\right)}{\partial x_{i_1 i_2}} \right\} \in R^{n_1 \times n_2}$.

Процес градієнтної оптимізації закінчується відшуканням m локальних екстремумів функції (3.3), 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 повинен бути модифікований для вирішення задачі кластеризації.

3.3 Нечітка модифікація методу піків щільності

Нечітка кластеризація на основі аналізу піків щільності розподілу є потужним методом, який дозволяє виявляти структуру даних без необхідності визначати чіткі кордони між кластерами. Цей підхід орієнтований на аналіз локальних піків у розподілі щільності даних, що надає можливість розпізнавати кластери на основі їх щільності, а не геометричних або топологічних характеристик. Основною перевагою цього методу є здатність виявляти класифікацію в умовах неоднорідності та складної структури даних,

де класичні методи кластеризації, такі як k-середні, можуть бути неефективними.

Замість того щоб намагатися чітко визначити межі між кластерами, методи нечіткої кластеризації на основі щільності фокусуються на виявленні інтенсивних зон або піків у щільності, що зазвичай вказують на скупчення даних, які можуть бути інтерпретовані як кластери. У контексті цього підходу кожен елемент даних оцінюється за рівнем щільності навколо нього, що дозволяє ідентифікувати області з високою концентрацією об'єктів, які ймовірно належать до одного кластера. З точки зору теоретичного обґрунтування, така кластеризація враховує природні характеристики даних і дозволяє отримати більш гнучкі та реалістичні результати, які часто краще відображають структуру даних у реальних умовах.

Аналіз піків щільності вимагає від алгоритмів здатності адаптуватися до локальних особливостей розподілу даних. Це означає, що в межах одного набору даних можуть бути виявлені множинні області з різними рівнями щільності, що дозволяє отримувати кластеризації, де кількість кластерів не є фіксованою, а варіюється залежно від змін у розподілі даних. У традиційних методах кластеризації, таких як алгоритм k-середніх, кількість кластерів часто потрібно задавати наперед, що може призвести до неправильних результатів у разі складних або змінних структур даних. Нечітка кластеризація на основі щільності дозволяє уникнути цієї проблеми, оскільки не вимагає попереднього визначення кількості кластерів, а визначає їх лише на основі властивостей самих даних.

Процес виявлення піків у щільності зазвичай передбачає використання різноманітних математичних та статистичних методів, таких як методи ядрового згладжування або класифікації через класифікаційні функції щільності. Ці методи допомагають виявляти найбільш виразні й значущі скупчення, навіть якщо вони не є ідеально ізольованими чи мають неправильну форму. Завдяки такій гнучкості, алгоритми нечіткої кластеризації можуть працювати з даними різних типів, включаючи неструктуровані, неявно

організовані набори даних, де звичні методи виявлення кластерів можуть бути неефективними.

Важливо зазначити, що нечітка кластеризація на основі щільності також володіє певними обмеженнями та викликами. Наприклад, високий рівень чутливості до параметрів алгоритму, таких як вибір порогу для виявлення піків щільності або кількості точок, що мають бути включені до складу кожного кластеру, може призвести до неточних або нестабільних результатів. Крім того, необхідність враховувати масштаби та рівні щільності для кожної пари точок може збільшити обчислювальні витрати, що може бути особливо важливим при обробці великих наборів даних.

Відтак, застосування нечіткої кластеризації на основі аналізу піків щільності розподілу відкриває нові можливості для класифікації складних структурованих і неструктурованих даних, дозволяючи значно покращити точність кластеризації у випадках, коли традиційні методи не дають задовільних результатів. Її здатність адаптуватися до різноманітних типів даних та виявляти природні структури без необхідності чітко визначати кордони між кластерами робить цей підхід важливим інструментом у сучасних задачах класифікації та аналізу 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, $-1 \le x_i(k) \le 1$, $x_i(k), i = 1, 2, ..., n$.

В процесі кластеризації на основі аналізу піків щільності розподілу даних аналізується два параметри: ρ_k - локальна щільність і δ_k - відстань до точки з більш високою щільністю. Окрім того вводиться єдиний вільний

параметр d_c - відстань зрізу, яка задається і варіюється користувачем для отримання необхідної точності рішення задачі.

Роботу методу можна сформулювати як наступну послідовність 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 $\rho = \{\rho_k\} \in R^N$:

$$\rho_k = \sum_{k=1}^{N} \chi(d_{kl} - d_c),$$

де

$$\chi(d) = \begin{cases} 1, \text{ якщо } d < 0, \\ 0, \text{ в іншому випадку.} \end{cases}$$

Відстань зрізу обирається з суто емпіричних міркувань, при цьому автори методу [145, 154, 156-165] рекомендують вибирати його так, щоб в околі, який формується, потрапляло $0,01N - 0,02N$ спостережень вибірки, що оброблюється.

Крок 3. Розрахунок вектора мінімальних відстаней $\delta = \{\delta_k\} \in P^N$ до точок з більш високою щільністю

$$\delta_k = \min_{\forall l, \rho_l > \rho_k} \{d_{kl}\},$$

а для точки з максимальною щільністю δ_k^* розраховується:

$$\delta_k^* = \max_l \{d_{kl}\}.$$

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

$$\min(d_{kl}) \equiv d_{ql}.$$

Зауважимо також, що в [123, 124] в якості центроїдів пропонується використовувати значення $c_q = x(k)$ з максимальним значенням добутків $\rho_k \cdot \delta_k$.

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 [1, 2, 8, 11, 12].

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 $x(k) \neq c_q$ розраховуються рівні нечіткої належності в стандартній формі [2]

$$\mu_q(k) = \frac{d_{qk}^{-2}}{\sum_{l=1}^{m} d_{lk}^{-2}}, \tag{3.4}$$

або на основі функції щільності розподілу Коші [121, 122, 146, 166]

$$\mu_q(k) = \left(1 + \frac{d_{qk}^{-2}}{\sigma_q^2}\right), \tag{3.5}$$

де

$$\sigma_q^2 = \left(\sum_{\substack{l=1 \\ l \neq k}}^{m} d_{lk}^{-2}\right)^{-1}.$$

Крок 6. На основі оцінок імовірнісної нечіткої належності (3.4), (3.5) розраховується рівень довіри до отриманих результатів на основі стандартного правдоподібного підходу [97-107]

$$Cred_q(k) = \frac{1}{2}\left(\mu_q^*(k) + 1 - \sup \mu_q^*(k)\right), \tag{3.6}$$

де

$$\mu_q^*(k) = \frac{\mu_q(k)}{\sup \mu_l(k)}.$$

Крок 7. Завершення процедури нечіткої кластеризації шляхом оцінки якості результатів за допомогою будь-якого з критеріїв, що застосовуються в нечіткій кластеризації [3, 5-44], хоча оцінка (3.6) вже сама по собі надає наскільки можна довіряти правдоподібності отриманих результатів.

3.4 Нечітка правдоподібна кластеризація даних на основі аналізу щільності розподілу даних та їх піків

Нечітка правдоподібна кластеризація даних, основана на аналізі щільності розподілу даних та їхніх піків, є інноваційним підходом, який поєднує в собі переваги нечіткої логіки та статистичних методів для виявлення складних структур у даних. Цей метод дозволяє вирішувати проблеми, пов'язані з класифікацією даних, у випадках, коли традиційні алгоритми кластеризації не можуть точно відобразити реальну структуру даних, зокрема у випадку, коли дані мають неявно виражену структуру або невизначену кількість кластерів.

Основним принципом нечіткої правдоподібної кластеризації є обробка даних на основі їхньої щільності та пошук локальних піків у цьому розподілі. Піки щільності вказують на зони з високою концентрацією точок даних, що може вказувати на наявність кластера. Однак, на відміну від класичних методів кластеризації, таких як k-середніх або ієрархічних методів, цей підхід не передбачає чіткої границі між кластерами, а дозволяє класифікувати об'єкти на основі їхньої належності до кожного з можливих кластерів за допомогою нечітких значень.

Нечітка логіка є ключовою складовою цього підходу, оскільки вона дозволяє моделювати невизначеність і нечіткість в процесі класифікації. У традиційних методах кластеризації кожен елемент або точно належить до одного кластера, або не належить взагалі. Однак у реальних ситуаціях часто буває, що елементи мають часткову належність до кількох кластерів, що є характерним для нечіткої класифікації. Враховуючи це, метод нечіткої правдоподібної кластеризації може застосовувати нечіткі множини для представлення ступеня належності кожної точки до кількох кластерів, що значно покращує точність класифікації в складних або двозначних випадках.

Аналіз щільності розподілу даних дозволяє виявляти пік щільності, який представляє собою локальні максимуми у розподілі даних. Ці максимуми

можуть вказувати на наявність груп об'єктів, що тісно пов'язані між собою. Однак для точної ідентифікації таких пік важливо враховувати масштаби та розподіли щільності, оскільки дані можуть мати різні рівні інтенсивності та можуть бути дуже розрідженими або навпаки, густо населеними в певних областях простору.

Процес кластеризації включає в себе кілька етапів. Першим етапом є оцінка щільності точок у різних частинах простору, що дозволяє виділити області з високою щільністю. Далі, для кожного елемента даних визначається ступінь його належності до різних кластерів на основі аналізу цих піків щільності. Це дозволяє врахувати всі можливі варіанти належності точки до кількох кластерів, а також дозволяє ігнорувати точки, які не входять до жодного значного кластеру, або знаходяться у зонах низької щільності.

Застосування правдоподібних методів у цьому контексті дозволяє підвищити стійкість кластеризації до шуму та інших факторів невизначеності в даних. Вони можуть включати ймовірнісні моделі, які дають змогу оцінити ступінь правдоподібності для кожної точки належати до конкретного кластера, зважаючи на ймовірність її знаходження в певній області щільності. Це дозволяє отримати більш надійні результати кластеризації, де кожна точка даних має своє "правдоподібне" представлення у кластері, а не просто жорстку належність.

Проте, на практиці методи нечіткої правдоподібної кластеризації мають деякі складнощі. Однією з основних проблем є вибір параметрів, зокрема параметрів, що контролюють чутливість до піків щільності та масштаби кластерів. Некоректний вибір цих параметрів може призвести до неточних результатів, що, у свою чергу, потребує постійної адаптації алгоритмів до конкретних наборів даних.

Таким чином, нечітка правдоподібна кластеризація на основі аналізу щільності та піків розподілу є перспективним напрямом у класифікації складних даних, дозволяючи отримувати більш гнучкі та адаптивні рішення в умовах невизначеності і складних структур.

Цей метод особливо корисний у випадках, коли дані не відповідають умовам класичних методів кластеризації і потребують більш гнучкої, адаптивної оцінки кластерів.

Для будь якої точки $\tilde{x}$ з масиву X її базова функція впливу $f_B^{\tilde{x}}(x) = f(x, \tilde{x})$ є деякою ядерною дзвонуватою функцією, серед яких автори методу [154, 156-165] відзначають, так звану, прямокутну хвильову функцію впливу (Рисунок 3.1).

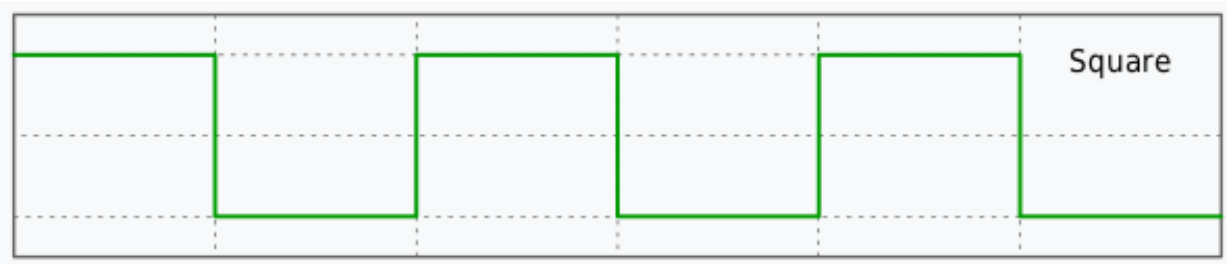


Рисунок 3.1 – Прямокутна хвильова функція

Прямокутна хвиля – це несинусоїдальна періодична форма хвилі, амплітуда якої змінюється на постійній частоті між фіксованими мінімальним і максимальним значеннями, з однаковою тривалістю в мінімальному та максимальному значеннях. В ідеальній квадратній хвилі переходи між мінімумом і максимумом є миттєвими.

Прямокутна хвиля є окремим випадком пульсової хвилі , яка допускає довільну тривалість при мінімальній і максимальній амплітудах. Відношення високого періоду до загального періоду пульсової хвилі називається робочим циклом. Справжня прямокутна хвиля має 50% робочий цикл (рівні високий і низький періоди).

$$f_S^{\tilde{x}}(x) = \begin{cases} 0, \text{ якщо } d(x, \tilde{x}) > \sigma, \\ 1 \text{ інакше} \end{cases}$$

(тут d - відстань у прийнятій метриці, зазвичай евклідовій, σ - параметр ширини - відстань зрізу в прийнятій метриці функції впливу) та гаусівську функцію впливу (Рисунок 3.2)

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

що є найбільш популярною, завдяки зручності обчислення її градієнта.

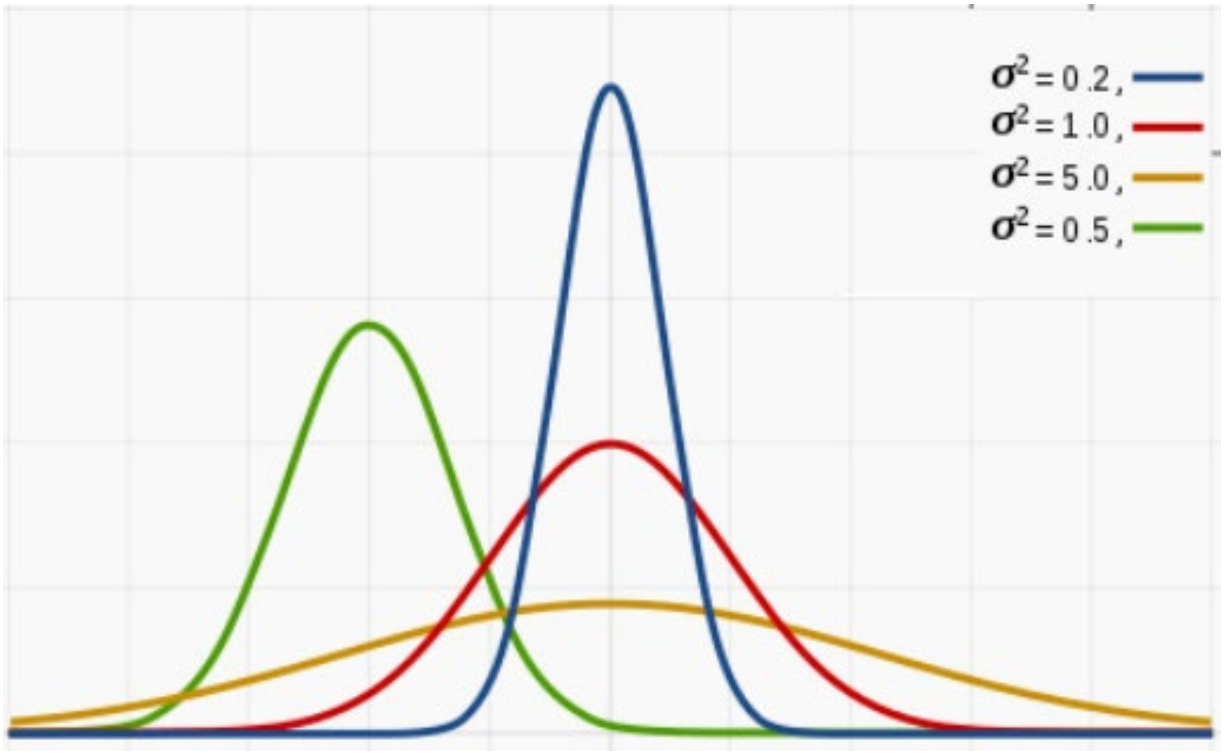


Рисунок 3.2 - Гаусівська функція з різною щільністю

Нескладно бачити, що в якості функцій впливу можуть також бути використана функція Коші, що часто виникає у задачах нечіткої кластеризації [166]

$$f_C^{\tilde{x}}(x) = \left(1+\frac{d^2(x,\tilde{x})}{\sigma^2}\right)^{-1} = \left(1+\frac{\|x-\tilde{x}\|^2}{\sigma^2}\right)^{-1},$$

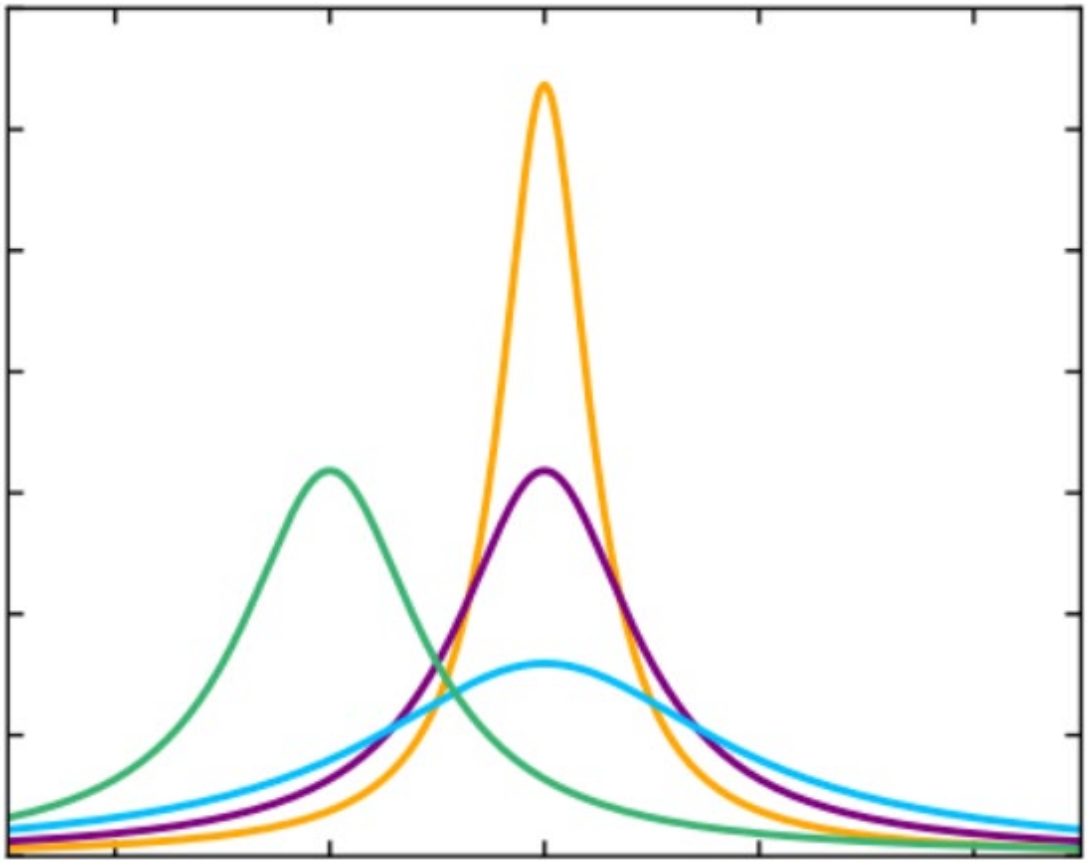


Рисунок 3.3 – Розподіл функції Коші

та функція Єпанечнікова [179]

$$f_E^{\tilde{x}}(x) = \left[1 - \frac{d^2(x,\tilde{x})}{2\sigma^2}\right]_+ = \left[1 - \frac{\|x-\tilde{x}\|^2}{2\sigma^2}\right]_+$$

(тут $[\bullet]_+ = \max\{0,\bullet\}$), цей градієнт має просту форму

$$\nabla_x f_E^{\tilde{x}}(x) = \left[\frac{\tilde{x}-x}{\sigma^2}\right]_+,$$

де операція проектування не достатній ортант $[\bullet]_+$ реалізується покомпонентно.

Функція щільності розподілу даних у масиві X, що містить N спостережень, формується на основі N функцій впливу у вигляді

$$f^x(x) = \sum_{k=1}^{N} f\left(x, x(k)\right) \tag{3.7}$$

і є близькою за суттю до Парзенівських вікон та оцінок Надарая-Ватсона.

Власне процедура кластеризації полягає у відшуканні максимумів функції $f^x(x)$, що задовольняють умові

$$f^x\left(x, x^*\right) > \xi, \tag{3.8}$$

де ξ - деякий поріг, що визначає, який із відшуканих атракторів є значущим, тобто «фільтрує» окремі аномальні викиди у вибірці X та виключає із розгляду «міні-кластери», що містять занадто мало спостережень. Зрозуміло, що чим більше значення ξ, тим менша кількість значущих кластерів буде сформована.

Для відшукання атракторів-екстремумів-максимумів функції щільності розподілу даних $f^x(x)$ зазвичай використовується градієнтна процедура оптимізації (hill climbing) [169], що може бути записана у вигляді

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

де η^l - параметр кроку пошуку, що визначає швидкість збіжності алгоритму. Різні модифікації DENCLUE пов’язані саме з намаганнями пришвидшити процес оптимізації прийнятої функції щільності [146, 166, 170].

Тут же слід відзначити, що використання гаусіанів в якості функцій впливу пов'язане з простою формою їх градієнтів, оскільки

$$f_G^x\left(x,x^l\right)=\sum_{k=1}^{N}\left(x^l-x\right)f_G^{x^l}\left(x\right)=\sum_{k=1}^{N}\left(x^l-x\right)\exp\left(-\frac{\left\|x-x^l\right\|^2}{2\sigma^2}\right).$$

Помітимо також, що функції Єпанечнікова мають ще більш просту форму градієнта

$$\nabla f_E^x\left(x,x^l\right)=\sum_{k=1}^{N}\left[\frac{x^l-x}{2\sigma^2}\right]_+.$$

Процес оптимізації починається з кожної точки $x(k)$ масиву даних X і закінчується відшуканням всіх екстремумів-максимумів функції щільності (3.7), що задовольняють нерівності (3.8).

Зрозуміло, що чим більше обсяг вибірки X, тим більше разів N повинна запускатися процедура оптимізації - пошуку атракторів. Пришвидшити процес кластеризації можна, скоротивши кількість запусків цієї процедури.

Тому пропонується починати процес пошуку атракторів не з кожної точки масиву даних X, а з, так званих, піків щільності [66] цього масиву.

Для знаходження цих піків у розгляд вводиться два параметри: $\rho(k)$ - локальна щільність та $\delta(k)$ - відстань від спостереження $x(k)$ до точки з більш високою щільністю. Крім того, аналогічно DENCLUE використовується відстань зрізу (cutoff distance) σ, що зазвичай задається та варіюється користувачем для отримання потрібної точності вирішення задачі.

Процес пошуку піків щільності починається з того, що на основі вихідної $(n \times N)$ - матриці «об'єкт-властивість» формується $(N \times N)$ матриця відстаней між спостереженнями

$$D = \left\{ d\left(x(k), x(q)\right) \right\}, \; d\left(x(k), x(q)\right) = \left\| x(k) - x(q) \right\| \forall k, q.$$

На основі цієї матриці формується $(N \times 1)$ - вектор локальних щільностей $\rho = \left\{ \rho(k) \right\} \in R^N$:

$$\rho(k) = \sum_{k=1}^{N} \chi\left(d\left(x(k), x(q)\right) - \sigma \right),$$

де

$$\chi(d) = \begin{cases} 1, \text{ if } d < 0, \\ 0, \text{ else.} \end{cases}$$

Тут відстань зрізу σ є найбільш впливовим параметром, що визначає якість кластеризації та обирається з суто емпіричних міркувань. Тут слід відмітити, що автори пікового алгоритму [145-165] радять обирати цю відстань так, щоб вона «накривала» $0,01N - 0,02N$ спостережень з масиву, що аналізується.

Після цього розраховується вектор мінімальних відстаней

$$\delta(k) = \min_{\forall q, \rho(q) > \rho(k)} \left\{ d\left(x(k), x(q)\right) \right\},$$

а для спостереження з мінімальною щільністю $\delta^*(k)$ покладається

$$\delta^*(k) = \max \left\{ d\left(x(k), x(q)\right) \right\}.$$

На базі цієї інформації формуються піки-центроїди кластерів $x_q^P, q = 1,2,...,m$, при цьому в якості цих піків-центроїдів обираються спостереження з найбільш високою щільністю, тобто центроїди згідно з цим підходом є деякі із спостережень вихідної вибірки. В той же час в ситуаціях, коли кластери мають досить складну форму, центроїд може не співпадати з жодною із точок $x(k)$. Тому пропонується після знаходження всіх піків $x_q^P, q = 1,2,...,m$ запускати процедуру оптимізації не з точок $x(k), k = 1,2,...,N$, а тільки з піків $x_q^P, q = 1,2,...,m$, кількість яких є значно меншою ніж обсяг вибірки X, тобто

$$m << N.$$

У сучасній кластеризації даних одним з основних напрямків є розробка та використання алгоритмів для виявлення груп або кластерів, що можуть мати різну складність та структуру в залежності від природи самих даних. Традиційні алгоритми кластеризації, такі як DENCLUE та пікова кластеризація, відносяться до чітких методів, що припускають, що кластери є чітко відокремленими один від одного в просторі ознак. Ці алгоритми, зокрема, застосовують концепцію, що класи не перетинаються і не можуть належати одночасно кільком кластерам. Проте на практиці часто виникають ситуації, коли дані не підкоряються таким чітким припущенням, і кластери можуть частково перекриватися, створюючи складнішу структуру, що потребує використання нечітких (фаззі) методів кластеризації [1-95], що базуються на двох основних підходах: імовірнісному та можливісному. Кожен з цих підходів має свої переваги та недоліки, яких позбавлений довірчий підхід до нечіткої кластеризації [3-44].

Довірчий підхід до нечіткої кластеризації, що базується на аналізі ймовірності і можливості належності точок до кластерів, є більш стійким до перехресних кластерів і більш адаптивним до реальних даних. На відміну від

класичних імовірнісних або можливісних підходів, цей метод не передбачає жорстких припущень про точне розташування меж кластерів або точок, що належать до декількох кластерів одночасно. Він дозволяє враховувати ймовірнісні моделі з фокусом на достовірності належності до кластерів, забезпечуючи більш точну і надійну кластеризацію при складних умовах.

Довірчий підхід поєднує у собі переваги імовірнісних та можливісних підходів, мінімізуючи їхні недоліки, і є надзвичайно корисним для аналізу та класифікації складних, багатовимірних даних, де традиційні чіткі алгоритми не можуть впоратися із завданням через високий рівень перекриття між кластерами або неоднозначність належності точок.

Згідно з цим підходом для кожного з атракторів (або піків) $x_q^*, q = 1,2,...,m$ та спостережень $x(k), k = 1,2,...,N$ розраховуються рівні нечіткої належності або у загальній формі [1-44]

$$\mu_q(k) = \frac{d^{-2}\left(x_q^*, x(k)\right)}{\sum_{r=1}^{m} d^{-2}\left(x_q^*, x(k)\right)} = \frac{\left\|x_q^* - x(k)\right\|^{-2}}{\sum_{r=1}^{m} \left\|x_q^* - x(k)\right\|^{-2}}, \tag{3.9}$$

або після деяких елементарних перетворень

$$\mu_q(k) = \left(1 + \frac{d^{-2}\left(x_q^*, x(k)\right)}{\sigma_q^2}\right), \tag{3.10}$$

де

$$\sigma_q^2 = \left(\sum_{\substack{r=1 \\ r \neq j}}^{m} d^{-2}\left(x_q^*, x(k)\right)\right)^{-1},$$

тобто виникає функція щільності розподілу Коші, що може бути використана

в якості функції впливу у DENCLUE.

Оцінки (3.9), (3.10) пов'язані з так званою, ймовірнісною нечіткою кластеризацією. На основі оцінок можуть бути розраховані рівні довіри отриманих результатів за допомогою співвідношень [96-106]:

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

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.

3.5 Апробація адаптивного методу швидкої нечіткої правдоподібної кластеризації на основі аналізу піків щільності розподілу

Експериментальні дослідження методу швидкої нечіткої правдоподібної кластеризації на основі аналізу піків щільності розподілу (FCDP) даних був реалізований на трьох масивах даних (Таблиця 3.1).

Таблиця 3.1 - Зразки тестових вибірок

Назва вибірки	**Кількість спостережень**	**Кількість атрибутів**
Іриси	150	4
Вина	178	13
Ecoli	336	8

Порівняльний аналіз проведено з відомими методами кластеризації які використовують параметр піків щільності розподілу даних, а саме:

– *k*-середніх;

– DBSCAN, який для заданої множини точок у деякому просторі відносить в одну групу точки, які розташовані найбільш щільно та розмічає точки, які лежать в областях з невеликою щільністю;

– DENCLUE, в якому кластери визначаються локальними максимумами оцінки щільності;

– OPTICS - алгоритм знаходження щільності на основі кластерів у просторових даних, що вирішує проблему визначення значущих кластерів в наборах даних різної щільності. За допомогою цих алгоритмів було проведено аналіз якості кластеризації на основі цих вибірок.

Заздалегідь, для порівняльного аналізу із вибірок бралась частина спостережень і проводився аналіз якості кластеризації даних, яка виміряна показником нормалізованої взаємної інформації (приймає значення 1, якщо ідеальна кластеризація даних знайдена).

За результатами аналізу кластеризації була отримана інформація. Для кожної з вибірок перевірили, як розмір вибірки впливає на якість кластеризації.

Порівняльний аналіз запропонованого методу продемонстровано в Таблиці 3.2, в якій наведені значення показника нормалізованої взаємної інформації для різних даних та методів, перше число у трьох правих стовпцях показує розмір вибірки.

Таблиця 3.2 - Значення показника нормалізованої взаємної інформації для різних даних та методів

Назва вибірки		**k-means**	**FCDP**	**DBSCAN**	**DENCLUE**	**OPTICS**
Іриси	0,8	0,67±0,06	0,79±0,03	0,68±0,06	0,67±0,06	0,78±0,02
	0,4	0,65±0,07	0,68±0,18	0,60±0,06	0,67±0,06	0,72±0,08
	0,2	0,64±0,07	0,64±0,10	0,54±0,04	0,54±0,07	0,64±0,06
Вина	0,8	0,70±0,11	0,78±0,02	0,66±0,01	0,68±0,01	0,76±0,04
	0,4	0,70±0,05	0,78±0,04	0,62±0,00	0,72±0,00	0,72±0,08
	0,2	0,58±0,21	0,69±0,11	0,48±0,01	0,70±0,01	0,59±0,01
Ecoli	0,8	0,65±0,02	0,77±0,09	0,75±0,07	0,75±0,11	0,75±0,05
	0,4	0,65±0,04	0,75±0,05	0,63±0,01	0,73±0,05	0,73±0,07
	0,2	0,65±0,03	0,70±0,10	0,55±0,01	0,66±0,21	0,68±0,11

Аналізуючи Таблицю 3.2, можна зробити висновки, що якість кластеризації даних не втрачається від кількості наявних спостережень у вибірці, тобто, незалежно від 20%, 40% або 80% наявності вибірки, якість кластеризації не зменшується.

Як видно із порівняльної Таблиці 3.2, показник нормалізованої взаємної інформації (NMI), що приймає значення 1, якщо ідеальна кластеризація даних знайдена серед всіх запропонованих методів кластеризації найкращій результат демонструє, коли даних все-ж таки більше.

На рисунку 3.4 продемонстрована залежність нормалізованої взаємної інформації (NMI) від розміру навчальної вибірки, що дає змогу говорити про те, що розмір вибірки не впливає на якість кластеризації, а NMI не є лінійним.

Таким чином, якість кластеризації не втрачається навіть при 20% наявності вибірки.

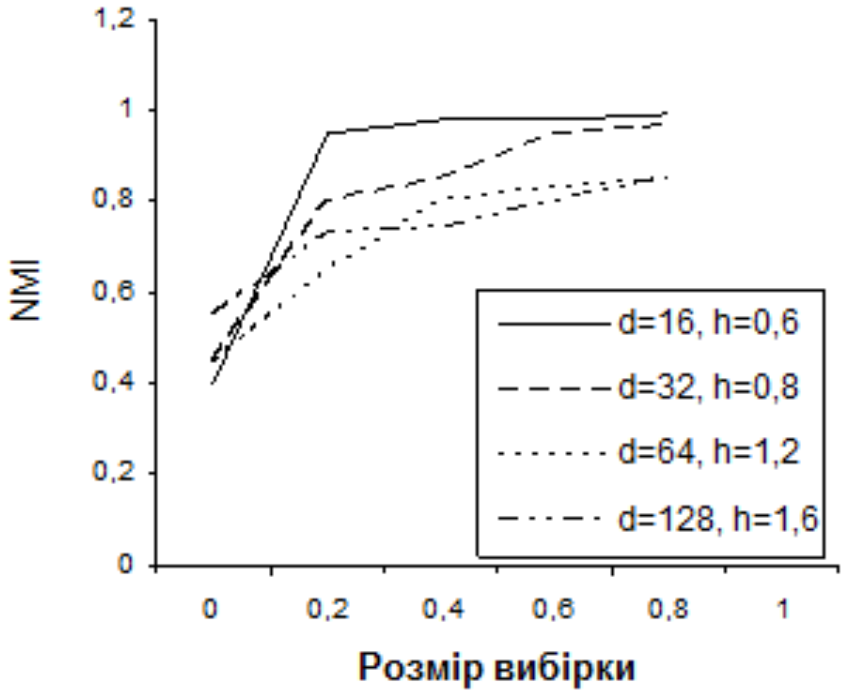


Рисунок 3.4 - Залежність показника нормалізованої взаємної інформації (NMI) від розміру навчальної вибірки

Якщо порівнювати якість кластеризації даних з відомими методами, можна зробити висновок, що запропонований метод швидкої нечіткої правдоподібної кластеризації на основі аналізу піків щільності розподілу даних (FCDP) демонструє значно вищі показники ніж *k*-means, DBSCAN і DENCLUE та майже однаково з методом OPTICS. Так, якщо більш детально проаналізувати результат роботи цих двох методів по кількості спостережень, показник нормалізованої взаємної інформації у методі FCDP вищій за OPTICS незалежно від виду вибірки, що подається на кластеризацію.

За результатами експериментальних досліджень та аналізу отриманих результатів, можна зробити висновок, що запропонований метод швидкої нечіткої правдоподібної кластеризації на основі аналізу піків щільності розподілу даних порівняно з методами, заснованими на використанні піків щільності, демонструє гарні результати роботи.

3.6 Апробація методу нечіткої правдоподібної кластеризації даних на основі аналізу щільності розподілу даних та їх піків

Дослідження методу нечіткої правдоподібної кластеризації даних на основі аналізу щільності розподілу даних та їх піків (NCrCP) проводились на двох навчальних вибірках UCI репозиторію Page Blocks та Spambase. В Таблиці 3.3 продемонстровані основні характеристики наборів даних.

Таблиця 3.3 - Зразки даних

Назва вибірки	**Кількість спостережень**	**Кількість атрибутів**	**Кількість кластерів**
Блоки сторінок	5472	10	5
База спаму	4601	57	2

На рисунках 3.5 та 3.6 продемонстровані набори вибірок даних, що аналізуються.

Робота запропонованого методу перевірялась за допомогою декількох класичних показників якості кластеризації, а саме індекс Дана (DI), індекс Девіса-Болдіна (DBI) та кластерна точність (CA).

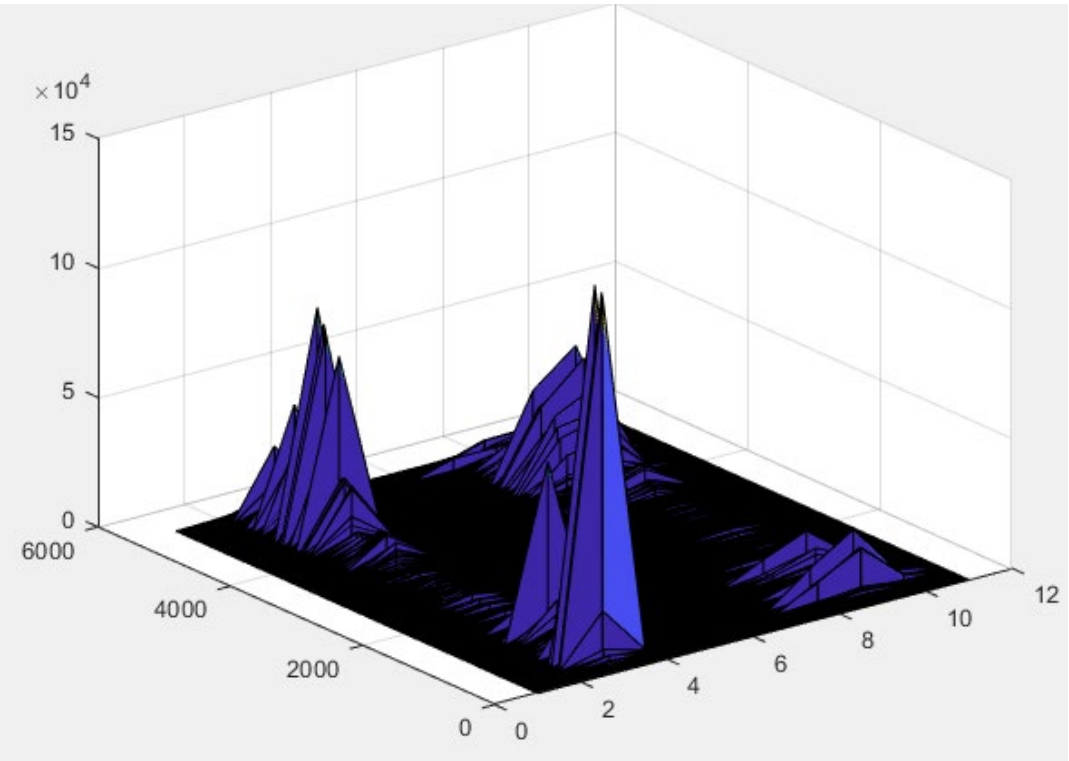


Рисунок 3.5 - Навчальна вибірка Page Blocks

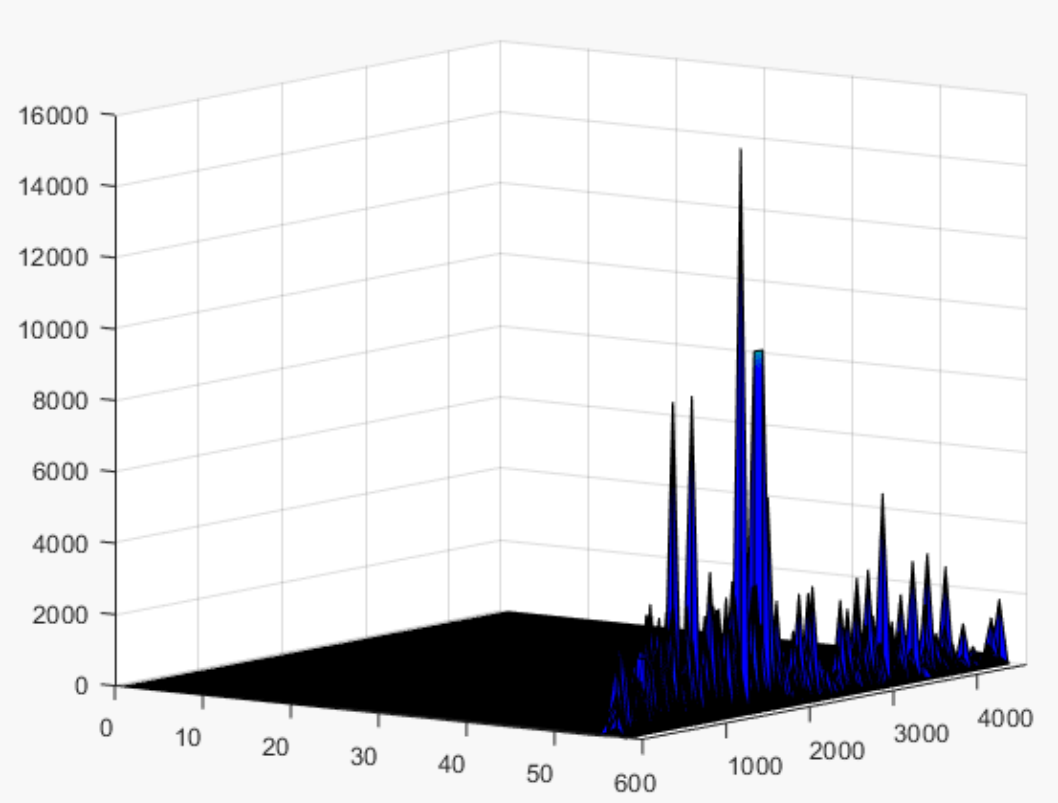


Рисунок 3.7 - Навчальна вибірка Spambase

Індекс Дана (DI) - цей індекс оцінює ступінь поділу між

спостереженнями одного кластера, тобто внутрішню схожість спостережень в кластері. Чим вище значення, тим краща кластеризація.

Індекс Девіса-Болдіна (DBI) - цей індекс, як DI, також оцінює ступінь поділу між кластерами (межкластерна несхожість), найменше значення вказує на кращу кластеризацію.

Кластерна точність (CA) - вимірює відсоток правильно класифікованих об’єктів у кластері на основі попередньо визначених міток класів. Цей індекс не працює з немаркованою базою даних, високе значення вказує на найкращу якість кластеризації.

Порівняльний аналіз проводився з більш відомими методами кластеризації даних, такими як DENCLUE-SA (імітований відпал), DENCLUE та DENCLUE-GA (генетичний алгоритм).

Результати кластеризації тестових даних різними методами кластеризації представлено на рисунках 3.8 і 3.9, які демонструють якісні характеристики кластеризації.

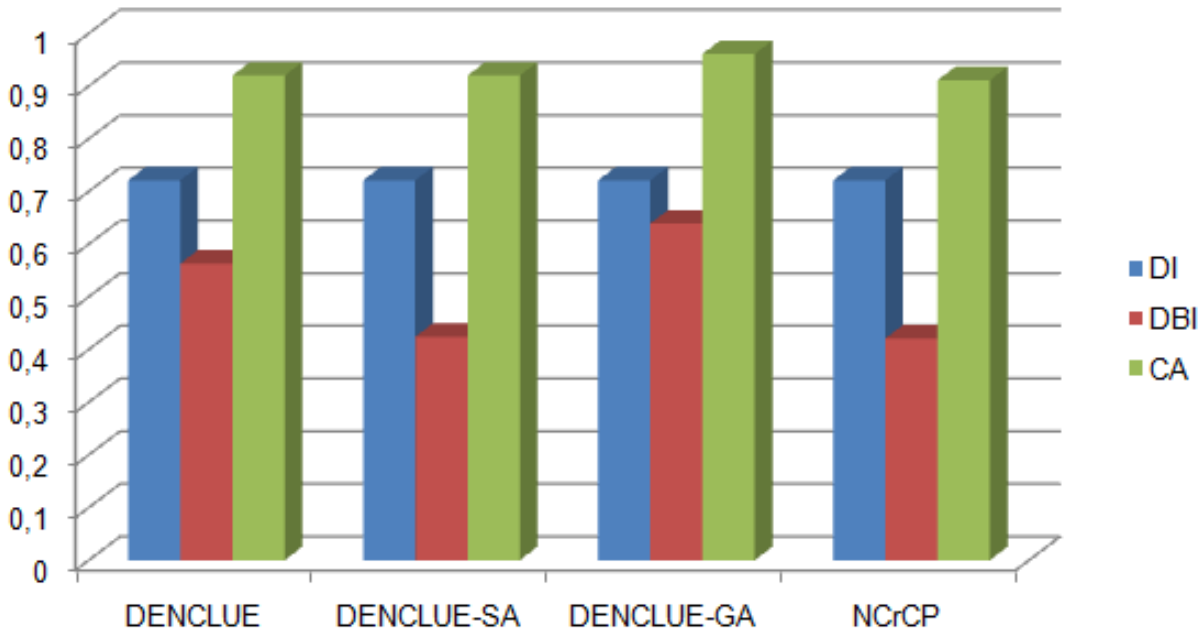


Рисунок 3.8 - Показники якості кластеризації Блоки сторінок

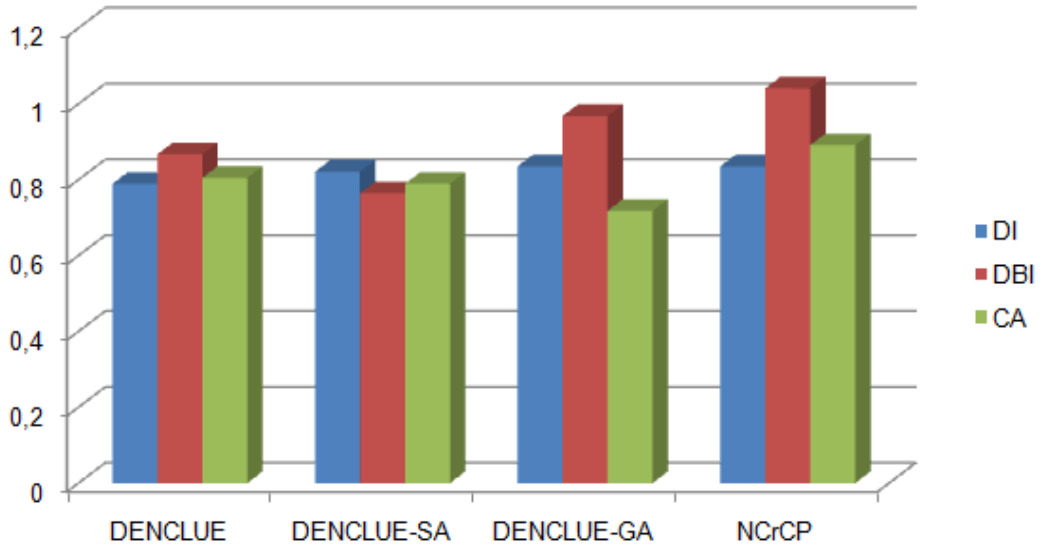


Рисунок 3.9 - Показники якості кластеризації База спаму

Як видно із гістограм, можна зробити висновки, що запропонований метод нечіткої правдоподібної кластеризації даних на основі аналізу щільності розподілу даних та їх піків (NCrCP) кластеризує дані якісніше за більшістю якісних характеристик кластеризації

Зокрема, якість методу кластеризації повинна відповідати вимогам не тільки якості кластеризації, а й швидкості і простоти з точки зору математичних розрахунків. Тому був проведений аналіз швидкості розрахунків кластеризації методу нечіткої правдоподібної кластеризації даних на основі аналізу щільності розподілу даних та їх піків та вище згаданих методів кластеризації.

В Таблиці 3.4 наведений порівняльний результат швидкості роботи методів кластеризації на основі щільності розподілу даних.

Таблиця 3.4 - Порівняння роботи алгоритмів за часом (с.)

Назва вибірки	**DENCLUE**	**DENCLUE-SA**	**DENCLUE-GA**	**NCrCP**
Page Blocks	71	107	158	**70**
Spambase	1285,9	1347	574	**440**

Аналіз результатів для різних варіантів кластеризації на вибірках Page Blocks і Spambase дозволяє зробити висновки про ефективність застосування різних модифікацій алгоритму DENCLUE та методів, таких як DENCLUE-SA, DENCLUE-GA та NCrCP. З результатів видно, як кожен з алгоритмів справляється з різними типами даних та показує свою здатність до кластеризації.

