Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Кваліфікаційна наукова
праця на правах рукопису

ЖЕРНОВА ПОЛІНА ЄВГЕНІЇВНА

УДК 004.032.26

ДИСЕРТАЦІЯ
НЕЧІТКА КЛАСТЕРИЗАЦІЯ ПОТОКІВ ДАНИХ ЗА УМОВ НЕВІДОМОЇ КІЛЬКОСТІ КЛАСТЕРІВ

05.13.23 – системи та засоби штучного інтелекту
Подається на здобуття наукового ступеня кандидата технічних наук

Дисертація містить результати власних досліджень. Використання ідей, результатів і текстів інших авторів мають посилання на відповідне джерело
_____П.Є. Жернова

Науковий керівник (консультант) Бодянський Євгеній Володимирович,
доктор технічних наук, професор

Цей примірник дисертаційної роботи
ідентичний за змістом з іншими,
поданими до спеціалізованої вченої ради.

Учений секретар спецради Д 64.052.01 Є. І. Литвинова

Харків – 2019

АНОТАЦІЯ

Жернова Поліна Євгеніївна. Нечітка кластеризація потоків даних за умов невідомої кількості кластерів. – Кваліфікаційна наукова робота на правах рукопису.

Дисертація на здобуття наукового ступеня кандидата технічних наук (доктора філософії) за спеціальністю 05.13.23 «Системи та засоби штучного інтелекту». – Харківський національний університет радіоелектроніки, Міністерство освіти и науки України, Харків, 2019.

Мета дослідження – розробка методів нечіткої кластеризації потоків даних високої розмірності з використанням ансамблевого підходу, коли кількість та форма кластерів заздалегідь не відома.

Задачі дослідження: 1) Провести аналіз існуючих методів та підходів для кластеризації потоків даних. 2) Розробити метод для кластеризації потоків даних з використанням самоорганізовних карт Т. Кохонена у випадку невідомої кількості кластерів. 3) Розробити архітектуру ансамблю нейронних мереж на основі самоорганізовних карт Т. Кохонена. 4) Розробити ансамбль нейронних мереж з використанням ядерних функцій для вирішення задачі за умов лінійної нероздільності класів. 5) Розробити ансамбль нейро-фаззі систем для кластеризації потоку даних за припущенням, що кількість та форма кластерів невідомі заздалегідь. 6) Розробити ансамбль нейро-фаззі мереж на основі імовірнісно-можливісного підходу для кластеризації потоків даних. 7) Провести експериментальні дослідження розроблених методів, вирішити за їх допомогою практичні задачі нечіткої кластеризації потоків даних високої розмірності.

Об’єкт дослідження – процес обробки даних високої розмірності, які надходять на обробку спостереження за спостереженням з використанням ансамблевого підходу.

Предмет дослідження – методи нечіткої кластеризації з використанням ансамблевого підходу у задачах нечіткої кластеризації даних.

В дисертаційній роботі запропоновано ансамбль самоорганізовних карт Т. Кохонена, який базується на використанні онлайн методу К-середніх. Даний підхід дозволяє обробляти інформацію, яка надходить на вхід системи спостереження за спостереженням. На відміну від існуючих методів кластеризації використання ансамблевого підходу дозволяє обійти проблему коли кількість класів заздалегідь невідома, оскільки кожна з мереж Кохонена налаштована на свою кількість кластерів.

Вдосконалено метод, заснований на ансамблевому підході, з використанням ядерних самоорганізовних карт Т. Кохонена, що дозволило завдяки додатковому прихованому ядерному шару нейромережі підвищити розмірність вхідного простору, що дає змогу кластеризувати дані, які є лінійно нероздільними.

Розроблено ансамбль нейро-фаззі самоорганізовних карт Т. Кохонена для кластеризації потоків даних, який за допомогою використання вдосконаленого методу С-середніх та додаткового ядерного шару здатний обробляти інформацію, яка є лінійно нероздільною, а також обробляти кластери довільної форми. Саме це дозволяє обробляти дані високої розмірності та уникнути ефект концентрації норм.

Вдосконалено ансамбль самоорганізовних карт Т. Кохонена для кластеризації потоків даних високої розмірності, який обробляє інформацію, що надходять на вхід системи з використанням декількох підходів: імовірнісного та можливісного.

Практична значимість отриманих результатів. Розроблені у роботі методи кластеризації даних на основі ансамблевого підходу та нейро-фаззі систем обчислювального інтелекту призначені для онлайн обробки потоків даних в умовах невизначеності про кількість та форму кластерів. Отриманий підхід є достатньо простим з обчислювальної точки зору та дозволяє вирішувати задачі інтелектуального аналізу даних та інтелектуального аналізу потоків даних. Використання методів кластеризації на основі ансамблевого підходу дозволяє підвищити ефективність вирішення задач обробки медичних даних. Ансамбль нейро-фаззі кластеризаційних мереж дозволяє підвищити точність аналізу потоків

даних. Цей підхід дозволив встановити закономірності формування відповідної реакції організму на сполучений вплив екологічних чинників; використання методичних підходів, щодо визначення гігієнічної значущості біологічних ефектів сполученої дії електромагнітного випромінювання та низьких температур при аналізі результатів НДР бюджетного фінансування «Встановити механізми адаптації до сполученої дії хімічних та фізичних чинників навколишнього середовища»

Також основні результати дисертаційної роботи використовуються в навчальному процесі Харківського національного університету радіоелектроніки на кафедрі системотехніки в курсі «Нейросистеми та генетичні алгоритми».

Публікації. Результати наукових досліджень опубліковані у 14 друкованих працях: 1 розділ у монографії (входить до наукометричної бази Scopus), 5 статей (включені до «Переліку наукових фахових видань України»), а також 8 наукових конференціях (з них 2 входять до наукометричної бази Scopus).

Ключові слова: кластеризація потоків даних, ансамбль нейронних мереж, самоорганізовні карти Т. Кохонена, нейро-фаззі кластеризаційні мережі.

СПИСОК ОПУБЛІКОВАНИХ РОБІТ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Список публікацій здобувача, в яких опубліковані основні наукові результати дисертації:

1. P. Zhernova, A. Deyneko, Z. Deyneko, I. Pliss and V. Ahafonov, "Data Stream Clustering in Conditions of an Unknown Amount of Classes," In: Hu Z., Petoukhov S., Dychka I., He M. (eds) *Advances in Computer Science for Engineering and Education. ICCSEEA 2018. Advances in Intelligent Systems and Computing*, vol 754. Springer, Cham, pp. 410-419, 2019.

2. Є. Бодянський, А. Дейнеко, П. Жернова, О. Золотухін та Я. Хаустова, «Послідовне ядерне нечітке кластерування великих масивів даних на основі гібридної системи обчислювального інтелекту,» *Вісник Національного*

університету "Львівська політехніка". Інформаційні системи та мережі, № 829, pp. 20-24, 2017.

3. Є. Бодянський, А. Дейнеко, П. Жернова та В. Репін, «Онлайн модифікація методу Х-середніх на основі ансамблю самоорганізовних мап Т. Когонена,» *Збірник наукових праць «Розвиток транспорту»,* № 1, pp. 96-107, 2017.

4. П. Жернова та Є. Бодянський, «Ядерна нечітка кластеризація потоків даних на основі ансамблю нейронних мереж,» *Сучасний стан наукових досліджень та технологій в промисловості,* № 4(6), pp. 42-49, 2018.

5. Y. Bodyanskiy, I. Perova and P. Zhernova, "Online fuzzy clustering of high - dimensional data based on ensembles in data stream mining tasks," *Innovative Technologles & Scientific Solutions for Industries*, no. 1(7), pp. 16-24, 2019.

6. П. Жернова та Є. Бодянський, «Нечітка імовірнісно-можливісна послідовна кластеризація даних на основі ансамблевого підходу,» *Науково-технічний журнал «Прикладна радіоелектроніка»,* № 1,2, pp. 40-45, 2019.

Результати, які засвідчують апробацію матеріалів дисертації:

7. Е. Бодянский, А. Дейнеко, П. Жернова и В. Репин, «Адаптивная модификация метода Х-средних на основе ансамбля кластеризующих нейронных сетей Т. Кохонена,» в *Матеріали VI Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології»*, Одесса, 2017.

8. Е. Бодянский, П. Жернова и А. Дейнеко, «Кластеризующий ансамбль нейронных сетей и его обучение в условиях неизвестного количества классов,» в *Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту»*, Залізний порт, Україна, 2018.

9. А. Дейнеко, П. Жернова, І. Плісс та О. Чала, «Модифікована нечітка ймовірнісна нейронна мережа,» в *Матеріали міжнародної наукової конференції*

«Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту», Залізний порт, Україна, 2018.

10. P. Zhernova, A. Deyneko, Y. Bodyanskiy and V. Riepin, "Adaptive kernel data streams clustering based on neural networks ensembles in conditions of uncertainty about amount and shapes of clusters," in *IEEE Second International Conference on Data Stream Mining & Processing*, Lviv, Ukraine, 2018.

11. Deineko, P. Zhernova, B. Gordon, O. Zayika, I. Pliss and N. Pabyrivska, "Data stream online clustering based on fuzzy expectation-maximization approaching formation on submission," in *IEEE Second International Conference on Data Stream Mining & Processing*, Lviv, Ukraine, 2018.

12. П. Жернова, «Вероятностно-возможностный подход для кластеризации потоков данных на основе ансамблей нейронных сетей,» в *Материалы международной научно-практической конференции «Информационные технологии и системы»*, Харьков, 2019.

13. П. Жернова та А. Лобинцев, «Кластеризація даних високої розмірності з використанням можливісного підходу,» в *Матеріали 23-го Міжнародного молодіжного форуму «Радіоелектроніка та молодь в 21 столітті»*, Харьков, 2019.

14. П. Жернова та Є. Бодянський, «Нейро-фаззі мережа та її навчання для кластеризації потоків даних високої розмірності,» в *Матеріали V міжнародної науково-практичної конференції «Обчислювальний інтелект (результати, проблеми, перспективи)»*, Ужгород, 2019.

ABSTRACT

Polina Zhernova. Fuzzy clustering of data stream in an unknown number of clusters. Qualifying scientific work as a manuscript

Thesis for the degree of candidate of technical sciences (Ph.D.) in specialty 05.13.23 "Systems and means of artificial intelligence". - Kharkiv National University of Radio Electronics, Ministry of Education and Science of Ukraine, Kharkiv, 2019.

The purpose of the research – development of fuzzy clustering methods for high-dimensional data streams using ensemble approach, when the number and shape of clusters are unknown in advance.

The research tasks: 1) Analyze existing methods and approaches for clustering streaming data. 2) Develop a method for clustering data flows using self-organizing Kohonen map in case of unknown number of clusters. 3) Develop ensemble architecture of neural networks based on self-organizing Kohonen map. 4) Develop a neural networks ensemble using kernel methods to solve the problem under conditions of classes linear inseparability. 5) Develop an ensemble of neuro-fuzzy systems for data flow clustering assuming that number and form of clusters are unknown in advance. 6) Develop an ensemble of neuro-fuzzy systems based on probabilistic-possibilistic approach for data flow clustering. 7) Conduct experimental studies with developed methods, solve practical tasks of fuzzy clustering of high-dimensional data streams with their help.

The object of the research – process of high dimensional data processing that arrive for processing in online mode with the use of ensemble approach.

The subject of the research – fuzzy clustering methods with the use of ensemble approach in data fuzzy clustering tasks.

In the dissertation work, self-organizing Kohonen map ensemble based on the use of online K-means method is proposed. This approach allows to process information that is fed to the system input observation by observation. Unlike existing clustering methods, the use of the ensemble approach allows to solve the problem when classes number is

unknown in advance, because each of the Kohonen networks is configured for its own clusters number.

The method based on ensemble approach is improved with the use of kernel self-organizing Kohonen map, which allowed the additional hidden kernel layer of neural network to increase the input space dimension, which makes it possible to cluster data that is linear inseparable.

A neuro-fuzzy ensemble of self-organizing Kohonen maps has been developed for data flows clustering, which by using an advanced fuzzy C-means and an extra kernel layer is capable to process an information that is linear inseparable as well as process arbitrary shape clusters. This is what allows to process high dimensional data and to avoid the effect of norms concentration.

The self-organizing Kohonen map ensemble has been improved for clustering of high-dimensional data streams that processes information entering the system using several approaches: probabilistic and possibilistic.

The practical significance of the thesis's results. Developed data clustering methods based on ensemble approach and neuro-fuzzy systems of computing intelligence are designed for online data streams processing in conditions of uncertainty about the number and form of clusters. The resulting approach is quite simple from a computational point of view and allows solving tasks of data mining analysis and data streams mining analysis. Using clustering methods based on ensemble approach can increase the efficiency of solving medical data processing tasks. The ensemble of neuro-fuzzy clustering networks improves accuracy of data streams analysis. This approach allowed to establish formation patterns of organism corresponding reaction to the combined effect of environmental factors; use of methodical approaches for determining the hygienic significance of combined action biological effects of electromagnetic radiation and low temperatures when analyzing the budget financed scientific research results «Set mechanisms for adaptation to combined effect of chemical and physical environment factors».

Also, dissertation work main results are used in the Kharkiv National University of Radio Electronics educational process at the System Engineering department in the "Neurosystems and Genetic Algorithms" course.

Publications. Scientific research results are published in 14 printed works: 1 section in the monograph (included to the scientometric base Scopus), 5 articles (included to the «List of professional scientific editions of Ukraine») and 8 scientific conferences (2 of these are included to the scientometric base Scopus).

Keywords: data streams clustering, neural networks ensemble, self-organizing Kohonen map, neuro-fuzzy clustering networks.

LIST OF PUBLICATIONS

The list of publications, which reflect the main scientific results of the thesis:

1. P. Zhernova, A. Deyneko, Z. Deyneko, I. Pliss and V. Ahafonov, "Data Stream Clustering in Conditions of an Unknown Amount of Classes," In: Hu Z., Petoukhov S., Dychka I., He M. (eds) *Advances in Computer Science for Engineering and Education. ICCSEEA 2018. Advances in Intelligent Systems and Computing*, vol 754. Springer, Cham, pp. 410-419, 2019.

2. Є. Бодянський, А. Дейнеко, П. Жернова, О. Золотухін та Я. Хаустова, «Послідовне ядерне нечітке кластерування великих масивів даних на основі гібридної системи обчислювального інтелекту,» *Вісник Національного університету "Львівська політехніка". Інформаційні системи та мережі,* № 829, pp. 20-24, 2017.

3. Є. Бодянський, А. Дейнеко, П. Жернова та В. Репін, «Онлайн модифікація методу X-середніх на основі ансамблю самоорганізовних мап Т. Когонена,» *Збірник наукових праць «Розвиток транспорту»,* № 1, pp. 96-107, 2017.

4. П. Жернова та Є. Бодянський, «Ядерна нечітка кластеризація потоків даних на основі ансамблю нейронних мереж,» *Сучасний стан наукових досліджень та технологій в промисловості,* № 4(6), pp. 42-49, 2018.

5. Y. Bodyanskiy, I. Perova and P. Zhernova, "Online fuzzy clustering of high - dimensional data based on ensembles in data stream mining tasks," *Innovative Technologles & Scientific Solutions for Industries*, no. 1(7), pp. 16-24, 2019.

6. П. Жернова та Є. Бодянський, «Нечітка імовірнісно-можливісна послідовна кластеризація даних на основі ансамблевого підходу,» *Науково-технічний журнал «Прикладна радіоелектроніка»,* № 1,2, pp. 40-45, 2019.

Results that confirm the approbation of the thesis:

7. Е. Бодянский, А. Дейнеко, П. Жернова и В. Репин, «Адаптивная модификация метода Х-средних на основе ансамбля кластеризующих нейронных сетей Т. Кохонена,» в *Матеріали VI Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології»*, Одесса, 2017.

8 Е. Бодянский, П. Жернова и А. Дейнеко, «Кластеризующий ансамбль нейронных сетей и его обучение в условиях неизвестного количества классов,» в *Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту»*, Залізний порт, Україна, 2018.

9. А. Дейнеко, П. Жернова, І. Плісс та О. Чала, «Модифікована нечітка ймовірнісна нейронна мережа,» в *Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту»*, Залізний порт, Україна, 2018.

10. P. Zhernova, A. Deyneko, Y. Bodyanskiy and V. Riepin, "Adaptive kernel data streams clustering based on neural networks ensembles in conditions of uncertainty about amount and shapes of clusters," in *IEEE Second International Conference on Data Stream Mining & Processing*, Lviv, Ukraine, 2018.

11. Deineko, P. Zhernova, B. Gordon, O. Zayika, I. Pliss and N. Pabyrivska, "Data stream online clustering based on fuzzy expectation-maximization approaching formation on submission," in *IEEE Second International Conference on Data Stream Mining & Processing*, Lviv, Ukraine, 2018.

12. П. Жернова, «Вероятностно-возможностный подход для кластеризации потоков данных на основе ансамблей нейронных сетей,» в *Материалы международной научно-практической конференции «Информационные технологии и системы»*, Харьков, 2019.

13. П. Жернова та А. Лобинцев, «Кластеризація даних високої розмірності з використанням можливісного підходу,» в *Матеріали 23-го Міжнародного молодіжного форуму «Радіоелектроніка та молодь в 21 столітті»*, Харьков, 2019.

14. П. Жернова та Є. Бодянський, «Нейро-фаззі мережа та її навчання для кластеризації потоків даних високої розмірності,» в *Матеріали V міжнародної науково-практичної конференції «Обчислювальний інтелект (результати, проблеми, перспективи)»*, Ужгород, 2019.

ЗМІСТ

ВСТУП

Актуальність дослідження. На сьогоднішній день для вирішення задач інтелектуального аналізу даних, насамперед кластеризації даних, що є невід'ємною частиною проблеми Data Mining, існує безліч підходів та методів які основуються на нейро-фаззі системах. Розроблений математичний апарат що дозволяє вирішувати задачі кластеризації даних у різних сферах таких, як медицина, наука, техніка та інші. Але більшість відомих методів за своєю суттю є чіткими процедурами, тобто кластери лінійно роздільні, а інформація обробляється в пакетному режимі. На сьогоднішній день існує також багато нечітких методів кластеризації, але всі вони використовують функцію належності та працюють с кластерами опуклої форми та лінійно роздільною інформацією. Але зараз на перший план виходять задачі, які пов'язані з Dynamic Data Mining, Data Stream Mining та Big Data, коли дані надходять у вигляді потоку та кластери мають довільну форму та перетинаються у просторі ознак, виникає потреба розробити методи нечіткої кластеризації для обробки потоків даних. Така задача може бути вирішена за допомогою нечітких методів та ядерного підходу згідно з гіпотезою Кавера: якщо задача лінійно нероздільна у вихідному просторі, вона може бути лінійно роздільною у просторі більш високої розмірності. Також однією з проблем кластеризації даних є те що заздалегідь невідома кількість кластерів, на яку будуть поділені вхідні дані, що надходять на обробку у вигляді потоку.

Таким чином, на сьогоднішній день є актуальною задача розробки нових методів для нечіткої кластеризації потоків даних високої розмірності призначених для обробки даних в онлайн режимі, коли кількість кластерів невідома заздалегідь та вони мають довільну форму та перетинаються у просторі ознак.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконана в рамках держбюджетних НДР: «Динамічний інтелектуальний аналіз послідовностей нечіткої інформації за умов суттєвої невизначеності на

основі гібридних систем обчислювального інтелекту» (№ГР 0116U002539); «Глибинні гібридні системи обчислювального інтелекту для аналізу потоків даних та їх швидке навчання» (№ГР 0119U001403) В рамках зазначених НДР здобувачем розроблені методи синтезу ансамблів нечіткої кластеризації, які призначені для обробки даних в онлайн режимі, коли дані надходять на обробку послідовно, одне за одним, а кластери можуть перетинатися у просторі ознак та мати довільну форму.

Мета і завдання дослідження. Розробка методів нечіткої кластеризації потоків даних високої розмірності з використанням ансамблевого підходу, коли кількість та форма кластерів заздалегідь не відома.

Відповідно до поставленої мети у дисертаційній роботі необхідно вирішити такі завдання:

1) Провести аналіз існуючих методів та підходів для кластеризації потоків даних;

2) Розробити метод для кластеризації потоків даних у випадку невідомої кількості кластерів;

3) Розробити архітектуру ансамблю нейронних мереж для кластеризації потоків даних;

4) Розробити ансамбль нейронних мереж з використанням ядерних функцій для вирішення задачі за умов лінійної нероздільності класів;

5) Розробити ансамбль нейро-фаззі систем для кластеризації потоку даних за припущенням, що кількість та форма кластерів невідомі заздалегідь;

6) Розробити ансамбль нейро-фаззі мереж на основі імовірнісно-можливісного підходу для кластеризації потоків даних;

7) Провести експериментальні дослідження розроблених методів, вирішити за їх допомогою практичні задачі нечіткої кластеризації потоків даних високої розмірності.

Об’єкт дослідження – процес обробки даних високої розмірності, які надходять на обробку спостереження за спостереженням за умов невизначеної кількості та форми кластерів.

Предмет дослідження – методи нечіткої кластеризації на основі ансамблевого підходу у задачах коли дані обробляються послідовно за умов невизначеної кількості та форми кластерів.