У вибірці Page Blocks, яка складається з даних про сторінки текстових документів (тобто структурованих даних про блоки тексту), найкращі результати досягнуті при використанні DENCLUE-GA, з показником 158. Це свідчить про те, що застосування генетичних алгоритмів у поєднанні з основною технікою DENCLUE дозволяє досягти більш точних кластерів, можливо, завдяки здатності генетичних алгоритмів оптимізувати параметри і покращити класифікацію навіть у складніших умовах. Тобто DENCLUE-GA має перевагу, оскільки генетичний алгоритм здатний ефективно налаштовувати параметри кластеризації під специфічні характеристики даних.

На другому місці знаходиться DENCLUE-SA з результатом 107, що свідчить про наявність переваг у застосуванні симуляції відпалу (Simulated Annealing). Хоча цей метод також дає добрі результати, генетичні алгоритми здатні краще працювати з більш складними структурованими даними, такими як Page Blocks.

NCrCP показав результат 70, що є близьким до стандартної реалізації DENCLUE (71). Це може свідчити про те, що в контексті цієї вибірки NCrCP має схожі характеристики до DENCLUE, але не дає суттєвого покращення в точності кластеризації.

У вибірці Spambase, яка містить дані для класифікації електронних листів на спам і не-спам, результати значно відрізняються від тих, що були отримані для Page Blocks. Тут DENCLUE показує найкращі результати з показником 1285,9, що може вказувати на високу ефективність цього методу в задачах, що вимагають високої гнучкості у визначенні кластерів, а також

можуть бути пов’язані з особливістю цієї вибірки, де дані можуть бути більш варіативними і менш структурованими.

DENCLUE-SA з результатом 1347 має дещо гірші результати, ніж базова версія DENCLUE, що свідчить про те, що симуляція відпалу в даному випадку не принесла значних переваг. Це може бути пов'язано з тим, що Spambase не вимагає сильної адаптації параметрів кластеризації або що інші аспекти даних краще вирішуються через інші техніки.

DENCLUE-GA має результат 574, що значно нижче за показник базового DENCLUE. Це може бути зумовлено тим, що генетичний алгоритм у цьому контексті не дав значних переваг у пошуку оптимальних параметрів для задачі кластеризації, оскільки генетичні алгоритми можуть мати більш високі вимоги до обчислювальних ресурсів і не завжди швидко конвертуються в покращення якості кластеризації для даних, таких як Spambase.

NCrCP виявився найбільш ефективним з результатом 440, що вказує на те, що цей метод може бути оптимальним для кластеризації в умовах більш «шумних» даних, як у випадку з Spambase. Метод NCrCP дає значно кращі результати порівняно з іншими підходами, особливо у випадку складніших, неструктурованих даних, що потребують кращої адаптації.

Отже, результати свідчать, що вибір алгоритму кластеризації та його модифікацій значною мірою залежить від характеру даних: для структурованих даних ефективніше використовувати методи з оптимізацією параметрів (як DENCLUE-GA), тоді як для більш складних та варіативних даних (як Spambase) кращі результати дає NCrCP.

3.7 Висновок до 3 розділу

1. 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.

2. 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.

3. 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.

4. 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] (Додаток А).

РОЗДІЛ 4
ЕВОЛЮЦІЙНІ МЕТОДИ ОПТИМІЗАЦІЇ В ЗАДАЧАХ КЛАСТЕРИЗАЦІЇ МАСИВІВ ДАНИХ РІЗНОЇ ПРИРОДИ

Запропоновані у розділі 3 методи адаптивної нечіткої кластеризації потоків даних з різною щільністю розподілу, забезпечують можливість швидше обробляти потоки даних та більш точними в порівнянні з методами на основі щільностей (DBSCAN, OPTICS, DENCLUE).

В основі цих методів полягає пошук екстремумів максимумів функції щільності розподілу даних у масиві, що аналізується (багатоестремальна оптимізація), при цьому ця функція формується, як суперпозиція ядерних (дзвонуватих) функцій, пов’язаних з кожним спостереженням.

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 [177, 189, 192, 245]. 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.

У найзагальнішому випадку, для пошуку глобального оптимуму мультиекстремальних функцій в умовах невизначеності. Історично першими еволюційними алгоритмами були так звані генетичні алгоритми [183, 185], 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) [174]. Ці алгоритми надихаються поведінкою рою птахів чи інших соціальних організмів, які спільно працюють 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. Вони 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 [190, 191].

4.1 Види еволюційних алгоритмів

Еволюційні алгоритми відіграють значну роль у сучасних методах кластеризації, оскільки вони забезпечують ефективний підхід до розв’язання складних задач групування даних. Їх актуальність обумовлена зростаючими вимогами до аналізу великих обсягів інформації, що потребують гнучких та адаптивних методів, здатних працювати з високорозмірними, нерівномірно розподіленими та складно структурованими даними. Класичні алгоритми кластеризації, такі як метод k-середніх або ієрархічна кластеризація, часто демонструють недостатню ефективність у випадках, коли дані мають складні нелінійні залежності або містять значну кількість шуму. У таких умовах еволюційні алгоритми набувають особливого значення завдяки здатності проводити глобальний пошук у просторі рішень, що дозволяє уникати локальних мінімумів та забезпечує якісніші результати кластеризації.

Основним принципом еволюційних алгоритмів є використання біологічно натхненних механізмів природного відбору, схрещування, мутації

та пристосованості для знаходження оптимальних розбиттів даних. Вони працюють на основі популяційного підходу, де кожен індивідуум представляє потенційне рішення, а процес еволюції поступово вдосконалює якість отриманих кластерів. Однією з ключових переваг такого підходу є можливість ефективної обробки складних структур даних, включно з кластерами нестандартної форми або неоднорідної густини. Крім того, вони дозволяють враховувати різноманітні критерії якості кластеризації, що дає змогу отримати більш гнучкі та релевантні результати порівняно з жорстко детермінованими методами.

Різноманітність еволюційних алгоритмів включає кілька основних напрямів, серед яких генетичні алгоритми, диференціальна еволюція, еволюційні стратегії та алгоритми рою частинок.

Генетичні алгоритми є найбільш розповсюдженим підходом і характеризуються використанням операцій схрещування та мутації для створення нових рішень на основі найкращих особин поточного покоління. Вони застосовуються в кластеризації завдяки здатності адаптивно налаштовувати параметри моделей та зменшувати ймовірність передчасної конвергенції.

Диференціальна еволюція фокусується на поступовому оновленні популяції через комбінування різниць між випадковими індивідуумами, що робить її особливо ефективною при оптимізації параметрів кластеризації в умовах високої розмірності даних.

Еволюційні стратегії відрізняються від генетичних алгоритмів тим, що використовують адаптивний підхід до зміни параметрів мутації, що дозволяє краще регулювати процес навчання та уникати надмірної експлорації. Алгоритми рою частинок, хоча й не базуються безпосередньо на принципах еволюції, демонструють схожі підходи до глобального пошуку оптимальних рішень, моделюючи поведінку групи частинок у багатовимірному просторі.

Попри численні переваги, використання еволюційних алгоритмів у кластеризації супроводжується низкою викликів.

Однією з головних проблем є висока обчислювальна складність, адже еволюційні методи вимагають обробки значної кількості рішень протягом великої кількості поколінь. Це може стати обмежувальним фактором у випадках, коли необхідно швидко отримати результати або обробити надзвичайно великі обсяги даних.

Другою складністю є налаштування параметрів алгоритму, зокрема вибір розміру популяції, ймовірностей мутації та схрещування, а також критеріїв зупинки процесу еволюції. Недостатньо оптимальне налаштування може призвести до передчасної конвергенції або до надмірної випадковості у виборі рішень.

Третьою проблемою є необхідність збереження різноманітності популяції протягом процесу еволюції, адже її зниження може спричинити зменшення пошукових можливостей алгоритму. Для вирішення цих питань використовуються додаткові техніки, такі як нішування, катастрофічний відбір, адаптивні ймовірності мутацій та методи динамічного регулювання параметрів.

Актуальність еволюційних алгоритмів у задачах кластеризації зростає в умовах сучасних інформаційних технологій, де обробка великих даних стає критично важливим завданням. Комбінація еволюційних підходів з іншими методами машинного навчання, зокрема нейронними мережами або нечіткою логікою, відкриває нові перспективи для покращення якості кластеризації та автоматизації процесів прийняття рішень. Використання гібридних підходів, що поєднують еволюційні алгоритми з традиційними методами, дозволяє отримати ще ефективніші результати, комбінуючи глобальний пошук із локальною оптимізацією.

Загалом, еволюційні алгоритми є потужним інструментом для кластеризації даних, здатним розв’язувати складні завдання, з якими не завжди справляються класичні методи. Вони дозволяють досягати більш точних та адаптивних результатів, хоча потребують ретельного налаштування параметрів та значних обчислювальних ресурсів.

4.2 Базовий метод кластеризації на основі котячих зграй

Еволюційні алгоритми широко використовуються для розв'язання складних оптимізаційних задач, зокрема кластеризації, завдяки своїй здатності проводити глобальний пошук у просторі можливих рішень. До найбільш відомих методів належать генетичні алгоритми, еволюційні стратегії, диференціальна еволюція, роєві алгоритми, такі як алгоритм рою частинок (PSO) і штучні бджолині колонії (ABC). Ці методи ґрунтуються на біологічних принципах, що дозволяють адаптивно та ефективно знаходити оптимальні розбиття даних, зменшуючи ймовірність потрапляння у локальні мінімуми.

Одним із порівняно нових, але перспективних еволюційних алгоритмів є алгоритм котячих зграй (Cat Swarm Optimization, CSO), який імітує поведінку котів у природі.

Для пошуку глобального екстремуму скалярної функції векторного аргументу $x = (x_1, x_2, ..., x_n)^T \in R^n$ авторами [192, 193, 195] було 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.

На початковому етапі CSO створює популяцію котів, де кожен котячий агент представляє потенційне рішення. Кожен кіт має певні атрибути, такі як

швидкість, положення в просторі рішень і режим поведінки. Популяція поділяється між двома режимами: більша частина котів перебуває в режимі спостереження, а менша - у режимі переслідування. У режимі спостереження кожен агент використовує механізми випадкових збурень та оцінки придатності для вибору кращого положення. Це дозволяє зберігати різноманітність популяції та уникати передчасної конвергенції. У режимі переслідування коти рухаються у напрямку найкращого знайденого рішення, коригуючи свої швидкості відповідно до отриманих параметрів.

Процес відшукання екстремуму за допомогою котячої зграї може бути реалізований у вигляді наступної 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 наступної 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

$$R_p = \frac{f(x_p(\tau)) - f_{\min}(x_p(\tau))}{f_{\max}(x_p(\tau)) - f_{\min}(x_p(\tau))}, \tau = 1,2,...,T \tag{4.1}$$

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

$$\upsilon_{pi}(\tau+1) = \upsilon_{pi}(\tau) + r(\tau)\eta_{TM}(x_{best,i}(\tau) - x_{pi}(\tau)), \tag{4.2}$$

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 ТМ 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 ТМ 3. Змінити становище кожного кота в гонитві відповідно до співвідношення

$$x_{pi}(\tau+1) = x_{pi}(\tau) + x_{pi}(\tau). \qquad (4.3)$$

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

Однією з ключових переваг CSO є його здатність балансувати між глобальним пошуком і локальною оптимізацією, що робить його ефективним для складних завдань, зокрема для кластеризації даних. Завдяки спостереженню, алгоритм уникає пастки локальних мінімумів, а механізм переслідування сприяє швидкому зближенню до оптимального розв'язку. Це дає йому певні переваги над традиційними роєвими алгоритмами, такими як PSO, де агенти можуть передчасно збиратися навколо субоптимальних рішень.

CSO застосовується в широкому спектрі задач, включаючи кластеризацію великих обсягів даних, оптимізацію параметрів нейронних мереж, розв'язання комбінаторних задач і навіть управління ресурсами в розподілених системах. Його гнучкість і здатність до адаптації роблять його

ефективним інструментом для аналізу даних із складною структурою, особливо в умовах високої розмірності та наявності шумових компонентів.

У загальному випадку базовий алгоритм оптимізації на основі котячих зграй може бути представлений у вигляді, наведеному на рисунку 4.1.

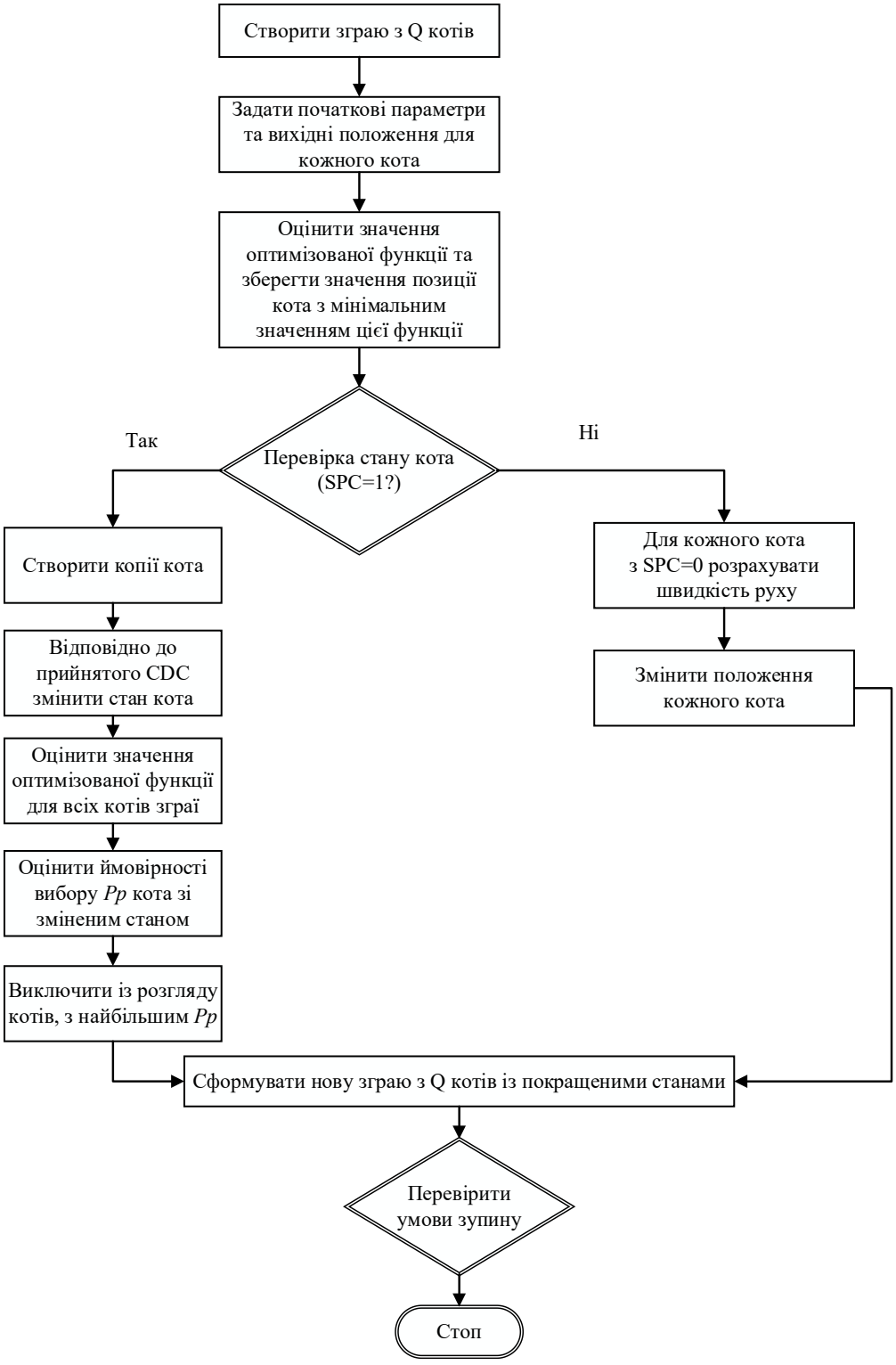


Рисунок 4.1 - Базовий алгоритм оптимізації на основі котячих зграй

Можна помітити, що розглянутий метод пошуку реалізує по суті покоординатний спуск (метод Гауса-Зейделя), що вимагає багаторазового оцінювання значень для оптимізації і характеризується низькою швидкістю збіжності. У режимі погоні 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 [202, 218, 222], що володіє цілою низкою переваг перед детермінованими процедурами пошуку екстремуму.

Незважаючи на перспективність, CSO, як і інші еволюційні алгоритми, має певні виклики. Основними проблемами є налаштування параметрів, зокрема визначення оптимального співвідношення котів у режимах спостереження та переслідування, а також вибір швидкості оновлення положень. Висока обчислювальна складність може стати обмеженням для задач із великими наборами даних. Однак застосування гібридних підходів, що поєднують CSO з іншими методами, такими як генетичні алгоритми або глибокі нейронні мережі, дозволяє підвищити його ефективність та розширити можливості практичного застосування.

Алгоритм котячих зграй є цікавим доповненням до сімейства еволюційних методів і демонструє високу ефективність у складних задачах оптимізації та кластеризації. Його здатність адаптуватися до змінних умов та комбінувати пошук з експлуатацією кращих рішень робить його конкурентоспроможним серед сучасних методів машинного навчання та аналізу даних.

4.3 Рандомізований метод оптимізації на основі котячих зграй

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

$$x_p(\tau+1) = x_p(\tau) - \eta_{SM}\hat{\nabla} f(x_p(\tau)), \qquad (4.4)$$

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 [196] в околі точки $x_p(\tau)$. Найбільш простим з обчислювальної точки зору є пошук з центральною пробою [191, 192], 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.

Визначивши $n+1$ значення функції $f(x_p(\tau))$, $f(x_p(\tau)+\eta_{SRD}e_2),...,f(x_p(\tau)+\eta_{SRD}e_n)$, замість градієнта

$\nabla f(x_p(\tau)) = \left(\frac{\partial f(x_p(\tau))}{\partial x_{p1}}, \frac{\partial f(x_p(\tau))}{\partial x_{p2}}, ..., \frac{\partial f(x_p(\tau))}{\partial x_{pn}} \right)^T$, можна запровадити його оцінку $\hat{\nabla} f(x_p(\tau))$ с компонентами

$$\frac{\partial \hat{f}(x_p(\tau))}{\partial x_{pi}} = \frac{1}{\eta_{SRD}} \left(f(x_p(\tau)) + \eta_{SRD} e_i \right) - f(x_p(\tau)), i = 1,2,...,n.$$

Реалізувавши далі крок у просторі P_x^n відповідно до (4.4), приходимо до нового стану cat_p у режимі пошуку з координатами

$$\begin{cases} x_{p1}(\tau+1) = x_{p1}(\tau) - \frac{\eta_{SM}}{\eta_{SRD}} (f(x_p(\tau)) + \eta_{SRD} e_1) - f(x_p(\tau)), \\ x_{p2}(\tau+1) = x_{p2}(\tau) - \frac{\eta_{SM}}{\eta_{SRD}} (f(x_p(\tau)) + \eta_{SRD} e_2) - f(x_p(\tau)), \\ x_{pn}(\tau+1) = x_{pn}(\tau) - \frac{\eta_{SM}}{\eta_{SRD}} (f(x_p(\tau)) + \eta_{SRD} e_n) - f(x_p(\tau)). \end{cases} \quad (4.5)$$

Можна помітити, 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, перевівши режим погоні.

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 [202-208, 218], 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

$$x_p(\tau+1)=x_p(\tau)-\eta_{SM}\Xi \tag{4.6}$$

(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. При цьому з вихідного стану робиться кілька випадкових проб оптимізованої функції $x_p(\tau)+\eta_{SRD}\Xi_l$ у випадкових напрямках $\Xi_l(l=1,2,...,n,...,L)$, при цьому фактор CDC може перевищувати значення n. 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

$$x_p(\tau+1)=x_p(\tau)+\eta_{SRD}\Xi^*. \tag{4.7}$$

Зауважимо також, що за $L=1$, процедури (4.6) і (4.7) збігаються.

Об’єднуючи процедури пошуку (4.4), (4.5), (4.7), можна ввести на розгляд пошук на основі статичного 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:

$$\hat{\nabla}f(x_p(\tau))=-\frac{\sum_{l=1}^{L}\Xi_l\left(f\left(x_p(\tau)+\eta_{SRD}\Xi_l\right)-\nabla f(x_p(\tau))\right)}{\left\|\sum_{l=1}^{L}\Xi_l\left(f\left(x_p(\tau)+\eta_{SRD}\Xi_l\right)-\nabla f(x_p(\tau))\right)\right\|}. \tag{4.8}$$

Підставляючи далі (4.8) в (4.7), 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 (4.2), (4.3) існують інші алгоритми, що володіють необхідними властивостями.

Одним із таких найбільш ефективних чисельно простих алгоритмів 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

$$x_p(\tau+1) = x_p(\tau) - \alpha(x_p(\tau) - x_p(\tau-1)) - \eta_{TM}\hat{\nabla} f(x_p(\tau)), \qquad (4.9)$$

де α - параметр, що визначає інерційні властивості процесу гонитви. При (4.9) повністю збігається з (4.4), відрізняючись лише кроком η_{SM}. При $\alpha = 1$ процес погоні 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 (4.9) може бути введена випадкова компонента, що вводить додаткове «гойдання» в процес погоні, що покращує глобальні властивості алгоритму. При цьому (4.9) 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.

4.4 Модифікований метод оптимізації на основі косяків риб

При використанні методів еволюційної оптимізації, що по суті 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

$$x_q^l = x_q^{l-1} + \eta_q^l Dir_q^l, \quad q = 1,2,...,Q,$$

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 [209-215]. 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

$$w_q^l = w_q^{l-1} + \frac{f^x\left(x_q^l\right) - f^x\left(x_q^{l-1}\right)}{\max\limits_p \left\{ f^x\left(x_q^l\right) - f^x\left(x_q^{l-1}\right) \right\}} \quad \forall q = 1,2,...,Q, \tag{4.10}$$

при цьому

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

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

$$x_{qi}^l = \begin{cases} x_{qi}^l + \eta_q^l Rand\{0,1\}, \text{ if } f^x\left(x_q^l\right) > f^a\left(x_q^{l-1}\right), \\ x_q^{l-1} \quad \text{інакше,} \end{cases} \tag{4.11}$$

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

$$x_q^l = x_q^{l-1} + \frac{\left(\sum_{p=1}^{Q}\left(x_p^l - x_p^{l-1}\right)\right)\left(f^x\left(x_q^l\right) - f^x\left(x_q^{l-1}\right)\right)}{\sum_{p=1}^{Q}\left(f^x\left(x_p^l\right) - f^x\left(x_p^{l-1}\right)\right)}. \quad (4.12)$$

Вводячи у розгляд зважений центр ваги косяка риб

$$Bar^l = \frac{\sum_{p=1}^{Q} x_p^l w_p^l}{\sum_{p=1}^{Q} w_p^l}, \quad (4.13)$$

можна записати цей рух у вигляді

$$x_q^l = \begin{cases} x_q^l - \eta_q^l Rand\{0,1\}\dfrac{x_q^{l-1} - Bar^{l-1}}{\left\|x_q^{l-1} - Bar^{l-1}\right\|}, \ if \ \ \sum_{p=1}^{Q} w_p^l > \sum_{p=1}^{Q} w_p^{l-1}, \\ x_q^l + \eta_q^l Rand\{0,1\}\dfrac{x_q^{l-1} - Bar^{l-1}}{\left\|x_q^{l-1} - Bar^{l-1}\right\|}, \ if \ \ \sum_{p=1}^{Q} w_p^l < \sum_{p=1}^{Q} w_p^{l-1}. \end{cases} \quad (4.14)$$

Для підвищення ефективності FSS у розгляд може бути введений 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 [179, 184, 187], серед яких з обчислювальної точки зору та ефективності - надійності відшукання екстремуму можна відзначити послідовний симплекс метод [216] та його модифікації [217].

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 $x_1^1, x_2^1, ..., x_Q^1$. Таким чином рух косяка-симплекса може бути описаний за допомогою співвідношень

$$\begin{cases} \overline{x}^{l-1} = \dfrac{1}{n}\sum\limits_{q=1}^{Q}\left(x_q^{l-1} - x_{qwost}^{l-1}\right), \\ x_q^l = \overline{x}^{l-1} + \alpha\left(\overline{x}_q^{l-1} - x_{qwost}^{l-1}\right), \end{cases} \tag{4.15}$$

що у загальному випадку є за своєю суттю алгоритмом оптимізації Нелдера-Міда [217]. 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.

4.5 Модифікований метод сірих вовків

Алгоритм оптимізації сірих вовків (Grey Wolf Optimizer, GWO) є одним із сучасних роївих методів оптимізації, який моделює соціальну поведінку вовчої зграї під час полювання. Він був запропонований у 2014 році і швидко здобув популярність завдяки своїй простоті, ефективності та здатності знаходити оптимальні рішення у складних багатовимірних задачах. Основною особливістю цього методу є те, що він імітує ієрархічну структуру зграї вовків та їхню стратегію колективного полювання, що дозволяє йому ефективно балансувати між глобальним пошуком та локальною оптимізацією.

За даними Мірджалілі [226], сірі вовки живуть разом і полюють групами. Процес пошуку та полювання можна описати так: (4.16) 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 алгоритм GWO. Алгоритм 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:

$$C_\alpha = \left| B_1 * GW_\alpha - X(t) \right|,$$

$$C_\beta = \left| B_2 * GW_\beta - X(t) \right|, \tag{4.16}$$

$$C_\delta = \left| B_3 * GW_\delta - X(t) \right|,$$

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:

$$GW_1 = GW_\alpha - A_1 * C_\alpha,$$

$$GW_2 = GW_\beta - A_2 * C_\beta, \tag{4.17}$$

$$GW_3 = GW_\delta - A_3 * C_\delta,$$

$$GW(t) = \frac{GW_1 + GW_2 + GW_3}{3}. \tag{4.18}$$

Параметри A та B є комбінаціями керуючого параметра α та випадкових чисел r_1 та r_2 [226, 228]:

$$A = 2\alpha r_1 - \alpha,$$

$$B = 2r_2. \tag{4.19}$$

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.

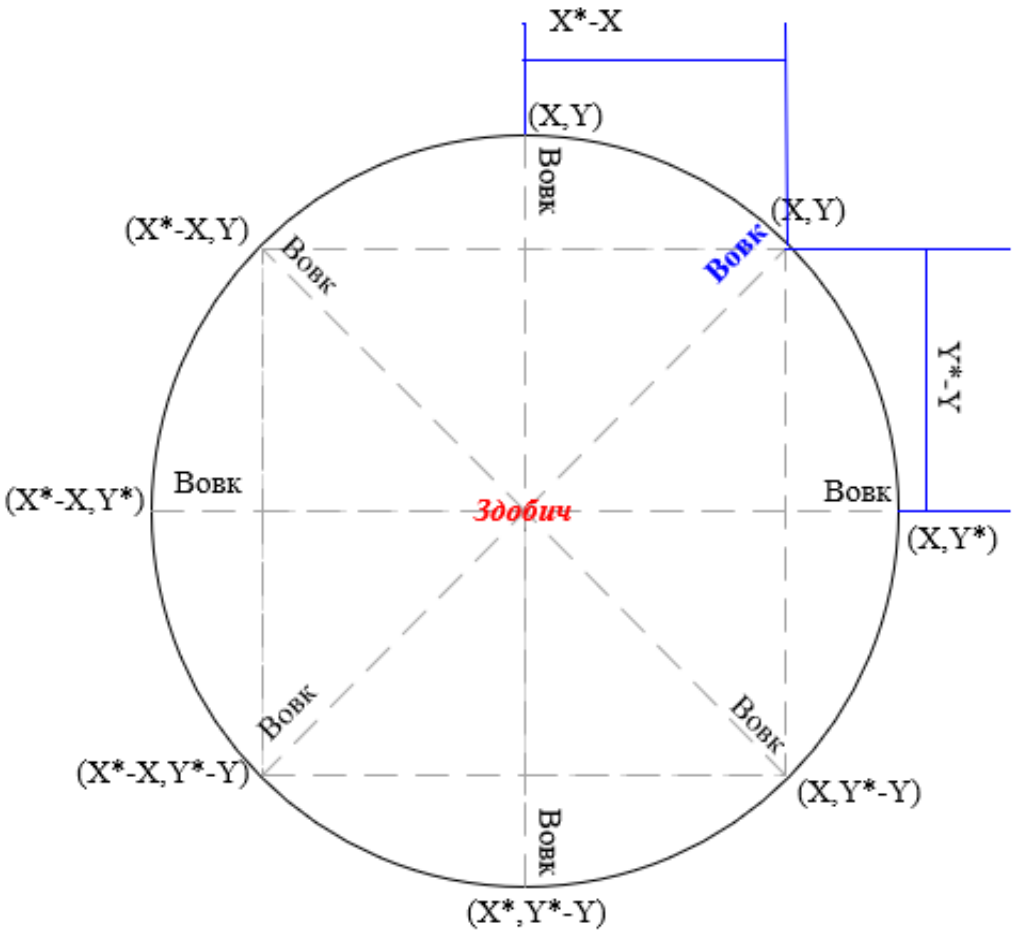


Рисунок 4.2 – Схема роботи алгоритму GWO

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 (4.18) 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 (4.18) слід враховувати лише положення альфа, або його вага має бути набагато більшою, ніж ваги інших домінант. Рівняння (4.18) можна переписати у вигляді:

$$GW(t+1) = \frac{w_1 GW_1 + w_2 GW_2 + w_3 GW_3}{3}, \tag{4.20}$$

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 $w_1 \geq w_2 \geq w_3$.

Псевдокод алгоритму сірих вовків зручно представити у вигляді таблиці, де описані основні початкові параметри налаштування алгоритму, ініціалізація початкових позицій агентів-вовків, та умови пошуку глобального екстремуму, в залежності від критеріїв зупинки, значень фітнес функцій та оновлення позицій вовків.

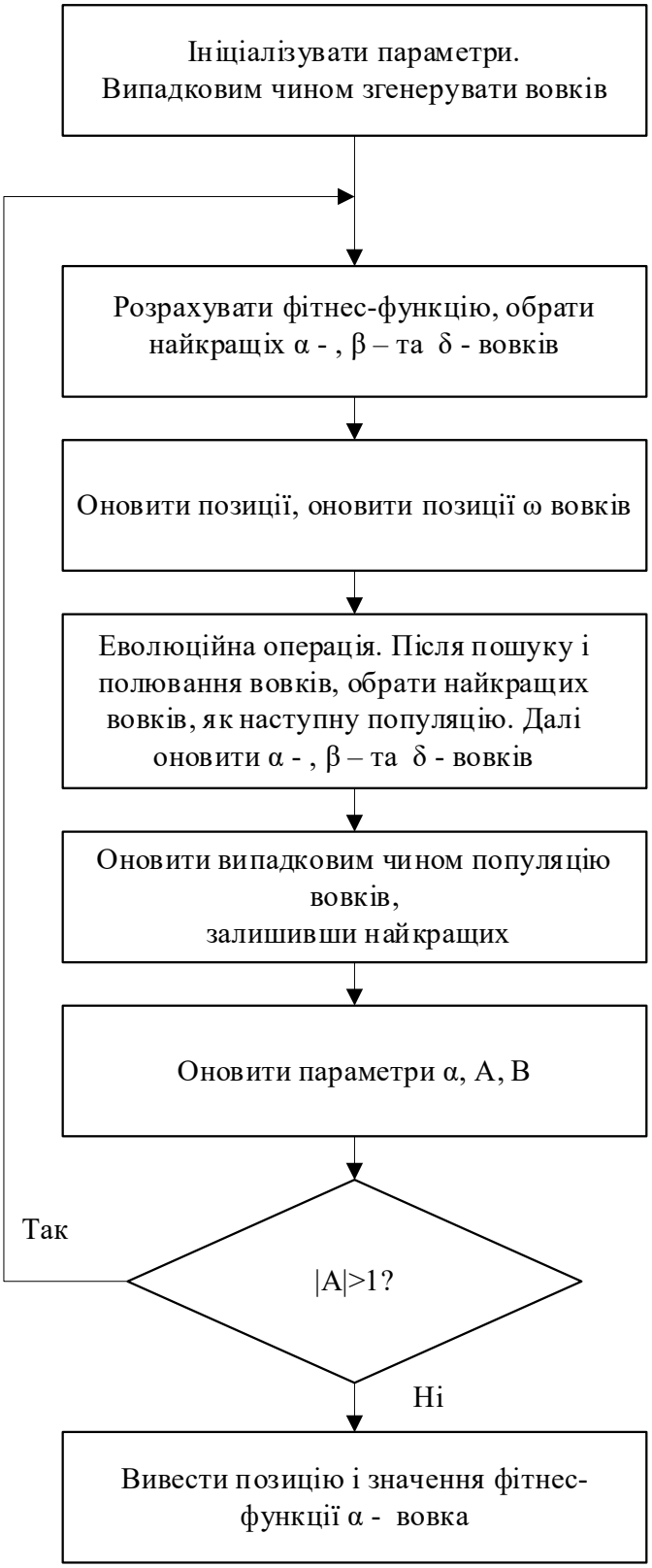


Рисунок 4.3 – Блок-схема алгоритму GWO

Блок-схема алгоритму сірих вовків наведена на Рисунку 4.3.

Таблиця 4.1 - Псевдокод роботи алгоритму сірих вовків

Опис	**Псевдокод**
Параметри налаштування	Вибірка даних Популяція зграї Контрольний параметр Критерій зупину
Ініціалізація	Початкові позиції сірих вовків-домінант
Пошук	Якщо це не критерій зупинки, то обчислення нового значення фітнес - функції Оновлення позиції Обмеження коло позицій вовків Оновити α, β і δ Оновити критерій зупинки. Кінець

Зміна позицій вовків описана наступними виразами:

$$C = \left| B * X_P(t) - X(t) \right|, \tag{4.21}$$

$$X\left(t+1\right) = X_P(t) - A * C, \tag{4.22}$$

де X_P - позиція здобичі,

X - позиція вовка.

Як видно з Рисунку 4.3, вовк у позиції (X, Y) може відтворити свою позицію навколо здобичі згідно з формулами оновлення (4.21) та (4.22).

Для цього, 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 метода можливісної нечіткої кластеризації.

Розглядаючи рівняння (4.18) 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 (4.18) слід враховувати лише положення альфа, або його вага має бути набагато більшою, ніж ваги інших домінант. Таким чином, рівняння (4.18) можна переписати у вигляді:

$$GW(t+1) = \frac{w_1 GW_1 + w_2 GW_2 + w_3 GW_3}{3}, \tag{4.23}$$

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 $w_1 \geq w_2 \geq w_3$.

На першій (або $t = 0$) ітерації пропонується задати ваги результатами алгоритму кластеризації, де:

$$c_1 = w_\alpha;$$

$$c_2 = w_\beta; \text{ при } t = 0 \tag{4.24}$$

$$c_3 = w_\delta;$$

Тоді можна визначити, що ваги змінних задовольняють гіпотезі про соціальну ієрархію функцій сірих вовків та їх пошукову поведінку.

Для підвищення надійності знаходження саме глобального екстремуму цільової функції можна використати ідею «божевільних котів» – оптимізацію [208], модифіковану введенням випадкового блукання, яка довела свою ефективність при розв'язанні мультиекстремальні проблеми. Вводячи, додаткове пошукове збурення, можна записати рух вовка у вигляді:

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

де γ - параметр корекції характеристик збурення,

$0 \le \delta \le 1$ - параметр швидкості самонавчання типу параметра інерції α,

σ^2 - дисперсія білого шуму $H(\tau)$.

Таким чином підвищується імовірність знаходження глобального екстремуму прийнятої цільової функції, що в кінцевому рахунку підвищує ефективність процесу нечіткої кластеризації.

4.6 Апробація рандомізованого методу оптимізації на основі котячих зграй

Експериментальні дослідження рандомізованого алгоритму оптимізації на основі котячих зграй проводилися з чотирма наборами даних: Iris, Cancer, Wine і Glass.

Кожен із наборів даних має ряд параметрів, представлених у таблиці 4.2. Було проведено порівняльний аналіз якості даних кластеризації за основними характеристиками рейтингів якості таких, як швидкість кластеризації даних та середня похибка.

Таблиця 4.2 – Характеристичні параметри вибірок

Назва вибірки	Число класів	Кількість атрибутів	Кількість спостережень
Іриси	3	4	150
Рак	2	9	683
Вина	3	13	178
Скло	6	8	214

Параметри налаштування запропонованого модифікованого методу нечіткої кластеризації на основі оптимізації зграї котів (FCMCSO) представлено у вигляді таблиці, де описані основні початкові параметри пошуку діапазону обраного виміру, пошуку пулу пам’яті, ініціалізація початкових позицій котів та кількість ітерацій.

Таблиця 4.3 – Параметри модифікованого методу нечіткої кластеризації на основі оптимізації зграї котів (FCMCSO)

Параметри	Значення
Пошук діапазону обраного виміру (SRD)	Випадково [0,1]
Пошук пулу пам’яті(SMP)	5
Розмір популяції	Кількість кластерів
r_1	Випадкове значення у диапазоні[0,1]
c_1	Константа
Самооцінка позиції (SPC)	Випадково в диапазоні [0,1]
Кількість ітерацій	Manually

Порівняльний аналіз модифікованого методу нечіткої кластеризації на основі оптимізації зграї котів (FCMCSO) за часовою складністю проводився з відомими методами та алгоритмами кластеризації даних: FCM, PSO, GSA, CSO та FCMCSO.

В таблиці 4.4 наведені результати експериментальних досліджень на різних вибірках даних для більш детального аналізу отриманих результатів.

Таблиця 4.4 – Порівняльні результати часової обробки методів кластеризації таких, як FCM, PSO, GSA, CSO та FCMCSO

Назва вибірки	**FCM**	**PSO**	**GSA**	**CSO**	**FCM CSO**
Іриси	0,008	0,020	0,022	0,043	0,006
Рак	0,009	0,138	0,204	0,026	0,007
Вина	0,009	0,282	0,098	0,076	0,008
Скло	0,010	0,431	0,431	0,021	0,015

Таблиця 4.4 порівнює результати часової обробки п’яти різних методів кластеризації – FCM, PSO, GSA, CSO та FCMCSO – для різних вибірок даних: «Іриси», «Рак», «Вина» і «Скло». Цей аналіз дозволяє оцінити ефективність методів кластеризації за часом обробки, що є важливим чинником при обранні оптимального методу для реальних задач.

Для вибірки «Іриси» метод FCMCSO показує найшвидший час обробки (0,006 с), що є суттєво меншим, ніж у решти методів. Метод CSO, хоча і має кращий час обробки, ніж інші (0,043 с), все ж є значно повільнішим, ніж FCMCSO. Інші методи – PSO, GSA, та FCM – мають трохи вищі значення часу обробки (0,020 с, 0,022 с і 0,008 с відповідно), але вони значно відстають від FCMCSO.

У вибірці «Рак» результат FCMCSO також є найшвидшим (0,007 с), тоді як PSO (0,138 с) і GSA (0,204 с) мають найбільші часи обробки. Метод CSO показує порівняно хороший результат (0,026 с), але все одно поступається FCMCSO. У цій вибірці видно значний розрив між швидкістю обробки методів.

Вибірка «Вина» демонструє схожу картину, де FCMCSO має найкращий час обробки (0,008 с), в той час як PSO (0,282 с) та GSA (0,098 с) значно перевищують цей показник.

Для вибірки «Скло» метод FCMCSO продовжує бути лідером за часом обробки (0,015 с), в той час як PSO і GSA обробляють дані найповільніше (0,431 с для обох методів). CSO має час обробки 0,021 с, що також є досить швидким порівняно з іншими методами, але не може конкурувати з FCMCSO.

Загалом, метод FCMCSO продемонстрував найкращі результати в плані часової обробки на всіх вибірках, що робить його найбільш ефективним за часом серед розглянутих методів. Метод PSO та GSA, з іншого боку, мають значно більший час обробки, що може бути важливим фактором у задачах, де важлива швидкість обробки даних.

В таблиці 4.5 наведено порівняльний аналіз отриманих результатів в залежності від кількості спостережень. Пропонується порівняти FCMCSO з класичним CSO за 3 параметрами: 50, 100, 150 ітерацій роботи методів.

Таблиця 4.5 – Результати кластеризації CSO та FCMCSO з різною кількістю ітерацій (середня похибка у %)

Назва вибірки	**Кількість ітерацій CSO**			**Кількість ітерацій FCMCSO**		
	50	**100**	**150**	**50**	**100**	**150**
Іриси	23,34	20,84	21,67	17,55	14,78	16,46
Рак	40,23	40,55	41,47	38,89	39,22	39,15
Вино	24,55	21,44	22,20	18,43	17,37	16,32
Скло	56,34	56,48	55,67	51,63	51,7	49,79

Таблиця 4.5 надає результати кластеризації за допомогою методів CSO та FCMCSO з різною кількістю ітерацій (50, 100 і 150) для чотирьох вибірок даних: «Іриси», «Рак», «Вино» та «Скло». У таблиці представлено середню похибку у відсотках для кожного методу та кожної кількості ітерацій.

Вибірка "Іриси" демонструє, що FCMCSO має кращі результати порівняно з CSO на всіх кількостях ітерацій. При 50 ітераціях FCMCSO показує похибку 17,55%, що є значно меншим, ніж 23,34% для CSO. Зі збільшенням кількості ітерацій, похибка у методі CSO зменшується до 20,84% при 100 ітераціях і 21,67% при 150 ітераціях, однак вона все ще значно перевищує похибку FCMCSO. У той час як FCMCSO також має зменшення похибки з 17,55% при 50 ітераціях до 14,78% при 100 і 16,46% при 150 ітераціях.

Вибірка "Рак" показує менші коливання між методами, але FCMCSO все одно має стабільно меншу похибку. Для CSO похибка становить 40,23% при 50 ітераціях, зменшуючись до 40,55% при 100 і 41,47% при 150 ітераціях. FCMCSO має похибку 38,89% при 50 ітераціях, що трохи покращується до 39,22% при 100 і 39,15% при 150 ітераціях.

Вибірка "Вино" також підтверджує тенденцію до кращих результатів FCMCSO. CSO має похибку 24,55% при 50 ітераціях, що зменшується до 21,44% при 100 і 22,20% при 150 ітераціях. FCMCSO показує кращу похибку: 18,43% при 50 ітераціях, 17,37% при 100 і 16,32% при 150 ітераціях.

Вибірка "Скло" показує найменше поліпшення з боку FCMCSO. Похибка для CSO становить 56,34% при 50 ітераціях, зменшуючись до 56,48% при 100 і 55,67% при 150 ітераціях. FCMCSO також показує деяке зменшення похибки з 51,63% при 50 ітераціях до 51,7% при 100 і 49,79% при 150 ітераціях. Хоча тут різниця між методами менша, FCMCSO все ж продовжує показувати деяке покращення.

Аналізуючи все вище наведено можна зробити висновки, що FCMCSO має кращу ефективність в плані зменшення середньої похибки порівняно з CSO для більшості вибірок, особливо для "Ірисів", "Вино" і "Скло". Кількість

ітерацій має певний вплив на похибку, але цей ефект різний для кожної вибірки. Загалом, збільшення ітерацій знижує похибку, однак перевага FCMCSO залишалася стабільною навіть на менших кількостях ітерацій. Вибірка "Скло" є винятком, де різниця між методами менша, але FCMCSO все одно показує кращі результати, хоч і з менш значним поліпшенням.

Крім проведеного аналізу швидкості роботи методів та похибки, необхідно також проаналізувати швидкість роботи методів в залежності від кількості спостережень з іншими методами кластеризації даних.

Пропонується порівняти запропоновану модифікацію нечіткої кластеризації даних на основі еволюційного методу котячих зграй з відомими методами кластеризації: FCM, Густафсон-Кессель (G-K), адаптивний можливісний метод нечіткої кластеризації та онлайн метод можливісної кластеризації даних на основі еволюційної оптимізації котячих зграй

В таблиці 4.6 та таблиці 4.7 наведені результати роботи методів, запропонованих для порівняння з різною кількістю спостережень.

Аналізуючи отримані результати, можна зробити висновок, що незалежно від розміру вихідної інформації, що подається на обробку запропонованим методом для порівняння працездатності та ефективності, запропонований підхід до можливісної кластеризації даних на основі еволюційного методу котячих зграй не поступається швидкодією та якістю кластерування у порівнянні з відомими алгоритмами.

Таблиця 4.6 - Порівняльні характеристики середньої похибки для вибірки Газ з різною кількістю спостережнь у відсотках

Метод	**50**	**Час**	**100**	**Час**	**150**	**Час**
FCM	1,62	1,19	1,35	2,55	0,98	3,03
GK	1,66	1,62	1,32	2,72	0,99	3,12
Адаптивний можливісний метод нечіткої кластеризації	1,22	1,15	1,02	2,02	0,75	2,10

Продовження таблиці 4.6

Онлайн метод можливісної кластеризації даних на основі еволюційної оптимізації котячих зграй	0,69	1,02	0,49	1,33	0,14	1,41

Таблиця 4.6 надає порівняльні характеристики середньої похибки для вибірки "Газ" з різною кількістю спостережень (50, 100, 150) для чотирьох методів кластеризації: FCM, GK (Густафсон-Кессель), адаптивного можливісного методу нечіткої кластеризації та онлайн методу можливісної кластеризації даних на основі еволюційної оптимізації котячих зграй. Окрім значень похибки, також надано час обробки для кожного методу та кількості спостережень.

Середня похибка для FCM зменшується з 1,62% для 50 спостережень до 0,98% для 150 спостережень. Час обробки збільшується з 1,19 с для 50 спостережень до 3,03 с для 150 спостережень. Це свідчить про те, що при збільшенні кількості спостережень FCM потребує більше часу для обробки, але його похибка знижується.

Похибка для GK також зменшується з 1,66% для 50 спостережень до 0,99% для 150 спостережень, що є подібним до результатів FCM. Час обробки для GK зростає з 1,62 с для 50 спостережень до 3,12 с для 150 спостережень. Як і в випадку з FCM, при більшій кількості спостережень час обробки збільшується, однак похибка зменшується.

Адаптивний можливісний метод нечіткої кластеризації має найменшу похибку серед всіх методів, зменшуючи її з 1,22% до 0,75% при збільшенні кількості спостережень. Час обробки цього методу також збільшується, але значення залишаються меншими порівняно з FCM і GK: 1,15 с для 50 спостережень до 2,10 с для 150 спостережень.

Онлайн метод можливісної кластеризації даних на основі еволюційної оптимізації котячих зграй показує найкращі результати за середньою

похибкою, особливо при 150 спостереженнях (0,14%), що значно нижче за всі інші методи. Час обробки для онлайн методу є суттєво нижчим, ніж для інших методів: 1,02 с для 50 спостережень, зростаючи до 1,41 с для 150 спостережень. Тобто, цей метод є не тільки найбільш ефективним за похибкою, але й швидким у виконанні.

Найкращі результати за похибкою показує онлайн метод можливісної кластеризації на основі еволюційної оптимізації котячих зграй, який має найменшу похибку при кожній кількості спостережень, і це зберігається навіть при зростанні числа спостережень. FCM та GK показують подібні результати в плані похибки, але FCM має дещо кращі результати з точки зору часу обробки, порівняно з GK.

Адаптивний метод нечіткої кластеризації має найменші часи обробки серед класичних методів, але його похибка все ж таки вища порівняно з онлайн методом котячих зграй.