Методи дослідження: базуються на теорії обчислювального інтелекту, а саме на методах теорії штучних нейронних мереж та теорії нечіткої логіки для побудови ансамблю нечітких нейро-фаззі мереж Т. Кохонена, що дозволяє провести нечітку кластеризацію потоків даних; теорії оптимізації для синтезу методів нечіткої кластеризації та методів самонавчання нейро-фаззі мереж. Імітаційне моделювання використовується для перевірки якості роботи ансамблю нейро-фаззі кластеризаційних мереж Т. Кохонена для потоку даних високої розмірності.

Наукова новизна результатів дослідження:

1. Вперше запропоновано ансамбль самоорганізовних карт Т. Кохонена, який відрізняється використанням онлайн методу К-середніх, що дозволяє кластеризувати дані за умов апріорі невідомої кількості класів.

2. Вперше запропоновано ансамбль нейро-фаззі самоорганізовних карт Т. Кохонена, який відрізняється використанням модифікованого онлайн методу нечітких С-середніх, коли апріорі невідома кількість та форма кластерів, що дозволяє кластеризувати потоки даних за умов лінійної нероздільності класів, які довільним чином перетинаються у просторі ознак.

3. Удосконалено ансамбль ядерних самоорганізовних карт Т. Кохонена, який характеризується введенням додаткового ядерного шару для підвищення розмірності вхідного простору, що дозволяє кластеризувати потоки даних за умов, коли кластери є лінійно нероздільними.

4. Удосконалено ансамбль самоорганізовних нечітких карт Т. Кохонена, який відрізняється одночасним використанням процедури імовірнісної та

можливісної кластеризації потоків даних, що дозволяє підвищити рівень якості кластеризації потоків даних.

Практична значимість отриманих результатів. Розроблені у роботі методи кластеризації даних на основі ансамблевого підходу та нейро-фаззі систем обчислювального інтелекту призначені для онлайн обробки потоку даних в умовах невизначеності про кількість та форму кластерів. Отриманий підхід є достатньо простим з обчислювальної точки зору та дозволяє вирішувати задачі інтелектуального аналізу даних та інтелектуального аналізу потоку даних. Використання методів кластеризації на основі ансамблевого підходу дозволяє підвищити ефективність вирішення задач обробки медичних даних. Ансамбль нейро-фаззі кластеризаційних мереж дозволяє підвищити точність аналізу потоків даних. Цей підхід дозволив встановити закономірності формування відповідної реакції організму на сполучений вплив екологічних чинників; використання методичних підходів, щодо визначення гігієнічної значущості біологічних ефектів сполученої дії електромагнітного випромінювання та позитивних низьких температур при аналізі результатів НДР бюджетного фінансування «Встановити механізми адаптації до сполученої дії хімічних та фізичних чинників навколишнього середовища»

Також основні результати дисертаційної роботи використовуються в навчальному процесі Харківського національного університету радіоелектроніки на кафедрі системотехніки в курсі «Нейросистеми та генетичні алгоритми».

Особистий внесок здобувача. Всі основні результати дисертаційної роботи, які виносяться на захист отримано автором особисто. У роботах, написаних зі співавторами, здобувачеві належить: [1] – ансамбль самоорганізовних карт Т. Кохонена для кластеризації потоків даних коли кількість кластерів апріорно невідома; [2] – архітектура ядерної кластерувальної мережі з використанням самоорганізовних карт Т. Кохонена; [3] – онлайн модифікація методу Х-середніх з використанням ансамблевого підходу для кластеризації потоків даних; [4] – ядерний ансамбль самоорганізовних карт Т. Кохонена для кластеризації потоків

даних коли кількість та форма кластерів заздалегідь невідома; [5] – ансамбль нейро-фаззі самоорганізовних карт Т. Кохонена для кластеризації потоків даних з використанням імовірнісного підходу; [6] – ансамбль нейро-фаззі самоорганізовних карт Т. Кохонена для кластеризації потоків даних з використанням імовірнісного-можливісного підходу; [7] – адаптивна модифікація методу Х-середніх для кластеризації потоків даних; [8] – ансамбль нейронних мереж та його навчання для кластеризації потоків даних; [9] – нечітка імовірнісна нейронна мережа для кластеризації даних; [10] – ансамбль ядерних самоорганізовних карт Т. Кохонена для кластеризації потоків даних у ситуаціях коли кластери є лінійно не роздільними; [11] – нечітка кластеризація потоків даних на основі методу С-середніх; [12]– ансамбль самоорганізовних карт Т. Кохонена для кластеризації потоків даних на основі імовірнісно-можливісного підходу; [13]– ансамбль нейро-фаззі мереж Кохонена для кластеризації потоків даних .

Апробація результатів роботи. Основні результати дисертаційної роботи доповідалися й обговорювалися на: VI Міжнародній науково-практичній конференції «Інформаційні управляючі системи та технології», Україна, м. Одеса, 2017 р.; Міжнародній науковій конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту», Україна, м. Залізний порт, 2018 р.; Міжнародній конференції «The 2 IEEE International Conference on Data Stream Mining & Processing», Україна, м. Львів, 2018 р.; Міжнародній науково-практичній конференції “Інформаційні технології та системи”, Україна, м. Харків, 2019 р.; XXIII-му Міжнародному молодіжному форумі «Радіоелектроніка та молодь у XXI столітті», Україна, м. Харків, 2019 р.; Міжнародному науковому симпозіумі "Інтелектуальні рішення", Україна, м. Ужгород, 2019 р..

Публікації. За результатами досліджень опубліковано 6 наукових праць: 5 статей у фахових періодичних виданнях України з технічних наук, 1 стаття в англомовному виданні, яке включено у міжнародну наукометричну базу Scopus.

Структура та обсяг роботи. Дисертація складається зі вступу, 6 розділів, висновків, списку використаних джерел, додатку. Загальний обсяг роботи складає

140 сторінок тексту, що містять 2 анотації на 10 сторінках, 38 рисунків, 9 таблиць, список використаних джерел з 117 найменувань на 12 сторінках, 2 додатки на 6 сторінках.

1 ОГЛЯД СТАНУ ПРОБЛЕМИ ТА ПОСТАНОВКА ЗАВДАННЯ ДОСЛІДЖЕННЯ

1.1 Класифікація та кластеризація

1.1.1 Класифікація

Класифікація є найбільш простим та одночасно найбільш часто розв'язуваним завданням Data Mining. З огляду на поширеність завдань класифікації необхідно чітке розуміння суті цього поняття.

Класифікація – впорядкована по деякому принципу множина об'єктів, які мають подібні класифікаційні ознаки (одна або кілька властивостей), обраних для визначення подібності або відмінності між цими об'єктами.

Класифікація вимагає дотримання наступних правил:

– на кожному кроці поділу необхідно застосовувати тільки одну основу;

– розподіл має бути пропорційним;

– члени поділу повинні взаємно виключати один одного, їх обсяги не повинні перехрещуватися;

– розподіл має бути послідовним.

Розрізняють:

– допоміжну (штучну) класифікацію, яка проводиться по зовнішній ознаці та служить для додання множини предметів (процесів, явищ) потрібного порядку;

– природну класифікацію, яка проводиться за істотними ознаками, що характеризує внутрішню спільність предметів та явищ. Вона є результатом та важливим засобом наукового дослідження тому, що передбачає та закріплює результати вивчення закономірностей класифікації об'єктів.

Залежно від обраних ознак, їх поєднання та процедури поділу понять класифікація може бути:

– простою – поділ родового поняття тільки за ознакою та тільки один раз до розкриття всіх видів. Прикладом такої класифікації є дихотомія, при якій членами поділу бувають тільки два поняття, кожне з яких є таким, що суперечить іншому (тобто дотримується принцип: "А та не А");

– складною – застосовується для поділу одного поняття з різних підстав і синтезу таких простих розподілів в єдине ціле. Прикладом такої класифікації є періодична система хімічних елементів.

Під класифікацією будемо розуміти віднесення об'єктів (спостережень, подій) до одного з заздалегідь відомих класів.

Класифікація – це закономірність, що дозволяє робити висновок щодо визначення характеристик конкретної групи. Таким чином, для проведення класифікації повинні бути присутні ознаки, що характеризують групу, до якої належить та чи інша подія або об'єкт (зазвичай при цьому на підставі аналізу вже класифікованих подій формулюються якісь правила).

Класифікація відноситься до стратегії навчання з вчителем (supervised learning), яке також називають контрольованим або керованим навчанням.

Завданням класифікації часто називають передбачення категоріальної залежної змінної (тобто залежною змінною, яка є категорією) на основі вибірки неперервних та / або категоріальних змінних.

Наприклад, можна передбачити, хто з клієнтів фірми є потенційним покупцем певного товару, а хто – ні, хто скористається послугою фірми, а хто – ні, та інше. Цей тип завдань відноситься до завдань бінарної класифікації, в них залежна змінна може приймати тільки два значення (наприклад, так чи ні, 0 або 1).

Інший варіант класифікації виникає, якщо залежна змінна може приймати значення з деякої множини визначених класів. Наприклад, коли необхідно передбачити, яку марку автомобіля захоче купити клієнт. У цих випадках розглядається множина класів для залежної змінної.

Класифікація може бути одновимірною (за однією ознакою) та багатовимірною (за двома та більше ознаками).

Багатовимірна класифікація була розроблена біологами при вирішенні проблем дискримінації для класифікації організмів. Однією з перших робіт, присвячених цьому напрямку, вважають роботу Р. Фішера (1930 р.), в якій організми поділялися на підвиди залежно від результатів вимірювань їх фізичних параметрів. Біологія була та залишається найбільш затребуваним та зручним середовищем для розробки багатовимірних методів класифікації.

Мета процесу класифікації полягає в тому, щоб побудувати модель, яка використовує атрибути, що вже прогнозовані в якості вхідних параметрів та отримує значення залежного атрибута. Процес класифікації полягає в розбитті множини об'єктів на класи за певним критерієм [14].

Класифікатором називається деяка сутність, яка визначає, якому з визначених класів належить об'єкт по вектору ознак.

Для проведення класифікації за допомогою математичних методів необхідно мати формальний опис об'єкта, яким можна оперувати, використовуючи математичний апарат класифікації. Таким описом в нашому випадку виступає база даних. Кожен об'єкт (запис бази даних) несе інформацію про деяку властивість об'єкта.

1.1.2 Кластеризація

Завдання кластеризації подібне до завдання класифікації, є її логічним продовженням, але її відмінність в тому, що класи досліджуваного набору даних заздалегідь не визначені [15].

Синонімами терміна "кластеризація" є "автоматична класифікація", "навчання без вчителя" та "таксономія" [16].

Кластеризація призначена для розбиття сукупності об'єктів на однорідні групи (кластери або класи). Якщо дані вибірки уявити як точки в просторі ознак, то задача кластеризації зводиться до визначення "згущень точок" [17].

Мета кластеризації – пошук існуючих структур.

Кластеризація є описовою процедурою, вона не робить ніяких статистичних висновків, але дає можливість провести розвідувальний аналіз та вивчити "структуру даних".

Саме поняття "кластер" визначено неоднозначно: в кожному дослідженні свої "кластери". Перекладається поняття кластер (cluster) як "скупчення", "гроно" [18] [19].

Кластер можна охарактеризувати як групу об'єктів, що мають спільні властивості [20].

Характеристиками кластера можна назвати дві ознаки:

– внутрішня однорідність;

– зовнішня ізольованість.

Питання, що задається аналітиками при вирішенні багатьох завдань, полягає в тому, як організувати дані в наочні структури, тобто розгорнути таксономії [21].

Найбільше застосування кластеризація спочатку отримала в таких науках як біологія, антропологія, психологія [22]. Для вирішення економічних завдань кластеризації тривалий час мало використовувалася через специфіку економічних даних та явищ [23].

У таблиці 1.1 наведено порівняння деяких параметрів завдань класифікації та кластеризації.

Таблиця 1.1 – Порівняння класифікації та кластеризаціі

Характеристика	Класифікація	Кластеризація
Контрольованість навчання	Контрольоване навчання	Неконтрольоване навчання
Стратегія	Навчання з вчителем	Навчання без учителя
Наявність позначки класу	Навчальна множина супроводжується міткою, яка вказує клас, до якого належить спостереження	Мітки класу навчальної множини невідомі

Продовження таблиці 1.1

Підстава для класифікації	Нові дані класифікуються на підставі навчальної множини	Наявність множини даних з метою встановлення існування класів або кластерів даних

На рисунку 1.1 схематично представлені завдання класифікації та кластеризації.

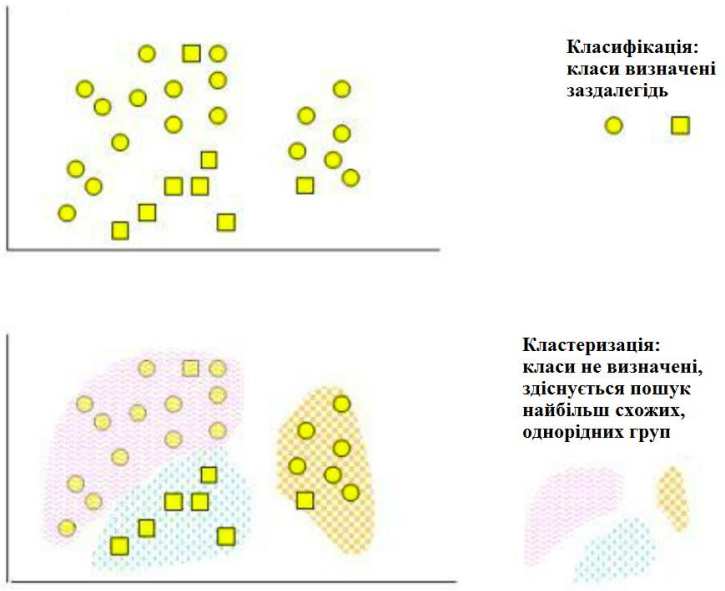

Рисунок 1.1 – Порівняння задач класифікації та кластеризації

Кластери можуть бути неперетинними, або ексклюзивними (non-overlapping, exclusive), та перетинними (overlapping) [24]. Схематичне зображення неперетинних та перетинних кластерів дано на рисунку 1.2

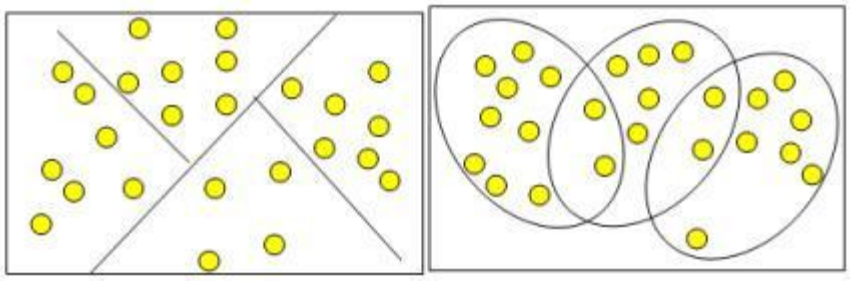

Рисунок 1.2 – Неперетинні та перетинні кластери

Слід зазначити, що в результаті застосування різних методів кластерного аналізу можуть бути отримані кластери різної форми. Наприклад, можливі кластери типу "ланцюжка", коли кластери представлені довгими "ланцюжками", кластери подовженої форми та інші, а деякі методи можуть створювати кластери довільної форми [25].

Різні методи можуть прагнути створювати кластери певних розмірів (наприклад, малих або великих) або припускати в наборі даних наявність кластерів різного розміру.

Деякі методи кластерного аналізу особливо чутливі до шумів або викидів, інші – менш.

В результаті застосування різних методів кластеризації можуть бути отримані неоднакові результати, це нормально та є особливістю роботи того чи іншого алгоритму.

Дані особливості слід враховувати при виборі методу кластеризації.

На сьогоднішній день розроблено більше сотні різних алгоритмів кластеризації [26].

Наведемо коротку характеристику підходів до кластеризації.

Алгоритми, засновані на поділі даних (Partitioning algorithms), у тому числі ітеративні:

– поділ об'єктів на k кластерів;
– ітеративний перерозподіл об'єктів для поліпшення кластеризації.

Ієрархічні алгоритми (Hierarchy algorithms):

– агломерація: кожен об'єкт спочатку є кластером, кластери, з'єднуючись один з одним, формують більший кластер і т.д.

Методи, засновані на концентрації об'єктів (Density-based methods):

– засновані на можливості з'єднання об'єктів;

– ігнорують шуми, знаходження кластерів довільної форми.

Грід-методи (Grid-based methods):

– квантування об'єктів в грід-структури.

Модельні методи (Model-based):

– використання моделі для знаходження кластерів, які найбільш відповідають даним.

Оцінка якості кластеризації може бути проведена на основі наступних процедур:

– ручна перевірка;

– встановлення контрольних точок та перевірка на отриманих кластерах;

– визначення стабільності кластеризації шляхом додавання в модель нових змінних;

– створення та порівняння кластерів з використанням різних методів. Різні методи кластеризації можуть створювати різні кластери, та це є нормальним явищем. Однак створення подібних кластерів різними методами вказує на правильність кластеризації.

Процес кластеризації залежить від обраного методу та майже завжди є ітеративним. Він може включати множину експериментів з вибору різноманітних параметрів, наприклад, міри відстані, типу стандартизації змінних, кількості кластерів та інші. Однак експерименти не повинні бути самоціллю – адже кінцевою метою кластеризації є отримання змістовних відомостей про структуру досліджуваних даних. Отримані результати вимагають подальшої інтерпретації, дослідження та вивчення властивостей та характеристик об'єктів для можливості точного опису сформованих кластерів.

1.2 Методи навчання без вчителя

У так званому неконтрольованому навчанні бажаний вихід моделі y невідомий або вважається невідомим. Метою неконтрольованих методів навчання є обробка або вилучення інформації з відомостями про вхідні дані $X=\{x(1),...,x(2),...,x(k),...,x(N),...\}\subset R^n \quad x(k)=(x_1(k),...,x_i(k),...,x_n(k))^T$, де $k=1,...,N$. Методи неконтрольованого навчання можуть бути дуже цікавими та корисними для попередньої обробки даних; див. рис. 1.3. Попередня обробка перетворює дані в іншу форму, яка може бути краще оброблена наступною моделлю [27]. У цьому контексті важливо мати на увазі, що бажаний вихід фактично доступний, і може існувати деякий ефективний спосіб включити ці знання навіть у фазу попередньої обробки.

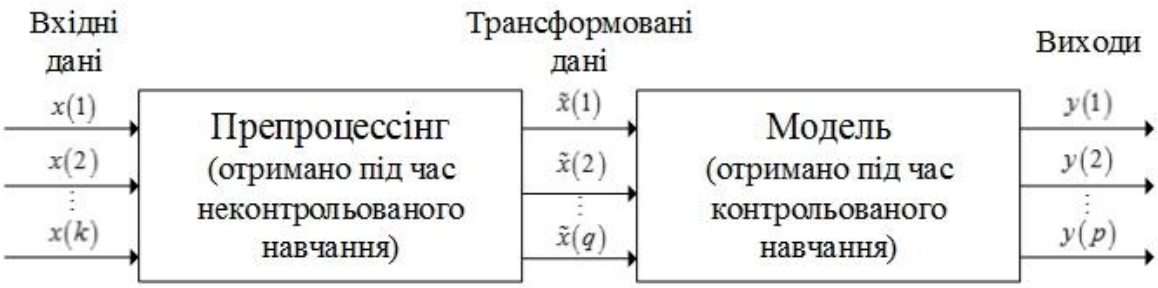

Рисунок 1.3 – Процес перетворення вхідних даних

Наступний приклад ілюструє типове використання методів неконтрольованого навчання для простої проблеми класифікації. На рисунку 1.4 показано розподіл вхідних даних у x_1-x_2 – вхідному просторі. Припустимо, що x_1 та x_2 представляють дві функції, які повинні бути зіставлені з класами, представленими цілими значеннями виводу y. Наприклад, правильна класифікація для XOR – подібної задачі – призначення верхньої лівої та нижньої груп даних (кластери) до класу 1 та двох інших кластерів до класу 0. У задачі з

контрольованого навчання кожен зразок навчальних даних буде складатися як з вхідних значень x_1, так і з x_2 та відповідного вихідного значення y, що дорівнює 1 або 0. При такому навчанні дані про класифікацію даних можуть бути легко вирішені, а для неконтрольованої проблеми навчання виходи невідомі, тобто навчальні дані складаються тільки з вхідних даних без будь-якої інформації про пов'язані класи.

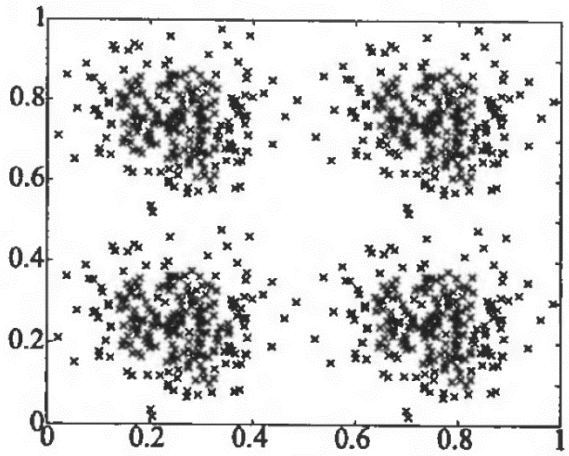

Рисунок 1.4 – Чотири кластери у двовимірному вхідному просторі.

Таким чином, найкраще, що може зробити методика без вчителя, групувати або кластеризувати вхідні дані якимось чином. Наприклад, алгоритм може знайти чотири кластера зі своїми центрами приблизно (0,25, 0,25), (0,75, 0,25), (0,25, 0,75), та (0,75, 0,75) у вигляді кіл з наближеним радіусом 0,25 кожен. Далі, на другому етапі, ці чотири кластери можуть бути відображені за допомогою методики контрольованого навчання для відповідних двох класів. Таким чином, кластеризація, що виконується методом навчання без вчителя, перетворила проблему відображення великої кількості вхідних даних до проблеми відображення чотирьох кластерів в двох класах. Тим самим складність другого етапу відображення значно зменшується.

Наведений вище приклад є типовий тим, що єдиний підхід навчанням з вчителем може бути замінений двоступеневою процедурою, що складається з фази

попередньої обробки без вчителя, за якою йде фаза навчання. Часто двоступеневий підхід є набагато менш вимогливим, ніж оригінальна одиночна проблема навчання. Перший етап такої двоступеневої стратегії можна розглядати як стиснення інформації або зменшення розмірності. Це звичайно найбільш перспективно та регулярно застосовується для проблем з великою кількістю даних та / або високої розмірності вхідних просторів.

Далі аналіз головних компонент вводиться як інструмент для перетворення координатних осей та зменшення розмірності.

1.3 Методи кластеризації

Кластер можна визначити як групу даних, які є більш схожими один на одного, ніж дані, що належать іншим кластерам [28]. На рисунку 1.5 показані чотири приклади [29]. Користувач повинен визначити, які типи кластерів буде шукати шляхом визначення міри подібності. Найпоширеніша форма кластерів це коло або (у вищих розмірностях) сфера, відповідно. Потім мірою подібності може бути відстань всіх зразків даних в межах кластера від центру кластера. У цьому випадку центр кластеру представляє кластер, та таким чином його називають прототипом кластера. Для інших прототипів кластерів міри подібності можуть бути різними, наприклад, 1.5 b або центр окружності/еліпсів та його радіус на рисунку 1.5 c та d.

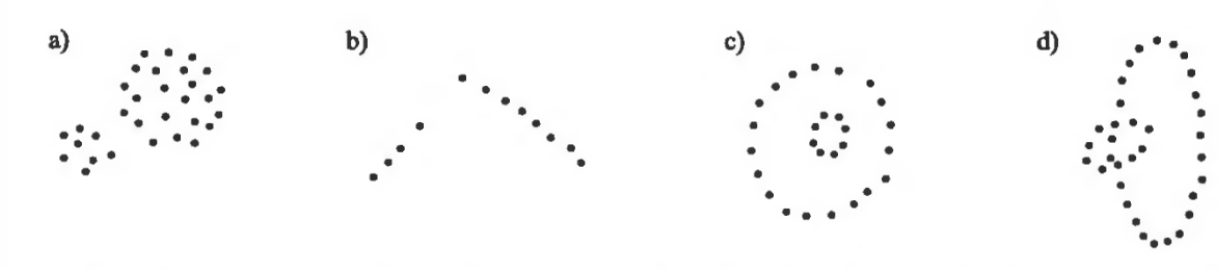

Рисунок 1.5 – Приклади різних форм кластерів: a) заповнені кола, b) лінії, c) порожні кола, d) порожні еліпси

Користувач визначає, як будуть виглядати кластери, та обирає відповідний алгоритм кластеризації для цього завдання. Методи кластеризації вимагають, щоб користувач обирав кількість кластерів апріорі. Більш просунуті методи можуть автоматично визначати кількість кластерів в залежності від міри деталізації наданої користувачем.

Класичні методи кластеризації, такі як К-середні [30], відносять кожен зразок даних одному кластеру (жорсткий або чіткий розділ). Сучасні методи кластеризації генерують нечіткий розділ. Це означає, що кожен зразок даних призначається кожному кластеру з певним ступенем належності $\mu(u)$. Для кожного зразка даних всі функції належності у сумі дають 1. Наприклад, один зразок даних u_i може бути призначений кластеру 1 з $\mu(u(i))=0.7$, для кластера 2 з $\mu(u(i))=0.2$, для кластера 3 з $\mu(u(i))=0.1$, а до всіх інших кластерів з $\mu(u(i))=0$.

Оскільки функції втрат, мінімізовані методами кластеризації, зазвичай є нелінійними, алгоритми працюють ітеративно, починаючи з початково вибраних кластерів. Загалом, збіжність до глобального оптимуму не може бути гарантована. Чутливість до початкових значень залежить від конкретного алгоритму. Якщо початкові кластери обираються обґрунтовано за попередніми знаннями, а не випадково, зазвичай можна уникнути збіжності до локальних оптимумів.

Іншим важливим питанням, пов'язаним з вибором міри подібності, є нормалізація даних. Більшість алгоритмів кластеризації дуже чутливі до масштабу даних. Наприклад, метод кластеризації, який шукає заповнені кола в даних, сильно залежить від масштабування осей, оскільки це змінює кола на еліпси. Ненормалізовані дані, наприклад, два входи в діапазонах $0<u_1<1$ та $0<u_2<1000$, можуть привести до погіршення роботи метода кластеризації, оскільки для обчислення відстаней величина u_1 майже не має значення в порівнянні з u_2. Таким чином, дані повинні бути нормалізовані або стандартизовані перед застосуванням методів кластеризації. Винятком з цього правила є методи з адаптивними властивостями, які автоматично масштабують дані, такі як алгоритм Густафсона-

Кесселя. Але навіть для цих алгоритмів нормалізовані дані дають чисельно кращі результати.

Методи кластеризації можна розрізнити за наступними властивостями:

– тип змінних, до яких вони можуть застосовуватися (безперервне, ціле, двійкове);

– показник належності;

– ієрархічна або неієрархічна;

– фіксоване або самоадаптивне число кластерів;

– жорсткі або нечіткі розподіли.

1.4 Метод К-середніх

В [31] [32] [33] автори описали, що алгоритми кластеризації поділяють набір даних на багато груп, які мають на меті розбити вхідний набір даних до кінцевої кількості груп за схожими параметрами. Ці алгоритми кластеризації можуть використовувати як нормалізовані, так і ненормалізовані дані. Якщо користувачі мають нормалізовані дані, то кількість ітерацій алгоритмів буде меншою. Тому більшість нормалізованих даних дає вищій результат порівняно з ненормалізованими даними. Серед цих багатьох алгоритмів кластеризації на основі щільності розподілення даних є найбільш популярним алгоритмом інтелектуального аналізу даних.

Алгоритми кластеризації також використовують формули відстані. Коли дані високої розмірності [34], то використовують метрику Мінковського

$$D_p\left(x_i, x_j\right) = \left(\sum_{k=1}^{d} \left| x_{i,k} - x_{j,k} \right|^p \right)^{\frac{1}{p}}, \tag{1.1}$$

де D – розмірність даних.

У випадку евклідової відстані, значення $p = 2$, та Манхеттенська відстань при значення $p = 1$.

Деякі алгоритми кластеризації працюють на нормалізованих даних, таких як розподілена кластеризація методом K-середніх [35]. Нормалізація даних є способом лінійного перетворення даних в єдиний діапазон.

Існує кілька підходів до нормалізації [36], розглянуті найбільш поширені методи Min-Max Normalization, нормалізація даних за допомогою Decimal Scaling та Z-score data Normalization. Нормалізація Min-Max виконує лінійне перетворення на вихідних даних. В [34] автори припускали, що ми маємо атрибут A та Max_a, Min_a – максимальні та мінімальні значення цього атрибута. Нормалізація Min-Max відображає значення в діапазоні (0, 1) шляхом обчислення

$$v' = \frac{v - \min_a}{\left(\max_a - \min_a\right)}. \tag{1.2}$$

При нормалізації Z-score значення атрибута (атрибута A) нормалізуються на основі середнього значення та стандартного відхилення атрибута (A). Значення атрибута A нормується на v шляхом обчислення:

$$v' = \frac{v - \bar{A}}{\sigma_A}, \tag{1.3}$$

де $\bar{A}$ – середнє значення,

σ_A – стандартне відхилення.

Цей метод ефективно працює в двох випадках: коли фактичне мінімальне значення та максимальне значення атрибута (A) невідомо та коли є шум, який диктує нормалізацію даних min-max.

У разі нормування за допомогою Decimal Scaling атрибута (припустимо, що A) нормалізується до V ', обчислюючи:

$$v' = \frac{v}{10j}, \tag{1.4}$$

де j – мале ціле число.

В області інтелектуального аналізу даних використовуються різні підходи до кластеризації. Але кожен метод кластеризації має певні переваги та недоліки. Кожна методика кластеризації не підходить для всіх умов [37] [38].

Метод кластеризації К-середніх. Алгоритм К-середніх, є найбільш поширеним та простим методом кластеризації. Його можна розглядати як основу для більш просунутих підходів. У назві кластеризації К-середніх, також відомому як кластеризація С-середніх [39], "К" або "С" означає фіксовану кількість кластерів, що задається користувачем апріорі. Кластеризація К-середніх мінімізує функцію втрат:

$$I = \sum_{j=1}^{C} \sum_{i \in S_i} \left\| x(i) - w_j \right\|^2 \to \min_{w_j}, \tag{1.5}$$

де індекс i проходить по усім елементам множини S_j, C – кількість кластерів, а w_j – центри кластерів (прототипи). Набори S_j містять всі індекси тих зразків даних (з усіх N), які належать до кластера j, тобто, які є найближчими до центру кластера c_j. Центри кластерів c_j є параметрами, які змінюють метод кластеризації для мінімізації (1.5).

Тому функція втрат (1.5) підсумовує всі квадратичні відстані від кожного центру кластера до зв'язаних з ними даних. Його також можна записати як

$$I = \sum_{j=1}^{C} \sum_{i=1}^{N} u_{ji} \left\| x(i) - w_j \right\|^2, \tag{1.6}$$

де $u_{ji}=1$, якщо спостереження $x(i)$ асоціюється (належить) кластеру j та $u_{ji}=0$ інакше.

Алгоритм К-середніх для мінімізації (1.6) працює наступним чином [39]:

1. Обрати початкові значення C для центрів w_j, $j=1,...,C$. Це можна зробити шляхом вибору випадкових C;

2. Віднести всі зразки даних до найближчого центру кластера;

3. Обчислити центроїд (середнє значення) кожного кластера. Встановити кожен центр кластера на центроїд кластера, тобто

$$w_j=\frac{\sum_{i\in S_j} x(i)}{N_j}, \tag{1.7}$$

де i проходить по спостереженням N_j, які належать до кластера j, тобто знаходяться в наборі S_j, а N_j – число елементів у множині S_j $\left(\sum_{j=1}^{C} N_j = N\right)$.

4. Якщо будь-який центр кластера переміщено на попередньому кроці, перейти до кроку 2; інакше зупиниться.

Рисунок 1.6 ілюструє поведінку збіжності К-середніх з $C=3$ кластерами. На рисунку 1.6 a показаний двовимірний набір даних. Рисунок 1.6 b зображує три центри кластера для п'яти ітерацій, необхідних для збіжності. Початкові центри кластерів були вибрані випадковим чином з набору даних, метод К-середніх швидко збігаються кінцевих центрів кластера.

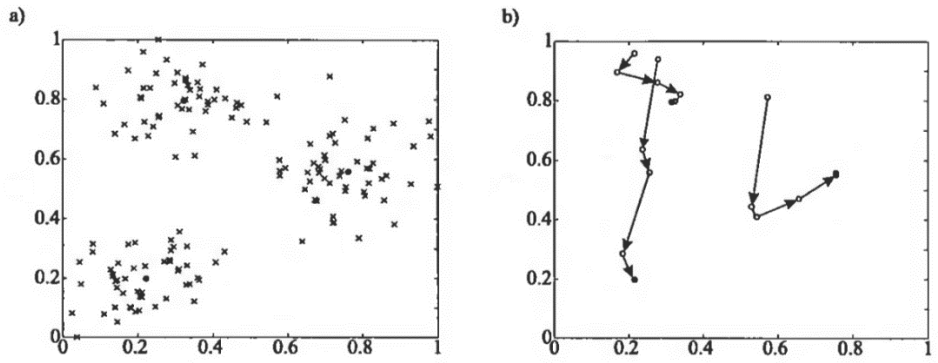

Рисунок 1.6 – a) кластеризація даних методом К-середніх призводить b) до збіжності в п'яти ітераціях.

Рисунок 1.7 ілюструє важливість нормалізації даних. На рисунку 1.7 a показано, які зразки даних належать до якого кластеру для нормалізованих даних з рис. 1.6 a. Рисунок 1.7 b показує кластери для ненормалізованого набору даних, де x_1 лежить між 0 та 100, а x_2 лежить між 0 та 1. Відстань у (1.6) домінує у x_1-розмірності, а межі кластера залежать майже виключно від x_1.

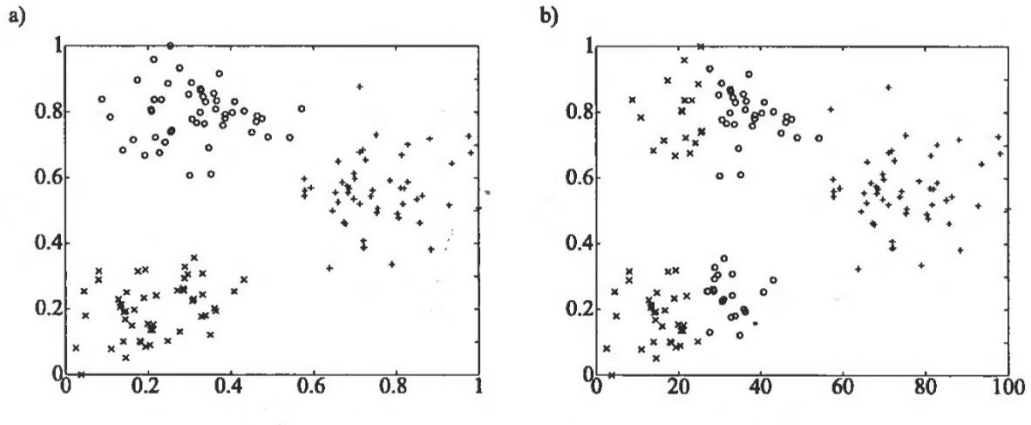

Рисунок 1.7 – a) порівняння нормалізованих даних та b) ненормалізованих даних та вплив нормалізації даних на кластеризацію методом К-середніх

Альтернативою нормалізації даних є зміна метрики відстані, що використовується в (1.6). Квадратична евклідова норма

$$D_{ij}^2 = \left\| x(i) - w_j \right\|^2 = \left(x(i) - w_j \right)^T \left(x(i) - w_j \right) \tag{1.8}$$

може поширюватися на квадратичну загальну норму Махаланобіса

$$D_{ij,\Sigma}^2 = \left\| x(i) - w_j \right\|_\Sigma^2 = \left(x(i) - w_j \right)^T \Sigma \left(x(i) - w_j \right). \tag{1.9}$$

Норма матриці Σ масштабує та обертає осі [40]. Для спеціального випадку, коли коваріаційна матриця дорівнює одиничній матриці $(\Sigma = I)$, норма Махаланобіса дорівнює евклідовій нормі. Для

$$\Sigma = \begin{bmatrix} 1/\sigma_1^2 & 0 & \cdots & 0 \\ 0 & 1/\sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1/\sigma_p^2 \end{bmatrix}, \tag{1.10}$$

де D позначає розмір вхідного простору, норма Махаланобіса дорівнює евклідовій нормі з масштабованими входами $x_l^{(scaled)} = x_l / \sigma_l$. У найбільш загальному випадку, матриця норм масштабує та обертає вхідні осі. Рисунок 1.8 підсумовує ці міри відстані. Зауважимо, що вибір у (1.6) еквівалентний евклідовій нормі з перетвореними вхідними осями. Одне обмеження у використанні метода К-середніх полягає в тому, що обрана норма фіксується для всього вхідного простору і, таким чином, для всіх кластерів. Це обмеження долається алгоритмом Густафсона-Кесселя, який використовує індивідуальні адаптивні міри схожості для кожного кластера.

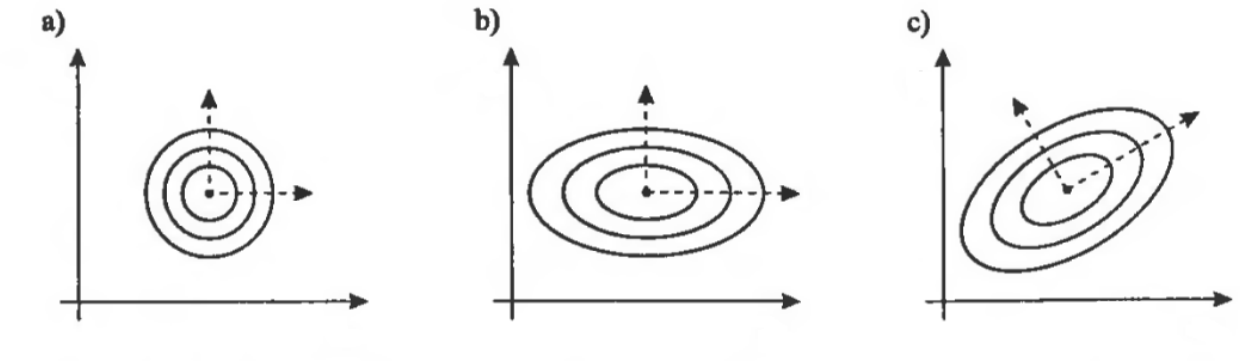

Рисунок 1.8 – Лінії з однаковою відстанню для різних норм: a) евклідова $(\Sigma = I)$, b) діагональна $(\Sigma = diagonal)$, та c) норма Махаланобіса $(\Sigma = general)$

1.5 Алгоритм нечітких C-середніх

Алгоритм нечітких C-середніх (FCM) є нечітким варіантом класичного алгоритму K-середніх [41], описаного вище. Функція мінімізації втрат майже однакова:

$$I = \sum_{j=1}^{C} \sum_{i=1}^{N} \mu_{ji}^{v} \left\| u(i) - w_j \right\|_{\Sigma}^{2} \tag{1.11}$$

з

$$\sum_{j=1}^{C} \mu_{ji} = 1.$$

Таким чином функція належності u_{ji} одного зразка даних $x(i)$ до кластера j не має бути рівного одиниці для одного кластера та нулю для всіх інших. Скоріше, кожен зразок даних може мати деяку ступінь належності між 0 та 1 до кожного кластера за умови, що всі ці функції належності у сумі дорівнюють 1, тобто 100% для кожного зразка даних [42]. Функція належності зводиться до ступеня v, що визначає розмитість кластерів. Цей ваговий показник v лежить в інтервалі

$(1,\infty)$, та частіше за все це значення дорівнює $v=2$. Якщо очікуються, що кластери легко відокремлювані, то значення для v, краще приймати близьким до 1, у такому випадку розмитість кластерів буде меншою, та ступень належності буде близько до 0 або 1. З іншого боку, якщо очікується, що кластери майже не розрізняються, слід вибрати високі значення для v.

Функція належності u_{ji} одного зразка даних $x(i)$ до кластера j визначається [29] [28] [43]

$$u_{ij}=\frac{1}{\sum_{l=1}^{C}\left(D_{ij,\Sigma}^2 \big/ D_{il,\Sigma}^2\right)^{\frac{1}{v-1}}} \tag{1.12}$$

де

$$D_{ij,\Sigma}^2=\left\|u(i)-w_j\right\|_{\Sigma}^2=\left(u(i)-w_j\right)^T\Sigma\left(u(i)-w_j\right) \tag{1.13}$$

Рисунок 1.9 ілюструє визначення відстаней для одного зразка даних та двох центрів кластера, а також евклідову міру відстані $\Sigma=I$. Тоді (1.12) стає

$$u_{ij}=\frac{1}{\left(D_{i1}^2 \big/ D_{i1}^2+D_{i1}^2 \big/ D_{i2}^2\right)^{\frac{1}{v-1}}}=\frac{1}{\left(1+D_{i1}^2 \big/ D_{i2}^2\right)^{\frac{1}{v-1}}} \tag{1.14}$$

з відстанями на рис. 1.9.

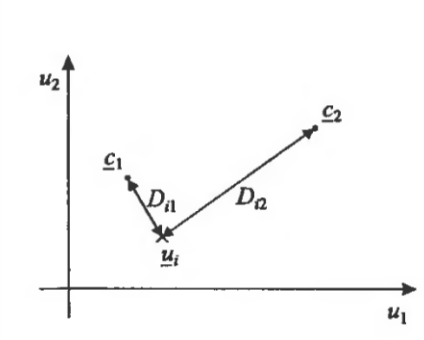

Рисунок 1.9 – Ілюстрація відстаней нечіткого алгоритму С-середніх

Очевидно, що при наближенні зразка даних до центру кластера $\left(D_{ij,\Sigma}^2 \to 0\right)$, ступінь належності до цього кластера наближається до 1 $\left(u_{ij} \to 1\right)$, а якщо $D_{ij,\Sigma}^2 \to \infty$, то $u_{ij} \to 0$. Ясно, що (1.12) автоматично виконує обмеження, що сума u_{ij} по всіх кластерах дорівнює 1 для кожного зразка даних.