Зі збільшенням кількості спостережень для всіх методів спостерігається збільшення часу обробки, але деякі методи, такі як онлайн метод, залишаються швидшими, що робить його оптимальним вибором для реальних задач з великими даними.

Таблиця 4.7 - Порівняльні характеристики середньої похибки для вибірки Скло з різною кількістю спостережень у відсотках

Метод	**50**	**Час**	**100**	**Час**	**150**	**Час**
FCM	1,74	1,21	1,44	2,40	0,86	3,10
GK	1,85	1,74	1,53	2,82	0,99	3,27
Адаптивний можливісний метод нечіткої кластеризації	1,43	1,36	1,22	2,55	0,65	2,60

Продовження таблиці 4.7

Онлайн метод можливісної кластеризації даних на основі еволюційної оптимізації котячих зграй	1,11	1,17	1,00	1,23	0,54	1,11

Таблиця 4.7 порівнює середню похибку для вибірки Glass при різній кількості спостережень (50, 100, 150) для чотирьох методів кластеризації: FCM, GK (Густафсон-Кессель), адаптивного можливісного методу нечіткої кластеризації та онлайн методу можливісної кластеризації на основі еволюційної оптимізації котячих зграй. Також для кожного методу зазначено час обробки.

У методі FCM середня похибка зменшується з 1,74% для 50 спостережень до 0,86% для 150 спостережень, що свідчить про поступове покращення точності методу при збільшенні кількості спостережень. Час обробки зростає з 1,21 с для 50 спостережень до 3,10 с для 150 спостережень, що є типовим для більшості методів кластеризації при збільшенні обсягу даних.

Похибка для GK зменшується з 1,85% для 50 спостережень до 0,99% для 150 спостережень, що демонструє подібну тенденцію до FCM. Час обробки для GK зростає з 1,74 с для 50 спостережень до 3,27 с для 150 спостережень, що також є звичайним явищем при збільшенні кількості спостережень.

Адаптивний можливісний метод нечіткої кластеризації має відносно низьку похибку, яка зменшується з 1,43% для 50 спостережень до 0,65% для 150 спостережень. Час обробки змінюється від 1,36 с для 50 спостережень до 2,60 с для 150 спостережень, що свідчить про помірне збільшення часу обробки.

Онлайн метод демонструє найкращі результати за середньою похибкою: 1,11% для 50 спостережень, зменшуючись до 0,54% для 150 спостережень. Це вказує на високу ефективність цього методу. Час обробки для онлайн методу є значно меншим порівняно з іншими методами: 1,17 с для 50 спостережень до 1,11 с для 150 спостережень. Це демонструє швидкість виконання та ефективність цього методу навіть при збільшенні кількості спостережень.

Аналізуючи все вище сказане, можна зробити висновок, що онлайн метод можливісної кластеризації даних на основі еволюційної оптимізації котячих зграй виявився найефективнішим за кількома показниками: він має найменшу похибку та швидкий час обробки на всіх етапах, що робить його оптимальним вибором для великих і складних наборів даних.

Адаптивний можливісний метод нечіткої кластеризації має кращу похибку та дещо вищий час обробки порівняно з іншими методами, проте його похибка все ж залишається значно нижчою, ніж у класичних методів, таких як FCM та GK.

FCM і GK демонструють схожі результати в плані похибки та часу обробки, з FCM виявляючи трохи кращі результати за похибкою, але з аналогічним часом виконання.

Всі методи показують покращення точності при збільшенні кількості спостережень, але онлайн метод відзначається найкращими показниками за ефективністю та швидкістю, що робить його перспективним для реальних завдань кластеризації.

Для зручного аналізу отриманих результатів, побудовано діаграми залежності похибки та часу від кількості спостережень для різних вибірок даних.

Рисунок 4.4 - Діаграма залежності похибки від кількості спостережень (50, 100, 150) для вибірки Gas

Рисунок 4.5 - Діаграма залежності похибки від кількості спостережень (50, 100, 150) для вибірки Glass

Як видно з діаграм, онлайн метод можливісної кластеризації даних на основі еволюційної оптимізації котячих зграй демонструє найкращі результати, незалежно від вибірки, кількості спостережень та ітерацій методу.

Для аналізу методів, побудовано також діаграми залежності часу кластеризації від кількості спостережень (50, 100, 150) для вибірок даних Газ та Скло.

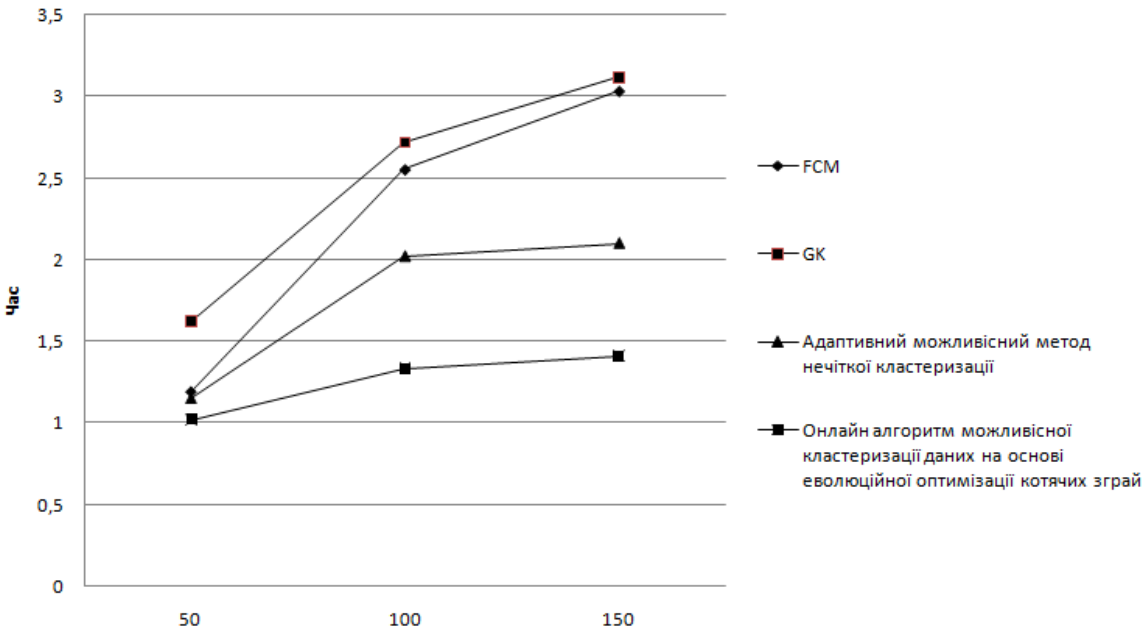


Рисунок 4.6 - Діаграма залежності часу кластеризації від кількості спостережень (50, 100, 150) для вибірки Газ

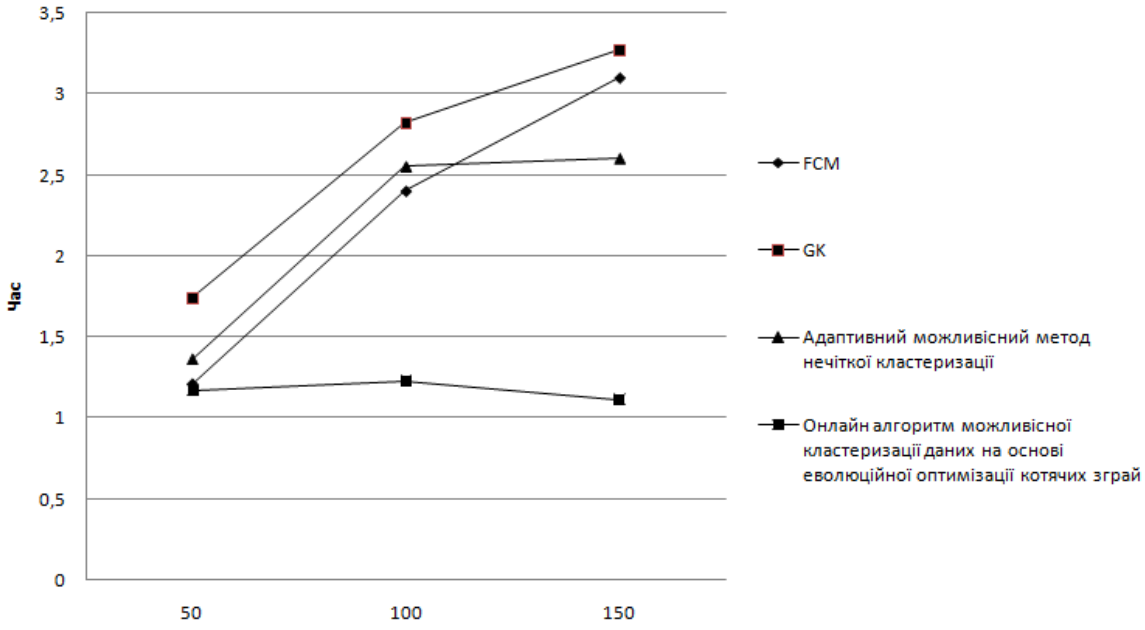


Рисунок 4.7 - Діаграма залежності часу кластеризації від кількості спостережень (50, 100, 150) для вибірки Скло

Як видно із діаграм, наведених на рисунках, швидкість та похибка в запропонованому методі можливісної кластеризації даних на основі еволюційного підходу демонструє достатньо високі показники.

Результати кластеризації наборів даних, наведених в таблицях, демонструють швидкодію та якість роботи методів кластеризації. Запропонований онлайн алгоритм можливісної кластеризації даних на основі еволюційної оптимізації котячих зграй демонструє гарні результати роботи.

Порівняльний аналіз запропонованого методу проводився з відомими на сьогодні алгоритмами кластеризації такими, як FCM, Густафсон-Кессель-алгоритм і адаптивний можливісний метод нечіткої кластеризації.

На рисунках, що наведені вище, продемонстрована робота алгоритмів у порівнянні із запропонованим онлайн алгоритмом можливісної кластеризації даних на основі еволюційної оптимізації котячих зграй.

Завдяки своїй адаптивності та функціям еволюційної оптимізації котячих зграй, алгоритм не потребує багато часу для обробки даних, що надходять у реальному часі, та не завантажує себе проміжними розрахунками за рахунок функцій адаптивності. Це досить яскраво демонструють діаграми залежності часу кластеризації від кількості спостережень та залежності похибки від кількості спостережень.

4.7 Апробація модифікованого методу оптимізації на основі косяків риб

Експериментальні дослідження проводилися на двох базах даних, таких як Блоки сторінок і База спаму, а також дві тестові мультиекстремальні функції.

Таблиця 4.8 - Опис наборів даних

Вибірка	**Кількість спостережень**	**Атрибути**	**Кластери**
Блоки сторінок	5472	10	5
База спаму	4601	57	2

Таблиця 4.9 - Тестові багатоекстремальні функції

Назва	Формули	Інтервал	Крок
Растрігін	$f(x) = 20 + x^2 + y^2 - -10\cos(2\pi x) + \cos(2\pi y)$	$[-5,12;5,12]$	$0,01$
Гріванг	$f(x) = \frac{1}{4000}x + \frac{1}{4000}y - -\cos(\frac{x}{\sqrt{1}})\cos(\frac{x}{\sqrt{2}}) + 1$	$[-30;30]$	$0,1$

У зв'язку з тим, що функції Растрігіна та Гріванга мають багато локальних екстремальних точок у своїй області пошуку, як показано на рисунках 4.8 (а) та (б), ми додаємо 514 агентів.

У наборі даних «Блоки сторінки» були представлені класифіковані блоки макета сторінки в документі, який був виявлений процесом сегментації. База даних спаму також витягується з репозиторію машинного навчання UCI і описує електронну пошту, яка класифікується як спам або не спам.

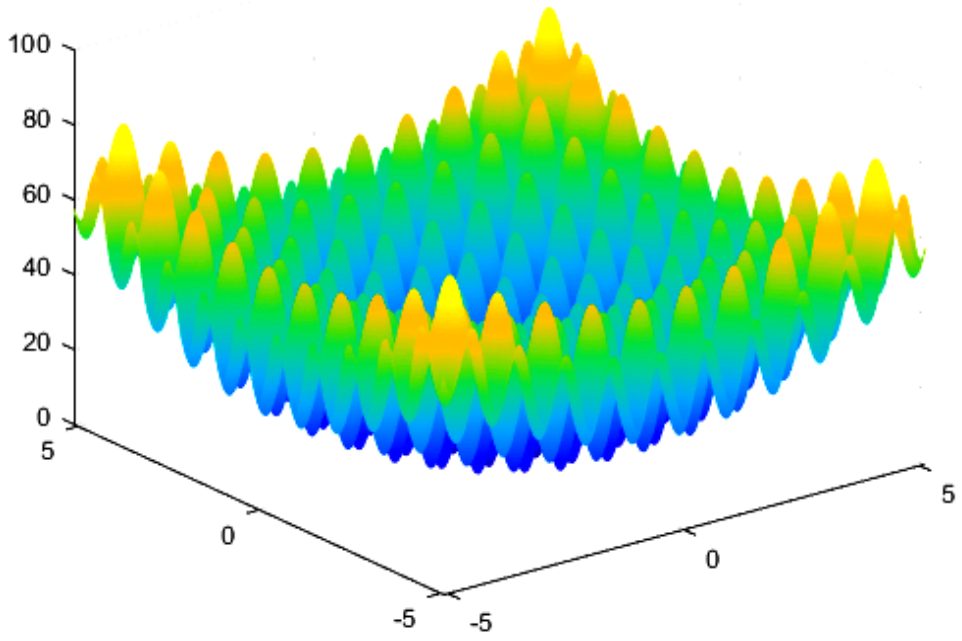


Рисунок 4.8 - Функція Растрігіна, що має багато екстремальних точок

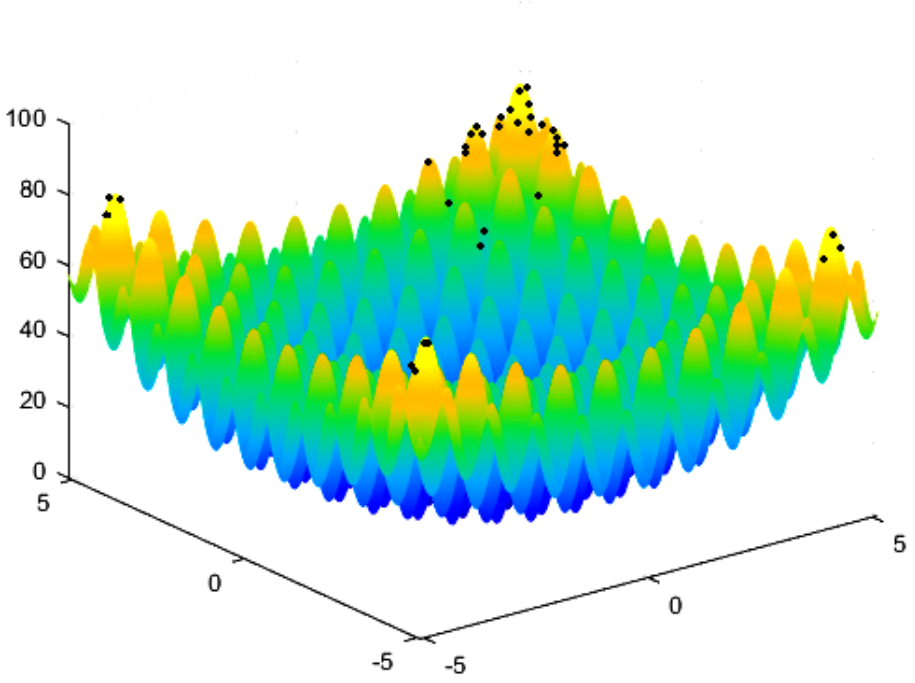


Рисунок 4.9 - Модифікований метод оптимізації на основі косяка риб на функції Растрігіна

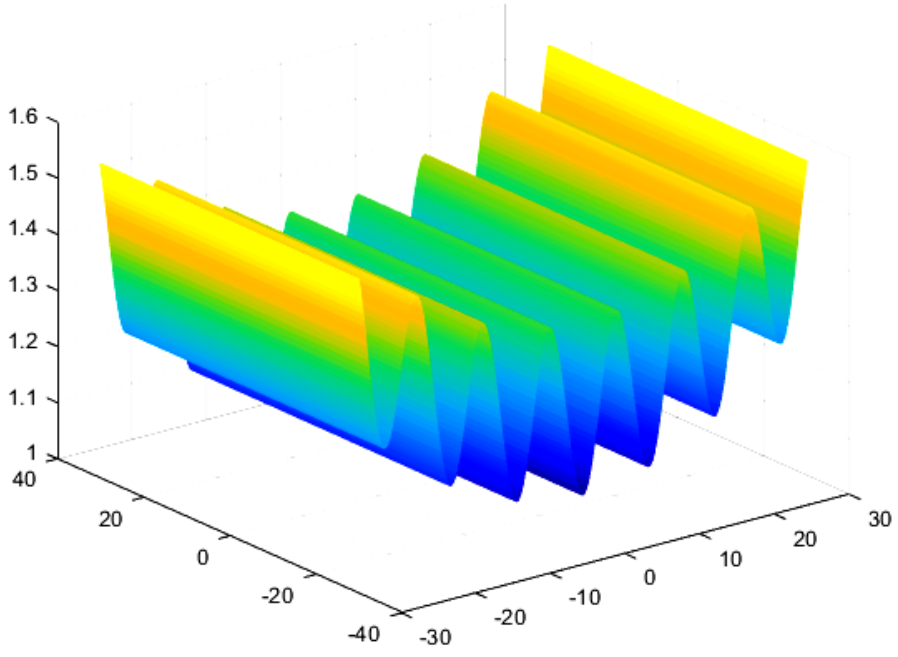


Рисунок 4.10 - Функція Гріванга, яка має багато екстремальних точок

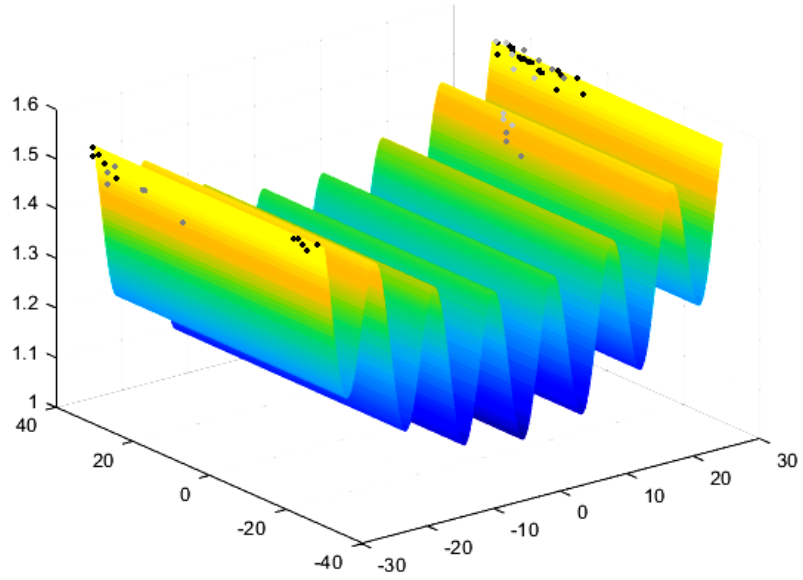


Рисунок 4.11 -Модифікований метод оптимізації на основі косяка риб на функцію Гріванга

Порівняння точності добре відомих алгоритмів оптимізації, таких як Fish School (FSS) і Cat Swarm (CSO), і запропонований модифікований метод оптимізації на основі Fish School (OMFS).

Таблиця 4.10 - Порівняння точності

Дані	**Точність**	**OMFS**	**FSS**	**CSO**
Растрігін	Середня Краща	**190,46** **195,83**	189,65 195,59	**190,46** **195,83**
Гріванг	Середня Краща	**3,65** **4,82**	3,41 4,12	**3,65** 4,81
Блоки сторінок	Середня Краща	**951,47** **959,64**	951,01 959,43	951,15 959,55
База спаму	Середня Краща	**291,77** **299,84**	291,17 299,48	**291,77** 299,64

OMFS та CSO демонструють дуже схожі результати для всіх наборів даних, з рівними або майже рівними значеннями середньої та найкращої точності. FSS також показує хороші результати, але, за винятком набору даних "Гріванг", зазвичай трохи поступається OMFS та CSO в середній точності.

У випадку набору даних Гріванг, FSS показує кращі результати за точністю, що може свідчити про його ефективність для конкретних типів даних.

В загальному, OMFS та CSO виглядають найбільш стабільними методами за точністю на всіх вибірках, хоча FSS може бути більш ефективним в окремих випадках.

Ці результати дозволяють зробити висновок, що вибір методу кластеризації залежить від конкретних характеристик набору даних, зокрема від їх розміру та структури. Процес збіжності гібридного алгоритму продемонстровано на рисунку 4.13 та рисунку 4.14.

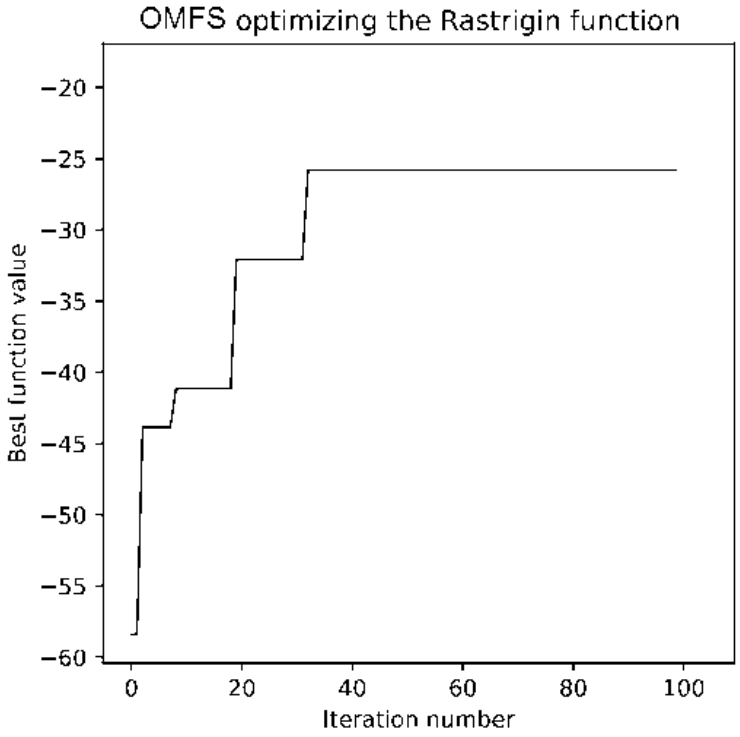


Рисунок 4.13- Модифікований метод оптимізації на основі тестової функції Fish School (функція Растрігіна)

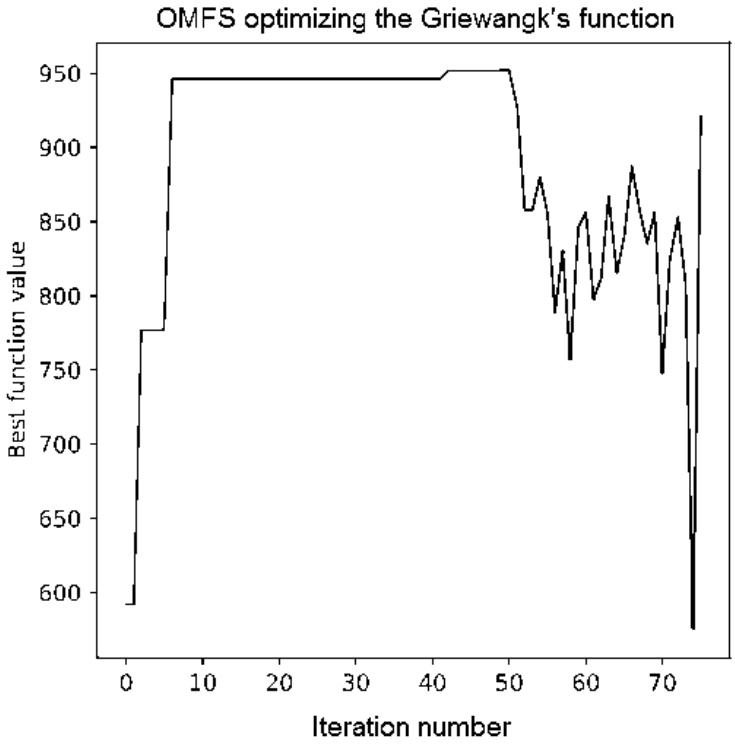


Рисунок 4.13 - Модифікований метод оптимізації на основі тестової функції Fish School (функція Гріванга)

Розглянуто задачу кластеризації масивів даних, що описано як у векторній, так і матричній формах. Для оптимізації цих функцій - пошуку локальних екстремумів запропонований метод, що є гібридом Fish School Search, випадкового пошуку та еволюційної оптимізації. Цей метод не потребує обчислення похідних функції, що оптимізується і у загальному випадку призначений для відшукання максимумів багатоекстремальних функцій матричного аргументу (зображень).

Запропонований підхід дозволяє скоротити кількість запусків процедури оптимізації, дозволяє знаходити екстремуми функцій складної форми та є простим в чисельній реалізації.

4.8 Апробація модифікованого методу сірих вовків

Експериментальні дослідження проводились на тестовій багатоекстремальній функції Швефеля (4.23)

$$f(x) = 418.9829N - \sum_{i=1}^{N} x_i \sin\left(\sqrt{x_i}\right). \quad (4.23)$$

Зовнішній вигляд цієї функції наведено на Рисунку 4.11 і Рисунку 4.12, де можна бачити велику кількість екстремумів, в яких звичайні оптимізаційні методи в’язнуть по принципу «важкої кулі».

Запропонований адаптивний підхід до нечіткої кластеризації на основі еволюційної оптимізації алгоритму сірих вовків виключає можливість «застрягання» в опосередкованих екстремумах за допомогою подвійної перевірки знаходження вовка-домінанта в екстремумі та порівнянні із заданою похибкою розрахунків.

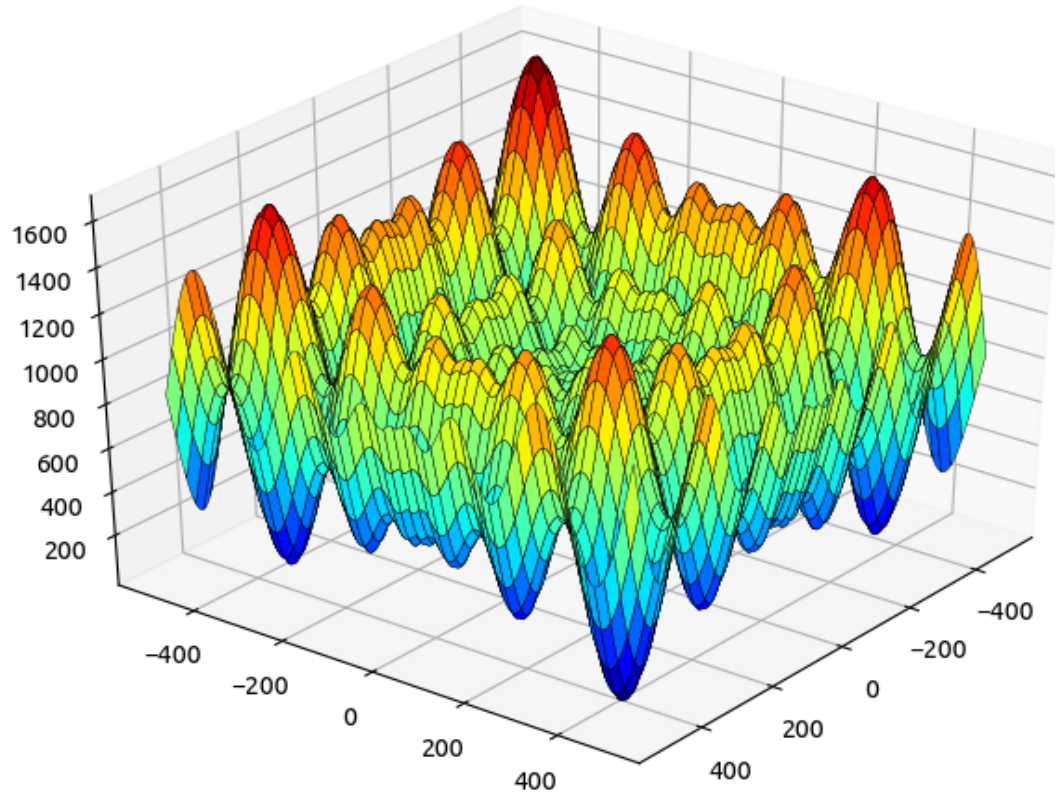


Рисунок 4.14 - Функція Швефеля

Таким чином швидкість роботи метода збільшується, а якість отриманих результатів покращується за допомогою так званого «ансамблю» двох методів: методу можливісної нечіткої кластеризації та оптимізаційного методу сірих вовків (PMGWO).

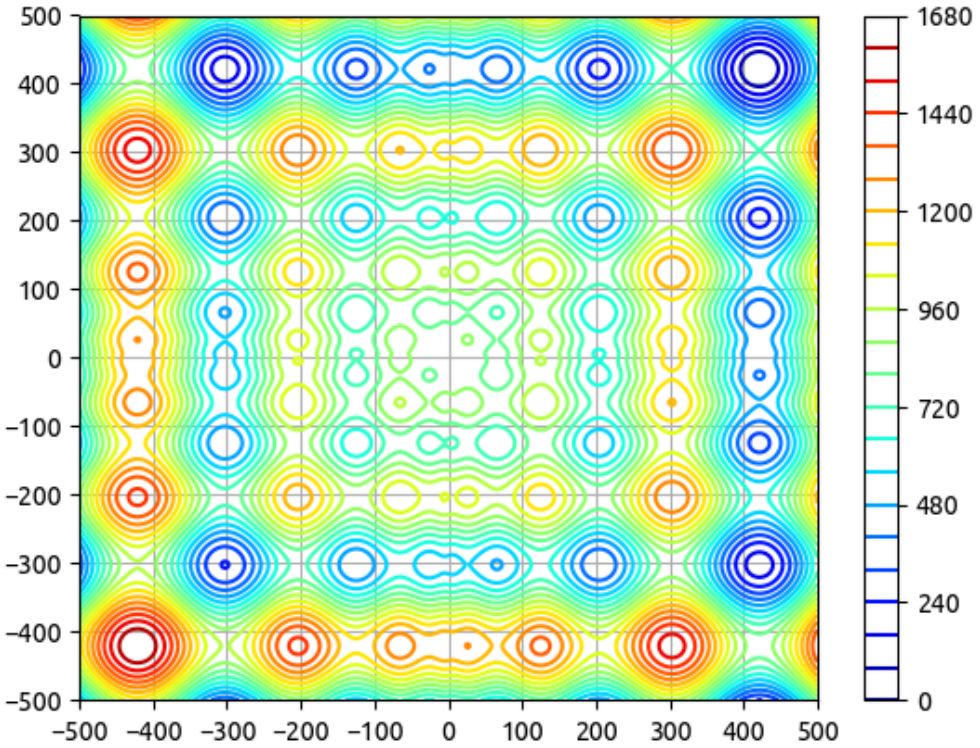


Рисунок 4.15 - Функція Швефеля вигляд зверху

Таблиця 4.11 - Порівняння точності

Функція	**Точність**	**PMGWO**	**GWO**	**FCM**
Швефель	Середня	**191,44**	189,65	190,46
	Краща	**195,75**	195,59	195,83

Таблиця 4.11 містить порівняльні результати для трьох методів кластеризації (PMGWO, GWO, FCM) за точністю на функції Швефель. Для кожного методу вказано середню та найкращу точність.

PMGWO демонструє досить високу точність як за середнім, так і за найкращим значенням серед інших методів.

Середня точність для GWO дорівнює 189,65, а найкраща точність — 195,59. Хоча середня точність GWO трохи нижча, ніж у PMGWO, найкраща точність майже така ж, що вказує на стабільність цього методу в пошуку оптимальних рішень.

Середня точність для FCM складає 190,46, а найкраща точність — 195,83. FCM демонструє дуже схожі результати з GWO, з невеликою перевагою за середньою точністю, хоча найкраща точність для FCM є найбільшою серед усіх методів.

Усі три методи (PMGWO, GWO, FCM) показують дуже схожі результати з невеликою різницею в середній точності. FCM має найвищу найкращу точність (195,83), що вказує на його потенціал у досягненні кращих результатів на функції Швефель. PMGWO і GWO мають подібні результати, причому PMGWO трохи перевершує GWO за середньою точністю.

У загальному, ці методи можуть бути використані в залежності від конкретних вимог до часу обробки і точності, але FCM є найбільш стабільним за найкращими результатами. Ці результати підтверджують, що вибір методу залежить від балансу між середньою точністю та досягненням найкращих результатів, що є важливим для практичних застосувань кластеризації в задачах оптимізації.

4.9 Висновок до 4 розділу

1. 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.

2. 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.

3. 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.

4. 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.

5. 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] (Додаток А).

РОЗДІЛ 5
ГІБРИДНІ ЕВОЛЮЦІЙНІ МЕТОДИ КЛАСТЕРИЗАЦІЇ МАСИВІВ ДАНИХ

Гібридні еволюційні методи кластеризації представляють собою поєднання технік еволюційного обчислення та методів кластеризації [152-161, 171, 186]. Ці методи спроектовані для покращення результатів кластеризації шляхом використання переваг обидвох підходів.

Загальні принципи гібридних еволюційних методів кластеризації:

– Початкова популяція кластерів: гібридні методи можуть використовувати еволюційні алгоритми для створення початкової популяції кластерів. Це може поліпшити якість кластеризації відразу з початку процесу.

– Оптимізація центрів кластерів: еволюційні алгоритми можуть використовуватися для оптимізації положень центрів кластерів під час ітераційного процесу. Це дозволяє враховувати сліди глобальної структури даних при розміщенні центрів.

– Визначення числа кластерів: гібридні методи можуть допомагати визначити оптимальну кількість кластерів, використовуючи еволюційні алгоритми для пошуку параметрів, таких як кількість кластерів.

– Адаптивність. Гібридні методи можуть бути адаптивними до змін у вхідних даних або структури кластерів, оновлюючи свої параметри через еволюційні процеси.

– Інтеграція з іншими методами: гібриди можуть поєднувати еволюційні алгоритми з іншими методами кластеризації, такими як ієрархічні, що дозволяє використовувати переваги різних підходів.

– Обмін інформацією між кластерами. Еволюційні алгоритми можуть використовувати обмін інформацією між кластерами для поліпшення загальної кластеризації.

Гібридні еволюційні методи намагаються використовувати сильні сторони обох підходів для досягнення кращих результатів у завданнях кластеризації даних.

5.1 Метод нечіткої кластеризації масивів даних на основі еволюційного методу оптимізації котячих зграй

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 в масиві даних A, або поточний дискретний час в задачах Data Stream Mining. Результатом кластеризації є розбиття цього масиву на m класів, що перетинаються з прототипами-центроїдами Cl_j, $c_j \in X^n, j = 1,2,...,m,$ при цьому поряд із знаходженням центроїдів c_j повинен бути оцінений рівень належності $0 < \mu_j(k) < 1$ кожного $x(k)$ до кожного з кластерів Cl_j. Передбачається також, що дані, які надходять на обробку, нормовані в гіперкуб $[-1;1]$ так, що $-1 \le x_i(k) \le 1$, де $x_i(k), i = 1,2,...,n$ - i-та компонента вектора спостережень $x(k)$.

В основі широко поширеного алгоритму ймовірнісної нечіткої кластеризації [1-22], лежить процедура мінімізації цільової функції

$$E(\mu_j(k), c_j) = \sum_{k=1}^{N} \sum_{j=1}^{m} \mu_j^{\beta}(k) \left\| x(k) - c_j \right\|^2 \qquad (5.1)$$

при обмеженнях

$$\sum_{j=1}^{m}\mu_j(k)=1,\quad 0\le\sum_{j=1}^{m}x_j(k)\le N, \tag{5.2}$$

Записавши функцію Лагранжа

$$L(\mu_j(k),c_j,\lambda(k))=\sum_{k=1}^{N}\sum_{j=1}^{m}\mu_j^{\beta}(k)\left\|x(k)-c_j\right\|^2+\sum_{k=1}^{N}\lambda(k)(\sum_{j=1}^{m}\mu_j(k)-1)$$

тут $\lambda(k)$ - невизначені множники Лагранжа) і вирішивши систему рівнянь Каруша-Куна-Таккера

$$\begin{cases}\dfrac{\partial L(\mu_j(k),c_j,\lambda(k))}{\partial\mu_j(k)}=0,\\ \nabla_{c_j}L(\mu_j(k),c_j,\lambda(k))=\vec{0},\\ \dfrac{\partial L(\mu_j(k),c_j,\lambda(k))}{\partial\lambda_j(k)}=0,\end{cases}$$

отримуємо шукане рішення виду

$$\begin{cases}\mu_j(k)=\dfrac{(\left\|x(k)-c_j\right\|^2)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}(\left\|x(k)-c_j\right\|^2)^{\frac{1}{1-\beta}}},\\ c_j=\dfrac{\sum\limits_{k=1}^{N}\mu_j^{\beta}(k)x(k)}{\sum\limits_{k=1}^{N}\mu_j^{\beta}(k)},\end{cases} \tag{5.3}$$

яке при $\beta=2$ співпадає з алгоритмом нечітких С – середніх (FCM) Дж. Бездека [57, 58]:

$$\begin{cases} \mu_j(k) = \dfrac{(\left\| x(k) - c_j \right\|^2)^{-1}}{\sum\limits_{l=1}^{m} (\left\| x(k) - c_j \right\|^2)^{-1}}, \\ c_j = \dfrac{\sum\limits_{k=1}^{N} \mu_j^2(k) x(k)}{\sum\limits_{k=1}^{N} \mu_j^2(k)}. \end{cases} \tag{5.4}$$

У [170] була доведена збіжність процедур (5.3), (5.4) до локального мінімуму, при цьому досягнення глобального екстремуму в загальному випадку не гарантується.

У роботах [202, 203-208] завдання умовної оптимізації (5.1), (5.2) було переформульовано на завдання безумовної оптимізації цільової функції виду

$$Goal\left(c_j\right) = \sum_{k=1}^{N} \left(\sum_{j=1}^{m} \left\| x(k) - c_j \right\|^{2(1-\beta)} \right)^{1-\beta}, \tag{5.5}$$

яка при $\beta = 2$ приймає вигляд

$$Goal\left(c_j\right) = \sum_{k=1}^{N} \left(\sum_{j=1}^{m} \left\| x(k) - c_j \right\|^{-2} \right)^{-1}, \tag{5.6}$$

при цьому цікаво відзначити, що в процесі мінімізації (5.5), (5.6) знаходяться тільки координати центроїдів $c_j, j = 1,2,...,m$, а для знаходження рівнів нечіткої належності можуть бути використані перші рівняння співвідношень (5.3), (5.4). Таким чином, завдання нечіткої кластеризації може бути зведена до пошуку глобального екстремуму цільових функцій (5.5), (5.6). Для вирішення завдання можуть бути використані інтенсивно розвиваються в даний час в рамках HSCI еволюційні біоінспіровані «ройові» процедури

оптимізації [174, -187] серед яких в якості одних з найбільш швидкодіючих можна відзначити, так звані, алгоритми котячих зграй [188, 189, 192-208]. Зауважимо, що саме котячі зграї з успіхом були використані для вирішення завдань чіткої кластеризації в рамках процедури c-середніх [1-22], що породжується цільовою функцією (5.1) 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 $Goal(c_j)$ (5.5) або (5.6).

В рамках класичного «котячого» алгоритму [192-208] 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 погоні визначається швидкими стрибками з великою амплітудою і дозволяє вивести кішку cat_p з локального екстремуму, якщо вона потрапила туди. Поєднання локального сканування та різких змін поточного стану дозволяє з більшою ймовірністю відшукати глобальний екстремум у порівнянні з традиційними методами багатоекстремальної оптимізації.

5.2 Онлайн метод для правдоподібної нечіткої кластеризації на основі еволюційної оптимізації котячої зграї

Дослідження гібридного методу правдоподібної кластеризації на основі еволюційного алгоритму є однією з актуальних задач сучасного аналізу даних, що охоплює такі напрями, як кластеризація, обробка великих обсягів

інформації в пакетному та онлайн-режимах, а також застосування ймовірнісних підходів до моделювання складних структур даних.

Ця задача є комплексною і передбачає інтеграцію методів машинного навчання та обчислювального інтелекту для створення адаптивного підходу до групування об’єктів у високорозмірному просторі даних.

Основна мета дослідження полягає у розробці ефективного механізму знаходження глобальних екстремумів цільової функції правдоподібної нечіткої кластеризації, використовуючи модифікований еволюційний алгоритм. Такий підхід дозволяє підвищити точність кластеризації, забезпечити адаптивність методу до складних типів даних і зменшити ймовірність потрапляння у локальні мінімуми, що є суттєвою проблемою для традиційних підходів.

Головною особливістю правдоподібного підходу до кластеризації є його орієнтація на максимізацію правдоподібності кластерів. На відміну від жорсткої кластеризації, де кожен об’єкт належить лише до одного кластера, нечітка кластеризація дозволяє об’єктам бути частково присутніми в різних групах із певними ступенями належності. Це особливо корисно в умовах реальних даних, які можуть містити невизначеність, шумові компоненти або складні нелінійні взаємозв’язки між спостереженнями.

Правдоподібна кластеризація дозволяє будувати більш точні моделі реальних процесів, оскільки бере до уваги ймовірнісний розподіл об’єктів між групами, що покращує інтерпретованість та гнучкість отриманих результатів.

Еволюційний алгоритм у розглянутій модифікації виступає як основний оптимізаційний механізм, що дозволяє знаходити глобальні екстремуми функції правдоподібності кластеризації. Традиційні методи кластеризації, такі як алгоритм k-середніх або ієрархічні методи, можуть демонструвати недостатню ефективність у випадках, коли дані мають складну структуру або містять значну кількість локальних мінімумів.

Еволюційний підхід, заснований на принципах природного відбору, схрещування та мутацій, дозволяє значно покращити пошукові характеристики алгоритму та уникнути проблеми передчасної конвергенції.

Однією з ключових модифікацій у розглянутому еволюційному алгоритмі є поєднання класичних операцій (відбір, мутація, кросовер) із механізмами глобального випадкового пошуку. Все це забезпечує не лише ефективне уточнення знайдених рішень, але й можливість адаптації алгоритму до різних типів задач кластеризації.

Динамічне регулювання параметрів, таких як ймовірності мутації та кросоверу, а також механізми контролю різноманітності популяції, дозволяють підтримувати баланс між глобальною експлорацією та локальною експлуатацією найкращих рішень.

Важливим аспектом дослідження є можливість використання гібридного методу у двох основних режимах - пакетному та онлайн. У пакетному режимі вся сукупність даних доступна для обробки одразу, що дозволяє застосовувати стандартні методи оптимізації, такі як генетичні алгоритми або диференціальна еволюція.

В онлайн-режимі нові дані надходять поступово, що потребує динамічного оновлення кластерної структури без необхідності повторного обчислення всіх параметрів моделі. Використання адаптивних еволюційних стратегій та механізмів самонавчання дозволяє розробленому підходу бути ефективним у реальних умовах із потоковими даними, що є важливим у таких сферах, як кібербезпека, аналіз поведінки користувачів або прогнозування динамічних систем.

Таким чином, гібридний метод правдоподібної кластеризації на основі еволюційного алгоритму є перспективним підходом до обробки складних даних, що поєднує адаптивність, точність і здатність працювати в різних режимах. Подальші дослідження можуть бути спрямовані на вдосконалення механізмів налаштування параметрів еволюційного алгоритму, зокрема на розробку адаптивних стратегій регулювання операторів мутації та

схрещування, що забезпечить ще ефективніше розв'язання задач кластеризації у високорозмірних просторах.

Цільова функція правдоподібної кластеризації має вигляд (2.21) за наявності обмежень (2.22).

Пакетний метод правдоподібної нечіткої кластеризації має вигляд (2.25), або його рекурентна версія може бути записана у формі (2.26) та (2.27), що дозволяє вирішувати задачу нечіткої кластеризації в онлайн-режимі.

Для знаходження глобального екстремуму (2.21) доцільно використовувати так звані еволюційні алгоритми оптимізації роїв частинок, серед яких є методи котячої зграї [192-208], 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:

$$c_p(\tau+1) = c_p(\tau) - \alpha(c_p(\tau) - c_p(\tau-1)) - \eta\hat{\nabla} Goal_M(c_p(\tau)) + \eta_\xi \Xi(\tau), \qquad (5.7)$$

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 (2.21) 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.

5.3 Метод глобальної оптимізації божевільної котячої зграї в 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 [192-208] і методів глобального випадкового пошуку [215-217].

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 погоні (TM), що ставить собі за мету відшукання глобального екстремуму. SM, як правило, реалізується на основі градієнтного пошуку з малим параметром навчання і таким чином сканує локальний окіл кожного з котів, що знаходиться в цьому режимі.

TM характеризується випадковими стрибками з великою амплітудою і ставить собі за мету «витягти» кота cat_p з локального екстремуму в разі його потрапляння туди.

В загальному випадку стандартний метод котячої зграї може бути представлений у вигляді послідовності ітерацій:

CS1: випадковим чином створюється зграя з Q котів cat_p, кожен з яких є по суті n-вимірним вектором $c_p(0)$ в області визначення функції, що оптимізується, та оцінюється значення функції в цій точці $Goal\left(c_p(0)\right)$.

CS2: вводиться параметр стану SPC, що приймає значення 0 або 1, за допомогою якого вихідна зграя розбивається випадковим чином на дві групи: якщо SPC = 1, то кіт знаходиться в режимі пошуку, якщо ж SPC = 0, то відповідний кіт знаходиться в режимі гонитви.

CS3: коти з SPC = 1 починають пошук локального екстремуму, а коти з SPC = 0 запускаються в режим гонитви.

CS4: оцінюються значення цільової функції для всіх котів і зберігаються всі $c_p(1)$ з найменшими значеннями цієї функції, коти з найбільшим значенням $Goal\left(c_p(1)\right)$ можуть бути видалені із зграї.

CS5: повернення до CS1 з новими значеннями, тобто починається новий етап з оновленою популяцією.

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.

При цьому рух кота в режимі пошуку може бути описано за допомогою рекурентної процедури

$$c_p(\tau+1) = c_p(\tau) - \eta_{SM}\,\hat{\nabla}\,Goal\left(c_p(\tau)\right),$$

де $\hat{\nabla} Goal\left(c_p(\tau)\right)$ - оцінка градієнта функції, що оптимізується в точці $c_p(\tau)$, одержувана або на основі пошуку з центральною пробою, або на основі випадкових проб (статистичний градієнт),

η_{SM} - малий крок пошуку в просторі A^n.

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

$$c_p(\tau+1) = c_p(\tau) - \alpha\left(c_p(\tau) - c_p(\tau-1)\right) - \eta_{TM}\,\hat{\nabla} Goal\left(c_p(\tau)\right) + \Xi(\tau), \quad (5.8)$$

де $0 < \alpha < 1$ - параметр інерції режиму погоні,

$\Xi(\tau)$ - випадкове збурення, що вводить додаткове сканування простору пошуку.

В [197, 199, 208, 222] 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 [190], який довів свою ефективність при вирішенні багатоекстремальних задач.

При цьому характеристики випадкового збурення змінюються відповідно до виразу:

$$\Xi(\tau) = \gamma\Xi(\tau-1) - \delta\left(E\left(w_p(\tau)\right) - E\left(w_p(\tau-1)\right)\right) + \sigma^2 H(k), \quad (5.9)$$

де γ - параметр корекції характеристик збурення,

$0 \le \delta \le 1$ - параметр швидкості самонавчання типу параметра інерції α у (5.9),

σ^2 - дисперсія білого шуму $H(\tau)$.

Таким чином весь процес оптимізації за допомогою підходу «божевільних котів» може бути описаний за допомогою рекурентних співвідношень (2.26), (2.27), (5.9), при цьому при $\alpha = \gamma = \delta = 0$ режим гонитви автоматично переходить в режим пошуку (рух по антиградієнту).

5.4 Метод кластеризації масивів даних на основі комбінованої оптимізації функцій щільності розподілу та еволюційного методу котячих зграй

Методи кластеризації відіграють ключову роль у сучасному аналізі даних, дозволяючи ефективно групувати об’єкти на основі схожості їхніх характеристик. Одним із перспективних напрямів у цій сфері є розробка комбінованих підходів, що об’єднують кілька стратегій оптимізації для досягнення більш точних і стійких результатів. Зокрема, метод кластеризації на основі комбінованої оптимізації функцій щільності розподілу та еволюційного методу є потужним інструментом для обробки складних структур даних, оскільки поєднує статистичні принципи аналізу розподілу об’єктів та механізми глобального пошуку оптимального розбиття.

Основна ідея даного підходу ґрунтується на використанні функцій щільності ймовірності для оцінки концентрації об’єктів у просторі ознак, що дозволяє визначати оптимальну кількість кластерів та їхні межі. На відміну від класичних методів, таких як k-середніх або ієрархічна кластеризація, що базуються на жорстко визначених метриках відстані, підхід на основі щільності дає змогу ефективно працювати з нерівномірно розподіленими та нелінійно відокремлюваними кластерами. Це особливо важливо при аналізі

реальних масивів даних, де об’єкти можуть формувати складні взаємозв’язки та розподілятися неоднорідно у просторі.

Процес кластеризації у запропонованому методі передбачає двоетапну процедуру оптимізації. На першому етапі здійснюється оцінка функції щільності розподілу об’єктів, що дозволяє виділити області з високою концентрацією точок і визначити потенційні центри кластерів. Цей підхід ґрунтується на методах ядерного оцінювання щільності (Kernel Density Estimation, KDE) або непараметричних статистичних моделях, які забезпечують адаптивний підхід до виявлення кластерних структур без потреби у попередньому визначенні кількості груп.

Другий етап реалізується за допомогою еволюційного алгоритму, який використовується для глобального пошуку оптимального розбиття даних. Еволюційні методи, такі як генетичні алгоритми, диференціальна еволюція або алгоритми роїв частинок, дозволяють ефективно досліджувати простір можливих рішень та адаптивно коригувати параметри кластеризації. Зокрема, в цьому методі використовується підхід, де кожен потенційний розв’язок представлений у вигляді сукупності параметрів, що визначають структуру кластерів (координати центрів, радіуси або вагові коефіцієнти належності об’єктів до груп). Оцінка якості кожного рішення здійснюється на основі функції правдоподібності кластеризації, яка включає в себе щільнісні характеристики розподілу об’єктів та мінімізує внутрішньокластерну варіацію.

Застосування еволюційних алгоритмів у такому контексті дає змогу уникати типових проблем традиційних методів, таких як потрапляння у локальні мінімуми або надмірна чутливість до початкових умов. За рахунок використання операторів схрещування, мутації та відбору, популяція рішень поступово покращується, наближаючись до оптимального розбиття. Важливим аспектом є також механізм підтримки різноманітності популяції, що запобігає передчасній конвергенції та дозволяє ефективніше знаходити глобальні екстремуми.

Ключовою перевагою комбінованого методу є його гнучкість та адаптивність до різних типів даних. На відміну від традиційних алгоритмів, що використовують фіксовані евклідові метрики або апріорно задану кількість кластерів, запропонований підхід дозволяє працювати з довільними розподілами та автоматично визначати оптимальні параметри групування. Це особливо корисно в задачах, де структура кластерів не є однорідною, як-от у фінансовому аналізі, біоінформатиці, аналізі соціальних мереж та комп'ютерному зорі.

Попри значні переваги, метод кластеризації на основі комбінованої оптимізації функцій щільності та еволюційного алгоритму має певні виклики. По-перше, він є обчислювально затратним через необхідність багаторазового обчислення оцінок щільності та застосування ітеративних еволюційних операторів. Це може створювати обмеження при роботі з великими наборами даних або в режимі реального часу. По-друге, ефективність методу залежить від правильного налаштування параметрів, таких як ширина ядерної функції у KDE або ймовірності мутацій та кросоверу в еволюційному алгоритмі. Неправильний вибір цих параметрів може призвести до надмірної фрагментації кластерів або, навпаки, до їхнього об'єднання у недостатньо чітко виражені групи.

Враховуючи ці особливості, подальші дослідження можуть бути спрямовані на розробку адаптивних стратегій налаштування параметрів, що ґрунтуються на принципах самонавчання та автоматичної корекції. Використання гібридних підходів, де еволюційні алгоритми поєднуються з нейронними мережами або байєсівськими моделями, може значно покращити продуктивність та точність методу.

Таким чином, метод кластеризації на основі комбінованої оптимізації функцій щільності розподілу та еволюційного підходу є перспективним напрямом у сфері аналізу даних, що поєднує статистичні та обчислювальні підходи для досягнення високої точності та гнучкості. Він демонструє значну ефективність у задачах з нерівномірно розподіленими, нелінійно

відокремлюваними та високорозмірними даними, що робить його універсальним інструментом для широкого спектра прикладних досліджень.

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

$$f_G^{x(\bullet)}(x) = \exp\left(-\frac{d^2(x, x(\bullet))}{2\sigma^2}\right) = \exp\left(-\frac{\|x - x(\bullet)\|^2}{2\sigma^2}\right) \qquad (5.10)$$

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 Флобеніуса, при цьому функція впливу набуває вигляду

$$f_G^{x(\bullet)}(x) = \exp\left(-\frac{d^2(x, x(\bullet))}{2\sigma^2}\right) = \exp\left(-\frac{Tr\left(x - x(\bullet)\right)\left(x - x(\bullet)\right)^T}{2\sigma^2}\right), \quad (5.11)$$

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

Нескладно бачити, що (5.11) є узагальненням (5.10).

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

$$f^x(x) = \sum_{k=1}^{N} f\left(x, x(k)\right), \tag{5.12}$$

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 (5.12) за допомогою градієнтної процедури

$$x^l = x^{l-1} + \eta^l \frac{\nabla f^x\left(x^l, x^{l-1}\right)}{\left\|\nabla f^x\left(x^l, x^{l-1}\right)\right\|}, \quad x_0 = x\left(k\right), l = 0,1,2,...; \forall k = 1,2,...,N, \tag{5.13}$$

тобто кількість запусків алгоритму (5.13) 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 (5.12) шляхом модифікації градієнтної процедури (5.13).

У випадку коли спостереження $x(k)$ у вибірці $X \in (n_1 \times n_2)$ - матрицями, нескладно ввести у розгляд матричний варіант процедури (5.13):

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

де $\Gamma^x \left(x, x^{l-1}\right) = \left\{\dfrac{\partial f^x \left(x, x^{l-1}\right)}{\partial x_{i_1 i_2}}\right\} \in X^{n_1 \times n_2}$.

Процес градієнтної оптимізації закінчується відшуканням m локальних екстремумів функції (5.12), 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.

5.5 Адаптивна нечітка кластеризація викривлених даних на основі стратегії найближчого прототипу-центроїда з використанням еволюційних процедур

Аналіз і обробка викривлених даних є однією з ключових проблем сучасного машинного навчання та інтелектуального аналізу даних. Викривлені дані можуть виникати внаслідок шуму, аномалій, нерівномірного розподілу або складних нелінійних залежностей, що ускладнює їх кластеризацію традиційними методами. Для вирішення цієї проблеми перспективним є застосування адаптивної нечіткої кластеризації, яка дозволяє гнучко моделювати невизначеність у даних та враховувати їхню неоднорідну

природу. Одним із ефективних підходів є стратегія кластеризації на основі найближчого прототипу-центроїда, що поєднується з еволюційними процедурами для оптимізації параметрів кластерної структури.

Основна ідея такого підходу полягає у використанні нечіткої кластеризації, де кожен об’єкт даних не належить жорстко до одного кластера, а має певний ступінь приналежності до різних груп. Це дозволяє забезпечити більш реалістичну інтерпретацію кластерних структур, особливо у випадках, коли дані мають складні викривлені розподіли. Одним із найефективніших способів визначення кластерних центрів є стратегія найближчого прототипу-центроїда, що передбачає поступове оновлення положення центрів кластерів на основі аналізу відстані між об’єктами та їхньої належності до кластерів.

Однак, у випадку викривлених даних, де традиційні метрики відстані, такі як евклідова або мангеттенська, можуть бути неефективними, стратегія найближчого прототипу-центроїда має бути адаптована шляхом використання гнучких механізмів оптимізації. Для цього застосовуються еволюційні процедури, які забезпечують адаптивний пошук оптимального розташування центрів кластерів у багатовимірному просторі. Використання еволюційних алгоритмів дозволяє уникнути проблеми передчасної конвергенції та забезпечити більш точне розбиття даних, особливо у випадках, коли викривлення кластерів робить їхню форму нерегулярною або анізотропною.

Еволюційні алгоритми, такі як генетичні алгоритми, диференціальна еволюція або ройові алгоритми, можуть ефективно використовуватися для налаштування параметрів нечіткої кластеризації. У цьому підході кожен потенційний набір кластерних центрів розглядається як особина популяції, а операції мутації, схрещування та відбору сприяють поступовому вдосконаленню кластерної структури. Функція пристосованості у такій системі зазвичай визначається на основі якості кластеризації, наприклад, за допомогою мінімізації сумарної внутрішньокластерної варіації або максимізації правдоподібності моделі.

Однією з головних проблем при кластеризації викривлених даних є вибір оптимального числа кластерів. У традиційних методах це число встановлюється апріорі, що може призвести до субоптимальних рішень. Використання еволюційного підходу дозволяє адаптивно визначати оптимальну кількість кластерів, аналізуючи зміну функції пристосованості при різних розмірах кластерної структури. Це робить підхід стійким до складних випадків, коли кластери можуть мати неоднорідну густину або перетинатися у просторі ознак.

Важливим аспектом запропонованого методу є його здатність до роботи з високорозмірними даними. Оскільки традиційні методи кластеризації часто страждають від так званого "прокляття розмірності", де відстані між об’єктами стають менш інформативними, запропонована стратегія використовує адаптивні метрики відстані, які можуть коригуватися в процесі еволюційного пошуку. Це дозволяє ефективно аналізувати складні багатовимірні розподіли та забезпечувати високу точність кластеризації навіть у випадках, коли дані містять значну кількість корельованих ознак.

Крім того, важливою перевагою такого підходу є його здатність працювати в динамічних умовах, тобто в умовах, коли нові дані надходять у реальному часі. Використання механізму еволюційної адаптації дозволяє оновлювати структуру кластерів без необхідності повного перерахунку всієї моделі, що робить цей метод придатним для застосування у потоковій кластеризації. Це є особливо актуальним для завдань кібербезпеки, аналізу поведінки користувачів, медичної діагностики та фінансового прогнозування, де необхідно обробляти великі обсяги даних в режимі реального часу.

Попри численні переваги, адаптивна нечітка кластеризація на основі стратегії найближчого прототипу-центроїда та еволюційних процедур має певні виклики. Один із них полягає у виборі оптимальних параметрів еволюційного алгоритму, таких як ймовірність мутації, розмір популяції та кількість поколінь. Неправильне налаштування цих параметрів може призвести до збільшення часу обчислень або зниження точності кластеризації.

Крім того, обчислювальна складність такого підходу є вищою, ніж у традиційних методів, через необхідність багаторазового виконання еволюційних операцій.

Отже, адаптивна нечітка кластеризація викривлених даних на основі стратегії найближчого прототипу-центроїда з використанням еволюційних процедур є потужним підходом, що забезпечує гнучкість, точність і адаптивність у розв'язанні задач групування складних даних. Поєднання методів правдоподібного оцінювання кластерної структури з глобальним пошуком оптимальних параметрів дозволяє значно підвищити ефективність кластеризації в умовах невизначеності, шуму та нелінійних розподілів. Подальші дослідження можуть бути зосереджені на оптимізації обчислювальної ефективності алгоритму, розробці гібридних стратегій адаптивного налаштування параметрів та застосуванні методів паралельних обчислень для обробки великих обсягів даних.

Вихідною інформацією є дані, що представлені у вигляді $(N \times n)$ таблиці «об'єкт-властивість» яка містить інформацію про N об'єктів, описаних у вигляді $(1 \times n)$ векторів-ознак. Результатом кластеризації вихідних даних є розбиття початкової вибірки на m класів з відповідним рівнем нечіткої належності $\mu_q(k)$ k-того вектора-спостереження до q-го кластера, де $1 \le q \le m$. Вихідні дані заздалегідь нормуються в гіперкуб $[-1;1]^n$.

Стратегія найближчого прототипу-центроїда може бути 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 спотворених спостережень шляхом знаходження прототипу $c_q^{(\tau+1)}$ найближчого до $a(k)$ в сенсі часткової відстані [86]

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

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 можна записати стратегію найближчого прототипу у рекурентній формі [32, 123]

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

Можливісна стратегія найближчого прототипу-центроїда у загублених спостереженнях може бути записана у вигляді послідовності 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:

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

3. Розрахунок центроїдів кластерів:

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

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

6. Розрахунок скалярного параметра відстані

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

7. Далі йти до кроку 2.

Аналогічно імовірнісній адаптивній кластеризації на основі стратегії найближчого центроїда можна організувати процес можливісної кластеризації у вигляді [32, 124, 135]

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

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 алгоритми рою частинок [174-212]. Одним з найшвидших алгоритмів рою частинок є, так званий, алгоритм 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.

5.6 Апробація онлайн методу для правдоподібної нечіткої кластеризації на основі еволюційної оптимізації котячої зграї

Онлайн-метод правдоподібної нечіткої кластеризації на основі еволюційної оптимізації котячого рою (OCrCSO) було виконано на 2 вибірках даних. Кластеризація вихідних даних за допомогою нечітких С-середніх, кластеризації даних адаптивного правдоподібного нечіткого методу, Густафсона-Кесселя та OCrCSO. Результати кластеризації представлені в таблиці 5.1

Таблиця 5.1 – Оцінка якості даних методів нечіткої кластеризації

Методи кластеризації	**PC**	**SC**	**XB**
Нечіткі С-середніх	0,50	1,62	0,19
Густафсона-Кесселя	0,27	1,66	1,62
Адаптивна нечітка правдоподібна кластеризація даних	0,26	1,22	**0,01**
Онлайн-метод правдоподібної нечіткої кластеризації на основі еволюційної оптимізації котячого рою (OCrCSO)	**0,24**	**0,69**	0,15

Також було проведено порівняльний аналіз якості даних кластеризації за основними характеристиками рейтингів якості, такими як: коефіцієнт розподілу (PC), індекс розподілу (SC), індекс Се та Бені (XB) існуючих методів кластеризації та запропонованого методу. Як видно з результатів експериментів, запропонований алгоритм показує досить хороші результати якості кластеризації.

5.7 Апробація методу глобальної оптимізації божевільної котячої зграї в задачі нечіткої кластеризації

Щоб перевірити ефективність та оцінити працездатність та спроможність якісно кластеризувати великі обсяги даних розробленого методу, а також довести його перевагу над аналогами, було проведено експериментальне дослідження за допомогою чотирьох різних навчальних вибірок, а саме Іриси, Рак, Вина та Скло [246-261].

В таблиці 5.2 наведено порівняльні результати показників ефективності таких алгоритмів, як PSO, CSO та запропонованого методу правдоподібної нечіткої кластеризація даних на основі еволюційного методу божевільних котів.

Таблиця 5.2 - Порівняльні результати показників ефективності таких алгоритмів, як PSO, CSO

Нав-чальна вибірка	**MSE**	**PSO**	**CSO**	**Правдоподібна нечітка кластеризація даних на основі еволюційного методу божевільних котів**
Іриси	Найкраще	4×10^{-7}	9×10^{-10}	$8{,}4\times10^{-11}$
	Середнє	7×10^{-6}	$7{,}3\times10^{-9}$	$5{,}2\times10^{-10}$
	Найгірше	$1{,}2\times10^{-5}$	$9{,}3\times10^{-9}$	$9{,}6\times10^{-10}$
Рак	Найкраще	$1{,}3\times10^{-7}$	8×10^{-10}	$8{,}5\times10^{-11}$
	Середнє	7×10^{-6}	$4{,}4\times10^{-9}$	$7{,}6\times10^{-11}$
	Найгірше	$2{,}02\times10^{-5}$	$6{,}8\times10^{-9}$	$7{,}8\times10^{-10}$
Вина	Найкраще	$1{,}4\times10^{-6}$	7×10^{-10}	$9{,}5\times10^{-11}$
	Середнє	5×10^{-5}	4×10^{-9}	$6{,}6\times10^{-11}$
	Найгірше	2×10^{-5}	6×10^{-9}	$6{,}8\times10^{-10}$
Скло	Найкраще	$2{,}5\times10^{-7}$	$7{,}9\times10^{-10}$	$8{,}7\times10^{-11}$
	Середнє	$6{,}9\times10^{-6}$	5×10^{-9}	$7{,}7\times10^{-11}$
	Найгірше	3×10^{-5}	$5{,}9\times10^{-9}$	$7{,}7\times10^{-10}$

Аналізуючи та оцінюючи отримані результати експериментальних досліджень, можна зробити висновки, що запропонований метод правдоподібної нечіткої кластеризації даних на основі еволюційного методу божевільних котів забезпечує достатньо якісні результати кластеризації, що підтверджується експериментально.

Аналізуючи Таблицю 5.2, можна сказати, що запропонований метод, демонструє найкраще рішення поставленої задачі, за рахунок використання теорії правдоподібності, нечіткої кластеризації та запропонованого методу глобальної оптимізації божевільної котячої зграї, а ефективність запропонованих методів на 7-8% краща, за відомі методи та алгоритми.

5.8 Апробація методу кластеризації масивів даних на основі комбінованої оптимізації функцій щільності розподілу та еволюційного методу котячих зграй

Дослідження методу кластеризації масивів даних на основі комбінованої оптимізації функцій щільності розподілу та еволюційного методу котячих зграй проводились на трьох навчальних вибірках: База спаму, Блоки сторінок, Іриси та Ecoli [261-266].

Таблиця 5.3 - Характеристики наборів даних

Вибірка	**Кількість спостережень**	**Кількість атрибутів**
База спаму	4601	57
Блоки сторінок	5472	10
Іриси	150	4
Вина	178	13
Ecoli	336	8

Якість роботи метода кластеризації даних на основі піків щільності розподілу даних та еволюційного методу котячих зграй (Proposed method - PM) перевірялась за допомогою основних оцінок якості кластеризації. Існує кілька метрик для оцінки якості кластеризації. Всі метрики, що використовуються для оцінки запропонованого методу базуються на так званому методі оцінювання за допомогою золотого стандарту(golden set).

1. Метрика чистоти кластеризації (purity - Pur). Для обчислення даного показника кожному кластеру присвоюється клас, з яким у кластера максимальне перекриття по привласненим об’єктам. Після присвоєння міток класів обчислюється правильність даної кластеризації як число об’єктів класу, з яким асоційований кластер, поділене на загальне число об'єктів в кластері. У цьому сенсі дана метрика схожа на показник точності класифікації.

2. Метрика нормованої взаємної інформації (normalized mutual information - NMI). Дана метрика заснована на понятті ентропії.

3. Коефіцієнт Ренда (rand index - RI). Даний підхід до оцінки якості кластеризації перегукується з методами оцінки якості алгоритмів класифікації.

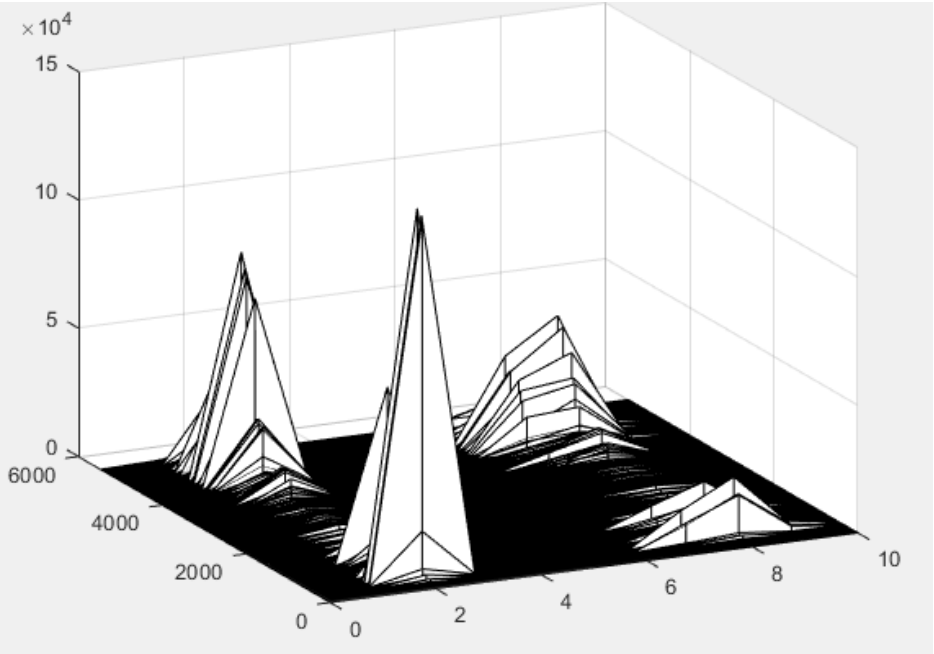


Рисунок 5.1 - Навчальна вибірка Page Blocks

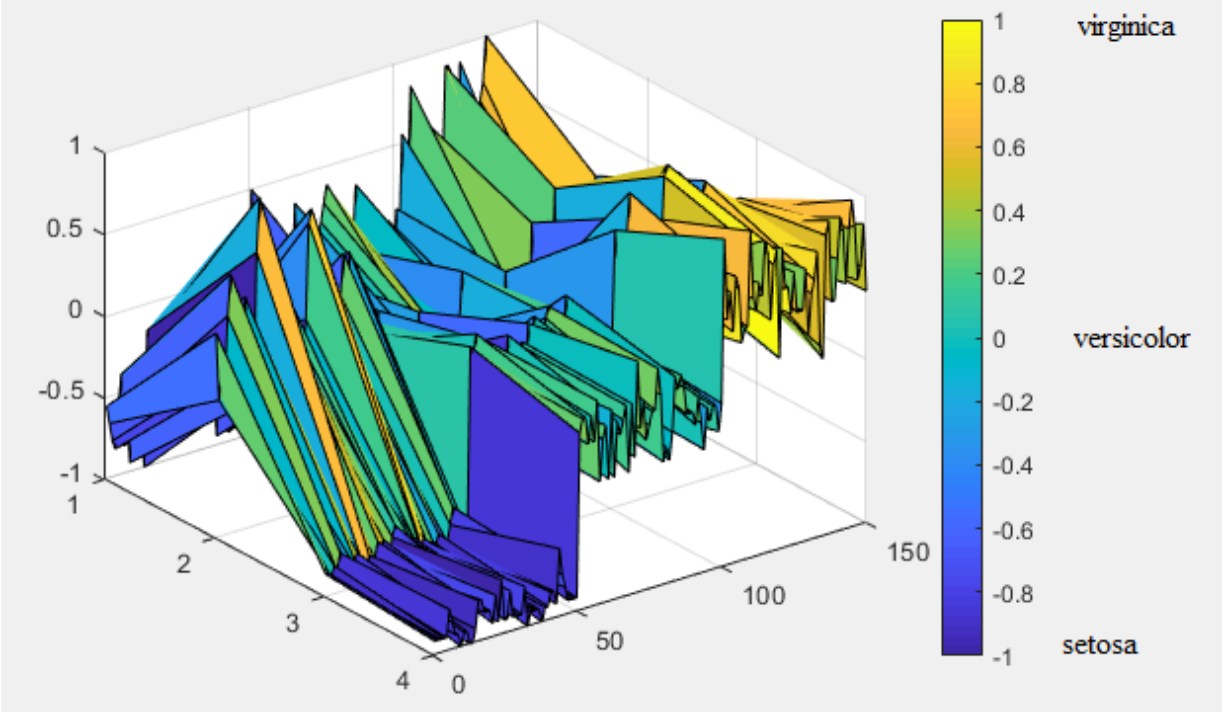


Рисунок 5.2 - Навчальна вибірка Iris

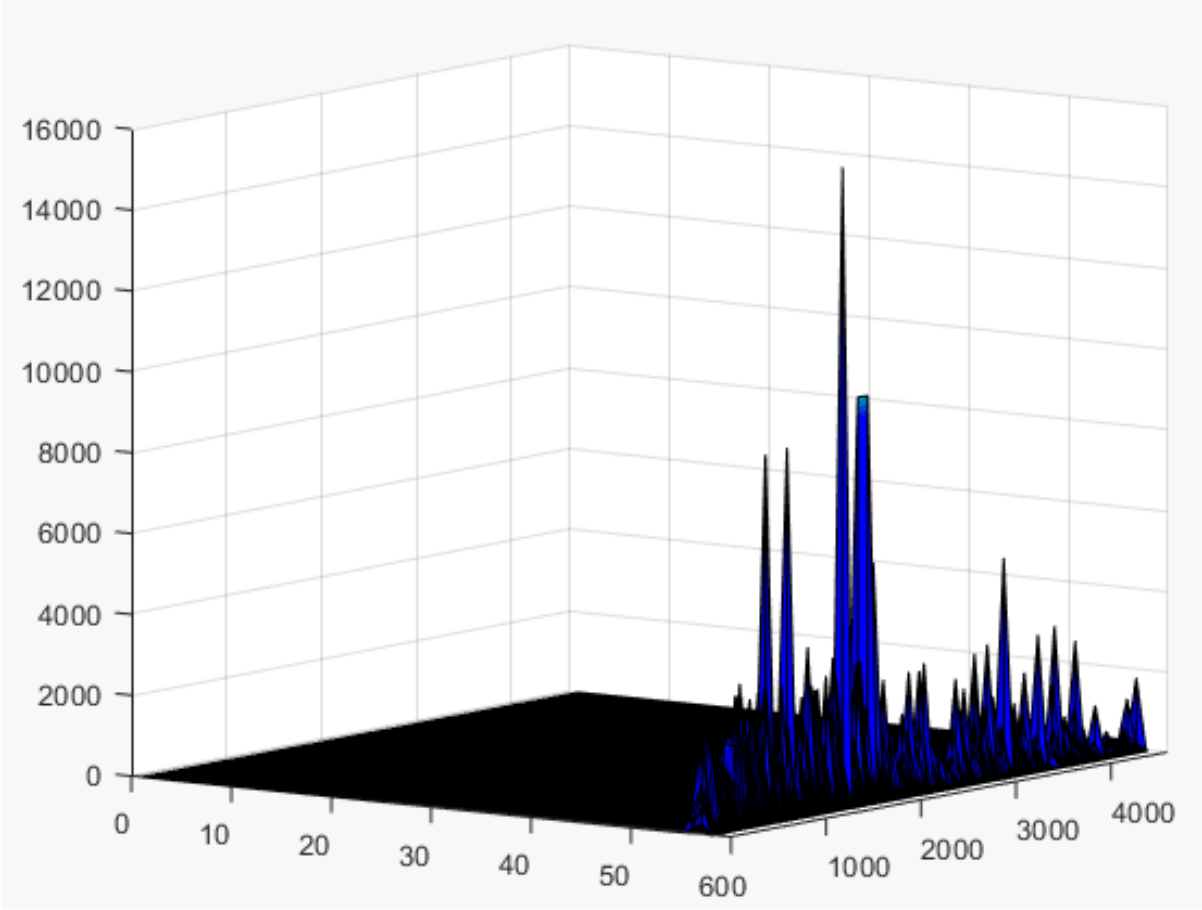


Рисунок 5.3 - Навчальна вибірка Spambase

Порівняльний аналіз проводився з відомими методами кластеризації даних, такими як FCM, DBSCAN та CLARA.

Результати кластерного аналізу даних на вибірках Page block, Wine, Iris, Ecoli та Spambase, за показниками оцінки якості кластеризації наведено на рисунках нижче.

На гістограмах продемонстровані результати кластерного аналізу за якими можна зробити висновок, що запропонований метод кластеризації даних на основі піків щільності розподілу даних та еволюційного методу котячих зграй дає оцінку кластеризації вище, ніж більш відомі методи кластеризації завдяки оптимізаційній процедурі еволюційного алгоритму.

Крім аналізу якості кластеризації даних, потрібно оцінити швидкість роботи методу. Якість методу кластеризації повинна відповідати швидкості і простоти з точки зору математичних розрахунків.

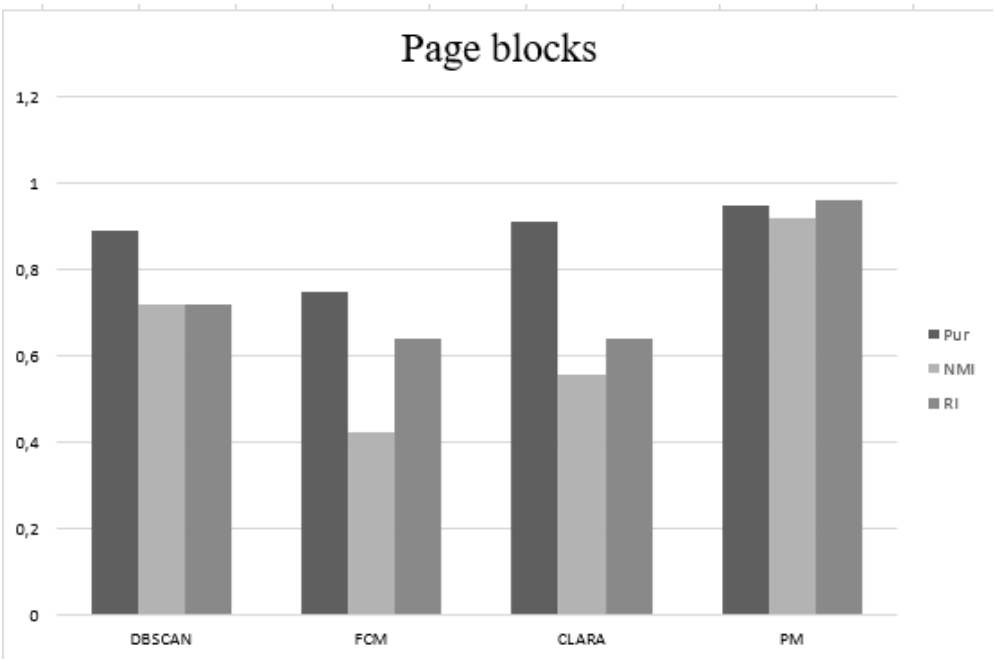


Рисунок 5.4 - Показники якості кластеризації вибірки Page Blocks

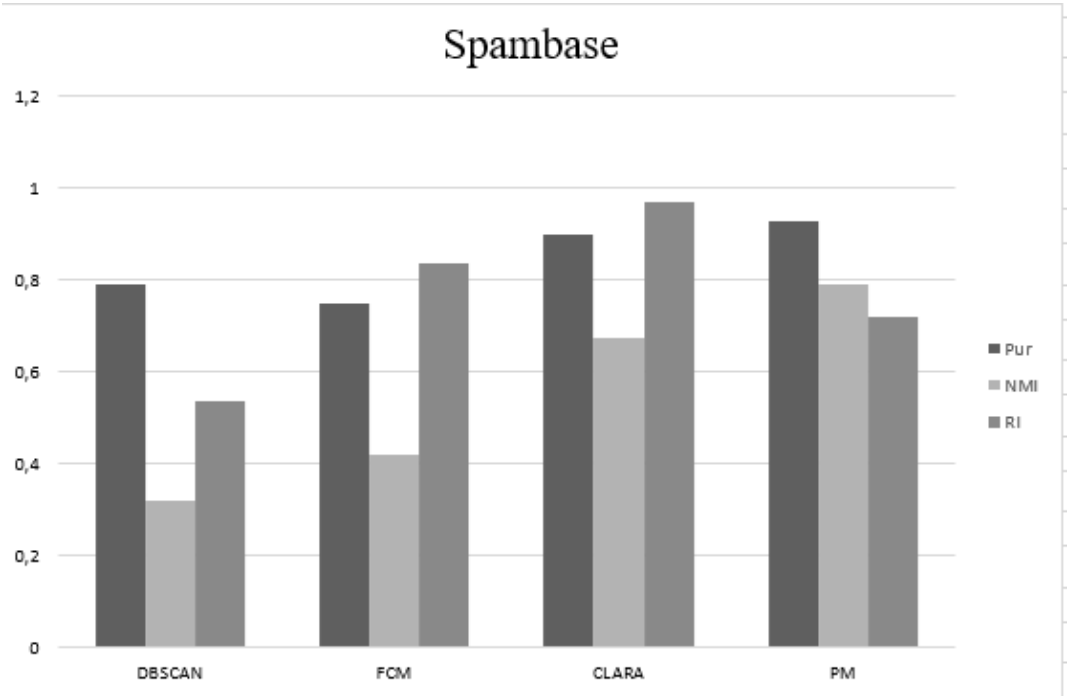


Рисунок 5.5 - Показники якості кластеризації вибірки База спаму

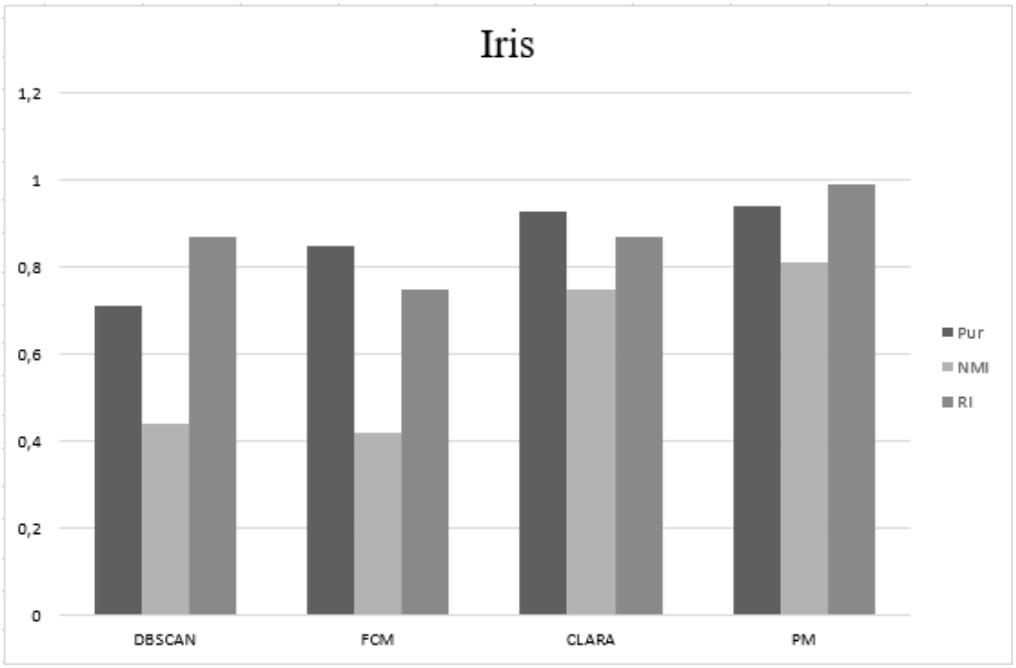


Рисунок 5.6 - Показники якості кластеризації вибірки Іриси

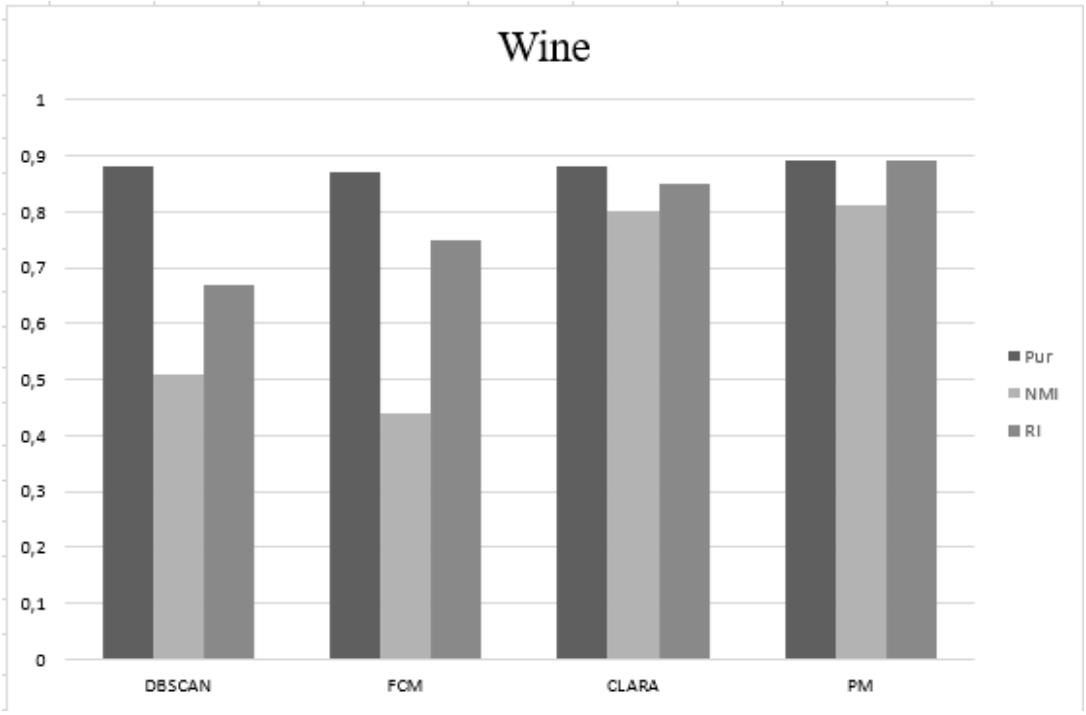


Рисунок 5.7 - Показники якості кластеризації вибірки Вина

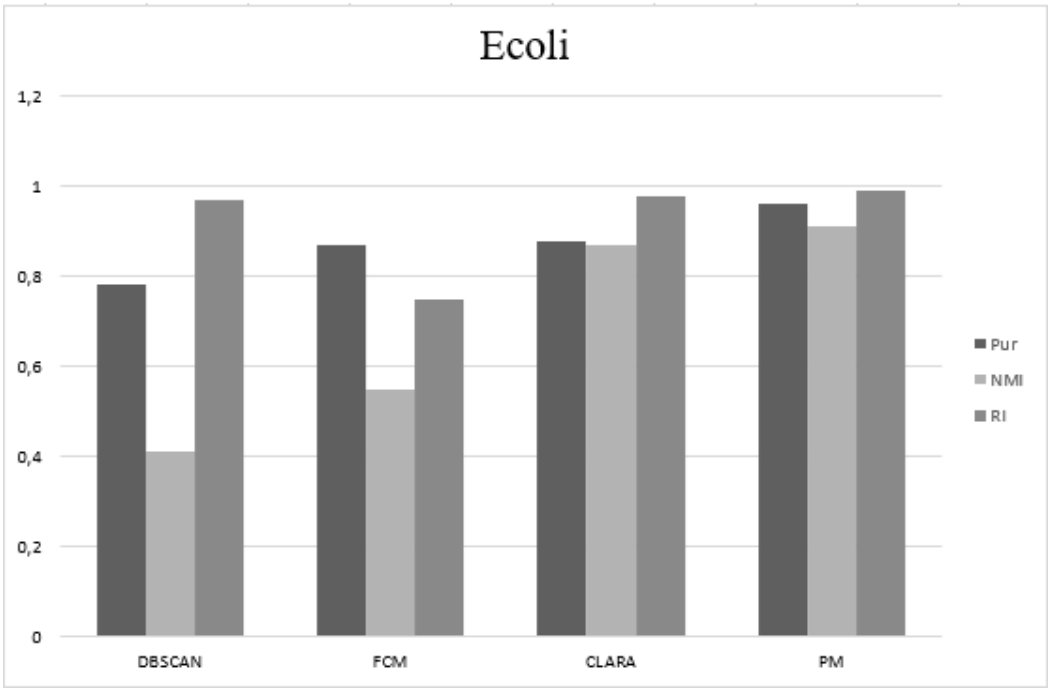


Рисунок 5.8 - Показники якості кластеризації вибірки Ecoli

Проведено аналіз методу кластеризації даних на основі піків щільності розподілу даних та еволюційного методу котячих зграй на 100 спостереженнях різних вибірок даних.

На рисунках, що представлені нижче наведений порівняльний результат швидкості роботи методів кластеризації.

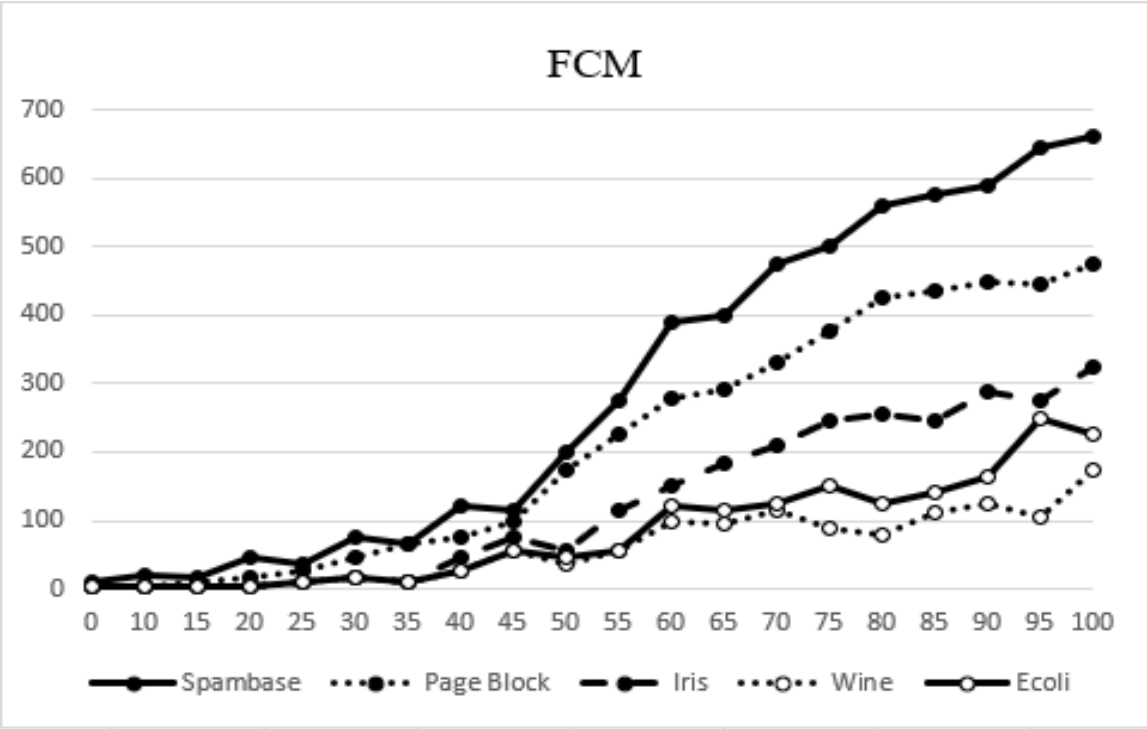


Рисунок 5.9 – Швидкість роботи FCM (в мсек.)

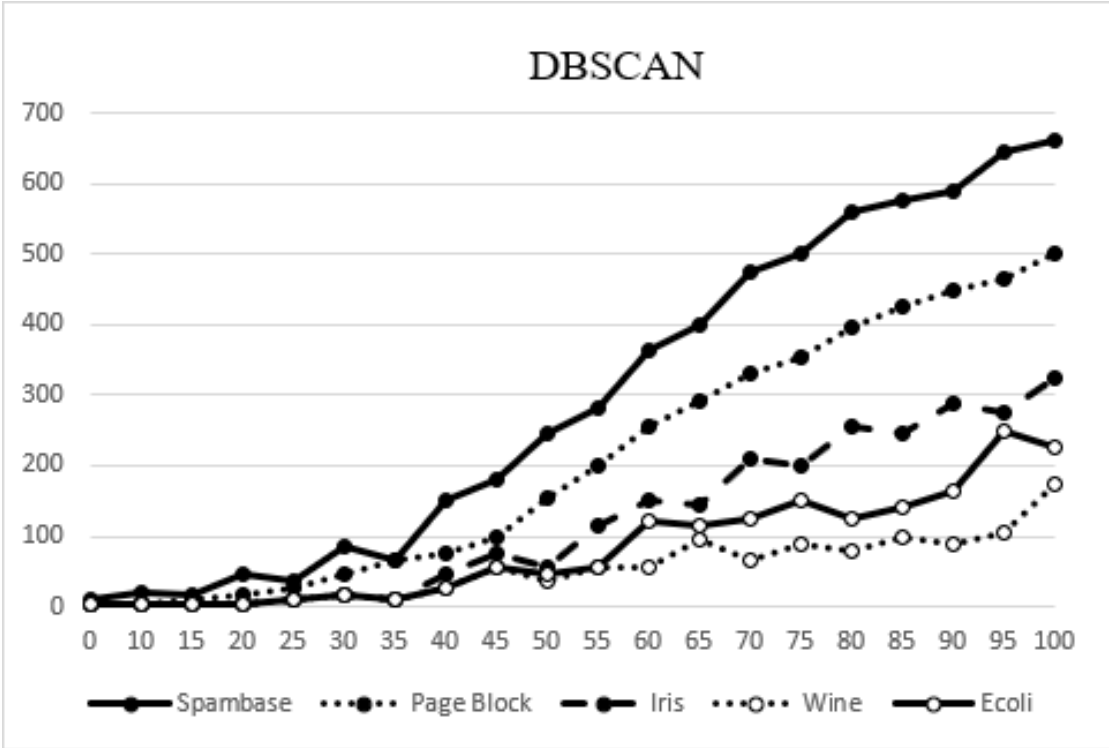


Рисунок 5.10 - Швидкість роботи DBSCAN (в мсек.)