При розрахунку функції належності відповідно до (1.12) необхідно враховувати наступні два особливі випадки:

– якщо в (1.2), (1.12) зразок даних $x(i)$ лежить точно на центрі кластера w_l, що не є кластером j $\left(l \neq j\right)$, то один знаменник у (1.12) $\left(D_{ij,\Sigma}^2\right)$ стає нулем та $u_{ij} = 0$;

– якщо в (1.12), (1.13), то зразок даних $x(i)$ лежить точно на центрі кластера w_j, то u_{ij} можна вибрати довільно, якщо виконано обмеження $\sum_{j=1}^{C} u_{ij} = 1.$

Алгоритм нечітких С-середніх, який мінімізує (1.11), працює наступним чином [29] [28] [43]:

1. Обрати початкові значення C для центрів кластерів w_j, $j = 1,...,\mathrm{C}$. Це може бути зроблено шляхом вибору випадкових C.

2. Розрахувати відстані $D_{ij,\Sigma}^2$ всі зразків даних $x(i)$ до кожного центру кластера w_j відповідно до (1.13).

3. Обчислити функцію належності для кожного зразка даних $x(i)$ до кожного кластера j відповідно до (1.12).

4. Знайти центроїд (середнє) кожного кластера. Використовуючи формулу (1.15)

$$w_j = \frac{\sum_{i=1}^{N} u_{ij}^{v} x_i}{\sum_{i=1}^{N} u_{ij}^{v}}. \qquad (1.15)$$

5. Якщо будь-який центр кластера був значно переміщений, скажімо, більше, ніж на ε, на кроці 4, то необхідно перейти до кроку 3, інакше зупинитися.

Подібно алгоритму K-середніх, ця фаззі-версія шукає заповнені кола$(\Sigma = I)$, осі-ортогональні еліпси $(\Sigma = diagonal)$ або довільно орієнтовані еліпси $(\Sigma = general)$. Форма кластера має фіксуватися користувачем апріорно. Вона не може бути адаптована до даних та не може бути різною для кожного окремого кластера.

1.6 Самоорганізовна карта Кохонена

Самоорганізовна карта Кохонена (SOM) [44] є найпопулярнішим нейромережевим підходом до кластеризації. Це розширення техніки векторного квантування (LVQ), яке також було розроблено Т. Кохоненом. Векторне квантування – це, в основному, спрощена версія кластеризації K-середніх при адаптації вибірки даних, тобто вона оновлює параметри не після одного проходу по всьому набору даних (пакетна адаптація) як це робить метод K-середніх, а після кожного вхідного зразка даних. Векторне квантування працює з "нейронами", які відповідають кластерам в K-середніх. Кожен нейрон має p (вхідний розмір) параметри або "ваги", що відповідають p компонентам кожного центрального вектора кластера w. Оскільки різниця між нейронами та кластером є

виключно термінологічною, параметри кожного нейрона також будуть позначені як w.

Алгоритм векторного квантування такий:

1. Обрати початкові значення C для векторів w_j, $j=1,...,C$. Це можна зробити, вибираючи випадково різні зразки даних C.

2. Вибрати один зразок для набору даних. Це можна зробити або випадковим чином, або послідовно, проходячи через весь набір даних (циклічний порядок).

3. Розрахувати відстань вибраного зразка даних до всіх нейронів-векторів. Як правило, використовується евклідова міра відстані. Нейрон з вектором, найближчим до зразка даних, називається нейроном-переможцем.

4. Оновити вектор нейрона-переможця таким чином, щоб перемістити його до обраного зразка даних x:

$$w_{win}^{(new)} = w_{win}^{(old)} + \eta\left(x - w_{win}^{(old)}\right). \tag{1.16}$$

5. Якщо будь-який нейрон-вектор був значно переміщений, скажімо більше, ніж на ε, на попередньому кроці перейти до кроку 2, інакше зупинитися.

На кроці 4 розмір кроку (швидкість навчання) η повинен бути обраний належним чином. Для більш швидкої збіжності рекомендується починати з великого кроку, скажімо 0,5, який зменшується в кожній ітерації алгоритму. Робота векторного квантування по суті є такою же, як і адаптивна версія кластеризації К-середніх [39] [45]. У К-середніх, однак, розмір кроку нормалізується на кількість зразків даних, які належать нейрону-переможцю. Це гарантує, що вектор нейрона-переможця збігається до центроїда (середнього) цих зразків даних.

Самоорганізовна карта Кохонена (SOM) є розширенням описаного вище алгоритму векторного квантування [46] [44] [47] [45]. У SOM нейрони – це не просто абстрактні структури, які представляють центр кластера, скоріше, нейрони організовані в одно-, дво-, а іноді і в більш вимірних топологіях, наведених на

рисунку 1.10. Для більшості застосувань використовується двовимірна топологія з гексагональною або прямокутною структурою (як на рис. 1.10 c).

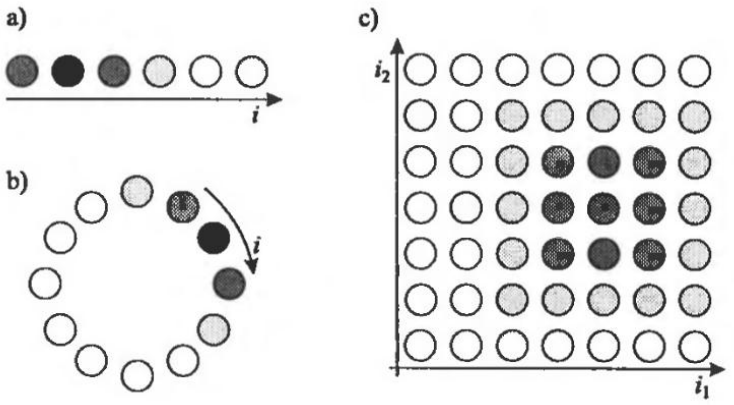

Рисунок 1.10 – Різні топології самоорганізовних карт: a) лінійна, b) кругова, c) двовимірна сітка

Ідея самоорганізовної карти полягає в тому, що нейрони, які є сусідами по топології мережі, повинні мати подібні вагові вектори (центри кластерів). Таким чином, відстань центрів різних кластерів в p-вимірному просторі представлена в меншому (зазвичай двовимірному) просторі. Звичайно, ця проекція високої розмірності на низьковимірний простір не може бути виконана бездоганно та включає в себе стиснення інформації [48]. Двовимірний SOM є відмінним інструментом для візуалізації розподілів даних високої розмірності. Введена функція сусідства, яка визначає активність тих нейронів, які є сусідами нейрона-переможця. На відміну від LVQ, не тільки нейрон-переможець оновлюється як у (1.16), але і його сусіди. Функція сусідства $\varphi(i)$ зазвичай дорівнює 1 для нейрона-переможця та зменшується зі зростанням відстані нейронів від переможця. Функція сусідства визначена в топології SOM. Наприклад, SOM на рис. 1.10 a та b мають одновимірні функції сусідства, наприклад,

$$\varphi(i) = \exp\left(-\frac{1}{2} \frac{\left(i^{(win)} - i \right)^2}{\sigma^2} \right), \tag{1.17}$$

де $i^{(win)}$ позначає індекс переможця-нейрона, а i позначає індекс будь-якого нейрона. SOM на рис. 1.10 с має двовимірну функцію сусідства, наприклад,

$$\varphi(i_1, i_2) = \exp\left(-\frac{1}{2} \frac{\left(i_1^{(win)} - i_1 \right)^2 + \left(i_2^{(win)} - i_2 \right)^2}{\sigma^2} \right), \tag{1.18}$$

де $i_1^{(win)}$ та $i_2^{(win)}$ позначають індекси нейрона-переможця, а i_1 та i_2 позначають індекси будь-якого нейрона. Поки функція сусідства має локальний характер, її точна форма не є вирішальною. Для алгоритму навчання SOM (1.16) на кроці 4 алгоритму LVQ розширений до

$$w_j^{(new)} = w_j^{(old)} + \eta \varphi(i) \left(x(i) - w_j^{(old)} \right), \tag{1.19}$$

де $\varphi(i)$ – функція сусідства розмірності топології SOM. Зауважимо, що (1.19) оцінюється для всіх активних нейронів $j = 1,...,C$, а не просто нейрона-переможця. В (1.19) ціла група сусідніх нейронів переміщується до вхідного зразка даних. Чим ближче до переможця є нейрон, тим більшим є $\varphi(i)$ отже, і розмір кроку.

Щоб проілюструвати вплив функції сусідства, доцільно розглянути два випадки надзвичайно гострої ($\sigma \to 0$ в (1.17) або (1.18)) та широкої ($\sigma \to \infty$) функції сусідства. У першому випадку мережа SOM зводиться до LVQ та не створюється зв'язок між сусідніми нейронами. У другому випадку всі нейрони ідентично визначають центроїди для всього набору даних. Алгоритм навчання SOM починається з широкої функції сусідства та зменшує його в кожній ітерації.

Завдяки цій стратегії в перших ітераціях мережа вивчає грубе представлення розподілу даних та уточнює її, оскільки функція сусідства стає все більш локальною. Це така ж стратегія, що й для розміру кроку η.

1.7 Висновки та постановка основних завдань дослідження

Як показує проведений аналіз, у рамках інтелектуального аналізу даних існує безліч методів для кластеризації даних, які відрізняються один від одного як математичним апаратом, так і результатами обробки інформації. Але більшість цих методів вимагає заздалегідь визначити кількість кластерів та щоб ці кластери мали опуклу або округлу форму, та є лінійно роздільними. Але у реальних задачах дотримуватись всіх цих вимог важко, оскільки кластери можуть буди довільної форми або бути лінійно нероздільними. Для вирішення даних задач існують методи, які засновані на ядерному підході для кластеризації даних [49] [50]. В рамках цього підходу передбачається, що кластери мають довільну форму, але вони не перетинаються у просторі ознак та вся вибірка даних задається заздалегідь, тобто працюють ці підходи у пакетному режимі. На сьогоднішній день існує безліч методів для кластеризації даних, але всі ці методи чіткі, тобто вони передбачають, що кластери є лінійно роздільними, та працюють у пакетному режимі. У реальних задачах кластери можуть перетинатись, а кожне спостереження може належати декільком кластерам з відповідною належністю. Для вирішення такої задачі існують методи нечіткої кластеризації [28] [51], проте сформовані кластери мають округлу форму. У ситуаціях коли кластери перетинаються та мають опуклу форму виникає необхідність розробки методів нечіткої кластеризації для потоків даних. Існує ще одна проблема, що кількість кластерів нам не відома, особливо, коли дані обробляються в онлайн режимі. Саме тому стає доцільним розробити метод, який обробляє інформацію у ситуаціях, коли нам невідома кількість класів.

Таким чином, на сьогоднішній день є актуальною задача розробки нечітких методів для кластеризації потоків даних великої розмірності, коли кількість кластерів заздалегідь невідома, та вони можуть перетинатися у просторі ознак та

мати довільну форму. Використання запропонованого ансамблевого підходу вирішує проблему заздалегідь невідомої кількості кластерів, а введені ядерні функції дозволяють підвищити простір ознак, що дозволяє обробляти дані у ситуації коли вони перетинаються.

Відповідно до поставленої мети у дисертаційній роботі необхідно вирішити такі завдання:

– аналіз існуючих методів та підходів для кластеризації потоків даних;

– розробка модифікації метода К-середніх для кластеризації даних, які послідовно надходять на обробку спостереження за спостереженням;

– розробка ансамблю нейронних мереж, який об’єднує у собі ідеї ядерних мереж та самонавчання;

– розробка архітектури ансамблю нейронних мереж на основі самоорганізовних карт Т. Кохонена;

– розробка ансамблю нейро-фаззі систем для кластеризації потоку даних за припущенням, що кількість та форма кластерів невідомі заздалегідь;

– розробка ансамблю нейро-фаззі мереж на основі імовірнісно-можливісного підходу для кластеризації потоків даних;

– імітаційне моделювання розроблених методів та моделей та рішення практичних задач нечіткої кластеризації потоків даних високої розмірності.

2 АНСАМБЛІ САМООРГАНІЗОВНИХ КАРТ КОХОНЕНА ДЛЯ КЛАСТЕРИЗАЦІІ ДАНИХ

Завдання кластеризації масивів даних є важливою частиною загальної проблеми Data Mining, а для її вирішення на сьогодні розроблено безліч різних методів [52] [50] [53] [54] [55]. При обробці великих обсягів інформації на перший план виходять вимоги по швидкодії і простоті чисельної реалізації використовуваних алгоритмів кластеризації. Одним з найбільш популярних алгоритмів є метод К-середніх, завдяки своїй простоті, наочності результатів і можливості їх ясної інтерпретації. Цей метод відноситься до алгоритмів, заснованих на обчисленні прототипів-центроїдів, в результаті чого масив вихідних даних $X = \{x(1), x(2), ..., x(k), ...x(N)\} \subset R^n$, $x(k) = ((x_1(k), ..., x_i(k), ..., x_u(k))^T$, $k = 1, 2...N$ розбивається на m кластерів, де їх кількість m задається апріорно або вибирається, як правило, виходячи з суто емпіричних міркувань.

Для формального знаходження числа кластерів m був розроблений метод Х-середніх [56] [57], заснований на статистичному аналізі розподілу даних у вихідному масиві X. Якщо при роботі з К-середніми число кластерів m було вибрано правильно, то отримані результати повністю збігаються з результатами Х-середніх.

Останні роки в зв'язку з інтенсивним розвитком Data Stream Mining [58] [59] природно виникла необхідність вирішення завдань кластеризації в online режимі, коли дані на обробку послідовно надходять спостереження за спостереженням, обсяг масиву N не обмежений і зростає з часом, а k набуває сенсу поточного дискретного часу. У подібній ситуації стандартні К-середні неефективні, проте з успіхом можуть бути використані кластерувальні нейронні мережі Т. Кохонена (SOM) [44] [60], що вирішують завдання в online режимі, а одержаний результат повністю збігається з К-середніми в силу використання загального критерію кластеризації-самонавчання, заснованого на евклідовій метриці. При цьому

проблема вибору m тут залишається відкритою, включення додаткових «мертвих» нейронів в мережу, як правило, її не вирішує, а використання X-середніх в online режимі в їх традиційній формі принципово неможливо.

Альтернативою стандартним X-середнім може бути використання ідеї кластерувальних ансамблів [61] [62] [63] [64], при цьому нами пропонується формувати ансамбль на основі паралельно з'єднаних входами SOM^m, кожна з яких апріорно орієнтована на різну кількість можливих кластерів $m = 2,3,...,M.$ Таким чином, перша кластерувальна мережа ансамблю працює в припущенні $m = 2$, тобто в шарі Кохонена містить всього два нейрони з синаптичними вагами-центроїдами w_1^2 і w_2^2. Другий елемент ансамблю містить три нейрони з векторами синаптичних ваг w_1^3, w_2^3, w_3^3 і, нарешті, остання SOM^M ансамблю працює в припущенні, що число можливих кластерів дорівнює M, тобто містить M нейронів - адаптивних лінійних асоціаторів.

2.1 Підготовка даних для навчання

При підготовці даних для навчання нейронної мережі необхідно звертати увагу на такі суттєві моменти.

Кількість спостережень в наборі даних. Слід враховувати той фактор, що чим більше розмірність даних, тим більше часу буде потрібно для навчання мережі.

Робота з викидами. Слід визначити наявність викидів та оцінити необхідність їх присутності в вибірці.

Навчальна вибірка повинна бути представницькою (репрезентативною).

Навчальна вибірка не повинна містити протиріч, оскільки нейронна мережа однозначно зіставляє вихідні значення вхідним.

Нейронна мережа працює тільки з числовими вхідними даними, тому важливим етапом при підготовці даних є перетворення та кодування даних.

Дані на вхід нейронної мережі слід подавати з того діапазону, на якому вона навчалася. Наприклад, якщо при навчанні нейронної мережі на один з її входів

подавалися значення від 0 до 10, то при її застосуванні на вхід слід подавати значення з цього ж діапазону.

Існує поняття нормалізації даних. Метою нормалізації значень є перетворення даних до виду, який найбільш підходить для обробки, тобто дані, що надходять на вхід, повинні мати числовий тип, а їх значення повинні бути розподілені в певному діапазоні. Нормализатор може призводити дискретні дані до набору індексів або перетворювати значення, що лежать в довільному діапазоні, в конкретний діапазон, наприклад, [0..1]. Нормалізація виконується шляхом ділення кожної компоненти вхідного вектора на норму вектора, що перетворює вхідний вектор в одиничний.

2.2 Налаштування нейронних мереж ансамблю

Для навчання кожної з окремих SOM^m можуть бути використані як стандартні кохоненівські WTA- і WTM-правила самонавчання [44], так і їх модифікації з використанням функції сусідства спеціального виду [65] [66] [67].

Розглянемо процес самонавчання m-ї мережі Кохонена SOM^m, що містить m нейронів з синаптичними вагами $\left\{w_1^m, w_2^m, ...w_m^m\right\} \subset R^n$. В основі алгоритму налаштування синаптичних ваг полягає принцип конкурентного самонавчання, який реалізується в три основних етапи (конкуренція, кооперація, синаптична адаптація) і починається з аналізу вхідного вектора-образу $x(k)$, що надходить з рецепторного (нульового) шару на всі нейрони шару Кохонена. Для кожного з нейронів обчислюється відстань

$$D\left(x(k), w_j^m(k-1)\right) = \left\| x(k) - w_j^m(k-1) \right\|,$$

де $j = 1,2,...m$, при цьому, якщо вхідні сигнали попередньо пронормовані за допомогою перетворення

$$x(k)=\frac{x(k)}{\|x(k)\|} \tag{2.1}$$

так, що $\|x(k)\|=1$, а в якості відстані використовується евклідова метрика, то мірою схожості (подібності) векторів $x(k), w_j^m(k-1)$ може служити скалярний добуток

$$sim\left(x(k), w_j^m(k-1)\right)=x^T(k)w_j^m(k-1)=\cos\left(x(k), w_j^m(k-1)\right). \tag{2.2}$$

Далі визначається нейрон-переможець «найближчий» до вхідного образу такий, що

$$sim\left(x(k), w^{m*}(k-1)\right)=\max_j sim\left(x(k), w_j^m(k-1)\right),$$

після чого, опускаючи тимчасово процес кооперації, можна уточнити синаптичні ваги переможця за допомогою рекурентного співвідношення

$$w_j^m(k)=\begin{cases} w_j^m(k-1)+\eta(k)\times \\ \times\left(x(k)-w_j^m(k-1)\right), \text{ якщо } w_j^m(k-1)=w^{m*}(k-1), \\ w_j^m(k-1) \text{ у протилежному випадку.} \end{cases} \tag{2.3}$$

Таким чином, процедура реалізує правило «переможець отримує все» (WTA), при цьому вектор синаптичних ваг переможця $w^{m*}(k-1)$ «підтягується» до вхідного образу на відстань, що визначається величиною кроку

$$0<\eta(k)<1.$$

Регулювання кроку $\eta(k)$ зазвичай проводиться, виходячи з емпіричних міркувань [44] [68] [69] [70] [71] [72], а загальна рекомендація полягає в тому, що він повинен монотонно зменшуватися в процесі самонавчання [73]. У найпростішому випадку для регулювання кроку можуть бути використані співвідношення

$$\begin{cases} \eta(k) = r^{-1}(k),\ r(k) = \alpha r(k-1) + \left\| x(k) \right\|^2,\ 0 \le \alpha \le 1 \\ r(k) = \alpha r(k-1) + 1,\ 0 \le \alpha \le 1 \end{cases} \tag{2.4}$$

для входів, нормованих відповідно до (2.1).

Зрозуміло, що при $\alpha = 1,\ \eta(k) = k^{-1}$, тобто задовольняє умовам стохастичної апроксимації.

Важливою особливістю нейронної мережі Кохонена є наявність етапу кооперації, коли нейрон-переможець $w^{m^*}(k-1)$ визначає локальну область топологічного сусідства, в якій збуджується не тільки він сам, але і його оточення, при цьому більш «схожі» на переможця нейрони збуджуються сильніше ніж більш віддалені «сусіди». Ця область описується функцією сусідства $\varphi(j,l),\ l = 1,2,...,m$, що залежить від відстані $D\left(w^{m^*}(k-1), w_l^m(k-1)\right) = D\left(w_j^m(k-1), w_l^m(k-1)\right)$, між переможцем і будь-яким з нейронів $w_l^m(k-1)$ шару Кохонена. Як правило $\varphi(j,l)$ – це ядерна функція симетрична щодо максимуму в точці з $D\left(w_j^m(k-1), w_j^m(k-1)\right)$ і приймаюча в ній одиничне значення $\varphi(j,l) = 1$. Зі збільшенням відстані $D\left(w_j^m(k-1), w_l^m(k-1)\right)$ ця функція монотонно зменшується. У переважній більшості випадків в якості функції сусідства використовується гауссіан [74], конус (трикутник) [68], параболоїд (перевернута квадратична функція) [75], «мексиканський капелюх» [72] і цілий ряд інших [66].

Аналіз збіжності процесів конкурентного самонавчання, проведений М. Котрелем і Дж. Фортом [76] [68], показав, що в процесі налаштування синаптичних ваг, повинен зменшуватися не тільки крок пошуку, але і параметр ширини функції сусідства, яка таким чином стає залежною від поточного часу.