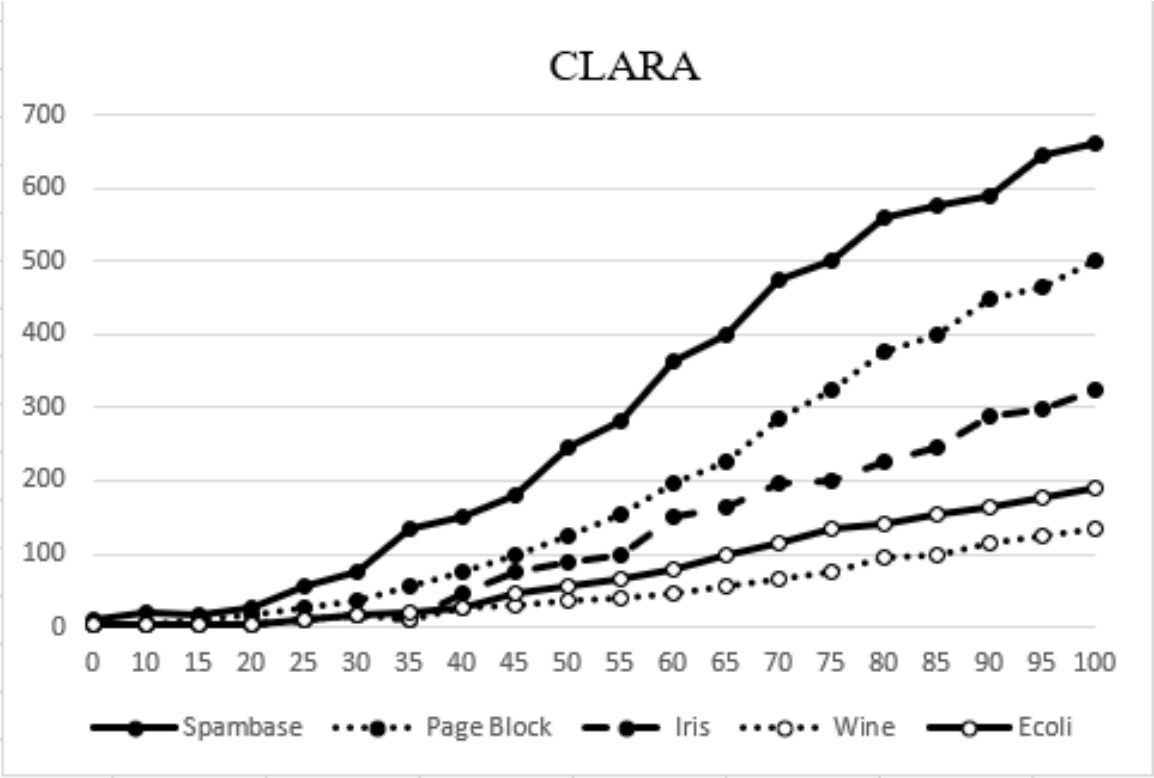


Рисунок 5.11 - Швидкість роботи CLARA (в мсек.)

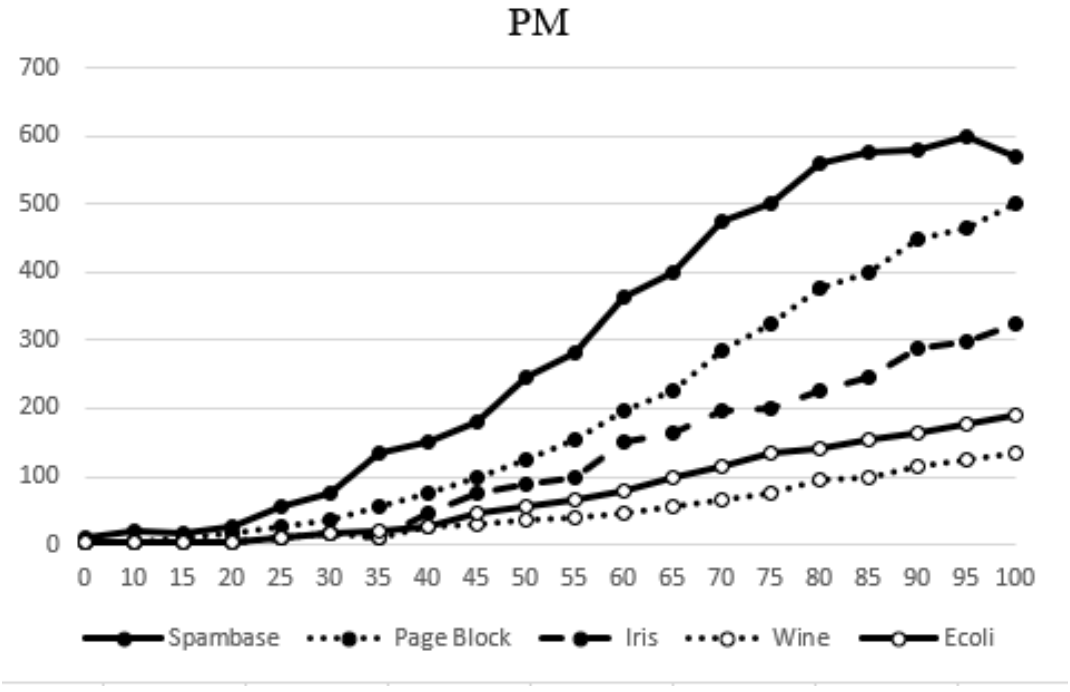


Рисунок 5.12 - Швидкість роботи PM (в мсек.)

Аналізуючи результати отриманих експериментальних досліджень, що проводились на п’яти різної природи даних із вибірок UCI репозиторію та

порівняльного аналізу роботи методу кластеризації даних на основі піків щільності розподілу даних та еволюційного методу котячих зграй із методами кластеризації, що базуються як на класичному підході до кластеризації даних, так і більш екзотичних: DBSCAN та CLARA, що використовують аналіз щільності даних які підлягають кластеризації, запропонований метод демонструє достатньо високі результати.

Основними перевагами запропонованого методу полягає в простоті математичних розрахунків, швидкості роботи з даними, незалежно від виду, розміру та якості вибірки, що аналізується. Порівняльний результат швидкості роботи методів кластеризації експериментальних досліджень наведений на графіках, які наочно демонструють швидкість роботи методів на різних вибірках. Слід відзначити точність роботи метода кластеризації даних на основі піків щільності розподілу даних та еволюційного методу котячих зграй та отриманих результатів кластеризації, що досягається за допомогою оптимізаційної процедури еволюційного алгоритму. Як видно із рисунків 4-8, показники якості кластеризації на різних вибірках запропонованого методу достатньо високий, незалежно від метрики, що використовуються для оцінки методів кластеризації.

5.9 Апробація методу адаптивної нечіткої кластеризації викривлених даних на основі стратегії найближчого прототипу-центроїда з використанням еволюційних процедур

Експериментальні дослідження запропонованого методу адаптивної нечіткої кластеризації викривлених пропусками та викидами даних на основі стратегії найближчого прототипу - центроїду з використанням еволюційних процедур було проведено на чотирьох різних вибірках даних, які були штучно пошкоджені викидами та пропусками. В Таблиці 5.4, наведено характеристики вибірок та кількість пошкоджених даних у відсотках (%).

Порівняльні експерименти запропонованого методу адаптивної нечіткої кластеризації викривлених пропусками та викидами даних на основі стратегії найближчого прототипу-центроїду з використанням еволюційних процедур проводились з більш відомими алгоритмами кластеризації такими як алгоритми *k*-середніх та *k*-прототипів та вимірялись за чотирма характеристиками: F-Measure, Rand Index, Jaccard Index і Ентропія.

Усі ці чотири показники мають значення від 0 до 1. В F-Measure, Rand Index та Jaccard Index значення одиниці вказує що кластери даних абсолютно однакові, а збільшення значень цих показників свідчить на кращу продуктивність.

Таблиця 5.4 - Характеристики вибірок та кількість пошкоджених даних у відсотках (%).

Вибірка	**Кількість кластерів**	**Кількість атрибутів**	**Кількість спостережень**	**Кількість викривлених даних (%)**
Hepatitis	2	19	155	10
Cancer	2	9	683	50
Stat Log Heart	2	13	270	25
Post Operative Patient	6	8	214	5

В Таблицях 5.5, 5.6, 5.7 наведено результати порівняльної роботи відомих методів кластеризації даних із запропонованим методом адаптивної нечіткої кластеризації викривлених пропусками та викидами даних основі стратегії найближчого прототипу - центроїду з використанням еволюційних процедур (AFC_PCEP).

Як видно із порівняльних таблиць, запропонований метод демонструє достатньо високі показники, незалежно від вибірки та якості даних на відміну

від більш відомих методів кластеризації даних, показник якого найближче до одиниці, що само по собі свідчить про високу якість кластеризації даних.

Таблиця 5.5 - Порівняльні результати методів за характеристикою F-Measure

Вибірка	**K-Means**	**K-Prototype**	**AFC_PCEP**
Hepatitis	0,75	0,86	0,88
Cancer	0,75	0,84	0,86
Stat Log Heart	0,77	0,88	0,89
Post Operative Patient	0,78	0,87	0,88

F-Measure (або F-score, F1-score) - це метрика оцінки якості кластеризації та класифікації, що поєднує точність (precision) і повноту (recall) у єдину числову характеристику. Ця метрика широко використовується для оцінки продуктивності алгоритмів машинного навчання, зокрема у випадках, коли важливо досягти балансу між правильністю класифікації та здатністю алгоритму виявляти всі релевантні об’єкти.

За результатами порівняльного аналізу за показником F-Measure, який враховує як точність, так і повноту кластеризації, усі три методи демонструють суттєві відмінності в якості кластеризації для різних вибірок даних. Для набору даних «Hepatitis» метод K-Means отримав F-Measure 0,75, що є нижчим за значення 0,86 для методу K-Prototype і 0,88 для методу AFC_PCEP. Схожа ситуація спостерігається і для інших вибірок: для «Cancer» відповідні значення складають 0,75 (K-Means), 0,84 (K-Prototype) та 0,86 (AFC_PCEP); для «Stat Log Heart» – 0,77, 0,88 і 0,89; а для «Post Operative Patient» – 0,78, 0,87 і 0,88. Таким чином, метод K-Means стабільно демонструє нижчі показники F-Measure, що свідчить про його менш ефективний поділ даних у порівнянні з іншими методами.

Методи K-Prototype і AFC_PCEP забезпечують значно кращу якість кластеризації, зокрема, метод AFC_PCEP майже завжди досягає найвищого значення F-Measure (від 0,86 до 0,89), що вказує на його здатність більш точно розподіляти дані між кластерами. Невеликі переваги AFC_PCEP над K-Prototype (покращення на 0,02–0,03) можуть свідчити про додаткові механізми адаптації або інтеграції ознак, які дозволяють цьому методу краще враховувати особливості даних. Загалом, результати демонструють, що використання адаптивних підходів у кластеризації (AFC_PCEP) є доцільним, оскільки вони суттєво перевершують традиційний K-Means за показником F-Measure, що є важливим критерієм для визначення якості кластерного розподілу даних.

Таким чином, аналізуючи отримані результати за відповідною метрикою можна зробити висновок, що метод адаптивної нечіткої кластеризації викривлених даних на основі стратегії найближчого прототипу-центроїда з використанням еволюційних процедур (AFC_PCEP) в порівнянні з відомими методами кластеризації даних демонструє кращій результат розбиття даних на кластери.

В таблиці 5.6 наведені порівняльні результати методів за характеристикою Rand Index.

Таблиця 5.6 - Порівняльні результати методів за характеристикою Rand Index

Вибірка	**K-Means**	**K-Prototype**	**AFC_PCEP**
Hepatitis	0,72	0,73	0,74
Cancer	0,53	0,56	0,62
Stat Log Heart	0,56	0,58	0,59
Post Operative Patient	0,41	0,45	0,48

Rand Index (RI) - це міра подібності між двома методами кластеризації одного набору даних. Вона використовується для оцінки якості алгоритму кластеризації шляхом порівняння отриманих кластерів із еталонним (істинним) розбиттям. Rand Index належить до категорії показників зовнішньої оцінки кластеризації, оскільки передбачає знання еталонного розбиття (ground truth).

Для вибірки Hepatitis значення Rand Index складають 0,72 для K-Means, 0,73 для K-Prototype та 0,74 для AFC_PCEP. Незначна перевага AFC_PCEP (0,74) над іншими методами свідчить про трохи кращу здатність цього алгоритму точно відтворювати структуру даних. У вибірці Cancer різниця стає більш вираженою: K-Means досягає 0,53, K-Prototype – 0,56, а AFC_PCEP – 0,62, що свідчить про значне покращення точності кластеризації при використанні AFC_PCEP. Для Stat Log Heart спостерігаються схожі тенденції – K-Means отримує 0,56, K-Prototype – 0,58, а AFC_PCEP – 0,59, тобто зростання значень є помірним. У випадку вибірки Post Operative Patient, яка має найнижчі абсолютні показники (K-Means – 0,41, K-Prototype – 0,45, AFC_PCEP – 0,48), перевага AFC_PCEP залишається, що свідчить про його здатність краще розділяти дані навіть при наявності складних або менш чітко виражених кластерів.

Таким чином, аналіз даних за характеристикою Rand Index вказує на те, що метод AFC_PCEP забезпечує найвищу точність кластеризації на всіх розглянутих вибірках, демонструючи переваги у порівнянні з традиційними методами K-Means та K-Prototype. Найбільш помітна перевага AFC_PCEP спостерігається для вибірки Cancer (0,62 проти 0,53–0,56) та для Post Operative Patient (0,48 проти 0,41–0,45), що свідчить про ефективність адаптивного підходу для даних, де структура кластерів є більш складною або нечіткою.

Таблиця 5.7 - Порівняльні результати методів за характеристикою Jaccard Index

Вибірка	**K-Means**	**K-Prototype**	**AFC_PCEP**
Hepatitis	0,62	0,63	0,65
Cancer	0,45	0,46	0,48
Stat Log Heart	0,54	0,56	0,71
Post Operative Patient	0,33	0,35	0,38

Наведена таблиця демонструє порівняльні результати трьох методів кластеризації – K-Means, K-Prototype та AFC_PCEP – за характеристикою Jaccard Index для вибірок даних «Hepatitis», «Cancer», «Stat Log Heart» та «Post Operative Patient». Jaccard Index вимірює ступінь відповідності між отриманими кластерами та еталонною класифікацією, причому вищі значення індексу свідчать про кращу якість кластеризації за критерієм подібності, оскільки відношення перетину до об’єднання кластерних розподілів є більшим.

За вибіркою «Hepatitis» значення Jaccard Index складають 0,62 для K-Means, 0,63 для K-Prototype та 0,65 для AFC_PCEP, що свідчить про поступове покращення якості кластеризації при переході від традиційних методів до адаптивного підходу. Схожа тенденція спостерігається для вибірки «Cancer», де показники становлять 0,45, 0,46 і 0,48 відповідно. Особливо виразну перевагу AFC_PCEP демонструє вибірка «Stat Log Heart», де індекс для цього методу досягає 0,71, що значно перевищує показники 0,54 (K-Means) та 0,56 (K-Prototype). Для вибірки «Post Operative Patient» значення Jaccard Index також послідовно зростають від 0,33 (K-Means) до 0,35 (K-Prototype) і 0,38 (AFC_PCEP). Загалом, результати таблиці свідчать про те, що адаптивний метод AFC_PCEP забезпечує більш точну класифікацію даних, відображену у

вищих значеннях Jaccard Index на всіх розглянутих вибірках, що підтверджує його переваги у порівнянні з традиційними підходами.

Таблиця 5.8 - Порівняльні результати методів за ентропією

Вибірка	***K*-Means**	***K*-Prototype**	**AFC_PCEP**
Hepatitis	0,52	0,52	0,52
Cancer	0,45	0,43	0,45
Stat Log Heart	0,45	0,44	0,43
Post Operative Patient	0,42	0,41	0,40

Наведена таблиця демонструє результати порівняльного аналізу трьох методів кластеризації – K-Means, K-Prototype та AFC_PCEP – за показником ентропії для вибірок даних «Hepatitis», «Cancer», «Stat Log Heart» та «Post Operative Patient». Ентропія в кластеризації відображає ступінь невизначеності розподілу об’єктів між кластерами: чим нижче значення, тим чіткіше сформовані кластери і тим менша неоднозначність при визначенні належності об’єктів до кластерів.

Для вибірки «Hepatitis» усі три методи демонструють однакове значення ентропії – 0,52, що свідчить про еквівалентну якість кластеризації з точки зору визначеності кластерних структур. У вибірці «Cancer» спостерігається невелика перевага методу K-Prototype, який має значення ентропії 0,43, тоді як K-Means та AFC_PCEP отримують трохи вищі значення – 0,45. Аналогічна тенденція простежується для вибірки «Stat Log Heart»: тут AFC_PCEP демонструє найнижче значення ентропії – 0,43, порівняно з 0,44 для K-Prototype та 0,45 для K-Means. У вибірці «Post Operative Patient» значення ентропії зменшуються – 0,40 для AFC_PCEP, 0,41 для K-Prototype і 0,42 для K-Means.

Таким чином, загальний аналіз показує, що всі розглянуті методи кластеризації працюють досить подібно за показником ентропії, проте

невеликі переваги AFC_PCEP, зокрема при вибірках «Stat Log Heart» і «Post Operative Patient», свідчать про його здатність формувати дещо більш визначені кластери. Це може бути важливим у тих випадках, коли критично необхідно мінімізувати невизначеність кластерного розподілу даних.

5.10 Апробація методу кластеризації масивів даних на основі модифікованого методу сірого вовка

Дослідження методу кластеризації масивів даних на основі покращеного методу сірого вовка (FGWO) проводились на двох багатоекстремальних функціях, наведених в Таблиці 1.

Якість роботи запропонованого метода (FGWO) порівнювалось із декількома класичними алгоритмами кластеризації, еволюційними процедурами, а також модифікованими методами кластеризації на основі оптимізаційних процедур, а саме алгоритм оптимізації рою частинок (PSO), алгоритм зграї котів (CSO), класичний метод сірого вовка (GWO) та модифікованого методу кластеризації на основі зграї котів (FCSO). Для кожного метода, задано 30 агентів, що шукають оптимум в багатоекстремальній функції. Перш за все перевіримо роботу запропонованого метода з його модифікацією, тобто використання ваг для кожного вовка.

Таблиця 5.14 - Тестові функції

Назва функції	**Формула**	**Інтервал**
Растрігін	$f(x) = 20 + x^2 + y^2 - $ $-10\cos(2\pi x) + \cos(2\pi y)$	$[-5{,}12; 5{,}12]$

Продовження таблиці 5.14

Гріванг	$f(x)=\frac{1}{4000}x+\frac{1}{4000}y-$ $-\cos\left(\frac{x}{\sqrt{1}}\right)\cos\left(\frac{x}{\sqrt{2}}\right)+1$	[–30;30]

Результат зміни ваг продемонстровано на Рисунку 5.12. Аналізючи отриманий графік залежності зміни ваг кожного вовка від кількості ітерацій, можна зробити висновок, що запропонований підхід є сприятливим для подальшого аналізу методу кластеризації масивів даних на основі покращеного алгоритму сірого вовка.

На Рисунку 5.13 та Рисунку 5.14 показане графічне по-рівняння методів та їх збіжності за функціями Растригіна та Грінварга відповідно.

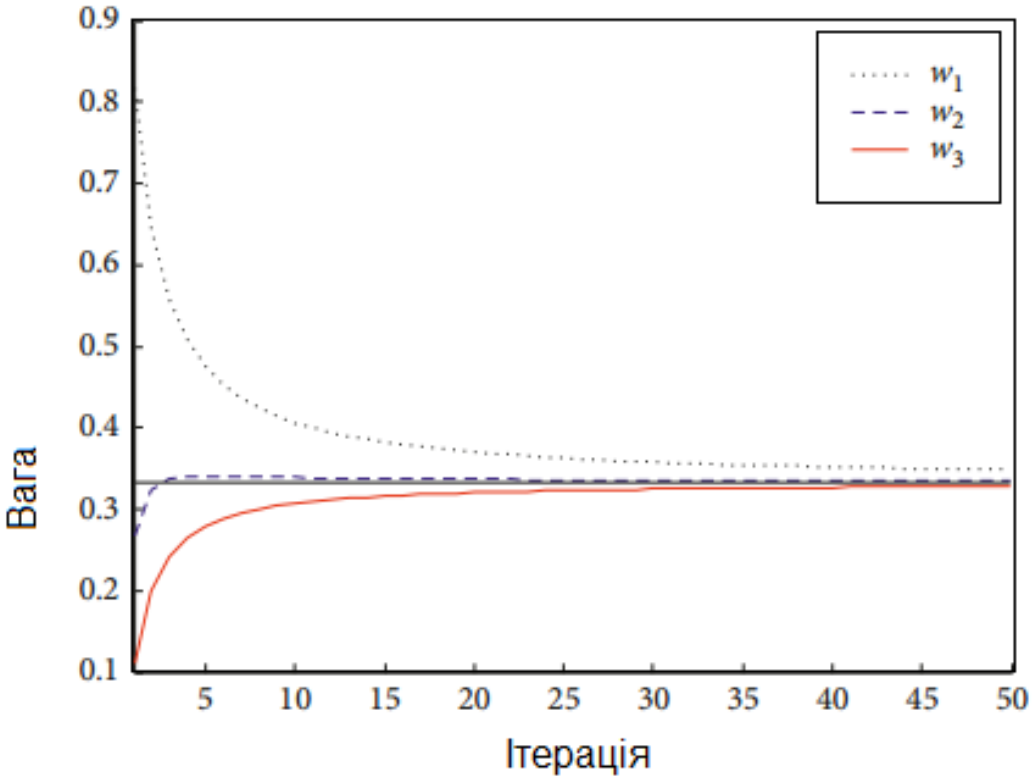


Рисунок 5.13 – Залежність зміни ваги вовків від кількості ітерацій

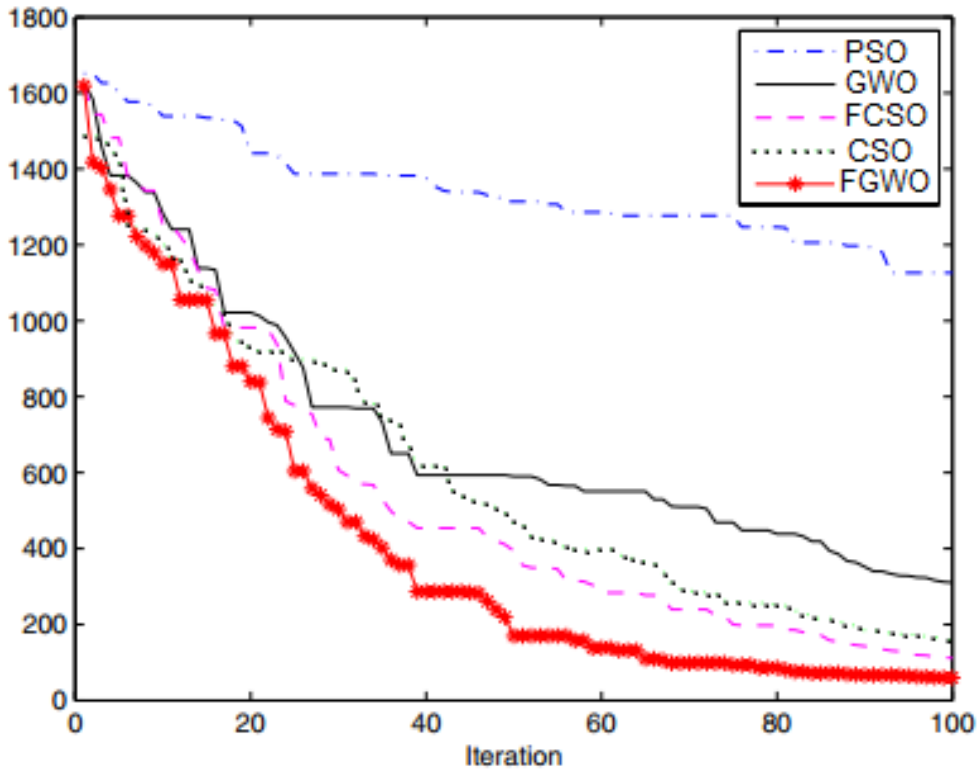


Рисунок 5.13 - Криві збіжності для функції Растригіна

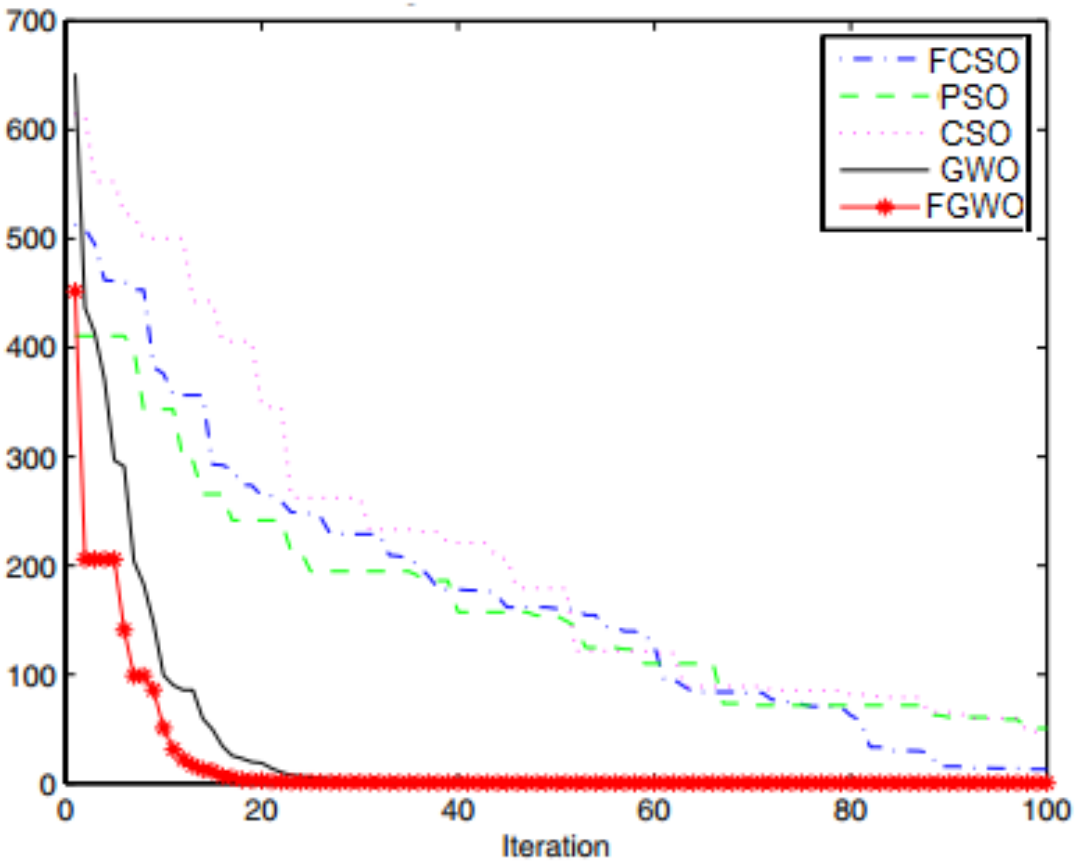


Рисунок 5.14 - Криві збіжності для функції Грінванга

Аналізуючи результати отриманих експериментальних досліджень та порівняльного аналізу роботи методу кластеризації масивів даних на основі покращеного алгоритму сірого вовка із методами кластеризації, що базуються як на класичному підході до кластеризації даних, так і більш екзотичних, запропонований метод демонструє достатньо високі результати.

Основними перевагами запропонованого методу полягає в простоті математичних розрахунків, швидкості роботи з даними, незалежно від виду, розміру та якості вибірки, що аналізується. Слід відзначити точність роботи метода кластеризації даних на основі покращеного алгоритму сірого та отриманих результатів кластеризації, що досягається за допомогою оптимізаційної процедури еволюційного алгоритму.

5.11 Апробація методу правдоподібної нечіткої кластеризації на основі еволюційного підходу божевільних вовків в режимі онлайн

Запропонований метод перевірено на відомій мультиекстремальній функції, такій як функція Еклі, представлена у формі (5.23), рисунок 5.15:

$$f(x) = -20\exp(-0.2\sqrt{0.5(x^2 + y^2)}) - \exp(0.5\cos(2\pi y)) + e + 20. \qquad (5.23)$$

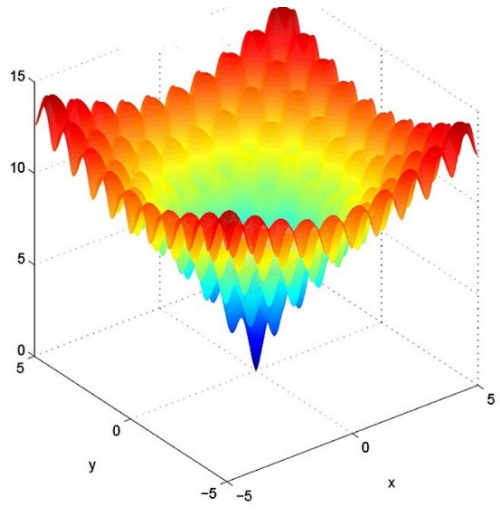


Рисунок 5.15 – Функція Еклі

Псевдокод запропонованого методу можна записати у вигляді наступних кроків, представлених у таблиці 5.15.

Таблиця 5.15 – Псевдокод методу божевільних вовків

Опис	**Псевдокод**
Параметри налаштування	Вибірка даних Населення зграї Контрольний параметр Критерій зупинки
Ініціалізація	Початкові позиції домінантів сірих вовків
Пошук	Якщо це не є критерієм зупинки, то виконується розрахунок нового значення функції придатності Оновлення позиції Обмеження посад вовків Оновити α, β і δ Оновити критерії зупинки Кінець

Поведінку методу божевільних вовків можна перевірити на мультиекстремальній функції Еклі. Історія пошуку глобального екстремуму наведена на рисунку 5.15.

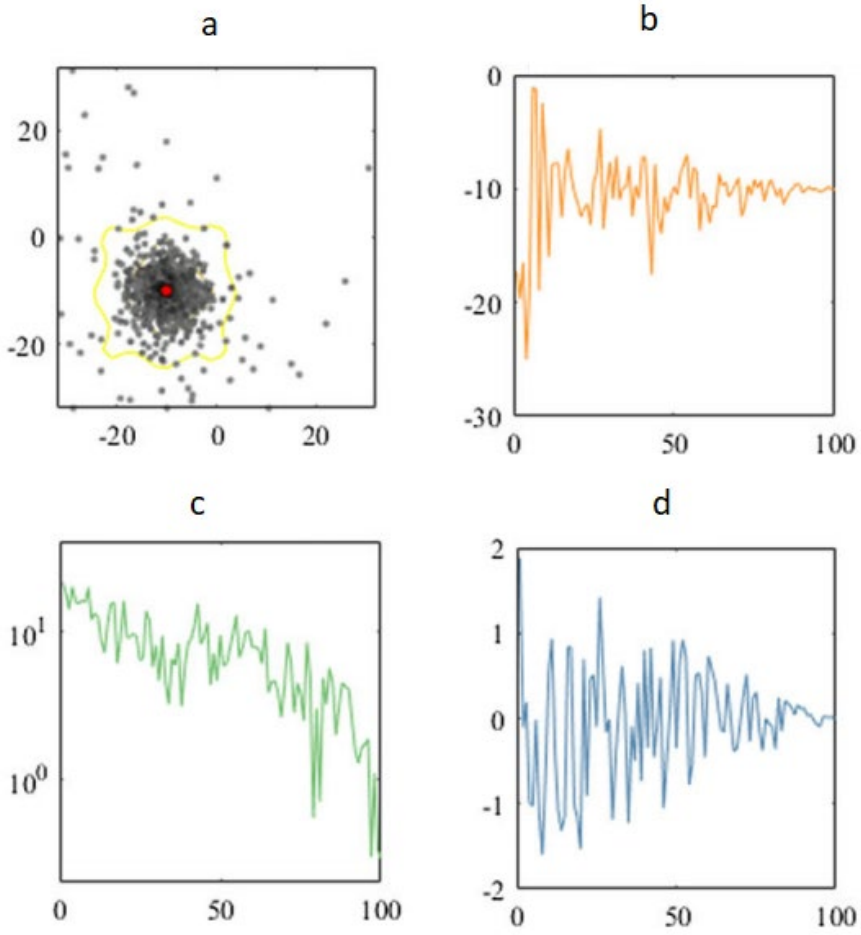


Рисунок 5.15 - Історія пошуку глобального екстремуму

Аналізи проводилися на 100 ітераціях. На рисунку 5.16 продемонстровано історію пошуку глобального екстремуму запропонованої функції (a), траєкторію першої змінної першого вовка (b), історію пристосованості всіх вовків (c) та параметр A (d). Для подальшої перевірки точності пошуку методу божевільних вовків (CGWO) з алгоритмами GWO та PSO.

Два методи ройового інтелекту (метод PSO та метод GWO) вибрано для проведення симуляційних контрастних експериментів із запропонованим CGWO, щоб перевірити його перевагу щодо швидкості конвергенції та точності пошуку. Результати збіжності моделювання на прийнятих функціях тестування показані на рисунку 5.16.

Таблиця 5.16 - Результати чисельної статистики

Точність	**CGWO**	**GWO**	**PSO**
Краща	**1,5099e − 017**	7,5495e − 013	0,0035
Середня	1,2204e − 016	1,0048e − 012	0,0055
Гірша	2,2204e − 014	1,4655e − 013	0,0082

Наведена таблиця демонструє результати чисельної статистики для трьох оптимізаційних алгоритмів, що оцінюються за показником «Точність». Значення наведені для трьох категорій – «Краща» (найкращий результат), «Середня» (середнє значення) та «Гірша» (найгірший результат) – що дозволяє оцінити стабільність та ефективність кожного алгоритму.

За отриманими даними, алгоритм CGWO демонструє надзвичайно високий рівень точності: його найкращий результат становить $1{,}51\times10^{-17}$, середнє значення – $1{,}22\times10^{-16}$, а найгірший – $2{,}22\times10^{-14}$. Ці наднизькі значення свідчать про майже ідеальну здатність алгоритму знаходити оптимальне рішення та мінімізувати похибку, що може бути пов’язано з використанням хаотичних механізмів для покращення пошуку у просторі рішень.

Алгоритм GWO показує результати в порядку 10^{-13}–10^{-12}: найкращий результат – $7{,}55\times10^{-13}$, середнє значення – $1{,}00\times10^{-12}$, а найгірший – $1{,}47\times10^{-13}$ (хоча варто відзначити, що представлення «Гірша» для GWO виглядає дещо неконсистентним, оскільки значення $1{,}47\times10^{-13}$ менше за найкращий результат; однак загальний масштаб характеристик залишається в межах 10^{-13}–10^{-12}). Незважаючи на цю неузгодженість, результати GWO все ж знаходяться на значно вищому рівні похибки, ніж для CGWO.

Алгоритм PSO демонструє значно вищі значення помилки: найкращий результат складає 0,0035, середнє – 0,0055, а найгірший – 0,0082, що говорить

про точність порядку 10^{-3}. Ці значення перевищують результати як CGWO, так і GWO на багато порядків.

Таким чином, порівняльний аналіз показників точності свідчить про те, що алгоритм CGWO є найточнішим та найбільш стабільним з-поміж розглянутих методів, досягаючи результатів з похибкою в межах 10^{-17}–10^{-14}, що суттєво перевищує точність GWO (10^{-13}–10^{-12}) і, особливо, PSO (10^{-3}). Різниця в порядках величин підкреслює переваги CGWO як засобу оптимізації, здатного досягати значно нижчих значень функції помилки, що є критично важливим для завдань з високими вимогами до точності.

Запропонований метод виключає можливість «застрягання» в локальних екстремумах шляхом подвійної перевірки знаходження домінантного вовка в екстремумі та, порівняно із заданою похибкою розрахунку, дозволяє зменшити кількість запусків процедур.

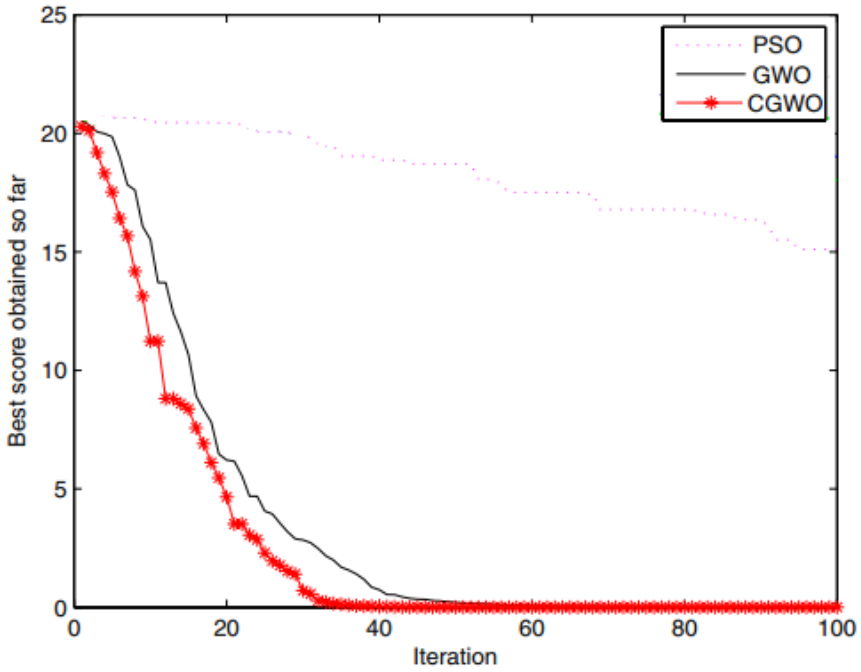


Рисунок 5.16 - Функція Еклі кривих збіжності

Особливістю запропонованої модифікованої процедури оптимізації є можливість керування випадковими збуреннями, що визначають властивості

модифікованого випадкового блукання, що підтвердило свою ефективність при розв’язуванні багатоекстремальних задач. Підхід досить швидкий та точніший на 5-7% за відомі методи, дозволяє знаходити глобальні екстремуми складних функцій, що підтверджено результатами експериментальних досліджень.

5.12 Висновок до розділу 5

1. Вирішено задачу розробки гібридних еволюційних методів нечіткої кластеризації масивів даних, які здатні ефективно справлятися з багатокомпонентними, складними просторами пошуку та мінімізувати залежність від початкових умов.

2. 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.

3. 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.

4. 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.

5. 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] (Додаток А).

РОЗДІЛ 6
РОЗВ'ЯЗАННЯ ПРАКТИЧНИХ ЗАДАЧ

Запропоновані у розділах 2-5 підходи та методи еволюційного самонавчання для адаптивної нечіткої кластеризації потоків викривлених даних в онлайн режимі за умов апріорної та поточної невизначеності, дають можливість підвищити ефективність 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.

6.1 Розв’язання задачі підвищення врожайності озимої пшениці за допомогою методу нечіткої правдоподібної кластеризації даних на основі аналізу щільності розподілу даних та їх піків

Розглядалось питання підвищення врожайності сільськогосподарських культур в сучасний час, що є надзвичайно актуальним. Результативні управлінські вирішення для збільшення цього показника виникають внаслідок аналізу, прогнозування, оптимізації, економічного обґрунтування та вибору оптимальної стратегії із численних варіантів [275-278].

В умовах невизначеності зовнішнього середовища важливо мати ефективний механізм прогнозування врожайності, оскільки збільшується потреба у точних прогнозах. Метою створення прогнозу є зменшення рівня невизначеності, в межах якого приймаються важливі управлінські рішення.

Аналізувались дані про врожайність у ТОВ НАУКОВО-ВИРОБНИЧІЙ ФІРМІ «ХЕЛП-АГРО» з 2012 року по 2022 рік. Дані за відповідні роки за гідрометеорологічними умовами у Близнюківському районі Харківської області отримані із сайтів http://meteo.gov.ua та https://www.gismeteo.ua.

Гідрометеорологічні фактори, які найбільше впливають на врожайність озимої пшениці, включають кількість опадів, температуру повітря та кількість сонячних днів [279-281].

На рисунках 6.1 і 6.2 наведені криві врожайності озимої пшениці у ТОВ НАУКОВО-ВИРОБНИЧІЙ ФІРМІ «ХЕЛП-АГРО» та криві вісьмох гідрометеорологічних умов у Близнюківському районі Харківської області. З графіків можна чітко спостерігати зв’язки між врожайністю та розглянутими гідрометеорологічними факторами.

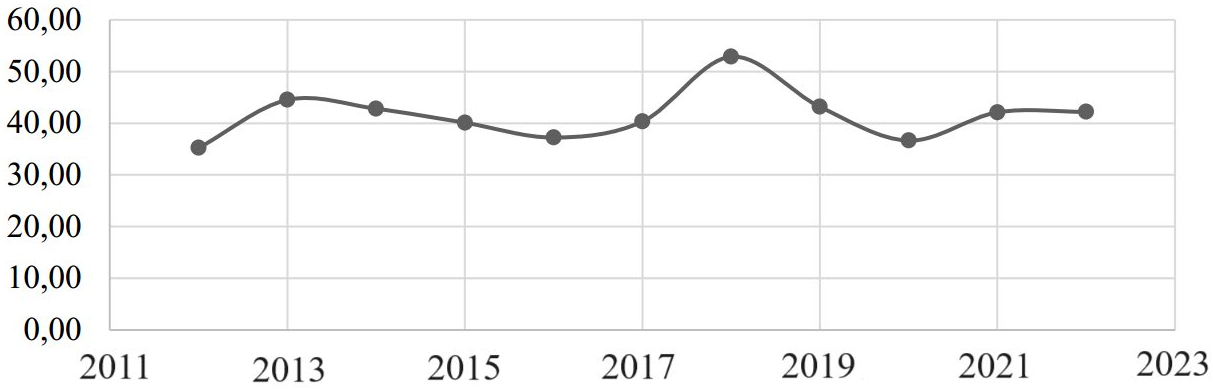


Рисунок 6.1 – Врожайність озимої пшениці ц/га

В таблиці 6.1 наведено дані про рівень врожайності озимої пшениці ТОВ НАУКОВО-ВИРОБНИЧІЙ ФІРМІ «ХЕЛП-АГРО» в залежності від гідрометеорологічних умов у Близнюківському районі Харківської області.

Отже, розглянемо наступні гідрометеорологічні фактори: середньомісячна кількість опадів у травні; середньомісячна кількість опадів у червні; середньомісячна температура повітря у квітні; середньомісячна температура повітря у червні; середньомісячна температура повітря у травні; середньомісячна температура повітря у липні; кількість сонячних днів у січні; кількість сонячних днів у квітні.

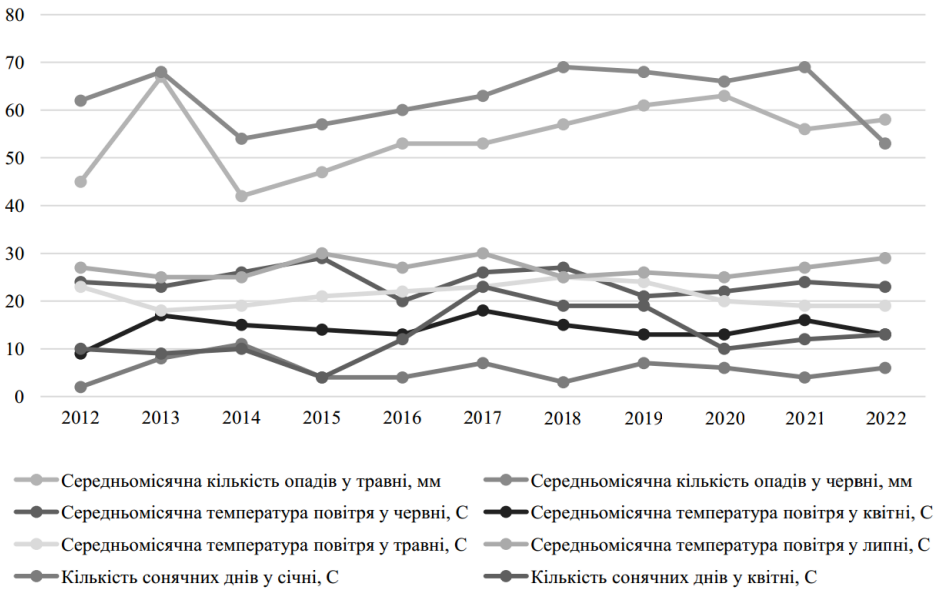


Рисунок 6.2 – Гідрометеорологічні умови в Близнюківському районі Харківської області з 2012 по 2022 роки

Зробимо аналіз таблиці 6.1 за допомогою методу нечіткої довірчої кластеризації даних на основі аналізу щільності розподілу даних та їх піків. Якщо взяти за цільову функцію врожайність озимої пшениці, яка прагне до максимуму, в залежності від гідрометеорологічних показників та обмежень у вигляді кількості сонячних днів, можна спрогнозувати врожайність на наступний рік.

Таблиця 6.1 – Рівень врожайності озимої пшениці в залежності від гідрометеорологічних умов

Роки	**Врожайність ц/га**	**Середньомісячна температура повітря у квітні, °С**	**Середньомісячна кількість опадів у травні, мм**	**Середньомісячна температура повітря у травні, °С**	**Середньомісячна кількість опадів у червні, мм**	**Середньомісячна температура повітря у червні, °С**	**Середньомісячна температура повітря у липні, °С**	**Середньомісячна кількість опадів у липні, мм**	**Кількість сонячних днів у січні, дні**	**Кількість сонячних днів у квітні, дні**
2012	35,20	9	45	23	62	24	27	61	2	10
2013	44,50	17	67	18	68	23	25	62	8	9
2014	42,8	15	42	19	54	26	25	55	11	10
2015	40,10	14	47	21	57	29	30	57	4	4
2016	37,23	13	53	22	60	20	27	59	4	12
2017	40,31	18	53	23	63	26	30	62	7	23
2018	52,86	15	57	25	69	27	25	65	3	19
2019	43,17	13	61	24	98	21	26	78	7	19
2020	36,64	13	63	20	66	22	25	64	6	10
2021	42,10	16	56	19	69	24	27	65	4	12
2022	42,20	13	58	19	53	23	29	55	6	13

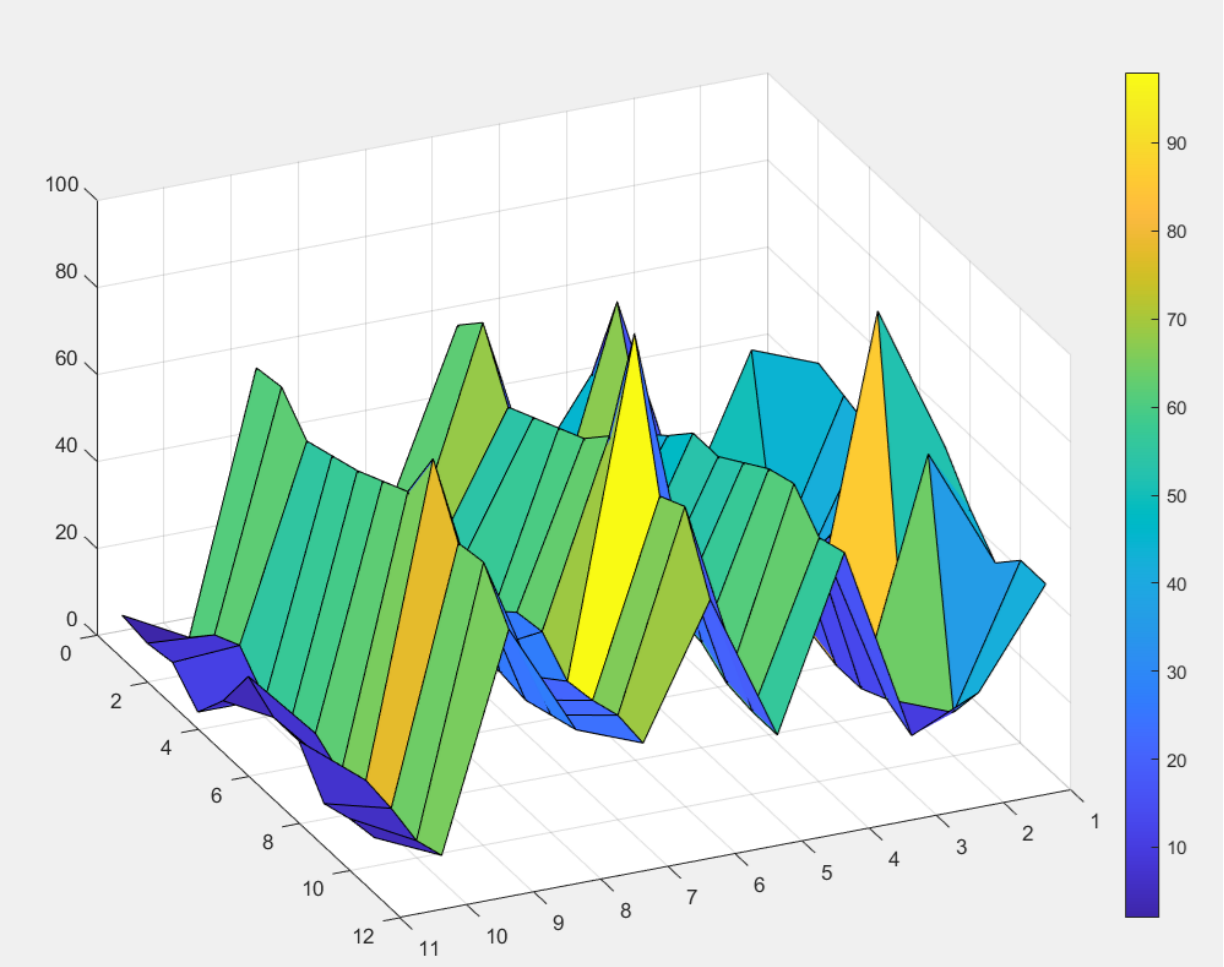


Рисунок 6.3 – Вибірка даних із таблиці 6.1

Процес пошуку атракторів починається не з кожної точки масиву даних, а з, так званих, піків щільності цього масиву (вибірки). 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.