Для гаусівської функції

$$\varphi(j,l)=\exp\left(-\frac{\left\|w_l^m(k-1)-w^{m*}(k-1)\right\|^2}{2\sigma^2}\right).$$

Г. Ріттером та К. Шультеном [77] [78] було запропоновано для налаштування параметра ширини використовувати процедуру

$$\sigma(k)=\sigma(0)\exp\left(-\frac{k}{\beta}\right),$$

де $\beta>0$ скалярний параметр, що визначає швидкість зменшення сили впливу нейрона переможця на своє оточення.

Використання функції сусідства призводить до алгоритму самонавчання

$$w_l^m(k)=w_l^m(k-1)+\eta(k)\varphi(j,l)\left(x(k)-w_l^m(k-1)\right)\forall l=1,2,...,m, \qquad (2.5)$$

що реалізує правило «переможець отримує більше» (WTM), при цьому при $l=j$ цей алгоритм збігається зі співвідношенням (2.3).

В принципі, можна взагалі відмовитися від етапу конкуренції та визначення переможця як такого. При цьому в ролі переможця в даному випадку виступає сам вхідний вектор-образ, а в якості функції сусідства використовується міра схожості (2.2).

При цьому алгоритм самонавчання *m*-го елемента ансамблю набуває вигляду

$$w_l^m(k) = w_l^m(k-1) + \eta(k)\left[\cos\left(x(k), w_l^m(k-1)\right)\right]_+ \left(x(k) - w_l^m(k-1)\right) =$$
$$= w_l^m(k-1) + \eta(k)\left[x^T(k) w_l^m(k-1)\right]_+ \left(x(k) - w_l^m(k-1)\right) = \quad (2.6)$$
$$= w_l^m(k-1) + \eta(k)\left[y_l^m(k)\right]_+ \left(x(k) - w_l^m(k-1)\right),$$

де $\left[y_l^m(k)\right]_+ = \max\left\{y_l^m(k), 0\right\}$ – невід'ємне значення *l*-го вихідного сигналу *m*-ої карти Кохонена ансамблю.

Зрозуміло, що процедура (2.6) з обчислювальної точки зору набагато простіша стандартних алгоритмів (2.3), (2.5), завдяки виключенню етапу конкуренції і має ясний фізичний зміст.

2.3 Визначення кількості кластерів

Методи оцінки якості кластерної моделі діляться на зовнішні, внутрішні і відносні. До зовнішніх відносяться метрики, які при оцінці якості використовують яку-небудь вже відому інформацію про структуру кластерів, яка існує в розглянутій множині. Як правило, такі метрики застосовуються при оцінці ефективності роботи алгоритму кластеризації, коли в якості тестової множини використовується будь-яку безліч даних з відомою структурою класів. До внутрішніх відносяться метрики (індекс Ренда, Жаккар, Folkes-Mallows index, F-міра), які при оцінці використовують тільки ту інформацію, яку можна отримати, спираючись на множену даних. Відносні методи (індекс Данна, Девіса-Булдена [79], індекс оцінки силуету, Maulik-Bandyopadhyay index [80], Calinski-Harabasz index) оцінюють якість, порівнюючи кілька кластерних структур між собою, не маючи апріорної інформації і беручи до уваги тільки відомості про кластерні структури та кластеризуючі множини [81]. В процесі роботи ансамблю постійно проводиться оцінка якості кластеризації за допомогою критерію Цаліньського-Харабаша [50] або в його стандартній формі, або за допомогою його online модифікації. При цьому критерій в загальному вигляді має форму

$$CH(m)=\frac{1}{m-1}TrS_B^m\left(\frac{1}{N-m}TrS_w^m\right)^{-1}, \qquad (2.7)$$

де $S_B^m=\frac{1}{N}\sum_{j=1}^{m}N_j^m\left(w_j^m-\overline{w}^m\right)\left(w_j-\overline{w}^m\right)^T$ – матриця міжкластерної відстані для m кластерів;

$\overline{w}^m=\frac{1}{N}\sum_{j=1}^{m}N_j^m w_j^m$ – центр ваги масиву даних X;

N_j^m – кількість спостережень, що відносяться до j-го кластеру, $j=1,2,...m$;

$S_w^m=\frac{1}{N}\sum_{j=1}^{m}\sum_{k=1}^{N}u_j(k)\left(x(k)-w_j^m\right)\left(x(k)-w_j^m\right)^T$ – матриця розсіяння m-го кластеру;

$u_j=\begin{cases}1, \text{ якщо } x(k) \text{ належить j-му кластеру,}\\ 0-\text{у протилежному випадку}\end{cases}$ – чітка функція належності k-го спостереження j-му кластеру.

Переписавши вираз для TrS_B^m у формі

$$TrS_B^m=\frac{1}{N}\sum_{j=1}^{m}N_j^m\left\|w_j^m-\overline{w}^m\right\|^2,$$

а TrS_w^m у формі

$$TrS_w^m=\frac{1}{N}\sum_{j=1}^{m}u_j(k)\left\|x(k)-w_j^m\right\|^2,$$

критерій (2.7) можна представити у вигляді

$$CH(m)=\frac{\frac{1}{m-1}\sum_{j=1}^{m}N_j^m\left\|w_j^m-\overline{w}^m\right\|^2}{\frac{1}{N-m}\sum_{j=1}^{m}\sum_{k=1}^{N}u_j(k)\left\|x(k)-w_j^m\right\|^2} \qquad (2.8)$$

більш зручному з точки зору обчислювальної реалізації.

При аналізі даних, що надходять на обробку в online режимі, розрахунок критерію (2.8) доцільно організувати на ковзному вікні розмірності s ($s = 1,2,...N$), при цьому в поточний момент часу k $CH(m)$ можна записати як

$$CH(m,k) = \frac{\frac{1}{m-1}\sum_{j=1}^{m}\sum_{\tau=k-s+1}^{k} N_j^m(\tau)\left\|w_j^m(\tau) - \overline{w}^m(\tau)\right\|^2}{\frac{1}{N-m}\sum_{j=1}^{m}\sum_{\tau=k-s+1}^{k} u_j(\tau)\left\|x(\tau) - w_j^m(\tau)\right\|^2},$$

де

$$\overline{w}^m(\tau) = \frac{1}{s}\sum_{\tau=k-s+1}^{k} x(\tau).$$

В якості оптимальної кількості кластерів у вибірці m^* приймається m, що забезпечує максимум значенню $CH(m)$, тобто

$$CH(m^*) = \max_m \{CH(2), CH(3), ..., CH(M)\}.$$

Запропонована процедура online кластеризації на основі ансамблю нейронних мереж Т. Кохонена є за суттю адаптивною модифікацією методу Х-середніх, орієнтованою на обробку потоків даних, досить проста в чисельній реалізації і дозволяє вирішити задачу чіткої кластеризації в умовах апріорно невідомого або змінного числа кластерів.

2.4 Експериментальне дослідження ансамблю нейронних мереж на основі карт Кохонена

Для підтвердження працездатності розробленого ансамблю самоорганізовних мап Т. Кохонена була вирішена задача кластеризації на основі штучно згенерованої вибірки і тестових вибірок з UCI-репозіторія. Було взято набори даних:

1. Штучна вибірка RandomMatrix, яка наочно відображає три лінійно розділимих кластери. Вибірка RandomMatrix, яка представлена на рис. 1, містить три лінійно розділимих класи, де кожен елемент вибірки має три випадкові параметри.

2. Тестова вибірка «Iris» [82]. Вибірка складається з даних про квітки ірису, по 50 примірників з трьох видів - Ірис щетиністий (Iris setosa), Ірис віргінський (Iris virginica) та Ірис різнокольоровий (Iris versicolor). Для кожного екземпляра вимірювалися чотири характеристики (в сантиметрах): довжина чашолистка (sepal length); ширина чашолистка (sepal width); довжина пелюстки (petal length); ширина пелюстки (petal width).

Всі дані були спочатку пронормовані на гіперкулю в інтервалі [-1,1] і відцентровані щодо середнього значення.

З метою оцінки ефективності ансамблю самоорганізовних мап Т. Кохонена $\left(SOM^{m}\right)$ результати кластеризації порівнювались зі стандартним методом кластеризації К-середніх. Для підтвердження якості кластеризації був взятий індекс Цаліньського-Харабаша який наведений в табл. 2.1 для вибірок RandomMatrix та Iris. Візуалізація результатів кластеризації наведена на рис. 2.1 – 2.5.

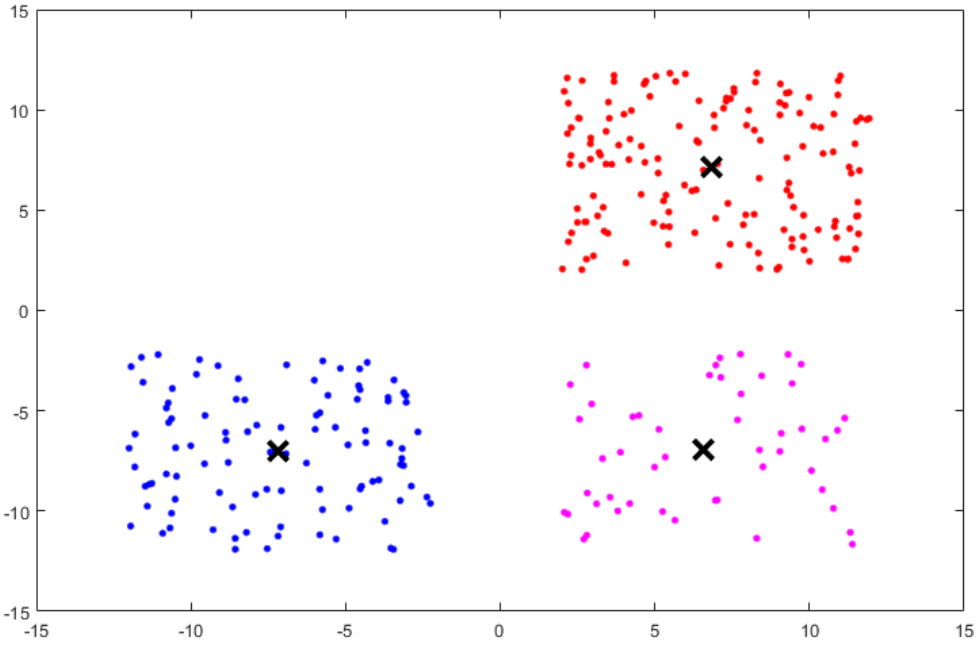

Рисунок 2.1 – Штучно згенерована лінійно розділима вибірка RandomMatrix

Таблиця 2.1 – Індекс Цаліньського-Харабаша для вибірок RandomMatrix та «Іриси Фішера»

Random matrix		
Метод	SOM^m	k-means
індекс CH для 2 кластерів	650,119137400817	49,3904185658740
індекс CH для 3 кластерів	782,603215072022	611,890289394461
індекс CH для 4 кластерів	585,205208331037	411,869958970689
«Іриси Фішера»		
Метод	SOM^m	k-means
індекс CH для 2 кластерів	506,384020879337	24,1478668157212
індекс CH для 3 кластерів	521,993404839107	95,9506726585689
індекс CH для 4 кластерів	463,871189144183	74,4873397342681

Як можна побачити на рисунку 2.2 та 2.3 відображено кластеризацію штучно згенерованої вибірки на три кластери, що було обґрунтовано індексом кластеризації.

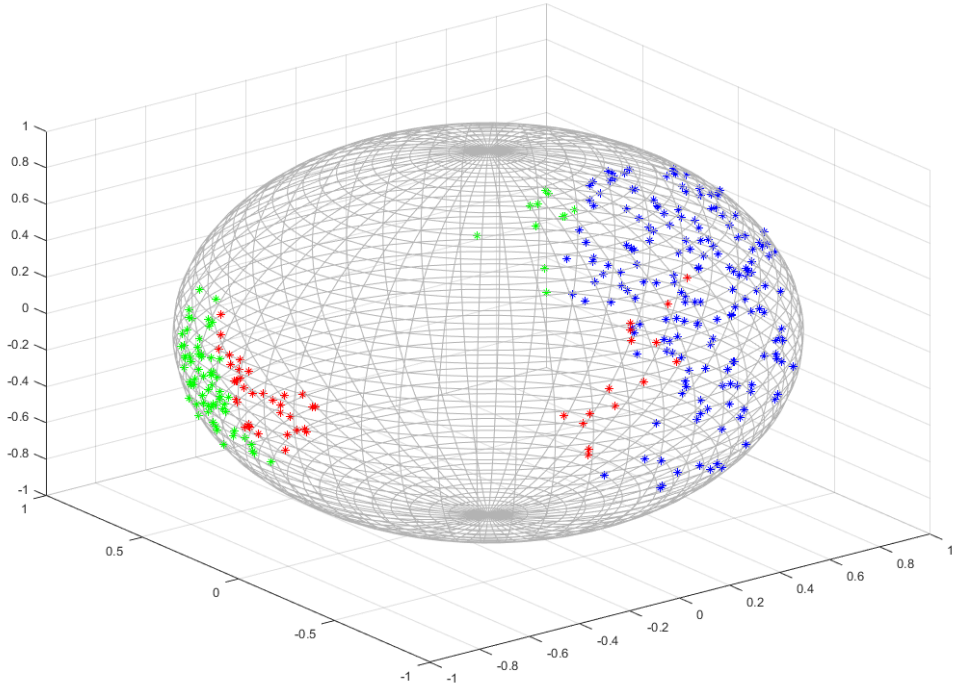

Рисунок 2.2 – Візуалізація вибірки RandomMatrix – вид збоку

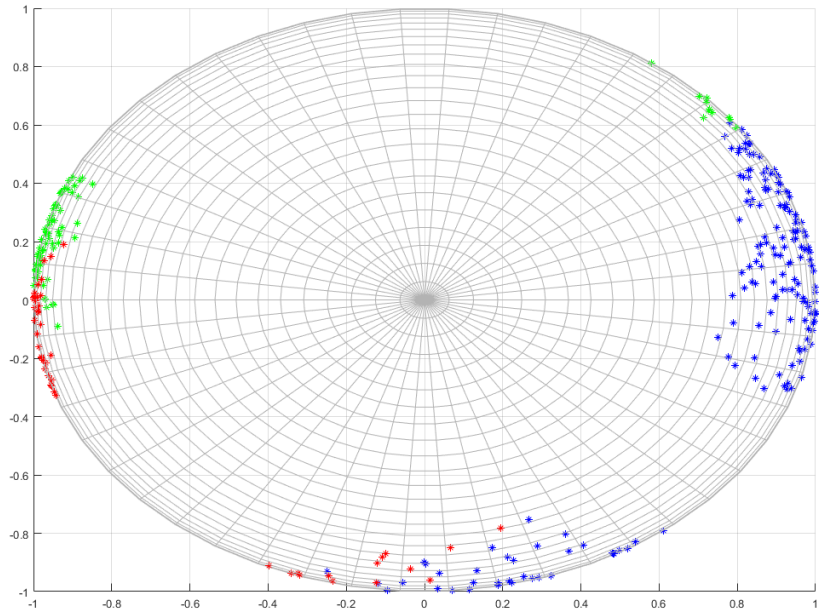

Рисунок 2.3 – Візуалізація вибірки RandomMatrix –вид зверху

На рисунках 2.4 – 2.5 візуалізація набору даних «Іриси Фішера», які було кластеризовано на три кластери, завдяки показникам індексу валідації, який дав максимальний результат при трьох кластерах.

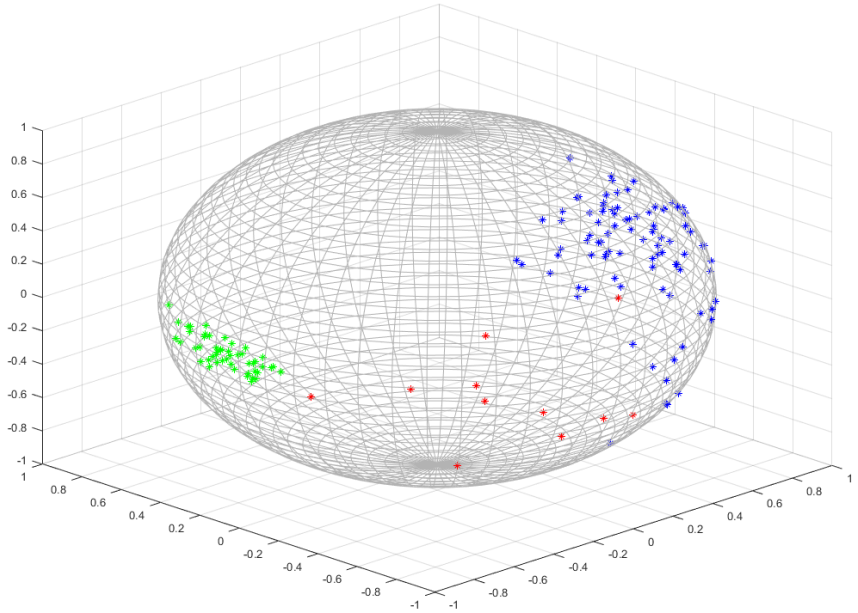

Рисунок 2.4 – Візуалізація вибірки «Іриси Фішера» – вид збоку

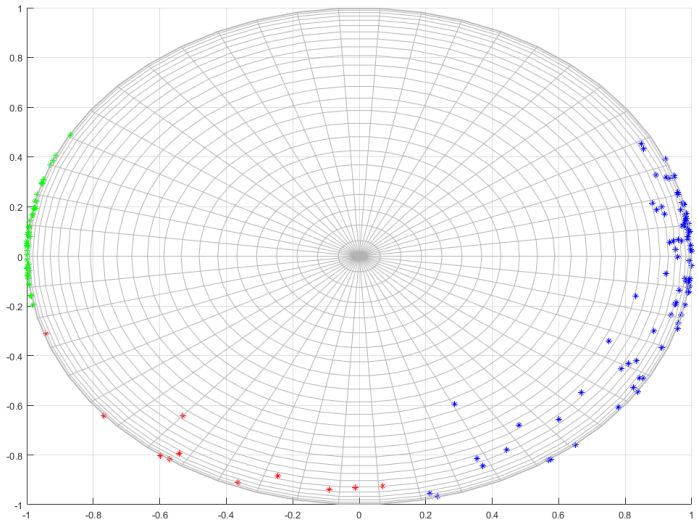

Рисунок 2.5 – Візуалізація вибірки «Іриси Фішера» –вид зверху

2.5 Висновки за розділом

1. Запропоновано ансамбль самоорганізовних карт Т. Кохонена для кластеризації даних за умови апріорно невідомої кількості кластерів. Цей підхід призначений для вирішення задач Data Stream Mining в умовах коли немає даних о кількості класів, але завдяки використанню карт Т. Кохонена даний підхід набагато простіший з обчислювальної точки зору.

2. Розроблено чисельно простий алгоритм кластеризації, заснований на метриці Евкліда, який дозволяє аналізувати потік даних, що послідовно надходять на обробку в on-line режимі.

3. Запропонований підхід базується на паралельно з’єднаних картах Кохонена, кожна з яких налаштована на свою кількість кластерів. Цей підхід дозволяє обробляти дані, які поступають на вхід незалежно від того, на яку кількість класів їх необхідно поділяти. Завдяки індексу валідації обирається та мережа, у якої критерій Цаліньського-Харабаша був максимальним. Тобто значення цього індексу визначає необхідну кількість класів, на яку поділяються вхідні данні.

3 АНСАМБЛІ ЯДЕРНИХ САМООРГАНІЗОВНИХ КАРТ КОХОНЕНА ДЛЯ КЛАСТЕРИЗАЦІЇ ПОТОКІВ ДАНИХ

У розділі 2 розглянуто ансамблевий підхід до кластеризації даних з використанням самоорганізовних карт Т. Кохонена або метода К-середніх. Ансамблі, що засновані на таких підходах, обробляють дані тільки з опуклими кластерами.

В процесі роботи ансамблю всі SOM^m функціонують паралельно, а в якості фінального результату обирається кластерувальна мережа-переможець, яка показала найкращий результат у сенсі застосовуваного критерію якості кластеризації [50] [83].

Відзначимо, що подібно до того, як в кожній з SOM^m на кожному кроці обробки інформації k обирається свій нейрон-переможець, так і в ансамблі на кожному кроці обирається нейронна мережа-переможець, яка забезпечує найкращий результат кластеризації.

Суттєвим обмеженням, що знижує можливості подібного підходу, є вимога лінійної роздільності та опуклості формованих кластерів, в той час як реальні дані можуть утворювати класи абсолютно довільної форми [84]. У подібних ситуаціях дуже корисним може виявитися використання теореми Ковера (T. Cover) про лінійну роздільність в просторах ознак підвищеної розмірності [85]. Нелінійне перетворення складного завдання класифікації образів в простір більш високої розмірності підвищує ймовірність лінійної роздільності образів, та ядер Мерсера (J. Mercer) [86][18], що забезпечують це підвищення.

Теорема Ковера про роздільність образів базується на двох моментах [85]:

1. Визначення нелінійної функції $\varphi_k(x)$, де x – вхідний вектор, а $i=1,2,...,K$, K – розмірність вхідного простору.

2. Висока розмірність вхідного простору в порівнянні з розмірністю вхідного. Ця розмірність визначається значенням, присвоюється K (тобто кількістю прихованих нейронів).

На основі такого підходу були розроблені, так звані, ядерні самоорганізовні мапи (KSOM) [87] [88] [89] [90], які показали гарні результати в умовах класів досить довільної форми при відомій їх кількості m і в умовах фіксованого обсягу оброблюваної вибірки N.

У зв'язку з цим є доцільною розробка ансамблю ядерних кластерувальних нейронних мереж, призначеного для online обробки потоків даних в умовах невідомої або змінної кількості класів.

3.1 Архітектура ансамблю ядерних кластерувальних нейронних мереж

На рис. 3.1 наведена архітектура ансамблю ядерних кластерувальних нейронних мереж, яка містить п'ять шарів обробки інформації.

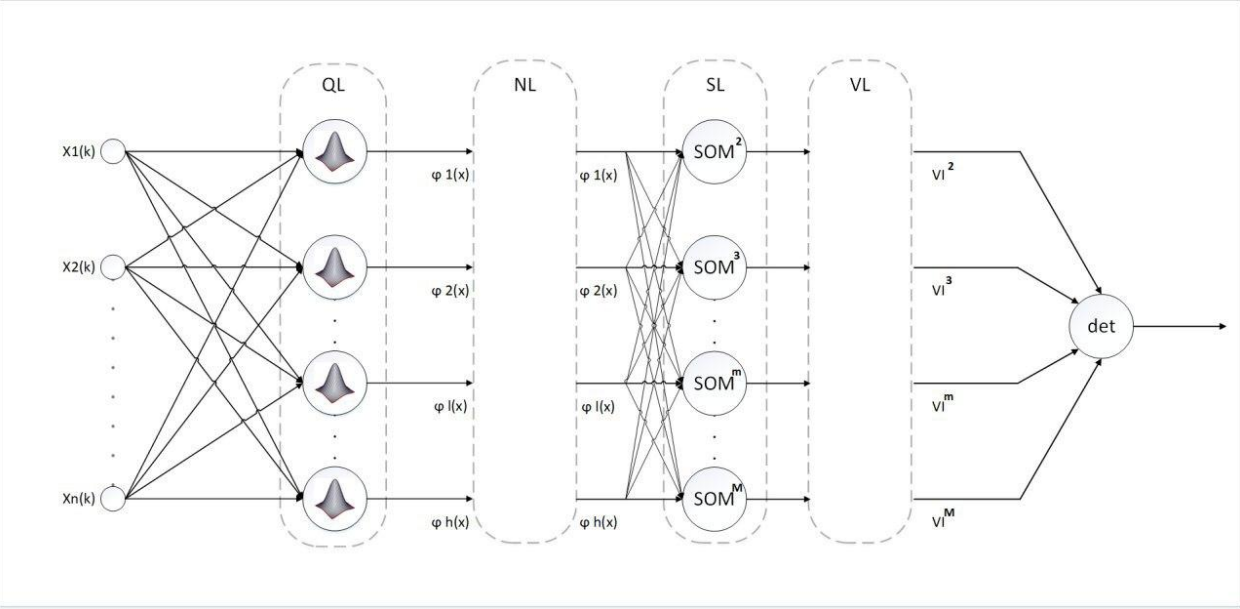

Рисунок 3.1. Архітектура ансамблю ядерних кластерувальних нейронних мереж

Вихідна інформація, яка підлягає кластеризації, подається на нульовий (вхідний) шар системи у вигляді послідовності $x(1), x(2), ..., x(k), ..., x(N), ...$, звідки надходить на перший прихований шар (RL) радіально-базисних функцій, утворений R-нейронами. Саме в цьому шарі відбувається підвищення розмірності вхідного простору за допомогою системи ядерних функцій $\varphi_1(x), \varphi_2(x), ..., \varphi_l(x), ..., \varphi_h(x)$, $h > n$, в якості яких використовуються або традиційні гаусіани, або інші дзвоноподібні функції, наприклад,

$$\varphi_l(x) = \left(1 + \frac{\|x - w_l\|^2}{\gamma_\varphi}\right)^{-1} = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - w_l\|^2},$$

де $w_l - (\text{n}\times 1)$ - вектор, що задає «центр» радіально-базисної функції $\varphi_l(x)$;

γ_φ – скалярний параметр, що визначає область рецепторного поля – «ширину» цієї функції.

На рисунках 3.2 – 3.4 показано, як змінюється ширина рецепторного поля зі збільшенням показника γ_φ.

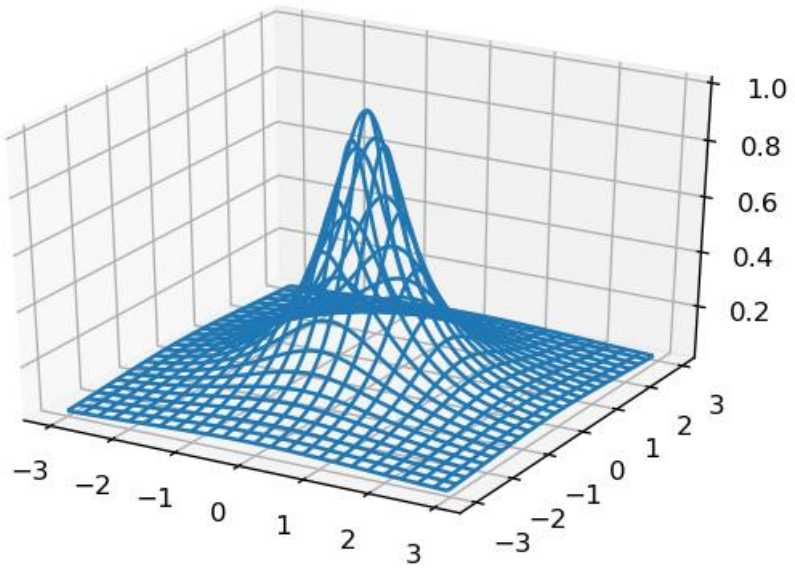

Рисунок 3.2 – Дзвоноподібна функція при $\gamma_\varphi = 0.5$

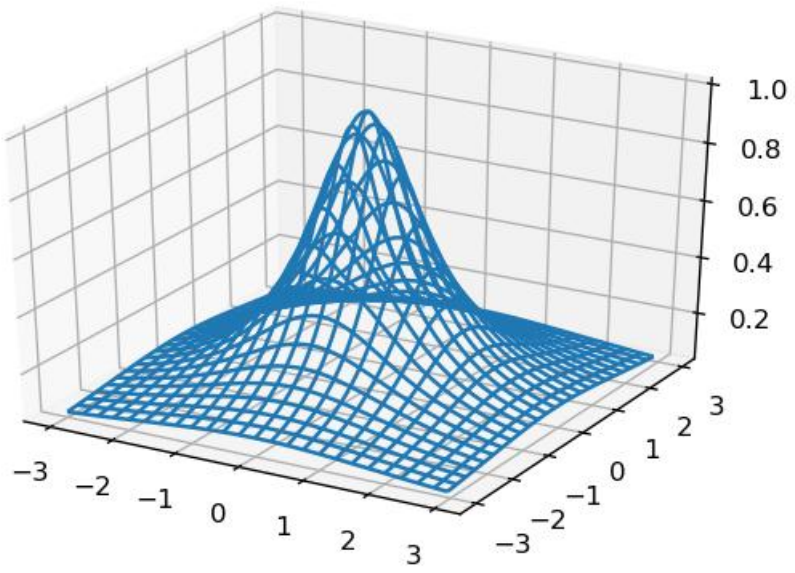

Рисунок 3.3 – Дзвоноподібна функція при $\gamma_\varphi = 1.0$

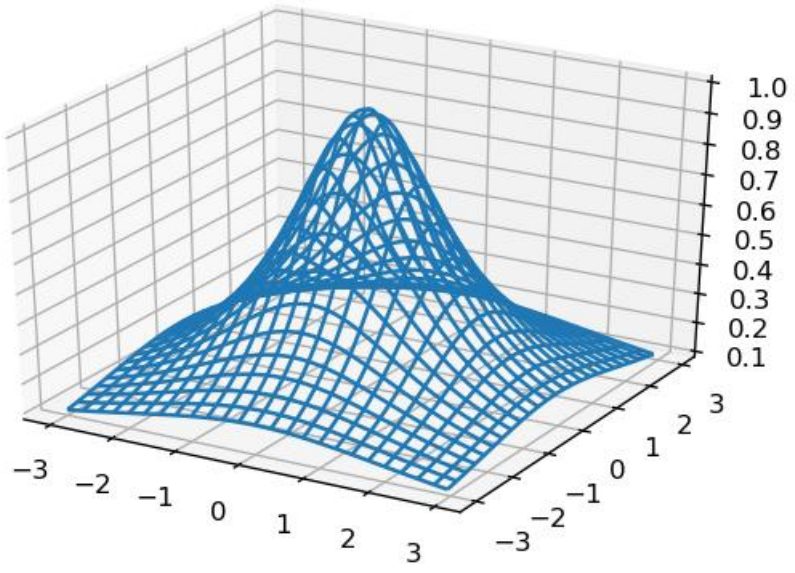

Рисунок 3.4 – Дзвоноподібна функція при $\gamma_\varphi = 2.0$

Таким чином, при надходженні на вхід системи векторного сигналу $x(k) = (x_1(k),...x_i(k),...,x_n(k))^T \in R^n$, на виході першого прихованого шару RL формується векторний сигнал $\varphi(x(k)) = (\varphi_1(x(k)),...,\varphi_l(x(k)),...,\varphi_h(x(k)))^T \in R^h$, $h > n$.

Другий прихований шар NL реалізує елементарну операцію нормалізації сигналу $\varphi(x(k))$ виду $\tilde{\varphi}(x(k)) = \frac{\varphi(x(k))}{\|\varphi(x(k))\|}$ необхідну для ефективної роботи третього прихованого шару SL, утвореного (M-1) самоорганізовними картами Кохонена SOM^m, кожна з яких працює в припущенні, що в оброблюваній виборці даних міститься m класів.

Якість кластеризації, що забезпечується кожною SOM^m, оцінюється за допомогою того чи іншого індексу валідації [50][2] в четвертому прихованому шарі VL, де обчислюються відповідні індекси $VI^2, VI^3, ..., VI^m, ..., VI^M$ для кожного з можливих $m = 2,3,..., M$.

І, нарешті, у вихідному шарі, що містить єдиний вузол - детектор оптимуму, визначається конкретна SOM^{m*}, що забезпечує найкращу якість кластеризації, при цьому вважається, що в аналізованому масиві даних міститься m^* кластерів.

3.2 Самонавчання ядерної кластерувальної системи на основі ансамблю нейронних мереж

Процес самонавчання даної системи реалізується на рівні першого шару RL, де налаштовуються центри w_l, $l = 1,2,..., h$ ядерних функцій $\varphi_l(x)$, і третього прихованого шару SL, де уточнюються синаптичні ваги w_j^m, $m = 2,3,..., M$, $j = 1,2..., m$ кожної нейронної мережі SOM^m ансамблю.

Розглянемо спочатку процес налаштування центрів ядерних функцій, що складається з послідовності наступних кроків [91] [21] [92]:

Крок 0: задати порогове значення Δ, що визначає рівень нерозрізненості двох сусідніх ядерних функцій, максимально можливу кількість цих функцій h і параметр рецепторного поля γ_φ.

Крок 1: при подаванні на вхід системи першого вектора - спостереження $x(1)$ формується перший центр w_1 і перша радіально-базисна функція

$$\varphi_1(x) = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - w_1\|^2},$$

де $w_1 = x(1)$.

Крок 2: при подаванні на вхід системи другого спостереження $x(2)$ перевіряється нерівність

$$\|x(2) - w_1\| \le \Delta$$

і якщо вона виконується, то $x(2)$ не формує новий центр, якщо ж виконується умова

$$\Delta < \|x(2) - w_1\| \le 2\Delta, \tag{3.1}$$

то w_1 коригується відповідно до правила самонавчання Т. Кохонена «Переможець отримує все» [44][10]:

$$w_1(2) = w_1(1) + \eta(2)(x(2) - w_1(1)), \tag{3.2}$$

де $w_1(1) = x(1)$, $0 < \eta(2) < 1$- параметр кроку навчання.

Якщо ж виконується умова

$$2\Delta < \|x(2) - w_1\|,$$

то формується нова ядерна функція

$$\varphi_2(x) = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - w_2\|^2} = \frac{\gamma_\varphi}{\gamma_\varphi + \|x - x(2)\|^2}.$$

Цей процес реалізується при надходженні кожного нового спостереження $x(k)$. Якщо ж на кроці N буде сформовано h радіально-базисних функцій, то в подальшому їх кількість не збільшується, а уточнення вже сформованих центрів w_l, $l = 1,2,...,h$ може проводитися тільки згідно з умовою (3.1) і правилом самонавчання (3.2).

Процес налаштування третього прихованого шару також складається з трьох етапів [44][10]: конкуренції, кооперації і синаптичної адаптації і реалізується для кожної SOM^m ансамблю, при цьому вектори синаптичних ваг w_j^m описують h-вимірні центроїди формованих кластерів.

На етапі конкуренції сигнал w виходу другого прихованого шару NL $\tilde{\varphi}(x(k)) \in R^h$ надходить на входи всіх SOM^m, де порівнюється з кожним з векторів синаптичних ваг $w_j^m(k-1)$ в сенсі відстані

$$D(\tilde{\varphi}(x(k)), w_j^m(k-1)) = \left\| \tilde{\varphi}(x(k)) - w_j^m(k-1) \right\|, \tag{3.3}$$

$j = 1,2,...\mathrm{m}$; $m = 2,3,...M$. Оскільки $\left\| \tilde{\varphi}(x(k)) \right\| = 1$, то замість евклідової метрики (3.3) набагато простіше використовувати косинусну міру подібності

$$sim(\tilde{\varphi}(x(k)), w_j^m(k-1)) = \tilde{\varphi}^T(x(k)) w_j^m(k-1),$$

за допомогою якої для кожної SOM^m визначається свій нейрон-переможець, для якого

$$\tilde{\varphi}^T(x(k)) w_j^{m*}(k-1) = \max_j \tilde{\varphi}^T(x(k)) w_j^m(k-1).$$

На етапі кооперації все $M-1$ нейронів-переможців ансамблю формують області топологічного сусідства, в яких налаштовуються не тільки ці переможці, а

й їхні найближчі сусіди. Ця область описується функціями сусідства $\varphi(j,l)$, в якості яких можуть бути використані ядерні функції аналогічні радіально-базисним функціям першого прихованого шару:

$$\varphi(j,l)=\frac{\gamma}{\gamma+\left\|w_l^m(k-1)-w_j^{m^*}(k-1)\right\|^2}.$$

На етапі синаптичної адаптації відбувається уточнення синаптичних ваг-центроїдів кожної з SOM^m за допомогою правила самонавчання Т. Кохонена «Переможець отримує більше»:

$$w_l^m(k)=w_l^m(k-1)+\eta(k)\varphi(j,l)(\tilde{\varphi}(x(k))-w_l^m(k-1)). \qquad (3.4)$$

Нескладно бачити, що для переможця $w_j^{m^*}$ (3.4) збігається з правилом навчання (3.2). Необхідно зауважити, що в правилі самонавчання (3.4) параметри кроку $\eta(k)$ та γ обираються, як правило, виходячи з емпіричних міркувань та повинні монотонно зменшуватися в процесі налаштування.

Цей процес зручно організувати за допомогою системи співвідношень:

$$\begin{cases}\eta(k)=r^{-1}(k);\ r(k)=\alpha r(k-1)+\left\|\tilde{\varphi}(x(k))\right\|^2=\alpha r(k-1)+1,\\ \gamma(k)=\eta(k)\gamma(k-1),\ 0<\alpha\le 1,\end{cases}$$

які при $\alpha=1$ автоматично перетворюються в процедуру стохастичної апроксимації.

Нескладно помітити, що перший і третій шари системи фактично навчаються згідно однотипним процедурам типу WTA і WTM [44].

3.2 Налаштування прихованих шарів

У четвертому прихованому шарі системи проводиться оцінка якості кластеризації за допомогою того чи іншого індексу валідації VI^m [52], при цьому цей індекс розраховується для кожної з мап Кохонена SOM^m, $m = 2,3,...M$. В якості такого індексу зручно використовувати критерій Девіса-Булдена (Davies DL, Bouldin DW) [79], за допомогою якого можна оцінювати якість кластеризації навіть у разі несферичних класів. Для випадку m кластерів цей індекс може бути записаний у вигляді

$$DB(m) = \sum_{j=1}^{m} \max_{\substack{1 \le q \le m \\ q \ne j}} \frac{s(w_j^m(k), u_j(k), \tilde{\varphi}(x(k)) - s(w_q^m(k), u_q(k), \tilde{\varphi}(x(k)))}{D(w_j^m(k), w_q^m(k))},$$

де $D(w_j^m(k), w_q^m(k))$ – відстань між центроїдами:

$$D(w_j^m(k), \mathrm{w}_q^m(k)) = \left\| w_j^m(k) - w_q^m(k) \right\|,$$

$s(w_j^m(k), u_j(k), \tilde{\varphi}(x(k)))$ – характеристики внутрішньокластерного розсіювання для j-го кластеру:

$$s(w_j^m(k), u_j(k), \tilde{\varphi}(x(\mathrm{k}))) = \left(\frac{\sum_{k=1}^{N} u_j(k) \left\| \tilde{\varphi}(x(k)) - \mathrm{w}_j^m(k) \right\|^2}{\sum_{k=1}^{N} u_j(k)} \right)^{\frac{1}{2}},$$

$u_j(k)$ – чітка функція належності вектора $\tilde{\varphi}(x(k))$ до j-го кластеру вигляду:

$$u_j(k)=\begin{cases}1, \text{ якщо } \tilde{\varphi}(x(k)) \text{ віднесен до } j\text{-го кластеру},\\ 0 \text{ в іншому випадку.}\end{cases}$$

В якості оптимальної кількості кластерів m^* обирається значення, що забезпечує мінімум $DB(m)$, тобто

$$DB(m^*)=\min_m\{DB(2),DB(3),...,DB(M)\},$$

який розраховується у вихідному шарі.

При обробці нестаціонарних даних, що надходять в online режимі, індекс $DB(m)$ доцільно модифікувати для роботи в режимі «ковзного вікна» розмірності $1<s<N$. При цьому модифікації піддаються тільки характеристики міжкластерної відстані, які розраховуються на «ковзному вікні» за допомогою виразу

$$\mathrm{s}(w_j^m(k),\mathrm{u}_j(k),\tilde{\varphi}(x(k)),\mathrm{s})=\left(\frac{\sum\limits_{\tau=k-s+1}^{k}u_j(\tau)\left\|\tilde{\varphi}(x(\tau))-w_j^m(k)\right\|^2}{\sum\limits_{\tau=k-s+1}^{k}u_j(\tau)}\right)^{\frac{1}{2}},$$

при цьому передбачається, що обсяг вибірки N необмежений, а зростає з плином часу $k=1,2,...,\mathrm{N},\mathrm{N}+1,...$

3.3 Експериментальне дослідження

Для підтвердження запропонованого методу була вирішена задача кластеризації із двома різними навчальними наборами даних. Перший набір даних штучно створено так, що він містить 3 кластери, 300 спостережень, кожне спостереження має 3 функції. Другий набір даних "Іриси Фішера" взято з UCI-репозиторію [82]. Цей набір даних складається з 150 спостережень, які поділяються

на 3 класи, де кожне спостереження має 3 випадкові функції. Кластери чітко видно в штучному створеному наборі даних та показані на рисунку 3.5.

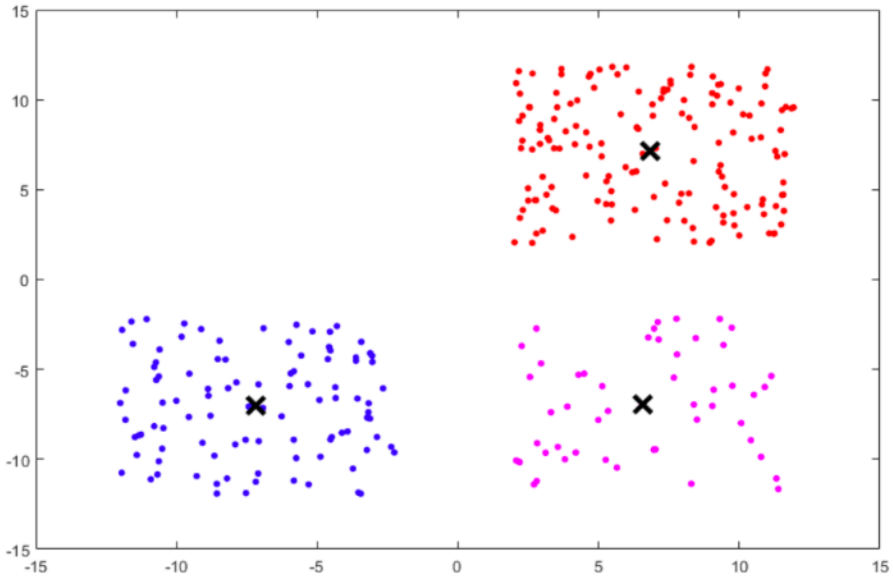

Рисунок 3.5 – Штучно згенерований набір даних

Обчислювальна точність запропонованого методу була порівняна з відомим алгоритмом K-середніх та наведена у Таблиці 3.1.