Використовуючи запропонований метод та прогнозні дані від гідрометеослужби, можна здійснювати прогноз значень врожайності на майбутній рік. Прогнозні дані можуть включати інформацію про очікувані погодні умови, такі як кількість опадів, температура повітря та кількість сонячних днів, які будуть впливати на умови вирощування культури.

Слід зауважити, що точність прогнозу може залежати від ряду факторів, включаючи точність прогнозів гідрометеорологічних умов і стабільність роботи методу у різних умовах.

Розглянувши прогнозні дані та використовуючи метод нечіткої правдоподібної кластеризації даних на основі аналізу щільності розподілу даних та їх піків, можемо визначити фактори, які мають найбільший вплив на рівень врожайності озимої пшениці в ТОВ НАУКОВО-ВИРОБНИЧІЙ ФІРМІ «ХЕЛП-АГРО». В таблиці 6.3 продемонстровані отримані результати.

Таблиця 6.3 – Прогноз врожайності озимої пшениці від гідрометеорологічних умов

Показники гідрометеорологічних умов	**Середнє значення**	**Прогнозне значення**
Середньомісячна кількість опадів у червні, мм	62,63	52
Середньомісячна кількість опадів у травні, мм	54,72	47
Середньомісячна температура повітря у квітні, °С	14,18	17,00
Середньомісячна температура повітря у травні, °С	21,18	22,00
Середньомісячна температура повітря у червні, °С	24,09	25,00
Середньомісячна температура повітря у липні, °С	26,90	28,0
Кількість сонячних днів у січні, дні	5,63	7,0
Кількість сонячних днів у квітні, дні	12,81	12,0
Врожайність, ц/га	41,55	47,12

Із досліджень видно, що за середніх погодних умов рівень врожайності становитиме 41,55 ц/га. Збільшення середньомісячної температури повітря у квітні на 1°С означає збільшення врожайності приблизно на 2,27 ц/га. При збільшення середньомісячної кількості опадів у травні на 1 мм несе

збільшення врожайності приблизно на 0,1 ц/га, а збільшення кількості сонячних днів у квітні на одиницю означає зменшення врожайності на 0,485 ц/га. Зокрема, за результатами аналізу визначено, що найбільший вплив на рівень врожайності має середньомісячна температура повітря у квітні.

Отримані результати можуть бути корисні для прогнозування врожайності на майбутні роки та прийняття управлінських рішень для оптимізації умов вирощування озимої пшениці в даному регіоні. Доцільно враховувати ці фактори при розробці стратегій управління сільськогосподарськими угіддями та культурами для підвищення врожайності та ефективності виробництва.

Також проведено порівняльний аналіз якості кластеризації даних за основними характеристиками оцінок якості, таких як: Partition Coefficient (PC), Silhouette Coefficient (SC), Xie and Beni's Index (XBI) вже існуючих методів кластеризації.

Таблиця 6.2 – Порівняльний аналіз якості кластеризації даних

Методи кластеризації	**PC**	**SC**	**XBI**	**Час (с)**
Fuzzy c-means	0,37	1,44	0,27	97
DENCLUE	0,28	**1,22**	0,18	71
Нечітка правдоподібна кластеризація даних на основі аналізу щільності розподілу даних та їх піків	**0,25**	**1,22**	**0,01**	**7**
Правдоподібна кластеризація даних	0,32	1,42	0,25	80

Порівняння методів кластеризації, дозволяють зробити кілька важливих спостережень щодо їх ефективності, точності, стабільності та швидкості виконання.

Метод Fuzzy C-means показав досить високу стабільність кластеризації (SC = 1,44), що свідчить про те, що алгоритм здатний зберігати сталі результати навіть при варіаціях у вхідних даних. Однак його точність кластеризації була

низькою (PC = 0,37), що вказує на слабку здатність до точного поділу кластерів, особливо в складних чи перекритих даних. Крім того, цей метод має значний час виконання (97 секунд), що робить його менш ефективним для великих вибірок даних або задач, де важлива швидкість обробки.

DENCLUE виявився більш швидким і менш обчислювальним у порівнянні з Fuzzy C-means (71 секунду), що є важливим для застосувань, де швидкість є критичним фактором. Проте точність кластеризації DENCLUE (PC = 0,28) також виявилася невисокою, що робить його менш ефективним для задач з високими вимогами до точності кластеризації. Незважаючи на це, метод забезпечує хорошу стабільність (SC = 1,22) і непоганий баланс між кластерами (XBI = 0,18).

Метод нечіткої правдоподібної кластеризації на основі аналізу щільності розподілу даних та їх піків продемонстрував найбільш швидкий час виконання (7 секунд), що є великим перевагою для задач, де потрібна висока швидкість обробки даних. Однак його точність (PC = 0,25) була найнижчою серед усіх методів, що свідчить про обмежену здатність цього методу до точного класифікування даних. Також він має дуже низький індекс балансу (XBI = 0,01), що може вказувати на погану рівномірність розподілу точок між кластерами.

Метод правдоподібної кластеризації даних виявився збалансованим за всіма параметрами. Його точність (PC = 0,32) була вищою, ніж у DENCLUE і нечіткої правдоподібної кластеризації, а також час виконання (80 секунд) знаходився між результатами Fuzzy C-means та DENCLUE. Стабільність цього методу також була досить високою (SC = 1,42), що робить його хорошим вибором для задач, де важлива надійність та якість результатів, зокрема у випадках складних або шумних даних.

Загалом, можна зробити висновок, що вибір методу кластеризації залежить від конкретних вимог до задачі. Якщо важлива швидкість обробки даних, то метод нечіткої правдоподібної кластеризації є найкращим вибором, хоча його точність залишає бажати кращого. Для задач, де важлива

стабільність і точність кластеризації, підходять методи Fuzzy C-means та правдоподібна кластеризація даних, зокрема, останній метод може забезпечити гармонійний баланс між різними критеріями. Метод DENCLUE виявився ефективним у контексті швидкості виконання, але його точність потребує покращення в порівнянні з іншими підходами.

Запропонований метод правдоподібної кластеризації даних на основі аналізу щільності розподілу даних та їх піків дозволяє спрогнозувати врожайність озимої пшениці, надає можливість визначати прогнозні значення урожайності, враховуючи основні гідрометеорологічні фактори. Важливо відзначити, що точність прогнозу залежить від точності передбачення гідрометеорологічних факторів. Таким чином, для поліпшення точності прогнозу слід звертати увагу на якість та достовірність прогнозів цих факторів.

Зазначений підхід може бути корисним для сільськогосподарських підприємств та фермерів, які прагнуть забезпечити ефективне управління виробництвом та оптимізувати умови вирощування озимої пшениці в умовах мінливого клімату, що підтверджено актом впровадження (акт від 27.02.2023р.).

6.2 Оцінка стану будинків для визначення готовності до експлуатації в зимових умовах за допомогою методу адаптивної нечіткої кластеризації даних різної природи

Оцінка стану будинків для визначення готовності до експлуатації в зимових умовах за допомогою методу адаптивної нечіткої кластеризації даних різної природи передбачає використання методу, який комбінує в собі аспекти нечіткої логіки та кластерного аналізу для здійснення комплексної оцінки стану будівель.

На підприємстві ТОВ «КОМУНСЕРВІС 2018» були взяті відомості стану будинків по вул. Мостобудівників м. Безлюдівка, так як стан покрівлі, горища,

сходів, підвалу, інженерного обладнання, прибирального і протипожежного інвентаря, що були сформовані у таблицю «об’єкт - властивість». В таблиці 6.4 продемонстрована частина даних щодо стану будинків по вул. Мостобудівників смт Безлюдівка наданих ТОВ «КОМУНСЕРВІС 2018».

Таблиця 6.4 – Інформація щодо стану будинків по вул. Мостобудівників, смт Безлюдівка (частина вибірки)

№	Перелік ремонтних робіт	Показник усередненої вартості одиниці обсягу, гривень з	Номер будинку					
			1	2	3	4	5	6
1.	Відновлення пошкодженої покрівлі, кв.м	512,15	50,2	140,7	90,60	40,40	56,65	41,00
2.	Відновлення пошкодженої покрівлі без заміни опорних конструкцій та стропильних систем, кв.м	2134,19	78,1	47,55	0	50,71	0	-
3.	Відновлення частини пошкодженого даху із ремонтом опорних конструкцій і крокв’яних системи та покрівлі із металочерепиці на площі до 25 % (кв.м)	2837,58	15,5	9,82	-	0	10,83	0
4.	Заміна пошкоджених міжкімнатних дверей (блоків)	9983,26	2,0	3,40	0	1,52	0	0
5.	Заміна металевих вхідних дверей	14414,89	1,75	0	2,00	1,82	0	0

Продовження таблиці 6.4

1	2	3	4	5	6	7	8	9
6.	Заміна пошкодженого скління/ склопакетів (кв.м)	4192,67	0,5	0,75	0	0	0,91	0
7.	Заміна віконного блока з урахуванням підвіконних дощок, відливів (кв.м)	5242,26	7,22	8,12	11,30	6,96	7,07	9,05
8.	Відновлення пошкоджених фасадів з урахуванням утеплення та зовнішнього оздоблення (кв.м)	2347,535	0	6,54	7,89	0	10,15	0
9.	Відновлення пошкодженого декоративного шару камінцевої штукатурки оздоблення утеплених фасадів без урахування утеплювача (кв.м)	684,8	0	10,00	-	0	0	0
10.	…	…	…	…	…	…	…	…

Набір даних містить інформацію про стан 50 будинків, які мають різний ступінь пошкоджень, зносу, тощо. Стан будинку описаний 57 характеристиками – атрибутами спостереження. В таблиці 6.4 продемонстрована лише частка вибірки.

Основною метою дослідження є оцінка стану будинків для визначення готовності до експлуатації в зимовий період, що дозволило б прискорити аналіз та прийняття рішень щодо першочерговості відновлення будинків, в залежності від категорії пошкоджень, зношеності та наявних ресурсів [282-285]. Для отримання категорії пошкоджень будівлі, пропонується розбити дані

на 3 кластери, кожен з яких буде вказувати ступінь зношеності будинку. Таким чином, буде проведений аналіз вулиці.

Для кластеризації даних обрано метод адаптивної кластеризації викривлених даних на основі достовірного підходу. Цей метод був обраний для можливості роботи з різними наборами даних, що містять пропуски. Як видно із таблиці 6.4, не всі спостереження заповнені повністю, є дані про будинки, інформація про які відсутня. Ці дані також необхідно аналізувати, та відносити до якогось кластера-категорії пошкоджень. Якість роботи методу перевірена з допомогою коефіцієнтів якості кластеризації та загальної точності. Також проведений порівняльний аналіз роботи запропонованого методу з більш відомими класичними алгоритмами кластеризації даних. В таблиці 6.5 наведені оцінки якості кластеризації методу адаптивної кластеризації викривлених даних на основі достовірного підходу, в порівнянні з методом Густафсона – Кесселя та алгоритму FCM.

Таблиця 6.5 – Оцінка якості кластеризації методів

Методи кластеризації даних	**PC**	**SC**	**XB**
FCM	0,51	1,63	0,18
Густафсон - Кесселя	0,26	**1,65**	1,63
Метод адаптивної кластеризації викривлених даних на основі достовірного підходу	**0,24**	0,64	**0,15**

Наведена таблиця містить результати кластеризації даних трьома методами, оціненими за трьома критеріями: точністю кластеризації (PC), стабільністю кластеризації (SC) та індексом балансу кластерів (XB). Дані числові значення дозволяють порівняти ефективність кожного алгоритму за різними аспектами роботи з даними.

Метод FCM (Fuzzy C-Means) демонструє найвищий показник точності кластеризації з PC рівним 0,51, що свідчить про його кращу здатність розділяти дані на окремі кластери у порівнянні з іншими методами, адже інші алгоритми мають значення PC 0,26 (Густафсон Кессель) та 0,24 (адаптивна кластеризація викривлених даних на основі достовірного підходу). При цьому стабільність FCM (SC = 1,63) є на високому рівні, що дозволяє вважати результати його кластеризації сталістю відносно незмінними при повторних запусках. Інтерпретуючи індекс балансу кластерів, значення XB у 0,18 для FCM вказує на певну рівномірність розподілу точок між кластерами, що є бажаною властивістю.

Метод Густафсона - Кесселя характеризується майже однаковою стабільністю (SC = 1,65), проте його точність кластеризації є значно нижчою (PC = 0,26), що свідчить про менш ефективний поділ даних. Крім того, значення XB, яке складає 1,63, вказує на суттєві відмінності у розподілі точок між кластерами, тобто кластери формуються менш збалансовано, що може негативно впливати на якість кластеризації.

Метод адаптивної кластеризації викривлених даних на основі достовірного підходу має найнижчий показник точності (PC = 0,24) та найнижчу стабільність (SC = 0,64) серед розглянутих методів, що свідчить про його слабку здатність до точного та сталого розподілу даних. Проте індекс балансу кластерів для цього методу (XB = 0,15) є найнижчим, що потенційно вказує на найбільш компактний або рівномірний розподіл даних у межах окремих кластерів. Таким чином, незважаючи на перевагу за балансом, низькі значення PC та SC свідчать про те, що цей метод може бути менш придатним для задач, де критично важлива висока точність та стабільність кластеризації.

Отже, якщо розглядати всі метрики комплексно, метод FCM видається найбільш ефективним з точки зору точності та стабільності, при цьому забезпечуючи прийнятний рівень балансу кластерів. Метод Густафсон Кессель має хорошу стабільність, але його низька точність та високий індекс XB свідчать про менш оптимальний розподіл даних між кластерами. Адаптивний

метод, хоч і демонструє найкращий баланс (XB), страждає від недостатньої точності та стабільності, що обмежує його застосування в задачах, де ці характеристики мають вирішальне значення.Для перевірки загальної точності роботи методу, вибірку штучно розбили на декілька частин: 10, 25 та 57 спостережень. Такий підхід дозволяє аналізувати вибірку в залежності від кількості спостережень та якості роботи методу. Зрозуміло, що чим більше спостережень, тим якісніше буде аналіз даних.

В таблиці 6.6 наведений аналіз точності роботи методу адаптивної кластеризації викривлених даних на основі правдоподібного підходу, в порівнянні з методом Густафсона – Кесселя та алгоритму FCM.

Таблиця 6.6 – Загальна точність методів кластеризації

Кількість спостережень	**Метод кластеризації**	**Загальна точність**	
		Highest	**Mean**
57	Метод адаптивної кластеризації викривлених даних на основі правдоподібного підходу	63,34	63,34
	Густафсон - Кессель	64,45	64,40
	FCM	64,28	64,28
25	Метод адаптивної кластеризації викривлених даних на основі правдоподібного підходу	62,45	62,24
	Густафсон Кессель	61,75	61,30
	FCM	61,68	60,98
10	Метод адаптивної кластеризації викривлених даних на основі правдоподібного підходу	58,54	58,54
	Густафсон Кессель	57,55	57,55
	FCM	57,48	57,48

Наведена таблиця демонструє загальну точність трьох методів кластеризації даних, оцінену за двома показниками при різній кількості спостережень (57, 25, 10). За результатами, при найбільшій вибірці з 57 спостережень метод Густафсон - Кесселя показує найвищу точність кластеризації – максимальне значення складає 64,45 %, а середнє – 64,40 %, що перевищує результати як FCM (64,28 %), так і адаптивного методу (63,34 % для обох показників). Цей факт свідчить про здатність методу Густафсона - Кесселя ефективно працювати з великою кількістю даних, забезпечуючи найкращий розподіл об'єктів за кластерами при насиченій вибірці.

Проте при зменшенні кількості спостережень до 25 спостережень спостерігається зниження точності для всіх методів, проте адаптивний метод демонструє кращий показник – Highest = 62,45 % та Mean = 62,24 % – порівняно з Густафсон Кессель (61,75 % – Highest, 61,30 % – Mean) та FCM (61,68 % – Highest, 60,98 % – Mean). При найменшій вибірці з 10 спостережень адаптивний метод знову лідирує, досягаючи точності 58,54 % (як для Highest, так і для Mean), тоді як метод Густафсона - Кесселя та FCM демонструють трохи нижчі значення – 57,55 % та 57,48 % відповідно.

Загальний результат свідчить про поступове зниження точності кластеризації із зменшенням кількості спостережень, а також про зміну ранжування методів: при великих вибірках найкращими є методи, орієнтовані на більш насичені дані (Густафсон-Кессель і FCM), тоді як при меншій кількості даних адаптивний метод на основі правдоподібного підходу забезпечує кращі результати. Це може свідчити про те, що адаптивний підхід краще пристосовується до обмежених даних, зберігаючи здатність виділяти суттєві особливості розподілу навіть при зменшеному обсязі вибірки.

На рисунку 6.4 та 6.5 наведені гістограми порівняння високої та середньої точності отриманих результатів кластеризації даних наданих ТОВ «КОМУНСЕРВІС 2018» відповідно.

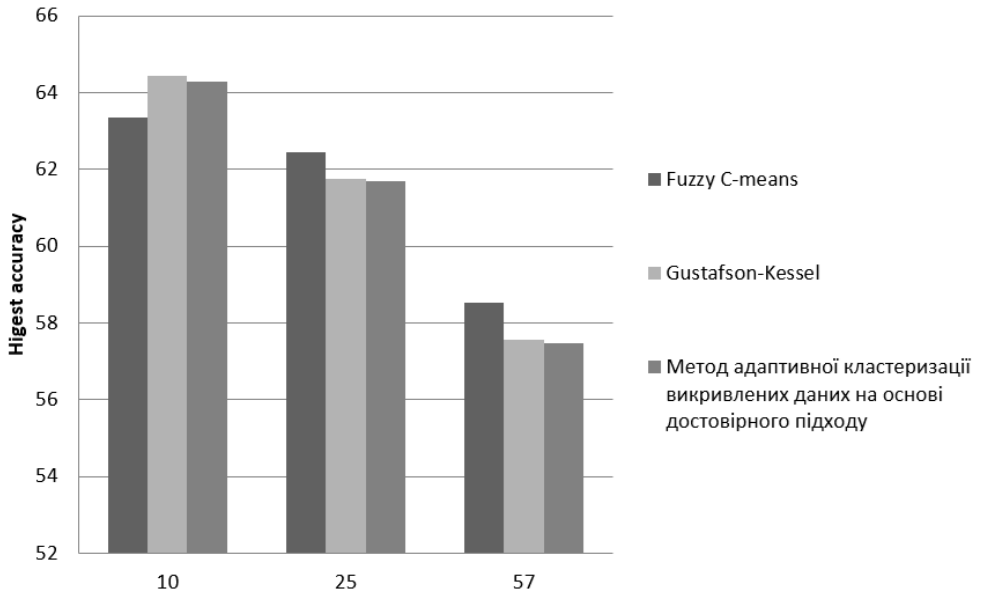


Рисунок 6.5 – Порівняння високої точності кластеризації даних методом FCM, Густафсона -Кесселя, та методом адаптивної кластеризації викривлених даних на основі достовірного (правдоподібного) підходу наданих ТОВ «КОМУНСЕРВІС 2018».

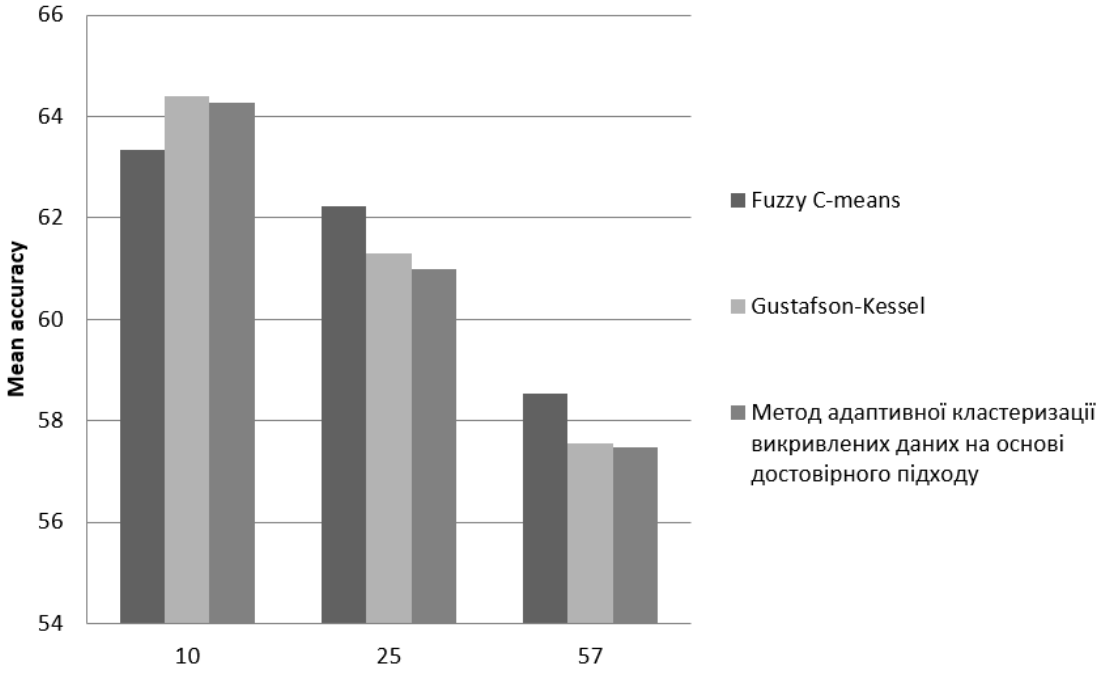


Рисунок 6.5 – Порівняння середньої точності кластеризації даних методом FCM, Густафсона -Кесселя, та методом адаптивної кластеризації викривлених даних на основі достовірного (правдоподібного) підходу наданих ТОВ «КОМУНСЕРВІС 2018».

З графіків видно, що робота методу демонструє достатньо гарні результати в аналізі реальних даних, надає максимально правдоподібні відповіді на запитання, що допомагає в прийнятті рішень.

Результати впровадження методу адаптивної кластеризації викривлених даних на основі правдоподібного підходу довели доцільність використання їх в аналізі та оцінці стану житлових будинків, що дозволяють прискорити аналіз та прийняття обґрунтованих рішень щодо першочерговості відновлення будинків, в залежності від категорії пошкоджень та зношеності.

Підтверджено актом впровадження (акт від 12.04.2023р.).

6.3 Вирішення практичної задачі класифікації технологічних процесів на будівництві за допомогою методу адаптивної нечіткої кластеризації даних

На ТОВ «Будівельно-монтажне підприємство - 168» були отримані дані будівельних та монтажних робіт загального призначення для отримання класифікації технологічних процесів на будівництві з метою підвищення їх ефективності.

Для класифікації були використані дані розрахунків об’єктів інженерної інфраструктури: монтаж внутрішніх інженерних мереж, систем, приладів, засобів вимірювання та іншого обладнання [285-289].

Ціллю впровадження методу адаптивної нечіткої кластеризації даних є надання можливості класифікувати технологічні процеси за класом наслідків (відповідальності), що належать до об’єктів із середніми та значними наслідками (СС2, СС3) та підвищує ефективність господарської діяльності з будівництва об’єктів.

Клас наслідків визначається відповідно до вимог будівельних норм і правил, затверджених згідно із законодавством. Клас наслідків визначається для кожного об’єкта - будинку, будівлі, споруди будь-якого призначення, їхніх

частин, лінійних об’єктів інженерно-транспортної інфраструктури, у тому числі тих, що належать до складу комплексу (будови).

До складу комплексу (будови) можуть належати об’єкти, будівництво яких здійснюється за єдиною проектно-кошторисною документацією.

Усі об’єкти поділяються за такими класами наслідків (відповідальності):

- незначні наслідки (СС1);
- середні наслідки (СС2);
- значні наслідки (СС3).

Під час здійснення державного архітектурно-будівельного контролю на об’єктах самочинного будівництва клас наслідків таких об’єктів визначається самостійно відповідними органами державного архітектурно-будівельного контролю або із залученням експертної організації чи експерта, який має відповідний кваліфікаційний сертифікат. При цьому пропонується наступна схема урахування класу наслідків (відповідальності) СС1; СС2; СС3:

- класу наслідків СС1 відповідають I та II категорія складності;
- класу наслідків СС2 відповідають III та IV категорія складності;
- класу наслідків СС3 відповідає V категорія складності.

Таблиця 6.7 - Характеристики можливих наслідків від відмови будівлі або споруди в залежності від класів наслідків (відповідальності) будівлі або споруди.

Клас наслідків (відповідальності) будівлі або споруди	**Характеристики можливих наслідків від відмови будівлі або споруди**					
	Можлива небезпека для здоров’я і життя людей, кількість осіб			**Обсяг можливого економічного збитку, мінімальний розмір заробітної плати**	**Втрата об’єктів культурної спадщини, категорії об’єктів**	**Припинення функціонування комунікацій транспорту, зв’язку, енергетики, інших інженерних мереж, рівень**
	які постійно перебувають на об’єкті	**які періодично перебувають на об’єкті**	**які перебувають поза об’єктом**			
СС3 значні наслідки	понад 300	понад 1000	понад 50000	понад 150000	національного значення	загальнодержавний

Продовження таблиці 6.7

СС2 середні наслідки	від 20 до 300	від 50 до 1000	від 100 до 50000	від 2000 до 150000	місцевого значення	регіональний, місцевий
СС1 незначні наслідки	до 20	до 50	до 100	до 2000	—	—

В таблиці 6.8 наведений перелік робіт із провадження господарської діяльності з будівництва об’єктів, що за класом наслідків (відповідальності) належать до об’єктів із середніми та значними наслідками (СС2, СС3) на ТОВ «Будівельно-монтажне підприємство - 168».

Таблиця 6.8 – Перелік видів робіт

Вид робіт	**Клас наслідків**
Улаштування основ та фундаментів збірних та монолітних	СС2
Улаштування фундаментів із застосуванням паль	СС2
Зведення металевих конструкцій	СС2
Зведення збірних бетонних та залізобетонних конструкцій	СС2
Зведення монолітних бетонних, залізобетонних конструкцій	СС2
Зведення кам’яних та армокам’яних конструкцій	СС2
Зведення дерев’яних конструкцій	СС2
Магістральні трубопроводи	СС2
Об’єкти водопроводу і каналізації (включаючи водонапірні башти, очисні споруди, водозабори) промислових підприємств і населених пунктів	СС2
….	

Продовження таблиці 6.8

Об’єкти нафто- і газодобувної, газопереробної, металургійної, хімічної та інших галузей промисловості, обладнані пожежо- і вибухонебезпечними ємкостями і сховищами рідкого палива, газу і газопродуктів, особливо при їх зберіганні під тиском (технологічні трубопроводи, апарати, котли, газгольдери, ізотермічні резервуари ємністю понад 10 тис. кубометрів, резервуари для зберігання нафти та нафтопродуктів ємністю 30 тис. кубометрів і більше, посудини високого тиску тощо)	СС3
Об’єкти хімічної, нафтохімічної, біотехнологічної, оборонної та інших галузей, що пов’язані з використанням, переробкою, виготовленням і зберіганням хімічно токсичних, вибухо- і пожежонебезпечних речовин і промислових вибухових матеріалів, біологічно небезпечних речовин тощо	СС3
Будівлі і споруди крупних залізничних вокзалів і аеровокзалів	СС3
Будівлі основних музеїв, державних архівів, сховищ національних історичних і культурних цінностей	СС3
Будівлі університетів, інститутів, шкіл, дошкільних закладів тощо	СС3
….	

Таким чином, взявши за основу базу даних об’єктів, які належить відбудувати, можна розбити за класами робіт, які необхідно виконати і віднести об’єкт за відповідною категорією технологічного процесу на будівництві. Таким чином, підвищити ефективність господарської діяльності з будівництва об’єктів.

Визначення класів наслідків (відповідальності) об’єктів житлового комплексу, який складається з трьох однакових односекційних 17-поверхових 102-квартирних житлових будинків, окремо розташованого продовольчого

магазину з дворівневим підземним паркінгом та трансформаторної підстанції. Проектування комплексу (будови), до складу якого входить кілька окремих об’єктів, виконують на підставі вихідних даних, зокрема містобудівних умов та обмежень на комплекс (будову) у цілому.

Клас наслідків (відповідальності) визначають окремо для кожного об’єкта, що входить до житлового комплексу. Кожен із житлових будинків має окреме підключення до інженерних мереж. Відповідно до сценарію аварії приймають імовірність настання таких подій:

– вихід з ладу та руйнування окремої несучої конструкції за рахунок її перевантаження понадпроектними сполученнями навантажень та впливів;

– виникнення великих просадок ґрунтових основ унаслідок аварійного замочування;

– вплив можливого карстового провалу, зсувів ґрунту тощо;

– можливість відмови конструкцій у разі виникнення пожежі;

– пошкодження будівельних конструкцій аварійними вибухами;

– вихід з ладу трансформаторної підстанції.

Таблиця 6.9 - Опис запланованих новобудов

Кількість кімнат у квартирі	**Площа квартир, м^2**	**Кількість квартир на будинок**	**Загальна площа квартир на будинок, м^2**	**Коефіцієнт розселення на квартиру**	**Розселення на будинок, осіб**
1	40,5 (30+10,5)	34	1 377	1,43	49
2	52,5 (42+10,5)	34	1 785	2	68
3	65,5 (55+10,5)	34	2 227	2,62	89

Визначення класу наслідків (відповідальності) житлового будинку починається з розрахунку кількості мешканців у житловому будинку, яка

залежить від площі квартири (за нормою 21 м2 на людину плюс 10,5 м2 на сім'ю). Сама база даних забудов, містить більше 150 об'єктів з відповідними характеристиками. Кожне спостереження – будинок забудовника, має свій план та складність виконання замовлення. Таким чином, необхідно проаналізувати вибірку даних, розбити об'єкти забудови на класи складності та об'єми виконання робіт.

Для реалізації поставленої задачі, був використаний метод адаптивної кластеризації даних. Дані, що надані по об'єкту запланованих новобудов підприємством ТОВ «Будівельно-монтажне підприємство - 168» були кластеризовані та розподілені по класам наслідків (відповідальності) будівлі або споруди відповідно до національного стандарту України ДСТУ 8855:2019 «Визначення класу наслідків (відповідальності)». Слід відзначити, що підприємство має ліцензії тільки на СС2 та СС3 класи наслідків. Тобто після кластеризації даних за класами наслідків, кластери-класи які відносяться до СС1 класу не аналізуються, роботи за цим класом будівлі не ведуться.

Якість кластеризації даних було перевірено за допомогою оцінок якості кластеризації PC, CE, SC, S, XB та DI.

Таблиця 6.10 – Оцінки якості кластеризації даних ТОВ «Будівельно-монтажне підприємство – 168»

Метод	**PC**	**CE**	**SC**	**S**	**XB**	**DI**
Адаптивний нечіткий метод кластеризації даних	**0,16**	**1,56**	7,38	2,71	**5,71**	**0,20**
FCM	0,79	0,38	7,33	**-6,84**	5,61	0,01
Густафсон-Кессель	0,55	0,63	**8,59**	0,04	1,07	0,10

Аналізуючи таблицю 6.10, можна зробити висновки, що якість кластеризації даних методом адаптивної нечіткої кластеризації надають більш

якісні рішення щодо відношення конкретної забудови до відповідного класу відповідальності, що на поточний момент відіграє значну роль в плануванні робіт на об’єктах. Слід відзначити основні коефіцієнти якості кластеризації, такі як PC, CE та XB, що більш ґрунтовно відносять відповідний об’єкт до кластеру-класу, аналізують точність віднесення конкретного об’єкта-спостереження до класу-кластера, таким чином знімаючи неоднозначність відношення об’єкта до кластера. Якщо подивитись на порівняльні методи кластеризації, жоден з класичних методів не надав такої якості кластеризації даних, ніж адаптивний метод нечіткої кластеризації.

Підтверджено актом впровадження (акт від 21.12.2023р.).

6.4 Імплементація методу відновлення та фільтрації потоків даних за умов перетинних кластерів для задач покращення якості води

На КП «Санітарно-екологічний центр» впроваджено метод відновлення та фільтрації потоків даних за умов перетинних кластерів для задач покращення якості води. Для аналізу були використані хімічні та електрофізичні показники води [290-300].

Щоби визначити ступінь придатності води для агротехнічного використання, тип засолення, характер та вірогідність засолення при тривалому зрошенні та надати рекомендацію щодо поліпшення характеристик води та/або зниження негативного впливу від використання було проаналізовані проби води Харківської області. Хімічні показники аналізу проб води наведені в таблиці 6.11 (частково). Залежно від якості води та необхідного ступеня обробки для доведення її до показників «Вода питна» водні об’єкти, придатні як джерела господарсько-питного водопостачання, ділять на 3 класи згідно з нормативами для питної води ДСанПіН 2.2.4-171-10.

Дані, щодо властивостей води були проаналізовані протягом 2023 року.

Таблиця 6.11 – Протокол випробувань води в Харківській області

№	Найменування показників	Результати лабораторних випробувань					Нормативи Для води	
		1	2	3	4	5	Безпечність та якість	Фізіологічна повноцінність
1	Активна реакція (pH)	6,50	6,64	6,67	6,42	6,59	6,5-8,5	
2	Смак та присмак, (бали)	-	-	-	-	-	≤3	
3	Запах, (бали)	1	1	1	1	1	≤3	
4	Хлориди, мг/дм^2	48,80	38,15	55,89	202,28	95,82	≤350	
5	Сульфати, мг/дм^2	296,87	287,69	230,15	226,32	163,02	≤500	
6	Нітрит-іон, мг/дм^2	н/в	н/в	0,001	0,001	0,001	≤3,3	
7	Нітрат-іон, мг/дм^2	15,91	19,55	15,68	53,69	11,53	≤50	
8	Кальцій, мг/дм^2	79,92	77,92	113,89	109,89	97,90		25-75
9	Магній, мг/дм^2	20,67	36,48	29,18	42,56	30,40		10-50
10	Амоній, мг/дм^2	0,17	0,25	0,48	0,57	0,55	≤2,6	
11	Залізо загальне, мг/дм^2	н/в	н/в	н/в	0,010	0,020	≤1,0	
12	Сухий залишок, мг/дм^2	710,00	715,0	720,0	900,0	690,0	≤1500	200-500
13	Загальна жорсткість, ммоль/дм^2	5,7	6,9	8,1	9,0	7,4	≤10,0	1,5-7,0
14	Загальна лужність, ммоль/дм^2	4,2	5,0	6,1	5,6	6,3	≤10,0	1,5-7,0
15	Поліфосфати, мг/дм^2	0,20	0,22	0,19	1,24	0,69	≤0,6 (та відсутність)	
16	Окиснюваність перманганатна, мгО/ дм^2	0,50	0,55	0,50	0,65	0,65	≤5,0	

Не дивлячись на те, що останнім часом спостерігається тенденція до зниження обсягів використання води на потреби галузей народного господарства (рисунок 6.6), а отже, відповідно і зменшення обсягів загального водовідведення, частка забруднених стоків у зворотних водах є досить високою, що викликає в кінцевому рахунку суттєве забруднення водойм стічними водами (рисунок 6.7).

Рисунок 6.6 – Динаміка збору води з природних водних джерел

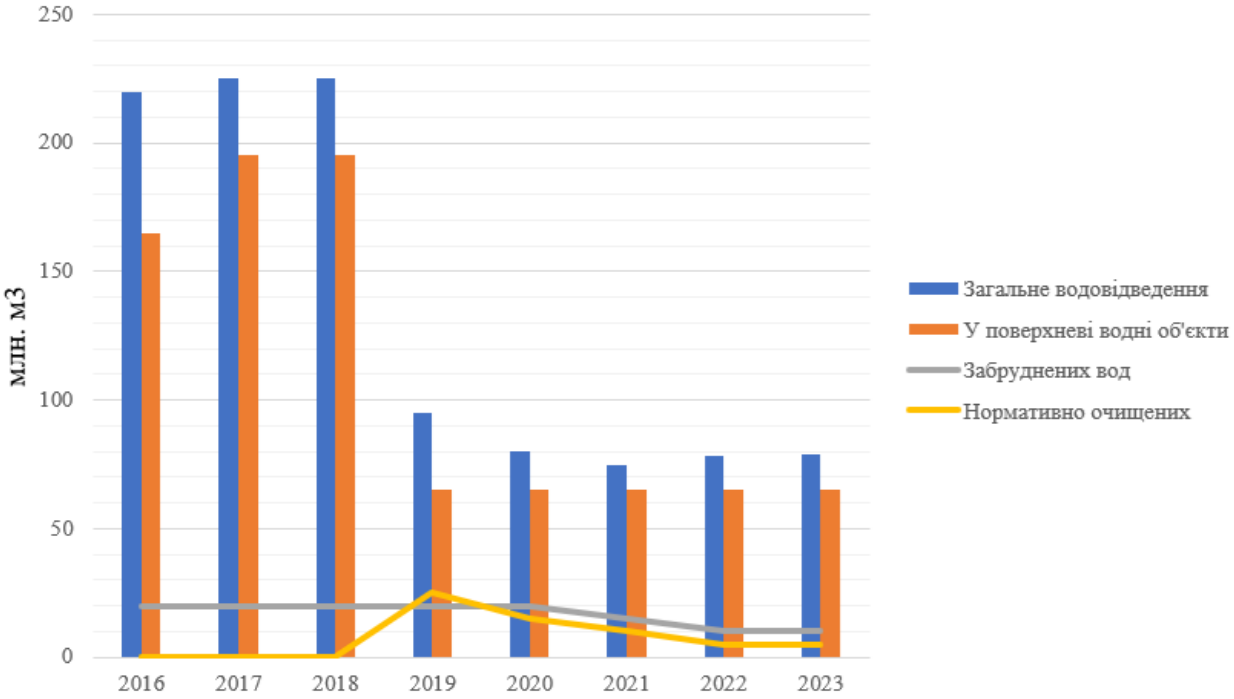
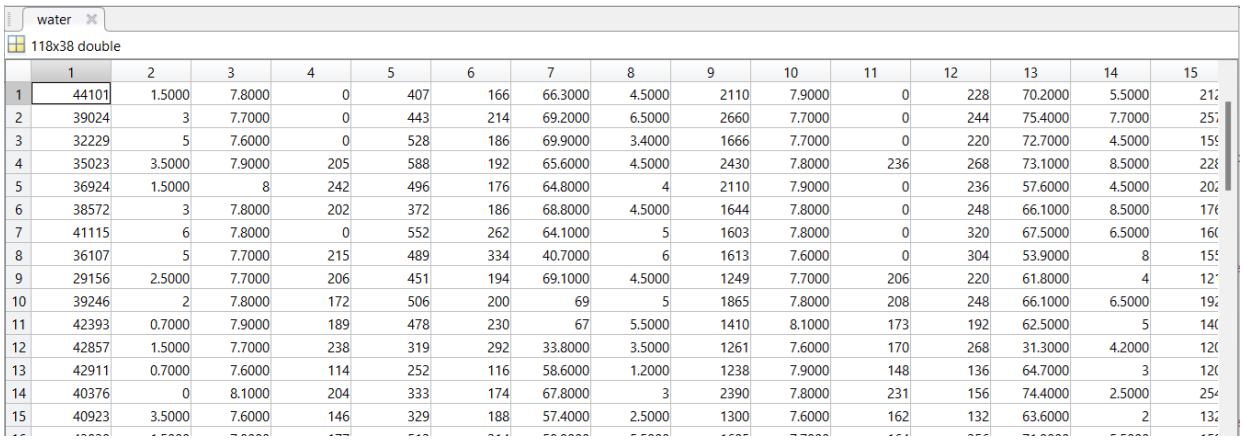


Рисунок 6.7 - Динаміка скиду зворотних вод за досліджуваний період

Інформація була сформульована у вигляді таблиці «об’єкт-властивість» (рисунок 6.8), попередньо нормована в гіперкуб [-1;1] (рисунок 6.9), та проведений кластерний аналіз за основними трьома класами.

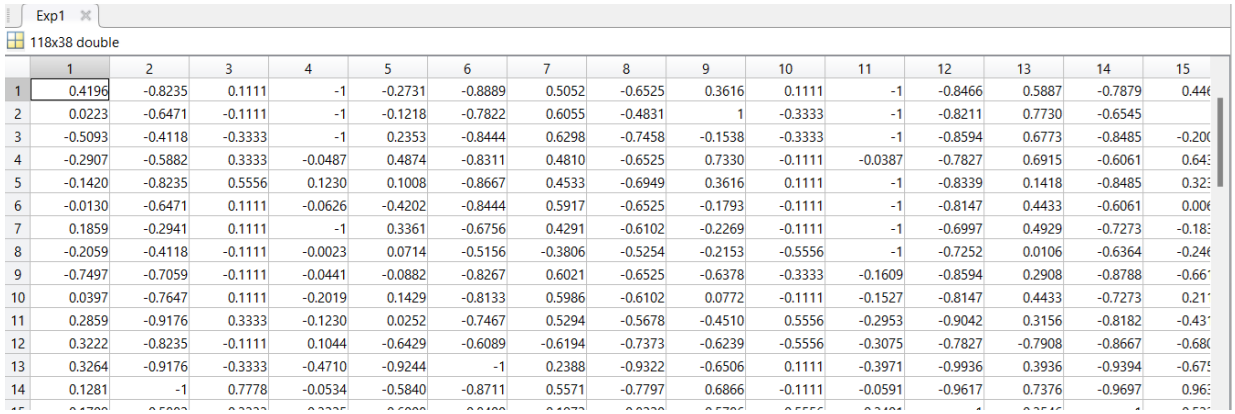


water

118x38 double

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	44101	1.5000	7.8000	0	407	166	66.3000	4.5000	2110	7.9000	0	228	70.2000	5.5000	21
2	39024	3	7.7000	0	443	214	69.2000	6.5000	2660	7.7000	0	244	75.4000	7.7000	25
3	32229	5	7.6000	0	528	186	69.9000	3.4000	1666	7.7000	0	220	72.7000	4.5000	15
4	35023	3.5000	7.9000	205	588	192	65.6000	4.5000	2430	7.8000	236	268	73.1000	8.5000	22
5	36924	1.5000	8	242	496	176	64.8000	4	2110	7.9000	0	236	57.6000	4.5000	20
6	38572	3	7.8000	202	372	186	68.8000	4.5000	1644	7.8000	0	248	66.1000	8.5000	17
7	41115	6	7.8000	0	552	262	64.1000	5	1603	7.8000	0	320	67.5000	6.5000	16
8	36107	5	7.7000	215	489	334	40.7000	6	1613	7.6000	0	304	53.9000	8	15
9	29156	2.5000	7.7000	206	451	194	69.1000	4.5000	1249	7.7000	206	220	61.8000	4	12
10	39246	2	7.8000	172	506	200	69	5	1865	7.8000	208	248	66.1000	6.5000	19
11	42393	0.7000	7.9000	189	478	230	67	5.5000	1410	8.1000	173	192	62.5000	5	14
12	42857	1.5000	7.7000	238	319	292	33.8000	3.5000	1261	7.6000	170	268	31.3000	4.2000	12
13	42911	0.7000	7.6000	114	252	116	58.6000	1.2000	1238	7.9000	148	136	64.7000	3	12
14	40376	0	8.1000	204	333	174	67.8000	3	2390	7.8000	231	156	74.4000	2.5000	25
15	40923	3.5000	7.6000	146	329	188	57.4000	2.5000	1300	7.6000	162	132	63.6000	2	13

Рисунок 6.8 – Вихідна вибірка аналізу води за основними хімічними характеристиками



Exp1

118x38 double

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0.4196	-0.8235	0.1111	-1	-0.2731	-0.8889	0.5052	-0.6525	0.3616	0.1111	-1	-0.8466	0.5887	-0.7879	0.44
2	0.0223	-0.6471	-0.1111	-1	-0.1218	-0.7822	0.6055	-0.4831	1	-0.3333	-1	-0.8211	0.7730	-0.6545	
3	-0.5093	-0.4118	-0.3333	-1	0.2353	-0.8444	0.6298	-0.7458	-0.1538	-0.3333	-1	-0.8594	0.6773	-0.8485	-0.20
4	-0.2907	-0.5882	0.3333	-0.0487	0.4874	-0.8311	0.4810	-0.6525	0.7330	-0.1111	-0.0387	-0.7827	0.6915	-0.6061	0.64
5	-0.1420	-0.8235	0.5556	0.1230	0.1008	-0.8667	0.4533	-0.6949	0.3616	0.1111	-1	-0.8339	0.1418	-0.8485	0.32
6	-0.0130	-0.6471	0.1111	-0.0626	-0.4202	-0.8444	0.5917	-0.6525	-0.1793	-0.1111	-1	-0.8147	0.4433	-0.6061	0.00
7	0.1859	-0.2941	0.1111	-1	0.3361	-0.6756	0.4291	-0.6102	-0.2269	-0.1111	-1	-0.6997	0.4929	-0.7273	-0.18
8	-0.2059	-0.4118	-0.1111	-0.0023	0.0714	-0.5156	-0.3806	-0.5254	-0.2153	-0.5556	-1	-0.7252	0.0106	-0.6364	-0.24
9	-0.7497	-0.7059	-0.1111	-0.0441	-0.0882	-0.8267	0.6021	-0.6525	-0.6378	-0.3333	-0.1609	-0.8594	0.2908	-0.8788	-0.66
10	0.0397	-0.7647	0.1111	-0.2019	0.1429	-0.8133	0.5986	-0.6102	0.0772	-0.1111	-0.1527	-0.8147	0.4433	-0.7273	0.21
11	0.2859	-0.9176	0.3333	-0.1230	0.0252	-0.7467	0.5294	-0.5678	-0.4510	0.5556	-0.2953	-0.9042	0.3156	-0.8182	-0.43
12	0.3222	-0.8235	-0.1111	0.1044	-0.6429	-0.6089	-0.6194	-0.7373	-0.6239	-0.5556	-0.3075	-0.7827	-0.7908	-0.8667	-0.68
13	0.3264	-0.9176	-0.3333	-0.4710	-0.9244	-1	0.2388	-0.9322	-0.6506	0.1111	-0.3971	-0.9936	0.3936	-0.9394	-0.67
14	0.1281	-1	0.7778	-0.0534	-0.5840	-0.8711	0.5571	-0.7797	0.6866	-0.1111	-0.0591	-0.9617	0.7376	-0.9697	0.96

Рисунок 6.9 – Вихідна вибірка аналізу води за основними хімічними характеристиками нормована в гіперкуб

Після того, як був проведений кластерний аналіз методом відновлення та фільтрації потоків даних за умов перетинних кластерів, надаються рекомендації щодо поліпшення характеристик та усунення негативного впливу від її використання в залежності від класу згідно з нормативами для питної води. Якість роботи методу була оцінена за допомогою індексів якості

кластеризації. В таблиці 6.12 наведені якісні показники кластеризації даних різними методами.