Таблиця 3.1 – Результати кластеризації для різної кількості кластерів

	SOM^m	k-means
Штучно згенерований набір даних		
точність кластерізації для 2-х кластерів	0,71	0,70
точність кластерізації для 3-х кластерів	**0,89**	0,76
точність кластерізації для 4-х кластерів	0,68	0,67
«Іриси Фішера»		
точність кластерізації для 2-х кластерів	0,84	0,83
точність кластерізації для 3-х кластерів	**0,91**	0,87
точність кластерізації для 4-х кластерів	0,72	0,73

Для візуалізації взяті набори даних проектували за допомогою методу PCA (аналіз основних компонент) на три основні компоненти. Результати роботи ансамблю нечіткої кластеризації потоків даних наведено на рисунку 3.6 та 3.7.

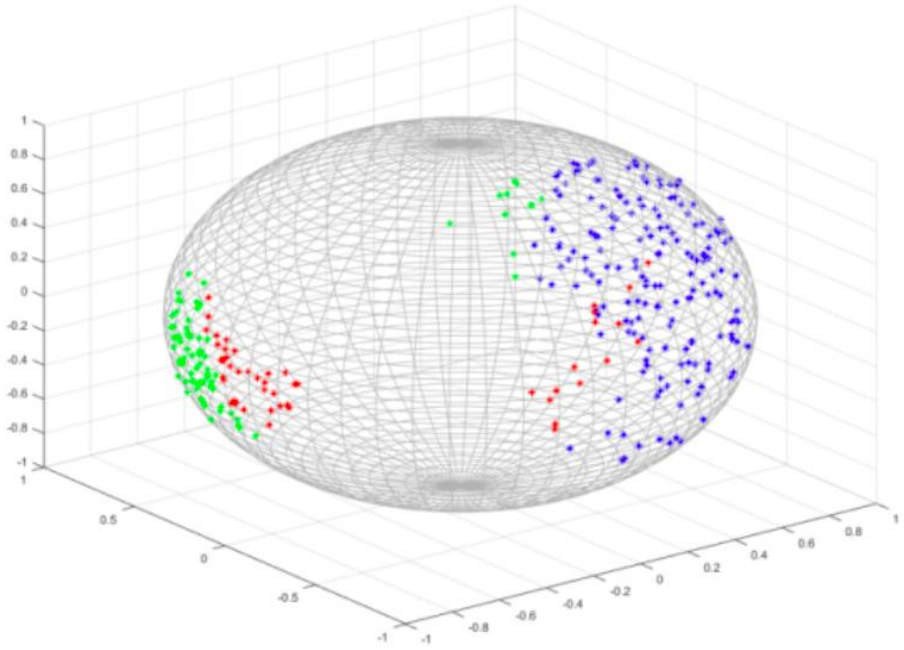

Рисунок 3.6 – Результати візуалізації запропонованого ансамблю для штучно згенерованого набору даних

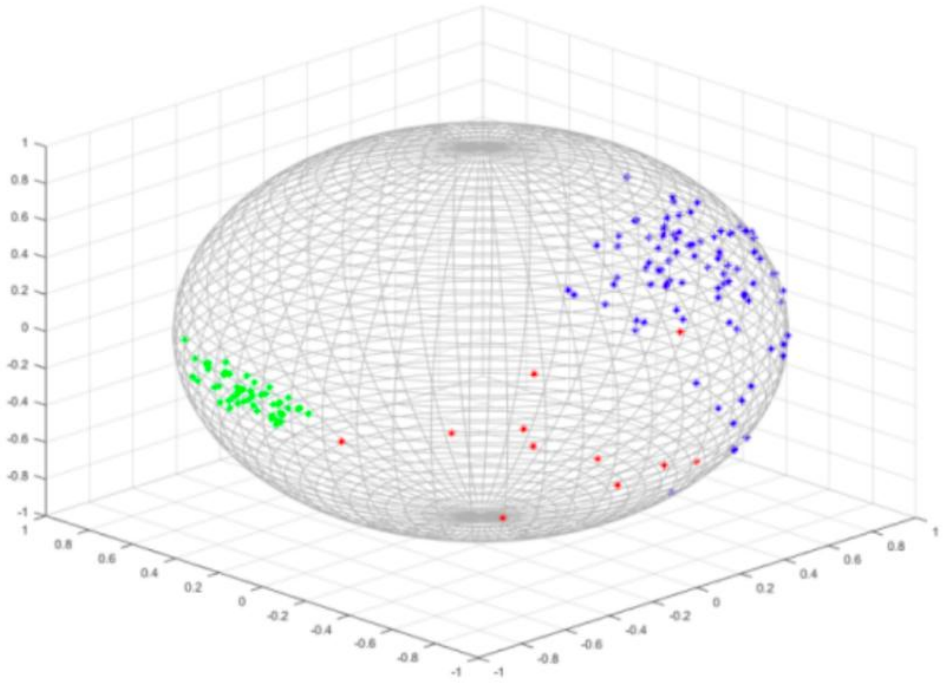

Рисунок 3.7 – Результати візуалізації запропонованого ансамблю на вибірці даних «Іриси Фішера»

3.4 Висновки за розділом

1. Запропоновано ансамбль ядерних самоорганізовних карт Т. Кохонена для кластеризації потоків даних. Завдяки ядерному шару запропонований метод має можливість обробляти дані коли кластери утворюють довільну форму.

2. Запропонована архітектура ансамблю нейронних мережі та метод її самонавчання, призначені для ядерної кластеризації потоку даних, коли спостереження надходять на обробку послідовно в on-line режимі. Запропонована система побудована на основі самоорганізовної карти Т.Кохонена.

3. Запропонована система дозволяє вирішувати завдання on-line кластеризації в умовах, коли утворені вихідними даними класи мають довільну форму. Введений ансамбль нейронних мереж простий у реалізації та дозволяє вирішувати досить широкий клас задач динамічного інтелектуального аналізу даних та інтелектуального аналізу потоків даних.

4 АНСАМБЛІ НЕЙРО-ФАЗЗІ САМООРГАНІЗОВНИХ КАРТ КОХОНЕНА ДЛЯ КЛАСТЕРУВАННЯ ПОТОКІВ ДАНИХ

В розділах 2 та 3 було вирішено проблему кластеризації потоків даних, коли інформація подається на входи системи спостереження за спостереженням, Вирішило проблему коли кластери не тільки сферичної форми, але і довільної.

Ситуація істотно ускладнюється, якщо кластери, які формуються, перетинаються у просторі ознак. Такі завдання вирішуються за допомогою методів нечіткої кластеризації [28] [83] [93], найбільш популярним з яких є алгоритм нечітких С-середніх (FCM). Для роботи в online режимі з успіхом можуть бути використані нечіткі карти Кохонена для кластеризації [94].

При цьому необхідно пам'ятати, що ефективність процедур нечіткої кластеризації обмежується, так званим, ефектом концентрації норм (concentration of norms – CoN) [43] [95], коли результати виявляються незадовільними при високих розмірностях простору ознак.

У зв'язку з цим є доцільною розробка online методу нечіткого кластеризації даних високої розмірності на основі ансамблів для кластеризації в умовах невідомої кількості класів у потоці оброблюваної інформації.

4.1 Нечітка кластерувальна нейронна мережа Т. Кохонена для обробки потоку даних високої розмірності

У класі процедур нечіткої кластеризації [96] з математичної точки зору найбільш коректними є алгоритми, засновані на цільових функціях [28] та які вирішують задачу їх оптимізації за наявності тих чи інших обмежень. Тут найбільш популярним є імовірнісний алгоритм нечіткої кластеризації, заснований на оптимізації цільової функції

$$E\left(u_j(k), w_j\right) = \sum_{k=1}^{N}\sum_{j=1}^{m} u_j^{\beta}(k)\left\|x(k) - w_j\right\|^2 = \sum_{k=1}^{N}\sum_{j=1}^{m} u_j^{\beta}(k)\sum_{i=1}^{n}\left(x_i(k) - w_{ji}\right)^2 \qquad (4.1)$$

за наявності обмежень

$$\sum_{j=1}^{m} u_j(k) = 1, \qquad (4.2)$$

$$0 \le \sum_{k=1}^{N} u_j(k) \le N. \qquad (4.3)$$

Тут $u_j(k) \in [0,1]$ – рівень нечіткої належності вектора спостережень $x(k)$ до j-го кластеру, w_j – центроїд-ваги j-го кластера, β – фаззіфікатор, що визначає розмитість границь між кластерами.

Рішення задачі оптимізації (4.1) за наявності обмежень (4.2), (4.3) за допомогою невизначених множників Лагранжа веде до результату

$$\begin{cases} u_j(k) = \dfrac{\left(\left\|x(k) - w_j\right\|^2\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(\left\|x(k) - w_l\right\|^2\right)^{\frac{1}{1-\beta}}}, \\ w_j = \dfrac{\sum\limits_{k=1}^{N} u_j^{\beta}(k)\, x(k)}{\sum\limits_{k=1}^{N} u_j^{\beta}(k)}, \end{cases} \qquad (4.4)$$

який при $\beta = 2$ повністю збігається з FCM Дж. Бездека [28].

Імовірнісний алгоритм нечіткої кластеризації (4.4) набув широкого поширення в Data Mining, однак, втрачає свою ефективність в задачах обробки даних високої розмірності через те, що виникає ефект концентрації норм. Для подолання цього недоліку в [43] [95] [97] [98] було запропоновано

використовувати, так званий, поліноміальний фаззіфікатор, що веде до процедури відомої як нечіткий метод C-середніх с поліноміальним фаззіфікатором (fuzzy C-means with polynomial fuzzifier (PFCM)). В [10] була введена адаптивна online версія PFCM, призначена для вирішення завдань Data Stream Mining.

В [99] [100] для вирішення завдань кластеризації даних високої розмірності була запропонована модифікація FCM зі зважуванням кожної з ознак $x_i(k)$, що утворюють вектор-образ $x(k) \in R^n$, $i = 1,2,...,n$.

Об'єднуючи ці два підходи, можна ввести в розгляд цільову функцію нечіткого кластеризації виду

$$E(u_j(k), w_j, \alpha, \gamma_{ji}) = \sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2(k) + (1-\alpha)u_j(k)\right)\left\|x(k) - w_j\right\|_{\Gamma_j^2}^2 = \\ = \sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2(k) + (1-\alpha)u_j(k)\right)\sum_{i=1}^{n}\gamma_{ji}^2\left(x_i(k) - w_{ji}\right)^2 \quad (4.5)$$

з обмеженнями (4.2), (4.3) і

$$\sum_{i=1}^{n}\gamma_{ji} = Tr\Gamma_j = 1 \;\forall\, j = 1,2,...m. \quad (4.6)$$

Тут $0 < \alpha \le 1$ – поліноміальний фаззіфікатор, γ_{ji} – вага i-ї ознаки в j-му кластері, $\Gamma_j = diag(\gamma_{j1}, \gamma_{j2}, ..., \gamma_{jm})$.

Оптимізація цільової функції (4.5) при обмеженнях (4.2), (4.3), (4.6) за допомогою невизначених множників Лагранжа веде до результату

$$\begin{cases} u_j(k) = \dfrac{\alpha-1}{2\alpha} + \dfrac{1-m\dfrac{\alpha-1}{2\alpha}}{\sum\limits_{l=1}^{m}\dfrac{\left\|x(k)-w_j\right\|^2_{\Gamma_j^2}}{\left\|x(k)-w_l\right\|^2_{\Gamma_l^2}}}, \\ \gamma_{ji} = \left(\sum\limits_{h=1}^{n}\left(\dfrac{\sum\limits_{k=1}^{N}\left(\alpha u_j^2(k)+(1-\alpha)u_j(k)\right)\left(x_i(k)-w_{ji}\right)^2}{\sum\limits_{k=1}^{N}\left(\alpha u_j^2(k)+(1-\alpha)u_j(k)\right)\left(x_h(k)-w_{ji}\right)^2}\right)\right)^{-1}, \\ w_{ji} = \dfrac{\sum\limits_{k=1}^{N}\left(\alpha u_j^2(k)+(1-\alpha)u_j(k)\right)\gamma_{ji}^2 x_i(k)}{\sum\limits_{k=1}^{N}\left(\alpha u_j^2(k)+(1-\alpha)u_j(k)\right)\gamma_{ji}^2}, \end{cases} \tag{4.7}$$

що є узагальненням (4.4) і збігається з ним при $\alpha=1, \gamma_{ji}=m^{-1}$.

Останнє співвідношення (4.7) для розрахунку центроїдів кластерів може бути переписано у рекурентній формі

$$\begin{aligned} w_j(k) = w_j(k-1) + \eta(k)\left(\alpha u_j^2(k-1)+(1-\alpha)u_j(k-1)\right)\times \\ \times\Gamma_j^2(k-1)\left(x(k)-w_j(k-1)\right), \end{aligned} \tag{4.8}$$

що є за суттю, WTM-правилом самонавчання Т. Кохонена [44], де співмножник $\left(\alpha u_j^2\ (k-1)+(1-\alpha)u_j(k-1)\right)\Gamma_j^2(k-1)$ задає функцію сусідства, а $0<\eta(k)<1$ параметр кроку навчання.

Таким чином, процес кластеризації даних високої розмірності (4.7), (4.8) зручно реалізувати за допомогою архітектури, наведеної на рис. 4.1, що є модифікацією нейро-фаззі мережі Т. Кохонена [101].

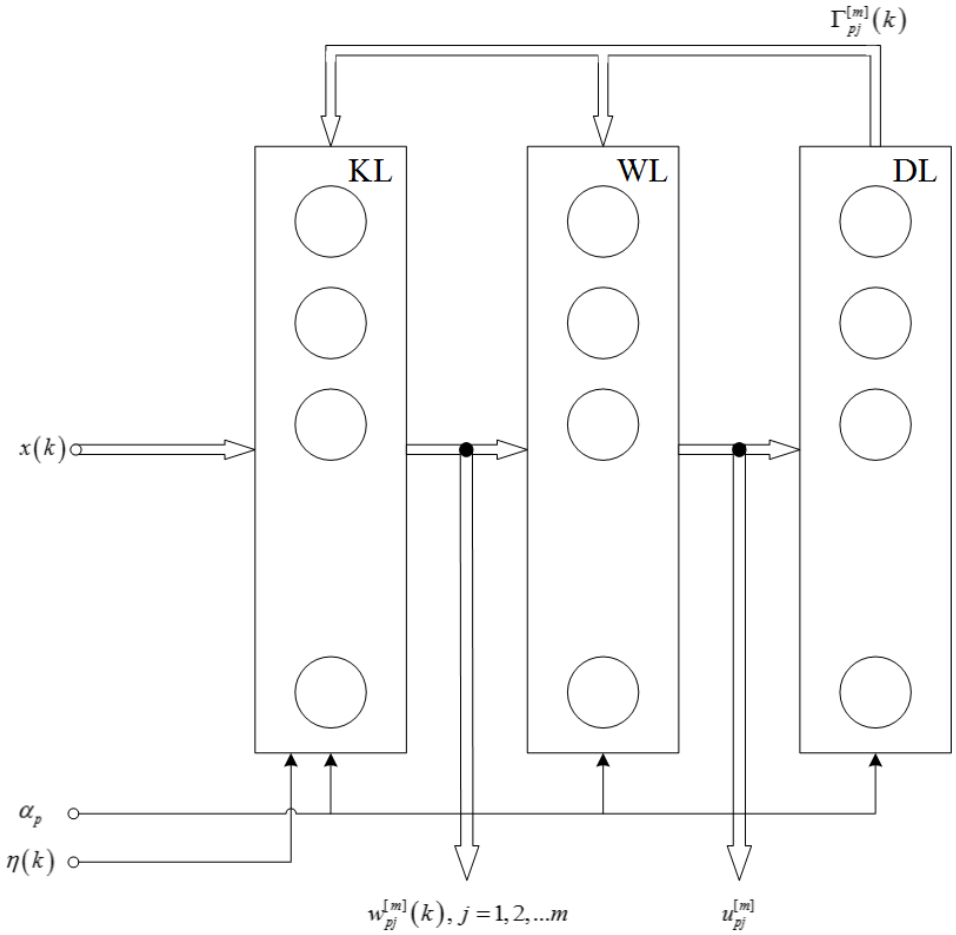

Рисунок 4.1 – Адаптивна нейро-фаззі мережа Т. Кохонена

Тут перший прихований шар KL є за суттю стандартною нейронною мережею SOM [44], що містить в шарі Кохонена m-нейронів, синаптичні ваги-центроїди, які настроюються за допомогою WTM-правила навчання (4.8), у другому прихованому шарі ML оцінюються рівні нечіткої належності k-го спостереження j-му кластеру $u_j(k)$ за допомогою першого співвідношення (4.7), а в вихідному шарі WL розраховуються значення ваг γ_{ji} за допомогою другого співвідношення (4.7).

На додаткові входи мережі подаються значення параметра швидкості навчання $\eta(k)$ і поліноміального фаззіфікатора з деякої апріорі заданої множини α_p, $0<\alpha_1,\alpha_2,...,\alpha_p,...\alpha_q=1$.

4.2 Архітектура кластерувального ансамблю

Для вирішення завдання кластеризації в умовах, коли кількість кластерів невідома, пропонується використовувати ансамбль кластерувальних нейро-фаззі мереж Кохонена архітектура якого наведена на рисунку 4.2. Даний ансамбль містить $(M-1)_q$ $FSOM_p^{[m]}$, де індекс $[m]$ означає кількість кластерів, на яку ця мережа розбиває оброблювану вибірку – тобто кількість нейронів в шарі Кохонена KL, а p – індекс конкретного фаззіфікатора, що приймає q значень. Всі $FSOM_p^{[m]}$ навчаються за допомогою однотипних процедур (4.7), (4.8), які відрізняються один від одного тільки значеннями m та α.

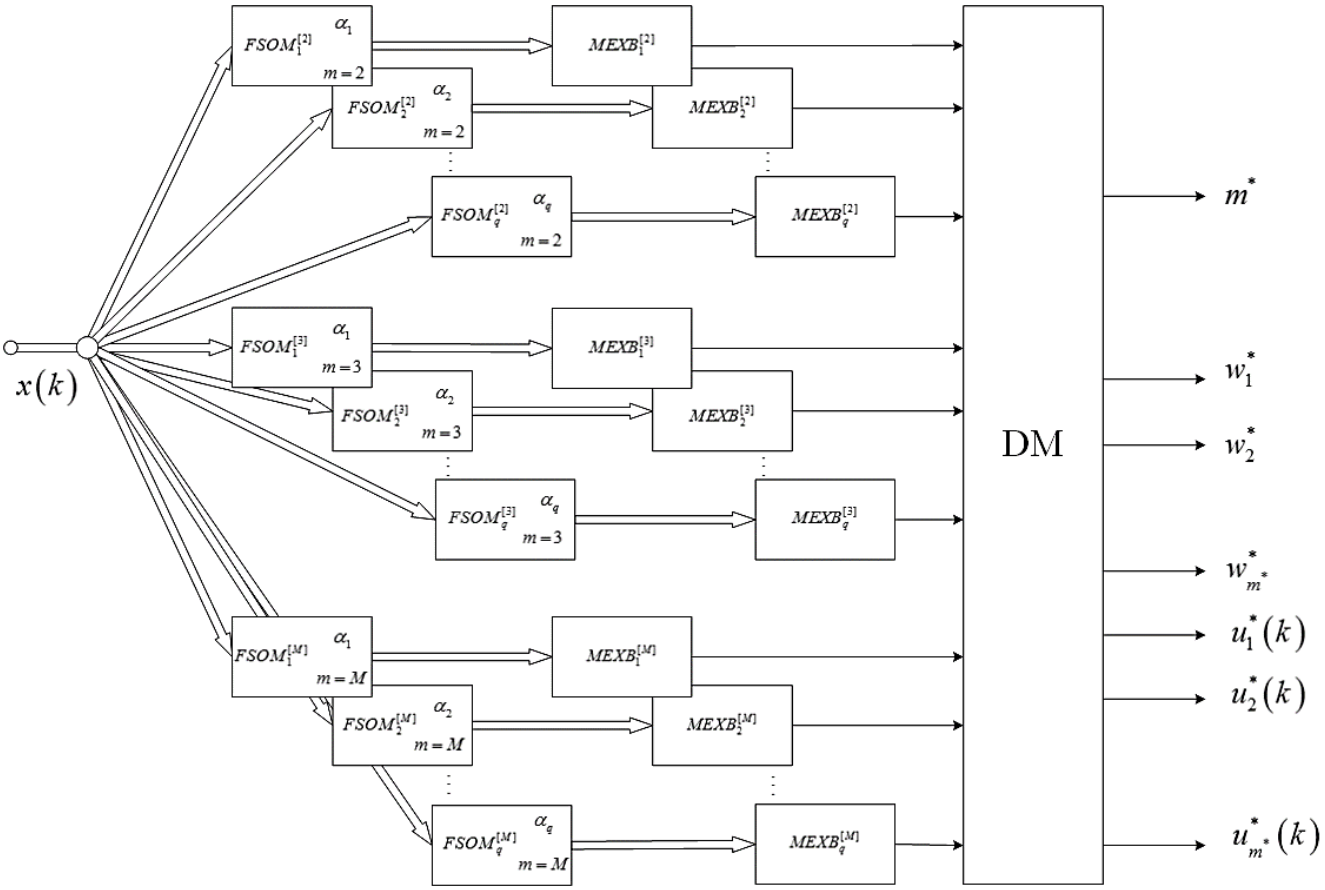

Рисунок 4.2 – Архітектура кластерувального ансамблю

У блоках $MEXB_p^{[m]}$ оцінюється якість кластеризації, що забезпечується конкретною FSOM, а вихідний шар ансамблю DM з $(M-1)_q$ результатів попередніх шарів виділяє найкращий, тобто кількість кластерів m^* в оброблюваних даних, центроїди сформованих кластерів $w_1^*, w_2^*, ..., w_{m^*}^*$ і рівні належності кожного спостереження $u_1^*(k), u_2^*(k), ..., u_{m^*}^*(k)$ до відповідного кластеру.