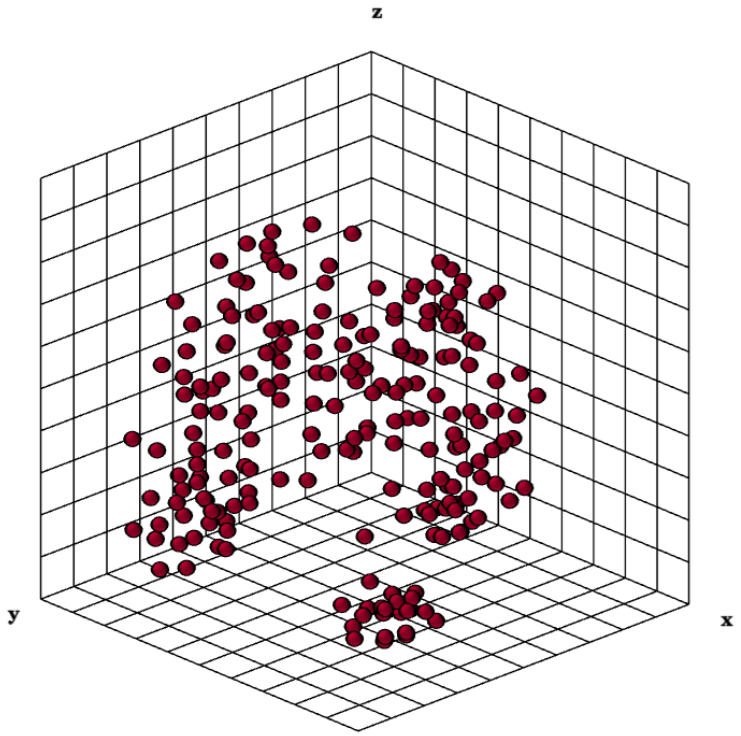


Рисунок 6.10 - Розкид спостережень в просторі гіперкуба

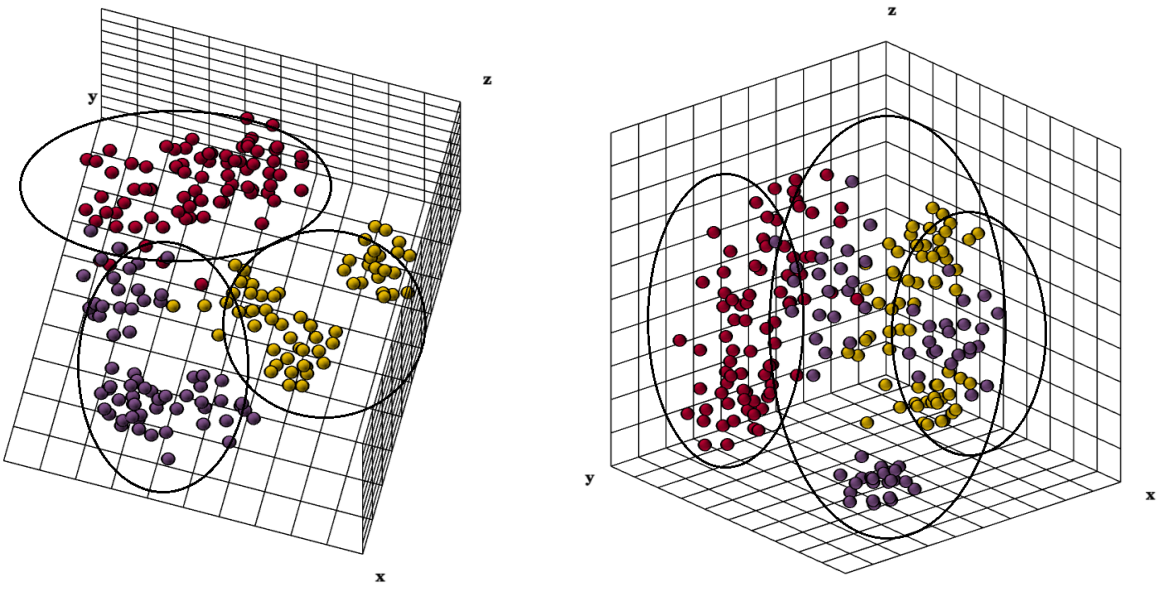


Рисунок 6.11 - Розбиття вибірки на 3 класи-кластери
(демонстрація з різних ракурсів)

Індекс суми квадратів помилок (SS) - низькі значення індексу відповідають кращій кластеризації.

Індекс Davies and Bouldin (DB) - індекс дорівнює нулю для тривіальної кластеризації, коли кожен об’єкт віднесено до різного класу. Крім того, кожен кластер повинен містити розумну кількість елементів. Індекс не визначено для випадку, коли об’єкти віднесені до одного кластеру.

Індекс Trace (TI) - найменше значення індексу для різної кількості класів *K* відповідає найкращому розбиття.

Індекс Calinski–Harabasz (CH) - високі значення індексу відповідають найкращій кластеризації.

Індекс Dunn (DI) - для даного розбиття на кластери, високе значення індексу означає найкращу кластеризацію.

Індекс PBM - чим більше значення індексу, тим кращі результати кластеризації.

Таблиця 6.12 – Оцінка якості кластеризації даних аналізу води в Харківській області

Метод	**SS**	**DB**	**TI**	**CH**	**DI**	**PBM**
K-means	4,891	66,214	344082,187	0,774	1,310	**428665,298**
Метод відновлення та фільтрації потоків даних за умов перетинних кластерів	**0,942**	19,311	**63926,297**	**6,985**	0,885	7399,157
FCM	1,770	21,639	102361,974	4,890	**2,730**	39028,109

Як видно з таблиці 6.12 метод відновлення та фільтрації потоків даних за умов перетинних кластерів з поставленою задачею впорався і демонструє достатні результати якості кластеризації за трьома показниками.

Підтверджено актом впровадження (акт від 29.06.2023р.).

6.5 Впровадження нечіткого методу кластеризації викривлених даних для класифікації пацієнтів з ознаками онкологічних захворювань

При виконанні спільних досліджень на КНП «Обласний центр онкології» було використані методи нечіткої кластеризації даних на основі оптимізаційних процедур для задач медичного діагностування хворих із ознаками онкологічних захворювань [301-312]. Для аналізу були взяті відомості діагностичних ознак ракових захворювань: текстура, площа, гладкість, компактність, вираженість, радіус, симетрія, фрактальна розмірність, кількість ділянок, тощо.

Інформація представлена у вигляді таблиці «об’єкт - властивість» (рисунок 6.12), для подальшого інтелектуального аналізу даних, вибірка була нормована в гіперкуб [-1;1] (рисунок 6.13).

Exp1 | Cancer

699x10 double

	1	2	3	4	5	6	7	8	9	10
1	5	1	1	1	2	1	3	1	1	2
2	5	4	4	5	7	10	3	2	1	2
3	3	1	1	1	2	2	3	1	1	2
4	6	8	8	1	3	4	3	7	1	2
5	4	1	1	3	2	1	3	1	1	2
6	8	10	10	8	7	10	9	7	1	4
7	1	1	1	1	2	10	3	1	1	2
8	2	1	2	1	2	1	3	1	1	2
9	2	1	1	1	2	1	1	1	5	2
10	4	2	1	1	2	1	2	1	1	2

Рисунок 6.12 – Діагностичні ознаки ракових захворювань

Exp1 | Cancer

699x10 double

	1	2	3	4	5	6	7	8	9	10
1	-0.1111	-1	-1	-1	-0.7778	-0.8000	-0.5556	-1	-1	-1
2	-0.1111	-0.3333	-0.3333	-0.1111	0.3333	1	-0.5556	-0.7778	-1	-1
3	-0.5556	-1	-1	-1	-0.7778	-0.6000	-0.5556	-1	-1	-1
4	0.1111	0.5556	0.5556	-1	-0.5556	-0.2000	-0.5556	0.3333	-1	-1
5	-0.3333	-1	-1	-0.5556	-0.7778	-0.8000	-0.5556	-1	-1	-1
6	0.5556	1	1	0.5556	0.3333	1	0.7778	0.3333	-1	1
7	-1	-1	-1	-1	-0.7778	1	-0.5556	-1	-1	-1
8	-0.7778	-1	-0.7778	-1	-0.7778	-0.8000	-0.5556	-1	-1	-1
9	-0.7778	-1	-1	-1	-0.7778	-0.8000	-1	-1	-0.1111	-1
10	-0.3333	-0.7778	-1	-1	-0.7778	-0.8000	-0.7778	-1	-1	-1
11	-1	-1	-1	-1	-1	-0.8000	-0.5556	-1	-1	-1
12	-0.7778	-1	-1	-1	-0.7778	-0.8000	-0.7778	-1	-1	-1
13	-0.1111	-0.5556	-0.5556	-0.5556	-0.7778	-0.4000	-0.3333	-0.3333	-1	1

Рисунок 6.13 – Нормовані дані

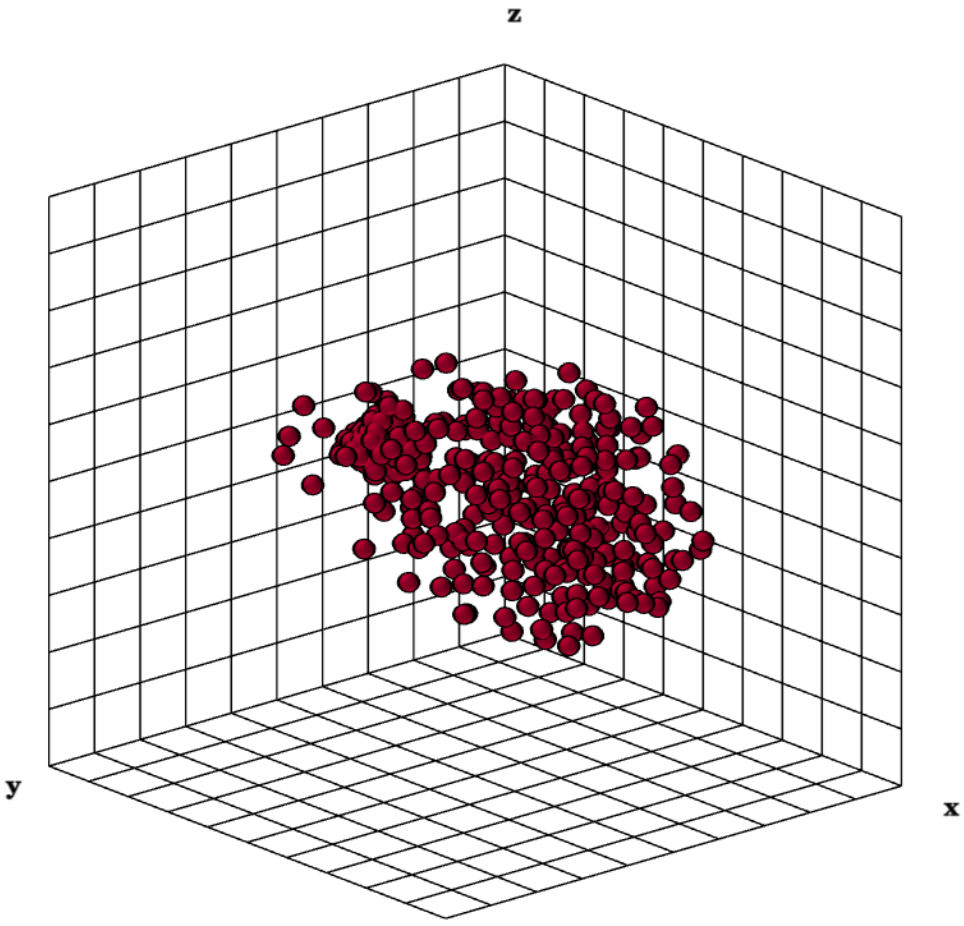


Рисунок 6.14 - Розкид спостережень в просторі гіперкуба

Попередньо була поставлена задача розбити спостереження про пацієнтів на 2 класи: доброякісні та злоякісні пухлини. На рисунку 6.15

продемонстрований кластерний аналіз даних, де ознаки діагностичних даних розбиті на 2 кластери. Якість кластеризації наведена в таблиці 6.13.

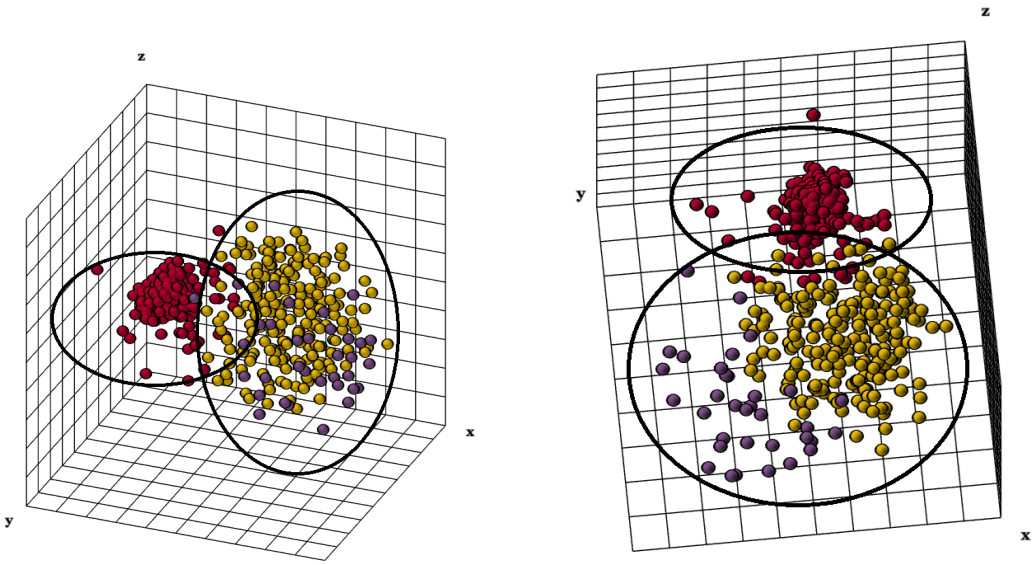


Рисунок 6.15 – Розбиття діагностичних даних на 2 кластери (демонстрація з різних ракурсів)

Таблиця 6.13 - Оцінка якості кластеризації даних

Метод	**SS**	**DB**	**TI**	**CH**	**DI**	**PBM**
K-means	5,406	57,683	2075591,347	**3,133**	**2,826**	13130110,514
FCM	5,411	56,728	2028958,754	3,119	2,665	**12503623,170**
Нечіткий метод кластеризації даних на основі оптимізаційних процедур	**4,712**	119,344	**1714284,385**	3,132	0,837	6667740,484

Аналізуючи вибірку методом нечіткої кластеризації даних на основі оптимізаційних процедур, які вміють працювати навіть з тими спостереженнями, які ще повністю заповнені (немає ознак, не проведений аналіз) та аналізувати їх. З допомогою запропонованого методу, виявили ще один клас-кластер. Пропонується розбити вибірку даних, на 3 кластери, перевірити якість кластеризації даних та проаналізувати всі 3 кластери окремо. На рисунку 6.16 продемонстрована кластеризація даних на 3 кластери, з урахуванням тих спостережень, інформація яких не аналізувалась.

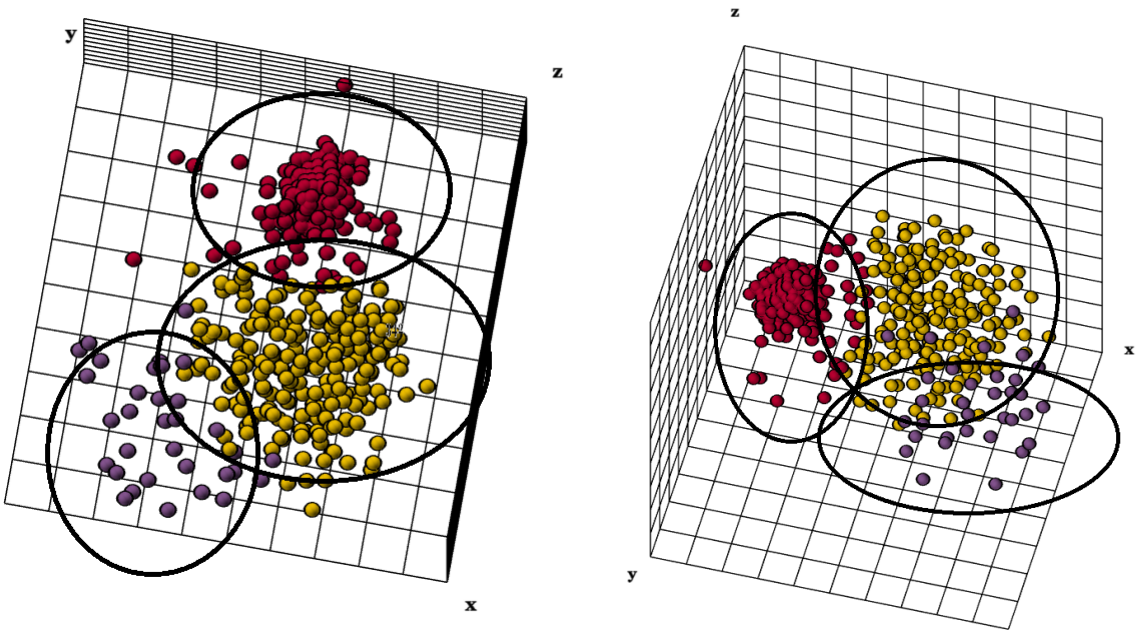


Рисунок 6.16 – Розбиття вибірки діагностичних даних на 3 кластери (демонстрація з різних ракурсів)

В таблиці 6.14 наведений аналіз якісних характеристик кластеризації за основними показниками.

Таблиця 6.14 - Оцінка якості кластеризації даних

Метод	SS	DB	TI	CH	DI	PBM
K-means	2,744	100,427	1052551,092	4,431	**0,154**	340954,404
FCM	**2,847**	85,064	1088165,657	4,446	0,056	**567262,454**
Нечіткий метод кластеризації даних на основі оптимізаційних процедур (FCO)	2,522	92,670	**1020380,863**	**5,020**	0,079	353858,179

Аналізуючи таблиці 6.13 та 6.14, можна зробити висновок, що зі збільшенням кількості кластерів та з урахуванням всіх спостережень, якість розбиття спостережень підвищується. Таким чином, можна зробити висновок, що крім 2 класів ракових пухлин (злоякісна, доброякісна), є ще один клас. Цей клас можна віднести до початкових ознак захворюваності пацієнта (ранні стадії), або ті пацієнти, які почали лікування.

Підтверджено актом впровадження (акт від 14.11.2023р.).

6.6 Висновок до розділу 6

1. 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.).

2. Проведена оцінка стану будинків для визначення готовності до експлуатації в зимових умовах за допомогою методу адаптивної нечіткої кластеризації даних різної природи на основі даних підприємства ТОВ

«КОМУНСЕРВІС 2018». Результати впровадження методу адаптивної кластеризації викривлених даних на основі правдоподібного підходу довели доцільність використання їх в аналізі та оцінці стану житлових будинків, що дозволяють прискорити аналіз та прийняття обґрунтованих рішень щодо першочерговості відновлення будинків, в залежності від категорії пошкоджень та зношеності (акт впровадження 12.04.2023р.)

3. Вирішення практичної задачі класифікації технологічних процесів на будівництві за допомогою методу адаптивної нечіткої кластеризації даних будівельних та монтажних робіт загального призначення для отримання класифікації технологічних процесів на будівництві з метою підвищення їх ефективності на ТОВ «Будівельно-монтажне підприємство - 168» (акт впровадження 21.12.2023р.).

4. Була проведена імплементація методу 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 29.06.2023р.).

5. 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 14.11.2023р.).

6. Запропоновані методи дозволяють:

- 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.

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

ВИСНОВКИ

У дисертаційній роботі 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.

Наукова новизна отриманих особисто здобувачкою полягає у такому:

1. Уперше запропоновано адаптивні ймовірнісні, можливісні та правдоподібні методи нечіткої кластеризації потоків викривлених даних, що 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, що 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 дозволило 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 дозволило 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 дозволив уникнути застрягання в локальних екстремумах.

8. Уперше запропоновано ефективні підходи до вирішення багатоекстремальної задачі правдоподібної нечіткої кластеризації на основі модифікованих оптимізаційних процедур божевільних котів та сірих вовків, що дозволило скоротити час вирішення задачі.

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

10. Удосконалено еволюційний метод на основі косяків риб, що підтвердив свою ефективність у вирішенні 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, що 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 та введено рандомізовану модифікацію базової процедури 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.

Практична значущість результатів дослідження 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 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.

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).

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Hartigan, J. A. (1975). *Clustering algorithms*. John Wiley & Sons, Inc.
2. Xu, R., & Wunsch, D. (2008). *Clustering*. John Wiley & Sons.
3. Miyamoto, S., Ichihashi, H., Honda, K., & Ichihashi, H. (2008). *Algorithms for fuzzy clustering* (Vol. 10). Heidelberg: Springer.
4. Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: application and trends. *Artificial Intelligence Review*, *56*(7), 6439-6475.
5. Dol, S. M., & Jawandhiya, P. M. (2023). Classification technique and its combination with clustering and association rule mining in educational data mining—A survey. *Engineering Applications of Artificial Intelligence*, *122*, 106071.
6. D'Urso, P., De Giovanni, L., Federico, L., & Vitale, V. (2023). Fuzzy clustering of spatial interval-valued data. *Spatial Statistics*, *57*, 100764.
7. Gong, M., Zhao, Y., Li, H., Qin, A. K., Xing, L., Li, J., ... & Liu, Y. (2023). Deep fuzzy variable C-means clustering incorporated with curriculum learning. *IEEE Transactions on Fuzzy Systems*.
8. Rokach, L., & Maimon, O. (2005). Clustering methods. *Data mining and knowledge discovery handbook*, 321-352.
9. Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (Eds.). (2015). *Handbook of cluster analysis*. CRC press.
10. Gelbard, R., Goldman, O., & Spiegler, I. (2007). Investigating diversity of clustering methods: An empirical comparison. *Data & Knowledge Engineering*, *63*(1), 155-166.
11. Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2002). Cluster validity methods: part I. *ACM Sigmod Record*, *31*(2), 40-45.
12. Wierzchoń, S. T., & Kłopotek, M. A. (2018). *Modern algorithms of cluster analysis* (Vol. 34). Springer International Publishing.

13. Höppner, F., Klawonn, F., Kruse, R., & Runkler, T. (1999). *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*. John Wiley & Sons.

14. Sato-Ilic, M., & Jain, L. C. (2006). *Innovations in fuzzy clustering*. Heidelberg: Springer.

15. Ruspini, E. H. (1970). Numerical methods for fuzzy clustering. *Information Sciences*, *2*(3), 319-350.

16. Wu, Z., Zhao, Y., Wang, W., & Li, C. (2023). Adaptive weighted fuzzy clustering based on intra-cluster data divergence. *Neurocomputing*, *552*, 126550.

17. Chithaluru, P., Jena, L., Singh, D., & Ravi Teja, K. M. V. (2022). An adaptive fuzzy-based clustering model for healthcare wireless sensor networks. In *Ambient Intelligence in Health Care: Proceedings of ICAIHC 2022* (pp. 1-10). Singapore: Springer Nature Singapore.

18. Yishan, Z., Chenxuan, Z., Fuqiang, L., Zongxin, H., & Yanhua, L. (2023, June). An adaptive method of selecting typical days based on improved fuzzy clustering algorithm. In *Sixth International Conference on Intelligent Computing, Communication, and Devices (ICCD 2023)* (Vol. 12703, pp. 213-222). SPIE.

19. Zhou, J., Huang, C., Gao, C., Wang, Y., Shen, X., & Wu, X. (2024). Weighted Subspace Fuzzy Clustering with Adaptive Projection. *International Journal of Intelligent Systems*, *2024*.

20. Li, D., Zhou, S., & Pedrycz, W. (2023). Accelerated Fuzzy C-Means Clustering Based on New Affinity Filtering and Membership Scaling. *IEEE Transactions on Knowledge and Data Engineering*.

21. Klawonn, F., & Höppner, F. (2003). What is fuzzy about fuzzy clustering? Understanding and improving the concept of the fuzzifier. In *Advances in Intelligent Data Analysis V: 5th International Symposium on Intelligent Data Analysis, IDA 2003, Berlin, Germany, August 28-30, 2003. Proceedings 5* (pp. 254-264). Springer Berlin Heidelberg.

22. Krishnapuram, R., & Keller, J. M. (1993). A possibilistic approach to clustering. *IEEE transactions on fuzzy systems*, *1*(2), 98-110.

23. Nascimento, S., Mirkin, B., & Moura-Pires, F. (2000, May). A fuzzy clustering model of data and fuzzy c-means. In *Ninth IEEE International Conference on Fuzzy Systems. FUZZ-IEEE 2000 (Cat. No. 00CH37063)* (Vol. 1, pp. 302-307). IEEE.

24. Dave, R. N., & Sen, S. (1997, September). Noise clustering algorithm revisited. In *1997 Annual Meeting of the North American Fuzzy Information Processing Society-NAFIPS (Cat. No. 97TH8297)* (pp. 199-204). IEEE.

25. Ruspini, E. H., Bezdek, J. C., & Keller, J. M. (2019). Fuzzy clustering: A historical perspective. *IEEE Computational Intelligence Magazine*, *14*(1), 45-55.

26. Srivastava, A., & Nawfal, M. (2024). Parallelization of the K-Means Algorithm with Applications to Big Data Clustering. *arXiv preprint arXiv:2405.12052*.

27. Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464-1480.

28. Jang, J. S. (1993). ANFIS: adaptive-network-based fuzzy inference system. *IEEE transactions on systems, man, and cybernetics*, *23*(3), 665-685.

29. Gorshkov, Y., Kolodyazhniy, V., & Bodyanskiy, Y. (2009, June). New recursive learning algorithms for fuzzy Kohonen clustering network. In *Proc. 17th Int. Workshop on Nonlinear Dynamics of Electronic Systems* (pp. 58-61).

30. Kasabov, N. K. (2015). Evolving connectionist systems for adaptive learning and knowledge discovery: Trends and directions. *Knowledge-Based Systems*, *80*, 24-33.

31. Gan, G., Ma, C., & Wu, J. (2020). *Data clustering: theory, algorithms, and applications*. Society for Industrial and Applied Mathematics.

32. Шафроненко, А., Бодянський, Є., & Плісс, І. (2022). Нечіткі методи інтелектуального аналізу даних.

33. Gorban, A. N., Kégl, B., Wunsch, D. C., & Zinovyev, A. Y. (Eds.). (2008). *Principal manifolds for data visualization and dimension reduction* (Vol. 58, pp. 96-130). Berlin: Springer.

34. Marwala, T. (Ed.). (2009). *Computational Intelligence for Missing Data Imputation, Estimation, and Management: Knowledge Optimization Techniques: Knowledge Optimization Techniques*. IGI Global.

35. Abidin, N. Z., Ismail, A. R., & Emran, N. A. (2018). Performance analysis of machine learning algorithms for missing value imputation. *International Journal of Advanced Computer Science and Applications*, *9*(6).

36. Garciarena, U., & Santana, R. (2017). An extensive analysis of the interaction between missing data types, imputation methods, and supervised classifiers. *Expert Systems with Applications*, *89*, 52-65.

37. Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

38. Gorban, A. N., Rossiev, A. A., & Wunsch, D. C. (2000). Neural network modeling of data with gaps. *Радіоелектроніка, інформатика, управління*, (1 (3)), 47-55.

39. Agrawal, A. V., Soni, M., Keshta, I., Savithri, V., Abdinabievna, P. S., & Singh, S. (2023). A probability-based fuzzy algorithm for multi-attribute decision-analysis with application to aviation disaster decision-making. *Decision Analytics Journal*, *8*, 100310.

40. Tkacz, M. (2005). Artificial neural networks in incomplete data sets processing. In *Intelligent Information Processing and Web Mining: Proceedings of the International IIS: IIPWM'05 Conference held in Gdansk, Poland, June 13–16, 2005* (pp. 577-583). Springer Berlin Heidelberg.

41. Golden, R. M. (1996). *Mathematical methods for neural network analysis and design*. MIT Press.

42. Bodyanskiy, Y., & Shafronenko, A. Robust adaptive fuzzy clustering for data with missing values. Transformation, 1, 1.

43. D'Urso, P., & Leski, J. M. (2020). Fuzzy clustering of fuzzy data based on robust loss functions and ordered weighted averaging. *Fuzzy Sets and Systems*, *389*, 1-28.

44. Braun, H. (2013). Neuronale Netze: Optimierung durch Lernen und Evolution. Springer-Verlag.

45. Dracopoulos, D. C. (2013). *Evolutionary learning algorithms for neural adaptive control*. Springer.

46. Shepherd, A. J. (2012). Second-order methods for neural networks: Fast and reliable training methods for multi-layer perceptrons. Springer Science & Business Media.

47. Lyashenko, V., Kobylin, O., & Shafronenko, A. (2019, September). Wavelet analysis and decomposition into color spaces in researching of human fluorescently labeled images tissues. In *2019 IEEE 8th International Conference on Advanced Optoelectronics and Lasers (CAOL)* (pp. 618-621). IEEE. DOI: 10.1109/CAOL46282.2019.9019575

48. Haykin, S. (1998). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

49. Bodyanskiy, Y., Kolodyazhniy, V., & Stephan, A. (2001). An adaptive learning algorithm for a neuro-fuzzy network. In *Computational Intelligence. Theory and Applications: International Conference, 7th Fuzzy Days Dortmund, Germany, October 1–3, 2001 Proceedings 7* (pp. 68-75). Springer Berlin Heidelberg.

50. Yamakawa, T. (1992). A neo fuzzy neuron and its applications to system identification and prediction of the system behavior. In Proc. of the 2nd Int. Conf. on Fuzzy Logic & Neural Networks (pp. 477-483).

51. Uchino, E., & Yamakawa, T. (1997). Soft computing based signal prediction, restoration, and filtering. Intelligent hybrid systems: fuzzy logic, neural networks, and genetic algorithms, 331-351.

52. Kruse, R. (2008). Fuzzy neural network. *Scholarpedia*, *3*(11), 6043.

53. Itakura, F. (1975). Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on acoustics, speech, and signal processing,* 23(1), 67-72

54. Miki, T. S. U. T. O. M. U., & Yamakawa, T. A. K. E. S. H. I. (1999). Analog implementation of neo-fuzzy neuron and its on-board learning. Computational Intelligence and Applications, 144, 149.

55. Hathaway, R. J., & Bezdek, J. C. (2001). Fuzzy c-means clustering of incomplete data. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 31(5), 735-744.

56. Hashemi, S. E., Gholian-Jouybari, F., & Hajiaghaei-Keshteli, M. (2023). A fuzzy C-means algorithm for optimizing data clustering. *Expert Systems with Applications*, *227*, 120377.

57. Bezdek, J. C. (2013). *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.

58. Bezdek, J. C., Ehrlich, R., & Full, W. (1984). FCM: The fuzzy c-means clustering algorithm. *Computers & geosciences*, *10*(2-3), 191-203.

59. Zhang, Y., Chen, T., Jiang, Y., & Wang, J. (2023). Possibilistic c-means clustering based on the nearest-neighbour isolation similarity. *Journal of Intelligent & Fuzzy Systems*, *44*(2), 1781-1792.

60. Ja, H. (1979). A k-means clustering algorithm. *JR Stat. Soc. Ser. C-Appl. Stat.*, *28*, 100-108.

61. Wongkhuenkaew, R., Auephanwiriyakul, S., Theera-Umpon, N., Teeyapan, K., & Yeesarapat, U. (2023). Fuzzy K-nearest neighbor based dental fluorosis classification using multi-prototype unsupervised possibilistic fuzzy clustering via cuckoo search algorithm. *International Journal of Environmental Research and Public Health*, *20*(4), 3394.

62. Bezdek, J. C., Keller, J., Krisnapuram, R., & Pal, N. (1999). *Fuzzy models and algorithms for pattern recognition and image processing* (Vol. 4). Springer Science & Business Media.

63. Raghavan, V., & Hafez, A. (2000, June). Dynamic data mining. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems* (pp. 220-229). Berlin, Heidelberg: Springer Berlin Heidelberg.

64. Brin, S., & Page, L. (1998). Dynamic data mining: Exploring large rule spaces by sampling. *Manuscript, Department of Computer Science, Stanford University, Stanford, CA*.

65. Gaber, M. M. (2012). Advances in data stream mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(1), 79-85.

66. Gaber, M. M., Zaslavsky, A., & Krishnaswamy, S. (2010). Data stream mining. *Data mining and knowledge discovery handbook*, 759-787.

67. De Francisci Morales, G., Bifet, A., Khan, L., Gama, J., & Fan, W. (2016, August). Iot big data stream mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2119-2120).

68. Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)* (pp. 42-47). IEEE.

69. Fasel, D., & Meier, A. (2014). *Big data*. Springer Vieweg.

70. Oussous, A., Benjelloun, F. Z., Lahcen, A. A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University-Computer and Information Sciences*, *30*(4), 431-448.

71. Che, D., Safran, M., & Peng, Z. (2013). From big data to big data mining: challenges, issues, and opportunities. In *Database Systems for Advanced Applications: 18th International Conference, DASFAA 2013, International Workshops: BDMA, SNSM, SeCoP, Wuhan, China, April 22-25, 2013. Proceedings 18* (pp. 1-15). Springer Berlin Heidelberg.

72. Mitsa, T. (2010). *Temporal data mining*. Chapman and Hall/CRC.

73. Post, A. R., & Harrison Jr, J. H. (2008). Temporal data mining. *Clinics in Laboratory Medicine*, *28*(1), 83-100.

74. Shahnawaz, M., Ranjan, A., & Danish, M. (2011). Temporal data mining: an overview. *International Journal of Engineering and Advanced Technology*, *1*(1), 2249-8958.

75. Karlin, S., & Studden, W. J. (1966). Tchebycheff systems: With applications in analysis and statistics. (No Title).

76. Mahalanobis, P. C. (2018). On the generalized distance in statistics. *Sankhyā: The Indian Journal of Statistics, Series A (2008-)*, *80*, S1-S7.

77. Yang, S. S., & Tseng, C. S. (1996). An orthogonal neural network for function approximation. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 26(5), 779-785.

78. Lee, T. T., & Jeng, J. T. (1998). The Chebyshev-polynomials-based unified model neural networks for function approximation. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 28(6), 925-935.

79. Ebadzadeh, M. M., & Salimi-Badr, A. (2017). IC-FNN: a novel fuzzy neural network with interpretable, intuitive, and correlated-contours fuzzy rules for function approximation. *IEEE Transactions on Fuzzy Systems*, *26*(3), 1288-1302.

80. Andras, P. (1999). Orthogonal RBF neural network approximation. Neural Processing Letters, 9, 141-151.

81. Sher, C. F., Tseng, C. S., & Chen, C. S. (2001). Properties and performance of orthogonal neural network in function approximation. International Journal of intelligent systems, 16(12), 1377-1392.

82. Patra, J. C., & Kot, A. C. (2002). Nonlinear dynamic system identification using Chebyshev functional link artificial neural networks. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 32(4), 505-511.

83. Bodyanskiy, Y., Kolodyazhniy, V., & Slipchenko, O. (2004). Structural and synaptic adaptation in the artificial neural networks with orthogonal activation functions. In Sci. Proc. of Riga Technical University. Comp. Sci., Inf. Technology and Management Sci (No. 20, pp. 69-76).

84. Hinneburg, A., & Keim, D. A. (1998). *An efficient approach to clustering in large multimedia databases with noise* (Vol. 98, pp. 58-65). Konstanz, Germany: Bibliothek der Universität Konstanz.

85. Bodyanskiy, Y., Shafronenko, A., & Volkova, V. (2012). Adaptive clustering of incomplete data using neuro-fuzzy Kohonen network. Artificial Intelligence Methods and Techniques for Business and Engineering Applications"–Rzeszow-Sofia: ITHEA, 287-296.

86. 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.

87. Nikolova, M. (2004). A variational approach to remove outliers and impulse noise. Journal of Mathematical Imaging and Vision, 20(1), 99-120.

88. Hathaway, R. J., Bezdek, J. C., & Hu, Y. (2000). Generalized fuzzy c-means clustering strategies using L/sub p/norm distances. IEEE transactions on Fuzzy Systems, 8(5), 576-582.

89. Rutkowski, L. (2008). Computational. *Intelligence Methods and, Techniques; Springer: Berlin/Heidelberg, Germany*.

90. Mumford, C. L. (2009). Synergy in computational intelligence. In *Computational Intelligence: Collaboration, Fusion and Emergence* (pp. 3-21). Berlin, Heidelberg: Springer Berlin Heidelberg.

91. Kruse, B., & Klawonn, M. (2013). Held Computational Intelligence: A Methodological Introduction.

92. Kroll, A. (2013). *Computational Intelligence: Eine Einführung in Probleme, Methoden und technische Anwendungen*. Oldenbourg Wissenschaftsverlag Verlag.

93. Kacprzyk, J., & Pedrycz, W. (Eds.). (2015). *Springer handbook of computational intelligence*. Springer.

94. Gruber, H. E. (1988). The evolving systems approach to creative work. *Creativity Research Journal*, *1*(1), 27-51.

95. Bezdek, J. C. (1980). A convergence theorem for the fuzzy ISODATA clustering algorithms. *IEEE transactions on pattern analysis and machine intelligence*, (1), 1-8.

96. Zhou, J., & Hung, C. C. (2007). A generalized approach to possibilistic clustering algorithms. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *15*(supp02), 117-138.

97. Zhou, J., Wang, Q., Hung, C. C., & Yi, X. (2015). Credibilistic clustering: the model and algorithms. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *23*(04), 545-564.

98. Liu, B. (2006). A survey of credibility theory. *Fuzzy optimization and decision making*, *5*, 387-408.

99. Zhou, J., Wang, Q., Hung, C. C., & Yi, X. (2015). Credibilistic clustering: the model and algorithms. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *23*(04), 545-564.

100. Zhou, J., Wang, Q., Hung, C. C., & Yang, F. (2017). Credibilistic clustering algorithms via alternating cluster estimation. *Journal of Intelligent Manufacturing*, *28*, 727-738.

101. Tao, H., Hou, C., Liu, X., Liu, T., Yi, D., & Zhu, J. (2018, April). Reliable multi-view clustering. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 32, No. 1).

102. Shamir, O., & Tishby, N. (2008). On the reliability of clustering stability in the large sample regime. *Advances in neural information processing systems*, *21*.

103. Muruganandam, S., & Renjit, J. A. (2021). Real-time reliable clustering and secure transmission scheme for QoS development in MANET. *Peer-to-Peer Networking and Applications*, *14*(6), 3502-3517.

104. Bagherinia, A., Minaei-Bidgoli, B., Hosseinzadeh, M., & Parvin, H. (2021). Reliability-based fuzzy clustering ensemble. *Fuzzy Sets and Systems*, *413*, 1-28.

105. Ramamoorthy, C. V., Bhide, A., & Srivastava, J. (1987, January). Reliable clustering techniques for large, mobile packet radio networks. In *Proceedings-IEEE INFOCOM* (pp. 218-226). IEEE.

106. Kärkkäinen, I. (2006). *Methods for fast and reliable clustering* (Doctoral dissertation, Joensuun yliopisto).

107. Nigro, L., Cicirelli, F., & Fränti, P. (2022, September). Efficient and reliable clustering by parallel random swap algorithm. In *2022 IEEE/ACM 26th International Symposium on Distributed Simulation and Real Time Applications (DS-RT)* (pp. 25-28). IEEE.

108. Park, D. C., & Dagher, I. (1994, June). Gradient based fuzzy c-means (GBFCM) algorithm. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)* (Vol. 3, pp. 1626-1631). IEEE.

109. Hussain, I., Sinaga, K. P., & Yang, M. S. (2023). Unsupervised multiview fuzzy c-means clustering algorithm. *Electronics*, *12*(21), 4467.

110. Yang, M. S., & Benjamin, J. B. (2023). Sparse possibilistic c-means clustering with Lasso. *Pattern Recognition*, *138*, 109348.

111. Chung, F. L., & Lee, T. (1994). Fuzzy competitive learning. *Neural Networks*, *7*(3), 539-551.

112. Zhang, Y., Chen, T., Jiang, Y., & Wang, J. (2023). Possibilistic c-means clustering based on the nearest-neighbour isolation similarity. *Journal of Intelligent & Fuzzy Systems*, *44*(2), 1781-1792.

113. Gustafson, D. E., & Kessel, W. C. (1979, January). Fuzzy clustering with a fuzzy covariance matrix. In *1978 IEEE conference on decision and control including the 17th symposium on adaptive processes* (pp. 761-766). IEEE.

114. Sherman, J., & Morrison, W. J. (1950). Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, *21*(1), 124-127.

115. Harville, D. A. (1998). Matrix algebra from a statistician's perspective.

116. Bodyanskiy, Y., & Shafronenko, A. Robust adaptive fuzzy clustering for data with missing values. *Transformation*, *1*, 1.

117. 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_4

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

119. 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

120. 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

121. 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.

122. 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.

123. 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.

124. 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.

125. Bodyanskiy, Y. V., Shafronenko, A., & Rudenko, D. (2019). Online Neuro Fuzzy Clustering of Data with Omissions and Outliers based on Completion Strategy. In *CMIS* (pp. 18-27).

126. 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.

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

128. Shafronenko, A., Bodyanskiy, Y., Pliss, I., & Popov, S. (2020, September). Evolving neo-fuzzy system for distorted data online processing. In *2020 10th International Conference on Advanced Computer Information Technologies (ACIT)* (pp. 352-355). IEEE.

129. Bodyanskiy, Y. V., Shafronenko, A., & Klymova, I. (2021, April). Adaptive Recovery of Distorted Data Based on Credibilistic Fuzzy Clustering Approach. In *COLINS* (pp. 6-15).

130. 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.

131. Ji, Z., Liu, J., Cao, G., Sun, Q., & Chen, Q. (2014). Robust spatially constrained fuzzy c-means algorithm for brain MR image segmentation. *Pattern recognition*, *47*(7), 2454-2466.

132. 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.

133. Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, *1*(3), 231-240.

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

135. Шафроненко А.Ю., Суліма В.С. (2023) Нечітка правдоподібна кластеризація багатовимірних даних з гіпереліпсоїдальними класами. In *27-й міжнародний молодіжний форум «Радіоелектроніка і молодь у XXI столітті»*, 78-79.

136. Шафроненко А.Ю., Авлякулов Т.Е. (2023) Нечітка кластеризація даних з різною щільністю. In *27-й міжнародний молодіжний форум «Радіоелектроніка і молодь у XXI столітті»,* 80-81.

137. 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.

138. Cao, F., Estert, M., Qian, W., & Zhou, A. (2006, April). Density-based clustering over an evolving data stream with noise. In *Proceedings of the 2006 SIAM international conference on data mining* (pp. 328-339). Society for industrial and applied mathematics.

139. Amini, A., Saboohi, H., & Wah, T. Y. (2013, December). A multi density-based clustering algorithm for data stream with noise. In *2013 IEEE 13th International Conference on Data Mining Workshops* (pp. 1105-1112). IEEE.

140. Guan, C., Yuen, K. K. F., & Chen, Q. (2017, June). Towards a hybrid approach of k-means and density-based spatial clustering of applications with noise for image segmentation. In *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)* (pp. 396-399). IEEE.

141. Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, *33*(3), 1065-1076.

142. Nadaraya, E. A. (1965). On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, *10*(1), 186-190.

143. Watson, G. S. (1964). Smooth regression analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, 359-372.

144. Hinneburg, A., & Gabriel, H. H. (2007, September). Denclue 2.0: Fast clustering based on kernel density estimation. In *International symposium on intelligent data analysis* (pp. 70-80). Berlin, Heidelberg: Springer Berlin Heidelberg.

145. Rodriguez, A., & Laio, A. (2014). Clustering by fast search and find of density peaks. *science*, *344*(6191), 1492-1496.

146. Bodyanskiy, Y. V., Pliss, I. P., & Shafronenko, A. Y. (2022). Швидка нечітка правдоподібна кластеризація на основі аналізу піків щільності розподілу даних. *Radio Electronics, Computer Science, Control*, (1), 76-76.

147. Hinneburg, A., & Keim, D. A. (2003). A general approach to clustering in large databases with noise. *Knowledge and information systems*, *5*, 387-415.

148. Rehioui, H., Idrissi, A., Abourezq, M., & Zegrari, F. (2016). DENCLUE-IM: A new approach for big data clustering. *Procedia Computer Science*, *83*, 560-567.

149. Zgurovsky, M. Z., & Zaychenko, Y. P. (2020). *Big data: conceptual analysis and applications*. Springer International Publishing.Ghosh, S., & Mitra, S. (2013). Clustering large data with uncertainty. *Applied Soft Computing*, *13*(4), 1639-1645.

150. Ravi, V., Bin, M., & Ravi Kumar, P. (2006). Threshold accepting based fuzzy clustering algorithms. *International Journal of Uncertainty, Fuzziness and Knowledge-based systems*, *14*(05), 617-632.

151. Balkis, A., Yahia, S. B., & Bouzeghoub, A. (2012, November). A new algorithm for fuzzy clustering able to find the optimal number of clusters. In *2012*

IEEE 24th International Conference on Tools with Artificial Intelligence (Vol. 1, pp. 806-813). IEEE.

152. Begum, N., Ulanova, L., Wang, J., & Keogh, E. (2015, August). Accelerating dynamic time warping clustering with a novel admissible pruning strategy. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 49-58).

153. Bie, R., Mehmood, R., Ruan, S., Sun, Y., & Dawood, H. (2016). Adaptive fuzzy clustering by fast search and find of density peaks. *Personal and Ubiquitous Computing*, *20*, 785-793.

154. Ienco, D., & Bordogna, G. (2018). Fuzzy extensions of the DBScan clustering algorithm. *Soft Computing*, *22*(5), 1719-1730.

155. Bian, Z., Chung, F. L., & Wang, S. (2020). Fuzzy density peaks clustering. *IEEE Transactions on Fuzzy Systems*, *29*(7), 1725-1738.

156. Tong, W., Liu, S., & Gao, X. Z. (2021). A density-peak-based clustering algorithm of automatically determining the number of clusters. *Neurocomputing*, *458*, 655-666.

157. Yuan, X., Yu, H., Liang, J., & Xu, B. (2021). A novel density peaks clustering algorithm based on K nearest neighbors with adaptive merging strategy. *International Journal of Machine Learning and Cybernetics*, *12*(10), 2825-2841.

158. Wang, Y., & Yang, Y. (2021). Relative density-based clustering algorithm for identifying diverse density clusters effectively. *Neural Computing and Applications*, *33*(16), 10141-10157.

159. Yaohui, L., Zhengming, M., & Fang, Y. (2017). Adaptive density peak clustering based on K-nearest neighbors with aggregating strategy. *Knowledge-Based Systems*, *133*, 208-220.

160. Guo, W., Wang, W., Zhao, S., Niu, Y., Zhang, Z., & Liu, X. (2022). Density peak clustering with connectivity estimation. *Knowledge-Based Systems*, *243*, 108501.

161. Xu, X., Ding, S., & Shi, Z. (2018). An improved density peaks clustering algorithm with fast finding cluster centers. *Knowledge-Based Systems*, *158*, 65-74.

162. Li, C., Ding, S., Xu, X., Hou, H., & Ding, L. (2023). Fast density peaks clustering algorithm based on improved mutual K-nearest-neighbor and sub-cluster merging. *Information Sciences*, *647*, 119470.

163. Sun, L., Liu, R., Xu, J., & Zhang, S. (2019). An adaptive density peaks clustering method with Fisher linear discriminant. *IEEE Access*, *7*, 72936-72955.

164. Xu, X., Ding, S., Wang, Y., Wang, L., & Jia, W. (2021). A fast density peaks clustering algorithm with sparse search. *Information Sciences*, *554*, 61-83.

165. Shafronenko, A., Bodyanskiy, Y., Pliss, I., & Irina, K. (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).

166. Hinneburg, A., & Gabriel, H. H. (2007, September). Denclue 2.0: Fast clustering based on kernel density estimation. In *International symposium on intelligent data analysis* (pp. 70-80). Berlin, Heidelberg: Springer Berlin Heidelberg.

167. Hinneburg, A., & Keim, D. A. (2003). A general approach to clustering in large databases with noise. *Knowledge and information systems*, *5*, 387-415.

168. Fukunaga, K., & Hostetler, L. (1975). The estimation of the gradient of a density function, with applications in pattern recognition. *IEEE Transactions on information theory*, *21*(1), 32-40.

169. Bodyanskiy, Y. V., Pliss, I. P., Shafronenko, A. Y., & Kalynychenko, O. V. (2022). Нечітка довірча кластеризація даних на основі аналізу щільності розподілу даних та їх піків. *Radio Electronics, Computer Science, Control*, (3), 58-58.

170. Gruber, H. E. (1988). The evolving systems approach to creative work. *Creativity Research Journal*, *1*(1), 27-51.

171. Allen, P. M. (1988). Dynamic models of evolving systems. *System Dynamics Review*, *4* (1-2), 109-130.

172. Leite, D., Škrjanc, I., & Gomide, F. (2020). An overview on evolving systems and learning from stream data. *Evolving systems*, *11*(2), 181-198.

173. Kennedy, J., & Eberhart, R. (1995, November). Particle swarm optimization. In *Proceedings of ICNN'95-international conference on neural networks* (Vol. 4, pp. 1942-1948). ieee.

174. Eiben, A. E., & Smith, J. E. (2015). *Introduction to evolutionary computing*. Springer-Verlag Berlin Heidelberg.

175. Liu, J., & Lampinen, J. (2005). A fuzzy adaptive differential evolution algorithm. *Soft Computing*, *9*, 448-462.

176. Vikhar, P. A. (2016, December). Evolutionary algorithms: A critical review and its future prospects. In *2016 International conference on global trends in signal processing, information computing and communication (ICGTSPICC)* (pp. 261-265). IEEE.

177. Eiben, A. E., Smith, J. E., Eiben, A. E., & Smith, J. E. (2015). What is an evolutionary algorithm?. *Introduction to evolutionary computing*, 25-48.

178. Mühlenbein, H., Gorges-Schleuter, M., & Krämer, O. (1988). Evolution algorithms in combinatorial optimization. *Parallel computing*, *7*(1), 65-85.

179. Yu, X., & Gen, M. (2010). *Introduction to evolutionary algorithms*. Springer Science & Business Media.

180. Slowik, A., & Kwasnicka, H. (2020). Evolutionary algorithms and their applications to engineering problems. *Neural Computing and Applications*, *32*, 12363-12379.

181. Dasgupta, D., & Michalewicz, Z. (Eds.). (2013). *Evolutionary algorithms in engineering applications*. Springer Science & Business Media.

182. Winter, G., Périaux, J., Galán, M., & Cuesta, P. (1996). *Genetic algorithms in engineering and computer science*. John Wiley & Sons, Inc..

183. Yao, X. (1999). *Evolutionary computation: Theory and applications*. World scientific.

184. Zalzala, A. M., & Fleming, P. J. (Eds.). (1997). *Genetic algorithms in engineering systems* (Vol. 55). Iet.

185. Grosan, C., & Abraham, A. (2007). Hybrid evolutionary algorithms: methodologies, architectures, and reviews. In *Hybrid evolutionary algorithms* (pp. 1-17). Berlin, Heidelberg: Springer Berlin Heidelberg.

186. Karpenko, A. P. (2012). Population algorithms for global continious optimization. *Review of new and little-known algorithms. Supplement to the journal" Information Technologies*, (7), 32.

187. Chu, S. C., Tsai, P. W., & Pan, J. S. (2006). Cat swarm optimization. In *PRICAI 2006: Trends in Artificial Intelligence: 9th Pacific Rim International Conference on Artificial Intelligence Guilin, China, August 7-11, 2006 Proceedings 9* (pp. 854-858). Springer Berlin Heidelberg.

188. Chu, S. C., & Tsai, P. W. (2007). Computational intelligence based on the behavior of cats. *International Journal of Innovative Computing, Information and Control*, *3*(1), 163-173.

189. Кравець, П. О. (2005). Ігрові методи випадкового пошуку в умовах невизначеності. *Вісник Національного університету "Львівська політехніка". Інформаційні системи та мережі*, (549), 105-117.

190. Гожий, О. П. (2016). Інформаційні технології динамічного планування та прийняття рішень на основі ймовірнісно-статистичних методів. *Г ожий ОП-Львів: Нац. ун-т «Львівська політехніка*.

191. Chu, S. C., & Tsai, P. W. (2007). Computational intelligence based on the behavior of cats. *International Journal of Innovative Computing, Information and Control*, *3*(1), 163-173.

192. Bahrami, M., Bozorg-Haddad, O., & Chu, X. (2018). Cat swarm optimization (CSO) algorithm. *Advanced optimization by nature-inspired algorithms*, 9-18.

193. Yang, J., & Zhuang, Y. (2010). An improved ant colony optimization algorithm for solving a complex combinatorial optimization problem. *Applied soft computing*, *10*(2), 653-660.

194. Tsai, P. W., Pan, J. S., Chen, S. M., Liao, B. Y., & Hao, S. P. (2008, July). Parallel cat swarm optimization. In *2008 international conference on machine learning and cybernetics* (Vol. 6, pp. 3328-3333). IEEE.

195. Toksari, M. D. (2006). Ant colony optimization for finding the global minimum. *Applied Mathematics and computation*, *176*(1), 308-316.

196. Sarangi, A., Sarangi, S. K., Mukherjee, M., & Panigrahi, S. P. (2015, December). System identification by Crazy-cat swarm optimization. In *2015 International Conference on Microwave, Optical and Communication Engineering (ICMOCE)* (pp. 439-442). IEEE.

197. So, J., & Jenkins, W. K. (2013, November). Comparison of cat swarm optimization with particle swarm optimization for IIR system identification. In *2013 Asilomar Conference on Signals, Systems and Computers* (pp. 903-910). IEEE.

198. Upadhyay, P., Kar, R., Mandal, D., & Ghoshal, S. P. (2014). Craziness based particle swarm optimization algorithm for IIR system identification problem. *AEU-International Journal of Electronics and Communications*, *68*(5), 369-378.

199. Eswari, P., Ramalakshmanna, Y., & Durga Prasad, C. (2021). An Improved Particle Swarm Optimization-Based System Identification. In *Machine Learning, Deep Learning and Computational Intelligence for Wireless Communication: Proceedings of MDCWC 2020* (pp. 137-142). Springer Singapore.

200. Panda, G., Pradhan, P. M., & Majhi, B. (2011). IIR system identification using cat swarm optimization. *Expert Systems with Applications*, *38*(10), 12671-12683.

201. Бодянський, Є. В., Шафроненко, А. Ю., & Патлань, Е. 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

202. Bodyanskiy, Y. V., Shafronenko, A. Y., & Klymova, I. N. (2021). Онлайн метод можливісної кластеризації даних на основі еволюційної

оптимізації котячих зграй. *Radio Electronics, Computer Science, Control*, (2), 65-70.

203. Шафроненко, А. Ю., Свистунов, І. О., & Танянський, О. С. (2021, November). Адаптивна нечітка кластеризація даних на основі еволюційних процедур. In *The 5 th International scientific and practical conference - Topical issues of modern science, society and education (November 28-30, 2021) SPC - Sci-conf. com. ua, Kharkiv, Ukraine. 2021. 2101 p.* (p. 644).

204. 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).

205. 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.

206. Shafronenko, A., Bodyanskiy, Y., & Rudenko, D. (2020). *Neuro-fuzzy clustering of Distorted Data Using Cat Swarm Optimization*. LAP LAMBERT Academic Publishing.

207. Bodyanskiy, Y., Shafronenko, A., & Pliss, I. (2021). Правдоподібна нечітка кластеризація даних на основі еволюційного методу божевільних котів. *System research and information technologies*, (3), 110-119.

208. Bastos Filho, C. J., de Lima Neto, F. B., Lins, A. J. D. C. C., Nascimento, A. I., & Lima, M. P. (2009). Fish School Search. *Nature-inspired algorithms for optimisation*, *193*, 261-277.

209. Cavalcanti-Júnior, G. M., Bastos-Filho, C. J., Lima-Neto, F. B., & Castro, R. M. (2011). A hybrid algorithm based on fish school search and particle swarm optimization for dynamic problems. In *Advances in Swarm Intelligence:*

Second International Conference, ICSI 2011, Chongqing, China, June 12-15, 2011, Proceedings, Part II 2 (pp. 543-552). Springer Berlin Heidelberg.

210. Janecek, A., & Tan, Y. (2011). Feeding the fish–weight update strategies for the fish school search algorithm. In *Advances in Swarm Intelligence: Second International Conference, ICSI 2011, Chongqing, China, June 12-15, 2011, Proceedings, Part II 2* (pp. 553-562). Springer Berlin Heidelberg.

211. Box, G. E. (1957). Evolutionary operation: A method for increasing industrial productivity. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *6*(2), 81-101.

212. Bodyanskiy, Y., Shafronenko, A., & Pliss, I. (2022). Clusterization of vector and matrix data arrays using the combined evolutionary method of fish schools. *Системні дослідження та інформаційні технології: міжнародний науково-технічний журнал, № 4*.

213. Umanets, V., Voinyk, B., Pavlov, V., & Nastenko, I. (2018). Estimation of algorithms efficiency in the task of biological objects clustering.

214. Spendley, W. G. R. F. R., Hext, G. R., & Himsworth, F. R. (1962). Sequential application of simplex designs in optimisation and evolutionary operation. *Technometrics*, *4*(4), 441-461.

215. Ja, N. (1965). A simplex method for function minimization. *Computer journal*, *7*, 308-313.

216. Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The computer journal*, *7*(4), 308-313.

217. Бодянский, Е. В., & Шафроненко, А. Ю. (2018). Рандомізована модифікація методу оптимізації на основі котячих зграй. *Системи обробки інформації*, (1), 142-147.

218. 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.

219. Abidi, B., & Ben Yahia, S. (2014). A new algorithm for fuzzy clustering handling incomplete dataset. *International Journal on Artificial Intelligence Tools*, *23*(04), 1460012.

220. Melin, P., Castillo, O., Melin, P., & Castillo, O. (2005). Clustering with Intelligent Techniques. *Hybrid Intelligent Systems for Pattern Recognition Using Soft Computing: An Evolutionary Approach for Neural Networks and Fuzzy Systems*, 169-184.

221. 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).

222. Kureichik, V. V., Kursitys, I. O., Kuliev, E. V., & Gerasimenko, E. M. (2020, December). Application of bioinspired algorithms for solving transcomputational tasks. In *Journal of Physics: Conference Series* (Vol. 1703, No. 1, p. 012021). IOP Publishing.

223. Shafronenko, A., & Bodyanskiy, Y. V. (2020). Adaptive fuzzy clustering approach based on evolutionary cat swarm optimization. In *CMIS* (pp. 832-842).

224. Bodyanskiy, Y. V., Pliss, I. P., & Shafronenko, A. Y. (2022). Кластеризація масивів даних на основі комбінованої оптимізації функцій щільності розподілу та еволюційного методу котячих зграй. *Radio Electronics, Computer Science, Control*, (4), 61-61.

225. Mirjalili, S., Mirjalili, S. M., & Lewis, A. (2014). Grey wolf optimizer. *Advances in engineering software*, *69*, 46-61.

226. Khoshkbarchi, A., Kamali, A., Amjadi, M., & Haeri, M. A. (2016, September). A modified hybrid fuzzy clustering method for big data. In *2016 8th International Symposium on Telecommunications (IST)* (pp. 196-201). IEEE.

227. 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.

228. Шишацький, А. В., Налапко, О. Л., & Одарущенко, О. Б. (2021). Основні біоінспіровані алгоритми обробки різнотипних даних. Інтеграція інформаційних систем і інтелектуальних технологій в умовах трансформації інформаційного суспільства: тези доповідей IV Міжнародної науково-практичної конференції, що присвячена 50-ій річниці кафедри інформаційних систем та технологій. Полтава: ПДАУ, 2021. 109-114. In *Integration of information systems and intelligent technologies in the conditions of information society transformation. Abstracts of the IVth International scientific-practical conference dedicated to the 50th anniversary of the Department of Information Systems and Technologies. Poltava, Ukraine. 2021. 144 p.* (p. 109).

229. Шафроненко А. Ю., Бодянський Є. В. (2022). Адаптивна кластеризація багатоекстремальних масивів даних з використанням модифікованого алгоритму риб'ячої зграї. *АСУ і прилади*. №178. 33-37.

230. 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.

231. Trzciński, M., Kowalski, P. A., & Łukasik, S. (2022). Clustering with Nature-Inspired Algorithm Based on Territorial Behavior of Predatory Animals. *Algorithms*, *15*(2), 43.

232. 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

233. Jie, L., Liu, W., Sun, Z., & Teng, S. (2017). Hybrid fuzzy clustering methods based on improved self-adaptive cellular genetic algorithm and optimal-selection-based fuzzy c-means. *Neurocomputing*, *249*, 140-156.

234. Liu, S., Yu, Q., Lin, Q., & Tan, K. C. (2020). An adaptive clustering-based evolutionary algorithm for many-objective optimization problems. *Information Sciences*, *537*, 261-283.

235. Ding, Y., & Fu, X. (2016). Kernel-based fuzzy c-means clustering algorithm based on genetic algorithm. *Neurocomputing*, *188*, 233-238.

236. Mukhopadhyay, A., Maulik, U., & Bandyopadhyay, S. (2009). Multiobjective genetic algorithm-based fuzzy clustering of categorical attributes. *IEEE transactions on evolutionary computation*, *13*(5), 991-1005.

237. Horta, D., Naldi, M., Campello, R. J. G. B., Hruschka, E. R., & de Carvalho, A. C. P. L. F. (2009). Evolutionary fuzzy clustering: an overview and efficiency issues. *Foundations of Computational Intelligence Volume 4: Bio-Inspired Data Mining*, 167-195.

238. Sáez, D., Cortés, C. E., & Núñez, A. (2008). Hybrid adaptive predictive control for the multi-vehicle dynamic pick-up and delivery problem based on genetic algorithms and fuzzy clustering. *Computers & Operations Research*, *35*(11), 3412-3438.

239. Wikaisuksakul, S. (2014). A multi-objective genetic algorithm with fuzzy c-means for automatic data clustering. *Applied Soft Computing*, *24*, 679-691.

240. Li, C., Zhou, J., Kou, P., & Xiao, J. (2012). A novel chaotic particle swarm optimization based fuzzy clustering algorithm. *Neurocomputing*, *83*, 98-109.

241. Saha, I., Maulik, U., & Bandyopadhyay, S. (2009, March). A new differential evolution based fuzzy clustering for automatic cluster evolution. In *2009 IEEE International Advance Computing Conference* (pp. 706-711). IEEE.

242. Alata, M., Molhim, M., & Ramini, A. (2008). Optimizing of fuzzy c-means clustering algorithm using GA. *International Journal of Computer and Information Engineering*, *2*(3), 670-675.

243. 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.

244. Shafronenko, A., Bodyanskiy, Y. V., & Pliss, I. (2023). Credibilistic Fuzzy Clustering Method Based on Evolutionary Approach of Crazy Wolfs in Online Mode. In *CMIS* (pp. 141-150)

245. Srinivasan, S., Gunasekaran, S., Mathivanan, S. K., M. B, B. A. M., Jayagopal, P., & Dalu, G. T. (2023). An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database. Scientific Reports, 13(1), 13588.

246. Chicho, B. T., Abdulazeez, A. M., Zeebaree, D. Q., & Zebari, D. A. (2021). Machine learning classifiers based classification for IRIS recognition. Qubahan Academic Journal, 1(2), 106-118.

247. Devasena, C. L., Sumathi, T., Gomathi, V. V., & Hemalatha, M. (2011). Effectiveness evaluation of rule based classifiers for the classification of iris data set. *Bonfring International Journal of Man Machine Interface*, *1*, 5.

248. Ali, E. H., Jaber, H. A., & Kadhim, N. N. (2023). New algorithm for localization of iris recognition using deep learning neural networks. *Indonesian Journal of Electrical Engineering and Computer Science*, *29*(1), 110-119.

249. Choudhary, D., Tiwari, S., & Singh, A. K. (2012). A survey: Feature extraction methods for iris recognition. *International Journal of Electronics Communication and Computer Technology*, *2*(6), 275-279.

250. Agrawal, R. (2019). Predictive analysis of breast cancer using machine learning techniques. Ingeniería Solidaria, 15(3), 1-23.

251. Srivenkatesh, M. (2020). Prediction of breast cancer disease using machine learning algorithms. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, *9*(4), 2868-2878.

252. Jain, S., & Kumar, P. (2020). Prediction of breast cancer using machine learning. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, *13*(5), 901-908.

253. Rana, M., Chandorkar, P., Dsouza, A., & Kazi, N. (2015). Breast cancer diagnosis and recurrence prediction using machine learning

techniques. *International journal of research in Engineering and Technology*, *4*(4), 372-376.

254. Shahriar, R. N., Muhammad, A., Shakib, M. A., & Habib, M. A. A Machine Learning Application to Extricate the Red Wine Quality.

255. Negi, A., Sharma, P., & Rawat, H. S. (2021). Wine quality prediction using machine learning.

256. Darade, S., & Korade, N. Wine quality prediction. *Volume*, *3*, 1246-1252.

257. Singhal, P., Singh, P., Hazela, B., Singh, V., & Singh, V. (2021). Machine learning algorithm: wine quality prediction. *SPAST Abstracts*, *1*(01).

258. Schaffer, C. (1993). Selecting a classification method by cross-validation. Machine learning, 13, 135-143.

259. Wainer, J., & Cawley, G. (2021). Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications*, *182*, 115222.

260. Rao, R. B., Fung, G., & Rosales, R. (2008, April). On the dangers of cross-validation. An experimental evaluation. In *Proceedings of the 2008 SIAM international conference on data mining* (pp. 588-596). Society for Industrial and Applied Mathematics.

261. Bassiouni, M., Ali, M., & El-Dahshan, E. A. (2018). Ham and spam e-mails classification using machine learning techniques. Journal of Applied Security Research, 13(3), 315-331.

262. Ying, K. C., Lin, S. W., Lee, Z. J., & Lin, Y. T. (2010). An ensemble approach applied to classify spam e-mails. *Expert Systems with Applications*, *37*(3), 2197-2201.

263. Ghosh, A., & Senthilrajan, A. (2023). Comparison of machine learning techniques for spam detection. *Multimedia Tools and Applications*, *82*(19), 29227-29254.

264. Sharma, S., & Arora, A. (2013). Adaptive approach for spam detection. *International Journal of Computer Science Issues (IJCSI)*, *10*(4), 23.

265. Sharaff, A., Nagwani, N. K., & Dhadse, A. (2016). Comparative study of classification algorithms for spam email detection. In *Emerging Research in Computing, Information, Communication and Applications: ERCICA 2015, Volume 2* (pp. 237-244). Springer India.

266. Avuçlu, E. (2023). Automatically Finding the Biggest Fold Value for More Accurate Classification and Diagnosis in Machine Learning Algorithms. Iranian Journal of Science and Technology, Transactions of Electrical Engineering, 1-26.

267. Olawade, O. E., Onashoga, S. A., & Arogundade, O. T. (2020, March). Comparative analysis of machine learning techniques in health system. In *2020 international conference in mathematics, computer engineering and computer science (ICMCECS)* (pp. 1-6). IEEE.

268. Avuçlu, E. (2023). Determining the most accurate machine learning algorithms for medical diagnosis using the monk'problems database and statistical measurements. *Journal of Experimental & Theoretical Artificial Intelligence*, 1-20.

269. Pavithra, D., & Jayanthi, A. N. (2018). A study on machine learning algorithm in medical diagnosis. *International Journal of Advanced Research in Computer Science*, *9*(4).

270. Raval, D., Bhatt, D., Kumhar, M. K., Parikh, V., & Vyas, D. (2016). Medical diagnosis system using machine learning. *International Journal of Computer Science & Communication*, *7*(1), 177-182.

271. Latif, J., Xiao, C., Imran, A., & Tu, S. (2019, January). Medical imaging using machine learning and deep learning algorithms: a review. In *2019 2nd International conference on computing, mathematics and engineering technologies (iCoMET)* (pp. 1-5). IEEE.

272. Nyangaresi, V. O., El-Omari, N. K. T., & Nyakina, J. N. (2022). Efficient feature selection and ML algorithm for accurate diagnostics. *Journal of Computer Science Research*, *4*(1), 10-19.

273. Sheikh, M. H., Mittal, S., & Bashir, R. (2022). An Analysis of Various Machine Learning Techniques Used for Diseases Prediction: A Review. *Recent Innovations in Computing: Proceedings of ICRIC 2021, Volume 2*, 467-476.

274. Жемела, Г. П., & Кузнецова, О. А. (2012). Вплив сортових властивостей на продуктивність та якість зерна пшениці м'якої озимої. *Scientific Progress & Innovations*, (3), 23-25.

275. Господаренко, Г. М., Любич, В. В., & Калантир, В. В. (2021, March). Удобрення пшениці твердої озимої. In *XI International Scientific and Practical Conference «Topical issues of modern science and education». Tallinn, Estonia* (pp. 12-15).

276. Боцян, М. Ю. (2019). Моделювання розвитку озимої пшениці на основі логістичної функції. *Інформаційні технології та моделювання систем: матеріали всеукраїнської студентської науково-практичної конференції, м. Житомир, 25 квітня 2019 р. Житомир: ЖНАЕУ, 2019.-108 с.*

277. Тимощук, Т. М., Котельницька, Г. М., Дереча, І. М., & Овсійчук, Є. М. (2022). Оцінювання сортів пшениці озимої за продуктивністю. *Ефективність агротехнологій в зоні полісся україни*, 40.

278. Морозов, О. В., Безніцька, Н. В., Нестеренко, В. П., & Пічура, В. І. (2014). Формування урожайності озимої пшениці залежно від кліматичних змін (на прикладі Херсонської області). *Таврійський науковий вісник*, 146-152.

279. Черенков, А. В., Нестерець, В. Г., Солодушко, М. М., & Романенко, О. Л. (2010). Вирощування озимої пшениці в зв'язку з регіональними змінами погодних умов в Степу України. *Бюлетень Інституту зернового господарства*, (38), 9-16.

280. Божко, Л. Ю., & Крисак, О. В. (2018). Оцінка агрокліматичних ресурсів перезимівлі озимої пшениці в Степовій зоні України: колективна монографія.

281. Абєлєшов, В. І. Технічна експлуатація житлових будівель, готелів і туристичних комплексів: навч. посібник.

282. Стяжкіна, О. О. (2021). Посилення й відновлення експлуатаційної придатності конструкцій при реконструкції адміністративної будівлі.

283. Байрактар, А. О. (2022). Шляхи підвищення експлуатаційних якостей та надійності цивільної багатоповерхової будівлі з урахуванням інструментального контролю.

284. Русінко, М. І. (2014). Класифікація факторів впливу на інноваційний розвиток будівельного підприємства. *Науковий вісник Херсонського державного університету. Серія: Економічні науки*, (9-1), 113-117.

285. Оліховський, В. Я. (2014). Технологічні карти та можливості їх використання у податковому плануванні. *Вісник Національного університету Львівська політехніка. Менеджмент та підприємництво в Україні: етапи становлення і проблеми розвитку*, (794), 295-304.

286. Черенько, Л. М., Полякова, С. В., Шишкін, В. С., Заяць, В. С., Когатько, Ю. Л., Васильєв, О. А., ... & Новосільська, Т. В. (2020). Житлові умови населення: чинники, сучасний стан і політика регулювання. *Київ: Ін-т демогр. та соц. дослідж. ім. МВ Птухи*.

287. Гуцан, Т. Г. (2013). Умови проживання як складова рівня життя населення та шляхи їх покращення в України. *Збірник наукових праць Харківського національного педагогічного університету імені ГС Сковороди. Економіка*, (13), 27-38.

288. Шишкін, В. С. (2019). Демографічні фактори житлових умов населення. *Демографія та соціальна економіка*, (1), 152-165.

289. Бендерська, О. В., Шутюк, В. В., & Бессараб, О. С. Нітрати та якість питної води. *Міжнародна науково-практична конференція*, 153.Петренко, Н. Ф., Мокієнко, А. В., & Платов, С. М. (2018). Загальна гігієнічна оцінка якості питної води та стану питного водопостачання в Україні. *Актуальні проблеми транспортної медицини: навколишнє середовище; професійне здоров'я; патологія*, (4), 7-16.

290. Зоріна, О. В. (2018). Результати гігієнічної оцінки якості водопровідної питної води України та новий порядок інформування споживачів. *Актуальні проблеми транспортної медицини: навколишнє середовище; професійне здоров'я; патологія*, (1), 38-47.

291. Гущук, І. В., Брезецька, О. І., & Гущук, В. І. (2014). Еколого-гігієнічна оцінка якості питної води з джерел та мережі централізованих водопроводів Рівненської області. *Гігієна населених місць*, (64), 76-80.

292. Прокопов, В. О. (2014). Стан та якість питної води централізованих систем водопостачання України в сучасних умовах (погляд на проблему з позицій гігієни). *Гігієна населених місць*, (64), 56-67.

293. Туровська, Г. І. (2019). Науково-методичні аспекти аналізу безпеки питної води. *Екологічні науки*, (4), 120-123.

294. Прокопов, В. О., & Зоріна, О. В. (2019). Результати гігієнічного моніторингу питної води поліпшеної якості в Україні. *Гігієна населених місць: зб. наук. пр. К*, 72-78.

295. Олійник, О. О. (2019). *Дослідження якості питної води м. Дніпро та обґрунтування шляхів її покращення* (Doctoral dissertation, Національний технічний університет «Дніпровська політехніка»).

296. Походило, Є. В., & Мартинович, Н. В. (2010). Контроль твердості питної води за електричними параметрами. *Вісник Національного технічного університету «ХПІ». Серія: Нові рішення у сучасних технологіях*, (46), 122-125.

297. Ліжевський, В. (2018). Очищення стічної води після гальванічного нікелювання. In *Наукові розробки молоді на сучасному етапі*. Київський національний університет технологій та дизайну.

298. Гулевський, В. Б., Гулевский, В. Б., Постол, Ю. О., Постол, Ю. А., Журавель, Д. П., Журавель, Д. П., ... & Ковалев, А. В. (2019). Електрохімічні технології очищення стічних вод.

299. Бобрик, С. В. (2020). ГІДРОХІМІЯ СТІЧНИХ ВОД І ЗДОРОВ'Я НАРОДОНАСЕЛЕННЯ. In *Актуальні проблеми охорони рослинного світу та відновлення біорозмаїття* (pp. 38-39).

300. Boyko, N. I., & Kurylo, V. (2023). Алгоритм класифікації медичних даних для прогнозування онкології. *Systems and Technologies*, *66*(2), 21-31.

301. Сипливий, В. О., Гузь, А. Г., Євтушенко, Д. В., Доценко, В. В., Петренко, Г. Д., Петюнін, О. Г., ... & Євтушенко, О. В. (2020). Пухлини. Етіологія, патогенез. Доброякісні і злоякісні пухлини. Гістогенетична, морфологічна, клінічна і міжнародна (TNM) класифікації. Клінічні групи онкологічних хворих. Клінічні прояви. Методи діагностики. Принципи лікування: методичні вказівки до практичних занять та самостійної роботи студентів.

302. Готько, Є. С., & Сочка, А. В. (2007). Рак грудної залози у чоловіків: вплив категорії Т на прогноз захворювання.

303. Винниченко, І. О., Москаленко, Ю. В., & Винниченко, О. І. (2017). Збірник тестових завдань з онкології (Класифікація TNM, сьоме видання, 2009).

304. Вирва, О. Є. (2023). Хондросаркома у XXI сторіччі.

305. ПЛЕСКАЧ, Б. В. (2018). *Особливості внутрішньоособистісного конфлікту в онкогематологічних хворих* (Doctoral dissertation, дис.… канд. психол. наук).

306. Хоперія, В. Г., Харченко, О. І., Дудла, Д. І., Цема, Є. В., Сафонов, В. Є., Гриценко, О. М., & Малиновська, О. В. (2017). Випадкова діагностика раку прищитоподібної залози у військовослужбовців—учасників бойових дій на сході України. *Хірургія України*, (4), 104-107.

307. Гайсенко, А. В., Михайлович, Ю. Й., Журбенко, А. В., & Трет'якова, Т. М. (2012). Медико-соціальне обґрунтування доцільності скринінгу найбільш поширених злоякісних новоутворень в популяції України як практичний аспект удосконалення якості профілактики онкологічних захворювань. *Клінична онкологія*, (1), 6-10.

308. Костюченко, Л. В. (2014). Рання діагностика тяжких комбінованих імунодефіцитів. *Буковинський медичний вісник*, (18, № 4), 63-69.

309. Мухаровська, І. Р. (2015). Медико-психологічний паспорт захворювання онкологічного профілю. *Медична психологія*, (10,№ 4), 15-19.

310. Романів, М. П. (2017). Медико-статистична оцінка статево-вікової структури захворюваності та смертності від онкологічних захворювань в Україні. *Вісник наукових досліджень*, (1), 85-90.

311. Северин, Ю. М., Стриженок, В. П., Устенко, Р. Л., Северин, Ю. Н., Стриженок, В. П., & Устенко, Р. Л. (2018). *Етіологія онкологічних захворювань* (Doctoral dissertation, Українська медична стоматологічна академія).

ДОДАТОК А

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, 60p.

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

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).

ДОДАТОК Б

Акти про реалізацію і впровадження результатів дисертаційної роботи

ЗАТВЕРДЖУЮ

Генеральний директор
ТОВ НАУКОВО-ВИРОБНИЧА
ФІРМА «ХЕЛП-АГРО»
_____Я.І. Вакуленко
«17» липня 2023 р.

АКТ
про впровадження дисертаційної роботи на здобуття
наукового ступеня доктора технічних наук
Шафроненко Аліни Юріївни

1. **Найменування пропозиції**: «Прогнозування врожайності озимої пшениці з урахуванням впливу основних гідрометеорологічних факторів».
2. **Ким запропоновано:** доцентом, кандидатом технічних наук, доцентом кафедри інформатики Харківського національного університету радіоелектроніки Шафроненко Аліною Юріївною.
3. **Джерело інформації** (методичні рекомендації, інформаційний лист, звіт про НДР, дисертація, монографія, з'їзди, конференції, семінари та ін.): Bodyanskiy, Y. V., Pliss, I. P., Shafronenko, A. Y., Kalynychenko, O. V. (2022). CREDIBILISTIC FUZZY CLUSTERING BASED ON ANALYSIS OF DATA DISTRIBUTION DENSITY AND THEIR PEAKS. Radio Electronics, Computer Science, Control, (3), 58. https://doi.org/10.15588/1607-3274-2022-3-6.
4. **Де і коли впроваджено:** ТОВ НАУКОВО-ВИРОБНИЧА ФІРМА «ХЕЛП-АГРО».
5. **Форма впровадження:** аналіз врожайності за умов прогнозних даних гідрометеослужб за попередні роки та прогнозування значення врожайності на майбутній рік.
6. **Ефективність впровадження за критеріями, висловленими в джерелі інформації**: запропонований підхід дає можливість приймати ефективні управлінські рішення щодо підвищення врожайності сільськогосподарських культур в умовах невизначеності зовнішнього середовища.
7. **Зауваження, пропозиції:** немає.

Акт складений для пред'явлення до спеціалізованої вченої ради із захисту дисертацій і не є підставою для фінансових розрахунків.

Члени комісії _____ (Харченко А.В.)

_____ (Калинич О.Ю.)

Керівник ТОВ
«Будівельно-монтажне підприємство - 168»
_____О.В. КРИВША
«21» 12 2023р.

АКТ ПРО ВПРОВАДЖЕННЯ

результатів дисертаційної роботи
на здобуття наукового ступеня доктора технічних наук
ШАФРОНЕНКО АЛІНИ ЮРІЇВНИ

Комісія у складі:

голова:
керівник ТОВ «Будівельно-монтажне підприємство - 168» Кривша О.В.
члени комісії: Кривша О.О., Сенько В.В.

склала даний акт про те, що на ТОВ «Будівельно-монтажне підприємство - 168» при виконанні будівельних та монтажних робіт загального призначення були використані методи адаптивної нечіткої кластеризації даних для класифікації технологічних процесів на будівництві з метою підвищення їх ефективності. Для класифікації були використані дані розрахунків для об'єктів інженерної інфраструктури: монтаж внутрішніх інженерних мереж, систем, приладів і засобів вимірювання, іншого обладнання.

Результати впровадження дають можливість класифікувати технологічні процеси за класом наслідків (відповідальності), що належать до об'єктів із середніми та значними наслідками (СС2, СС3) та підвищує ефективність господарської діяльності з будівництва об'єктів.

Акт складений для пред'явлення до спеціалізованої вченої ради із захисту дисертацій і не є підставою для фінансових розрахунків.

Члени комісії _____ (О.О. Кривша)
_____ (В.В. Сенько)

ЗАТВЕРДЖУЮ

Генеральний директор
КОМУНАЛЬНОГО НЕКОМЕРЦІЙНОГО ПІДПРИЄМСТВА «ОБЛАСНИЙ ЦЕНТР ОНКОЛОГІЇ»

_____ В.М. Лихман
«__» _____ 2023 р.

АКТ ПРО ВПРОВАДЖЕННЯ

1. **Найменування пропозиції** (метод профілактики, діагностики, лікування, пристрій, форма організаційної роботи та ін.): «Нечітка класифікація пацієнтів з ознаками онкологічних захворювань та підготовка інформації до формулювання діагнозу».
2. **Ким і коли запропонований:** доцентом, кандидатом технічних наук, доцентом кафедри інформатики Харківського національного університету радіоелектроніки Шафроненко Аліною Юріївною.
3. **Джерело інформації** (методичні рекомендації, інформаційний лист, звіт про НДР, дисертація, монографія, з'їзди, конференції, семінари та ін.): Shafronenko, A., Bodyanskiy, Ye., Rudenko, D.: Neuro-fuzzy clustering of Distorted Data Using Cat Swarm Optimization. Saarbrücken, LAP LAMBERT Academic Publishing (2020); Бодянський, Є., Плісс, І., Шафроненко, А.: Нечіткі методи інтелектуального аналізу. GlobeEdit (2022).
4. **Де і коли впроваджено:** КОМУНАЛЬНЕ НЕКОМЕРЦІЙНЕ ПІДПРИЄМСТВО «ОБЛАСНИЙ ЦЕНТР ОНКОЛОГІЇ».
5. **Форма впровадження:** попередня обробка діагностичних ознак та відновлення викривлених та пропущених даних.
6. **Ефективність впровадження за критеріями, висловленими в джерелі інформації**: підвищення точності та об'єктивності процесу медичного діагностування онкологічних захворювань на ранніх стадіях.
7. **Зауваження, пропозиції:** немає.

Акт складений для пред'явлення до спеціалізованої вченої ради із захисту дисертацій і не є підставою для фінансових розрахунків.

Відповідальний(і) за впровадження

Завідувач організаційно
– методичним відділом ОЦО Ф.Л. Уразов

_____ _____
(дата) (підпис)

УКРАЇНА

ТОВАРИСТВО З ОБМЕЖЕНОЮ ВІДПОВІДАЛЬНІСТЮ
«КОМУНСЕРВІС 2018»

Адреса юридична: смт Безлюдівка, вул. Кооперативна 30, кв.2, Хар.р-н.,Хар.обл., 62489, тел. 057-749-66-56

від «12» 04 2023р.

АКТ

про впровадження результатів дисертаційної роботи
на здобуття наукового ступеня доктора технічних наук
ШАФРОНЕНКО Аліни Юріївни

Комісія у складі:
Голова: директор ТОВ «Комунсервіс 2018» Т.О. Мазнєва
Члени комісії: бухгалтер Ю.М. Усіченко,
інспектор Є.П. Литвин

Склала цей акт про те, що на ТОВ «Комунсервіс 2018» при оцінці стану будинків для визначення готовності до експлуатації в зимових умовах, були застосовані методи адаптивної нечіткої кластеризації даних різної природи, розроблені Шафроненко А.Ю., для задач аналізу пошкоджень та їх усунення.

Для аналізу були взяті відомості стану будинків по вул. Мостобудівників такі як стан покрівлі, горища, сходів, підвалу, інженерного обладнання, прибирального і протипожежного інвентаря, що були сформовані у таблицю «об'єкт - властивість», яка обробляється послідовно.

Результати впровадження довели доцільність використання вищезазначених методів для оцінки стану житлових будинків, що дозволяють прискорити аналіз та прийняття обґрунтованих рішень щодо першочерговості відновлення будинків, в залежності від категорії пошкоджень, зношеності та наявних ресурсів.

Акт складений для пред'явлення до спеціалізованої вченої ради із захисту дисертацій і не є підставою для фінансових розрахунків.

ТОВАРИСТВО З ОБМЕЖЕНОЮ ВІДПОВІДАЛЬНІСТЮ «КОМУНСЕРВІС 2018» №42425543 Харківська область Харківський район

Голова комісії _____ Т.О. Мазнєва

Члени комісії _____ Ю.М. Усіченко

_____ Є. П. Литвин

ЗАТВЕРДЖУЮ

Генеральний директор
КОМУНАЛЬНОГО НЕКОМЕРЦІЙНОГО ПІДПРИЄМСТВА «ОБЛАСНИЙ ЦЕНТР ОНКОЛОГІЇ»

_____ **В.М. Лихман**
«22» квітня 2024 **р.**

АКТ ПРО ВПРОВАДЖЕННЯ

1. **Найменування пропозиції** (метод профілактики, діагностики, лікування, пристрій, форма організаційної роботи та ін.): «Метод медичного діагностування в режимі еволюційного самонавчання, що дозволило проводити ранню медичну діагностику онкології у пацієнтів».
2. **Ким і коли запропонований:** доцентом, кандидатом технічних наук, доцентом кафедри інформатики Харківського національного університету радіоелектроніки Шафроненко Аліною Юріївною.
3. **Джерело інформації** (методичні рекомендації, інформаційний лист, звіт про НДР, дисертація, монографія, з'їзди, конференції, семінари та ін.): Shafronenko, A., Bodyanskiy, Y. V., & Pliss, I. (2023). Credibilistic Fuzzy Clustering Method Based on Evolutionary Approach of Crazy Wolfs in Online Mode. In *CMIS* (pp. 141-150)
4. **Де і коли впроваджено:** КОМУНАЛЬНЕ НЕКОМЕРЦІЙНЕ ПІДПРИЄМСТВО «ОБЛАСНИЙ ЦЕНТР ОНКОЛОГІЇ».
5. **Форма впровадження:** попередня обробка діагностичних ознак та відновлення викривлених та пропущених даних.
6. **Ефективність впровадження за критеріями, висловленими в джерелі інформації**: підвищення надійності та об'єктивності процесу медичного діагностування пацієнтів з умовно невідомим діагнозом.
7. **Зауваження, пропозиції:** немає.

Акт складений для пред'явлення до спеціалізованої вченої ради із захисту дисертацій і не є підставою для фінансових розрахунків.

Відповідальний(і) за впровадження

Завідувач організаційно
– методичним відділом ОЦО Ф.Л. Уразов

22.04.2024р.
(дата) (підпис)

Директор КП
«Санітарно-екологічний центр»
Харківської міської ради

_____ Ігор КОТЕНКО
«29» 06 2023р.

АКТ

про впровадження результатів дисертаційної роботи
на здобуття наукового ступеня доктора технічних наук

Шафроненко Аліни Юріївни

Комісія у складі:

голова: директор КП «Санітарно-екологічний центр» ХМР Котенко І.О.

члени комісії: Перепилиця І.В., Малюга Д.М.

склала даний акт про те, що на комунальному підприємстві «Санітарно-екологічний центр» Харківської міської ради при аналізі води, який проводиться для встановлення її придатності для зрошення, або використання у гідропоніці, був використаний метод відновлення та фільтрації потоків даних за умов перетинних кластерів для задач покращення якості води, розроблений Шафроненко А.Ю. Для аналізу були використані хімічні та електрофізичні показники води.

Результати впровадження дають можливість визначити ступінь придатності води для агротехнічного використання, тип засолення, характер та вірогідність засолення при тривалому зрошенні, давати рекомендації щодо поліпшення характеристик води та/або зниження негативного впливу від її використання.

Акт складений для пред'явлення до спеціалізованої вченої ради із захисту дисертацій і не є підставою для фінансових розрахунків.

Члени комісії _____ Перепилиця І.В.

_____ Малюга Д.М.

Директор КП
«Санітарно-екологічний центр»
Харківської міської ради

Ігор КОТЕНКО
«26» березня 20 4 р.

АКТ

про впровадження результатів дисертаційної роботи
на здобуття наукового ступеня доктора технічних наук
ШАФРОНЕНКО АЛІНИ ЮРІЇВНИ

Комісія у складі:

голова: Котенко І. О.
члени комісії: Перепилиця І.В., Малюга Д. М.

склала даний акт про те, що на комунальному підприємстві «Санітарно-екологічний центр» Харківської міської ради при аналізі питної води в Харкові, був використаний модифікований рекурентний метод достовірної нечіткої кластеризації даних з використанням оптимізаційної процедури, розроблений Шафроненко А.Ю. Для аналізу були використані хімічні та електрофізичні показники 144 проб води, які були взяті з 12 джерел, що включають Саржин яр, Манжосів яр, селище Олешки, парк «Юність», «Петренки-1», «Петренки-2», «Петренки-3», «Олексіївська балка», а також вулиці Мінераловодська, Владислава Зубенка та Бучми (два каптажі).

Результати впровадження дають можливість визначити ступінь придатності питної води. За результатами аналізу питну воду можна набирати тільки з джерела у Саржиному Яру. У ній не виявили відхилень від гігієнічних нормативів за показниками епідемічної безпеки. Всі інші джерела не рекомендовані до вживання питної води.

Акт складений для пред'явлення до спеціалізованої вченої ради із захисту дисертацій і не є підставою для фінансових розрахунків.

Члени комісії _____ Перепилиця І.В.

_____ Малюга Д.М.

«ЗАТВЕРДЖУЮ»
В.о. ректора ХНУРЕ
_____Ігор РУБАН
«21» березня 2024 р.

АКТ

про впровадження в освітній процес ХНУРЕ результатів дисертаційної роботи **Шафроненко Аліни Юріївни** «Адаптивні нейро-фаззі методи для обробки потоків даних з використанням еволюційного самонавчання» за спеціальністю 05.13.23 – системи та засоби штучного інтелекту

Комісія у складі завідувача кафедри ШІ д.т.н., **проф. Філатова В.О.**; к.т.н., проф. каф. ШІ, **проф. Рябової Н.В.**; к.т.н., доц. кафедри ШІ, **доц. Чалої Л.Е.**, к.т.н., доц. кафедри ШІ, **доц. Золотухіна О.В.** розглянула матеріали дисертаційної роботи к.т.н., доц. каф. Інф., **доц. Шафроненко А.Ю.**, які використовуються в освітньому процесі кафедри ШІ ХНУРЕ у 2023/2024 навчальному році і прийшла до наступного висновку.

Розроблені у дисертаційній роботі адаптивні нейро-фаззі методи для обробки потоків даних з використанням еволюційного самонавчання використовуються в підготовці та написанні магістерських кваліфікаційних роботах освітньої програми 122 «Системи штучного інтелекту».

Результати за висновками комісії внесено до протоколу № 8 від 20 березня 2024р. засідання кафедри Штучного інтелекту.

Зав. каф. ШІ, проф. Валентин ФІЛАТОВ

проф. каф. ШІ, Наталія РЯБОВА

доц. каф. ШІ, Лариса ЧАЛА

доц. каф. ШІ, Олег ЗОЛОТУХІН

«ЗАТВЕРДЖУЮ»
В.о. ректора ХНУРЕ
_____Ігор РУБАН
«26» квітня 2024 р.

АКТ

про впровадження в освітній процес ХНУРЕ результатів дисертаційної роботи Шафроненко Аліни Юріївни «Адаптивні нейро-фаззі методи для обробки потоків даних з використанням еволюційного самонавчання» за спеціальністю 05.13.23 – системи та засоби штучного інтелекту

Комісія у складі завідувача кафедри Інформатики к.т.н., доц. Кобиліна О.А., к.т.н., доц., доц. каф. Інформатики Тітової О.В., к.т.н., доц., доц. кафедри Інформатики Руденко Д.О. розглянула матеріали дисертаційної роботи доцента Шафроненко А.Ю., які використовуються в освітньому процесі кафедри Інформатики ХНУРЕ у 2023/2024 навчальному році і прийшла до наступного висновку.

Розроблені у дисертаційній роботі адаптивні нейро-фаззі методи для обробки потоків даних з використанням еволюційного самонавчання, а саме:

- рекурентні ймовірнісні, можливісні та достовірні методи кластеризації, які працюють із задачами 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 вектор за вектором в міру їх надходження в систему;
- метод нечіткої кластеризації масивів даних, що базується на ідеях аналізу щільностей розподілу цих даних, їх піків та довірчого нечіткого підходу;
- метод кластеризації масивів даних на основі комбінованої оптимізації функцій щільності розподілу та еволюційних методів;

впроваджені в освітній процес кафедри Інформатики ХНУРЕ та використовуються у таких навчальних дисциплінах:

- у навчальній дисципліні «Машинне навчання» для бакалаврів освітньої програми 122 «Інформатика» у лекційному матеріалі за темами «Види машинного навчання», «Навчання на прикладах», «Ансамблі класифікаторів»;

– у навчальній дисципліні «Методи оптимізації в машинному навчанні» для магістрів освітньої програми 122 «Інформатика» у лекційному матеріалі за темами «Основи машинного навчання», «Регуляризація в машинному навчанні», «Оптимізація», «Методи оптимізації для глибинного навчання»;

використовуються в підготовці та написанні бакалаврських та магістерських кваліфікаційних роботах освітньої програми 122 «Інформатика»;
використовуються в підготовці аспірантів.

Результати за висновками комісії внесено до протоколу № 13 від 25 квітня 2024р. засідання кафедри Інформатики.

_____ Зав. каф. Інформатики, доц. Олег КОБИЛІН

_____ доц. каф. Інформатики Олена ТІТОВА

_____ доц. каф. Інформатики Діана РУДЕНКО