Для оцінки якості кластеризації кожним з елементів ансамблю може бути використаний будь-який з індексів нечіткої кластеризації [50][2], де одним з найбільш популярних є індекс Ксі-Бені (Xie-Beni index) [102][21], який для FCM-процедури в разі m кластерів може бути записаний у формі

$$XB^{[m]} = \frac{\dfrac{\sum\limits_{k=1}^{N}\sum\limits_{j=1}^{m} u_j^2(k)\left\|x(k)-w_j\right\|^2}{N}}{\min\limits_{l\neq j}\left\|w_j - w_l\right\|^2} = \frac{NXB^{[m]}}{DXB^{[m]}}. \tag{4.9}$$

Для послідовної обробки можна ввести online версію XB-індексу у вигляді

$$XB^{[m]}(k) = \frac{NXB^{[m]}(k)}{DXB^{[m]}(k)} = \frac{NXB^{[m]}(k-1)}{\min\limits_{l\neq j}\left\|w_j(k)-w_l(k)\right\|^2} + \\ + \frac{\dfrac{1}{k}\left(\sum\limits_{j=1}^{m} u_j^2(k)\left\|x(k)-w_j(k)\right\|^2 - NXB^{[m]}(k-1)\right)}{\min\limits_{l\neq j}\left\|w_j(k)-w_l(k)\right\|^2}. \tag{4.10}$$

Чим менше значення (4.9), (4.10), тим вище якість кластеризації. Для процедури (4.4) може бути використаний розширений індекс Ксі-Бені (extended Xie-Beni index) [103]

$$EXB^{[m]} = \frac{\sum_{k=1}^{N}\sum_{j=1}^{m} u_j^{\beta}(k)\left\|x(k) - w_j\right\|^2}{\min\limits_{l \neq j}\left\|w_j - w_l\right\|^2} \tag{4.11}$$

або його online версія

$$EXB^{[m]}(k) = \frac{NEXB^{[m]}(k)}{DEXB^{[m]}(k)} = \\ = \frac{NEXB^{[m]}(k-1) + \frac{1}{k}\left(\sum_{j=1}^{m} u_j^{\beta}(k)\left\|x(k) - w_j(k)\right\|^2 - NEXB^{[m]}(k-1)\right)}{\min\limits_{l \neq j}\left\|w_j(k) - w_l(k)\right\|^2}. \tag{4.12}$$

За аналогією з (4.11), (4.12) можна ввести модифікацію EXB-index для цільової функції (4.5)

$$MEXB_p^{[m]} = \frac{\sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha_p\left(u_{pj}^{[m]}(k)\right)^2 + \left(1-\alpha_p\right)u_{pj}^{[m]}(k)\right)\left\|x(k) - w_{pj}^{[m]}\right\|^2_{\left(\Gamma_{pj}^{[m]}\right)^2}}{\min\limits_{l \neq j}\left\|w_{pj}^{[m]} - w_{pl}^{[m]}\right\|^2} = \\ = \frac{NMEXB_p^{[m]}}{DMEXB_p^{[m]}} \tag{4.13}$$

або в online версії

$$MEXB_p^{[m]}(k)=\frac{NMEXB_p^{[m]}(k)}{DMEXB_p^{[m]}(k)}=\frac{NMEXB_p^{[m]}(k-1)}{\min\limits_{l\neq j}\left\|w_{pj}^{[m]}(k)-w_{pl}^{[m]}(k)\right\|^2}+$$

$$+\left(\frac{1}{k}\left(\sum_{j=1}^{m}\left(\alpha_p\left(u_{pj}^{[m]}(k)\right)^2+(1-\alpha)u_{pj}^{[m]}(k)\right)\times\right.\right. \tag{4.14}$$

$$\left.\left.\times\left\|x(k)-w_{pj}^{[m]}(k)\right\|^2_{\left(\Gamma_{pj}^{[m]}(k)\right)^2}-NMEXB_p^{[m]}(k-1)\right)\right)\left(\min_{l\neq j}\left\|w_{pj}^{[m]}(k)-w_{pl}^{[m]}(k)\right\|^2\right)^{-1}.$$

В процесі обробки даних блок DM знаходить $FSOM_p^{[m^*]}$ з найкращим значенням $MEXB_p^{[m^*]}$ і результати роботи саме цієї нейро-фаззі мережі визначають кінцевий результат кластеризації.

4.3 Експериментальне дослідження

Для вирішення проблеми визначення оптимального числа кластерів у наборах даних ми використали запропонований ансамбль для онлайнової нечіткої кластеризації. Ми обрали набір даних Waveform Database Generator з UCI Machine Repository [82]. Він містить 5000 спостережень, 21 атрибут з безперервними значеннями від 0 до 6. Кожен клас генерується з комбінації 2-3 "базових" хвиль. Кожен екземпляр генерується додатковим шумом (середнє значення 0, дисперсія 1) у кожному атрибуті. Пропущених значень атрибутів немає. Розподіл класу: 33% для кожного з 3 класів.

На рисунку 4.3 представлена візуалізація з використанням аналізу основних компонентів (PCA-аналіз, три основні компоненти).

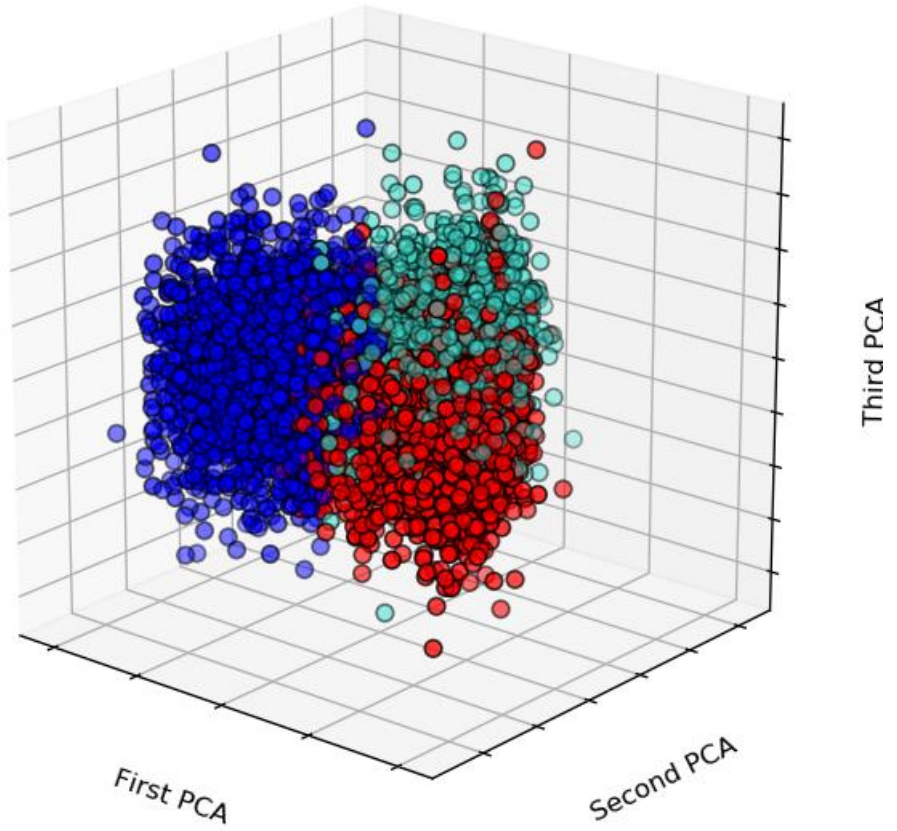

Рисунок 4.3 – Візуалізація набора даних Waveform Database Generator

Таблиця 4.1. Значення параметра індекс Ксі-Бені для набору даних для Waveform Database Generator

Алгоритм	Ксі-Бені індекс	Кількість кластерів
K-means	328803900538.9107	2
MHDFCM $\alpha = 0,5$	0,00003	
MHDFCM $\alpha = 1,0$	0,00002	
FCM	CoN	
K-means	182839765709655.22	3
MHDFCM $\alpha = 0,5$	0,00009	

Продовження таблиці 4.1

MHDFCM $\alpha = 1,0$	0,00004	
FCM	CoN	
K-means	870995049068624.5	4
MHDFCM $\alpha = 0,5$	0.0003	
MHDFCM $\alpha = 1,0$	0.0001	
FCM	CoN	
K-means	120017764730808.92	5
MHDFCM $\alpha = 0,5$	2.6266	
MHDFCM $\alpha = 1,0$	0.0310	
FCM	CoN	

У таблиці 4.1 представлені значення індексів Ксі-Бені для різної кількості кластерів. Мінімальне значення відповідає MHDFCM, коли параметр α дорівнює 0,5 або 1. Алгоритм K-середніх має дуже високі значення індексу Ксі-Бені, тому ми можемо говорити про неефективні результати кластеризації. Алгоритм FCM не може обробити запропонований набір даних через концентрацію норм (CoN).

Запропонована архітектура та алгоритм самонавчання нейро-фаззі системи, призначеної для вирішення завдання online кластеризації потоку даних високої розмірності в умовах, коли кластери які формуються перекриваються та їх число заздалегідь невідомо. Запропонована система є ансамблем нейро-фаззі самоорганізовних мап Т. Кохонена, кожна з яких відрізняється від інших кількістю нейронів і значенням поліноміального фаззіфікатора. Налаштування кожного з членів ансамблю відбувається за допомогою модифікованого WTM правила самонавчання, при цьому в процесі налаштування проводиться автоматичне зважування усіх компонент оброблюваних векторів.

Запропонований підхід є узагальненням ряду відомих процедур нечіткої ймовірнісної кластеризації і може бути використаний для вирішення задач обробки потоків даних.

4.4 Висновки за розділом

1. Вперше запропоновано ансамбль нейро-фаззі самоорганізовних карт Т. Кохонена для кластеризації потоків даних великої розмірності, що послідовно поступають на вхід системи спостереження за спостереженням, який дозволяє в процесі самонавчання налаштовувати не тільки свої параметри, а й архітектуру в on-line режимі та вирішувати завдання кластеризації потоку даних за умови апріорно невідомої форми та кількості кластерів.

2. Запропонована архітектура та алгоритми самонавчання ансамблю нейро-фаззі систем обчислювального інтелекту для кластеризації даних в умовах, коли кластери можуть мати довільну форму та взаємно перетинатися. В основі запропонованого ансамблю лежать нечітка нейронна мережа та нейро-фаззі мережа Т. Кохонена.

3. Введені процедури самонавчання досить прості в чисельній реалізації та призначені для обробки даних, які послідовно в on-line режимі надходять в систему.

5 АНСАМБЛЬ НЕЙРО-ФАЗЗІ МЕРЕЖ Т. КОХОНЕНА З ВИКОРИСТАННЯМ ІМОВІРНІСНО-МОЖЛИВІСНОГО ПІДХОДУ

5.1 Нечітка кластерувальна нейронна мережа Т. Кохонена для обробки потоку даних

У класі процедур нечіткої кластеризації з математичної точки зору найбільш коректними є алгоритми, засновані на цільових функціях [28] та які вирішують задачу їх оптимізації за наявності тих чи інших обмежень. Тут найбільш популярним є імовірнісний алгоритм нечіткої кластеризації, заснований на оптимізації цільової функції

$$E(u_j(k), w_j) = \sum_{k=1}^{N}\sum_{j=1}^{m} u_j^{\beta}(k) D^2(x(k), w_j) \tag{5.1}$$

за наявності обмежень

$$\sum_{j=1}^{m} u_j(k) = 1, \tag{5.2}$$

$$0 \le \sum_{k=1}^{N} u_j(k) \le N. \tag{5.3}$$

Тут $u_j(k) \in [0,1]$ – рівень нечіткої належності вектора спостережень $x(k)$ до j-го кластеру, w_j – центроїд j-го кластера, β – фаззіфікатор, що визначає розмитість границь між кластерами, $D(x(k), w_j)$ – відстань між $x(k)$ і w_j відповідно до обраної метрики.

Результатом кластеризації є $N \times m$ матриця нечіткого розбиття (fuzzy partitioning matrix)

$$W = \{u_j(k)\}. \tag{5.4}$$

Необхідно зазначити, що оскільки елементи матриці W можуть розглядатися як імовірності гіпотез належності векторів даних визначеним кластерам, процедури, що породжуються (5.1) за обмежень (5.2), (5.3), називаються імовірнісними методами кластеризації [104].

В якості функції відстані $D(x(k), w_j)$ зазвичай вибирається відстань Мінковського в L^p метриці [105]:

$$D^p(x(k), w_j) = \| x_i(k) - w_{ji} \|_{L_p}^p = \left(\sum_{i=1}^{n} | x_i(k) - w_{ji} |^p \right)^{\frac{1}{p}}, \quad p \geq 1, \tag{5.5}$$

де $x_i(k)$ – i-а компонента $(n \times 1)$-вектора $x(k)$; w_{ji} – i-а компонента $(n \times 1)$–вектора w_j.

Розглянемо функцію Лагранжа

$$L(u_j(k), w_j, \lambda(k)) = \sum_{k=1}^{N} \sum_{j=1}^{m} u_j^{\beta}(k) D^2(x(k), w_j) + \sum_{k=1}^{N} \lambda(k) \left(\sum_{j=1}^{m} u_j(k) - 1 \right) =$$

$$= \sum_{k=1}^{N} \left(\sum_{j=1}^{m} u_j^{\beta}(k) D^2(x(k), w_j) + \lambda(k) \left(\sum_{j=1}^{m} u_j(k) - 1 \right) \right), \tag{5.6}$$

де $\lambda(k)$ – невизначений множник Лагранжа, що забезпечує умови (5.2), (5.3).

Розв’язуючи систему рівнянь Куна-Таккера

$$\begin{cases} \dfrac{\partial L(u_j(k), w_j, \lambda(k))}{\partial w_j(k)} = 0, \\ \dfrac{\partial L(u_j(k), w_j, \lambda(k))}{\partial \lambda(k)} = 0, \\ \\ \nabla_{w_j} L(u_j(k), w_j, \lambda(k)) = \vec{0}, \end{cases} \tag{5.7}$$

легко отримати шуканий розв'язок у вигляді [106]:

$$u_j^{pr}(k) = \frac{(D^2(x(k), w_j))^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m}(D^2(x(k), w_l))^{\frac{1}{1-\beta}}}, \tag{5.8}$$

$$\lambda(k) = -\left(\sum_{l=1}^{m}(\beta D^2(x(k), w_l))^{\frac{1}{1-\beta}}\right)^{1-\beta} \tag{5.9}$$

$$w_j^{pr} = \frac{\sum_{k=1}^{N} u_j^{\beta}(k) x(k)}{\sum_{k=1}^{N} u_j^{\beta}(k)}. \tag{5.10}$$

Рівняння (5.8)-(5.10) породжують широкий клас процедур кластеризації.

При $\beta = p = 2$, тобто в евклідовому просторі

$$D^E(x(k), w_j) = \| x(k) - w_j \| = \sqrt{(x(k) - w_j)^T (x(k) - w_j)}, \tag{5.11}$$

отримуємо достатньо простую та ефективну процедуру кластеризації нечітких С-середніх Бездека [28]:

$$u_j^{pr}(k) = \frac{\| x(k) - w_j \|^{-2}}{\sum_{l=1}^{m} \|x(k) - w_l\|^{-2}}, \tag{5.12}$$

$$w_j^{pr} = \frac{\sum_{k=1}^{N} u_j^2(k) x(k)}{\sum_{k=1}^{N} u_j^2(k)}, \tag{5.13}$$

$$\lambda(k) = -\sum_{l=1}^{m} \left(\frac{\| x(k) - w_l \|^{-2}}{2} \right)^{-1}. \tag{5.14}$$

До імовірнісних методів кластеризації відносяться також процедури Ґустафсона-Кесселя [106], Ґата-Ґеви [107] та багато інших. Не дивлячись на незначну обчислювану складність, процедура (5.12) – (5.13) має недолік, який виражений в необхідності виконання загальної для всіх імовірнісних методів нечіткої кластеризації умови (5.2).

В найпростішому випадку двох кластерів $(m = 2)$ легко бачити, що спостереження $x(k)$, що рівноправно належить до обох кластерів, і спостереження $x(p)$, що не належить жодному з них, можуть мати однакові рівні належності $u_1^{pr}(k) = u_2^{pr}(k) = u_1^{pr}(p) = u_2^{pr}(p) = 0.5$. Очевидно, що ця особливість може суттєво знизити якість класифікації. В той же час можливісний підхід до нечіткої кластеризації [108] [109] [110] допомагає уникнути відзначеної вище ситуації і тим самим покращити якість класифікації.

Для можливісного підходу до кластеризації критерій, що мінімізується, має вигляд

$$E(u_j(k), w_j) = \sum_{k=1}^{N} \sum_{j=1}^{m} u_j^{\beta}(k) D^2(x(k), w_j) + \sum_{j=1}^{m} \mu_j \sum_{k=1}^{N} (1 - u_j(k))^{\beta}, \tag{5.15}$$

де скалярний параметр $\mu_j > 0$ визначає відстань, на якій рівень належності приймає значення 0.5, тобто якщо $D^2(x(k), w_j) = \mu_j$, то $u_j(k) = 0.5$.

Мінімізація критерію (5.14) за параметрами $u_j(k)$, w_j, μ_j призводить до системи рівнянь:

$$\begin{cases} \dfrac{\partial E(u_j(k), w_j)}{\partial u_j(k)} = 0, \\ \dfrac{\partial E(u_j(k), w_j)}{\partial \mu(k)} = 0, \\ \\ \nabla_{w_j} E(u_j(k), w_j) = \vec{0}. \end{cases} \tag{5.16}$$

Розв’язок перших двох рівнянь дає відомий результат:

$$u_j^{pos}(k) = \left(1 + \left(\frac{D^2(x(k), w_j)}{\mu_j} \right)^{\frac{1}{\beta - 1}} \right)^{-1}, \tag{5.17}$$

$$\mu_j = \frac{\sum_{k=1}^{N} u_j^{\beta}(k) D^2(x(k), w_j)}{\sum_{k=1}^{N} u_j^{\beta}(k)}. \tag{5.18}$$

Розв’язок третього рівняння системи (5.16) для евклідової норми (5.11) має вигляд:

$$w_j^{pos} = \frac{\sum_{k=1}^{N} u_j^{\beta}(k) x(k)}{\sum_{k=1}^{N} u_j^{\beta}(k)}. \tag{5.19}$$

Можна бачити, що можливісні та імовірнісні методи достатньо схожі та переходять один в інший шляхом заміни виразу (5.17) на формулу (5.8), і навпаки. Загальним недоліком розглянутих методів є неможливість роботи в реальному часі, коли дані надходять, наприклад, у формі потока відео.

Робота процедури (5.8)–(5.9) розпочинається із завдання початкової (зазвичай випадкової) матриці розбиття W^0. На основі її значень обчислюється початковий набір прототипів w_j^0, які потім використовуються для уточнення нової матриці W^1. Наступним кроком в пакетному режимі є обчислення w_j^1, $W^2,\ldots,W^t,w_j^t,W^{t+1}$ і так далі, доки різниця $\|W^{t+1}-W^t\|$ не стане меншою за деяке наперед задане порогове значення ε. Таким чином, вся вибірка даних оброблюється багатократно.

Розв’язок, що може бути отриманий за допомогою імовірнісного метода, рекомендується використовувати в якості початкових умов для можливісного метода (5.17)–(5.19) [110] [111], при чому початкові значення параметрів відстані μ_j^t обираються у відповідності до (5.18) за результатами роботи імовірнісної процедури.

Аналіз рівняння (5.8) показує, що для обчислення рівнів належності $u_j(k)$ замість функції Лагранжа (5.6) можна використовувати її локальну модифікацію:

$$L_k(u_j(k),w_j,\lambda(k))=\sum_{j=1}^{m}u_j^{\beta}(k)D^2(x(k),w_j)+\lambda(k)\left(\sum_{j=1}^{m}u_j(k)-1\right). \qquad (5.20)$$

Оптимізація виразу (5.20) за допомогою процедури Ерроу–Гурвіца–Удзави [110] [112] веде до процедури:

$$u_j^{pr}(k) = \frac{(D^2(x(k), w_j(k)))^{\frac{1}{1-\beta}}}{\sum_{l=1}^{m}(D^2(x(k), w_l(k)))^{\frac{1}{1-\beta}}}, \tag{5.21}$$

$$w_j^{pr}(k+1) = w_j^{pr}(k) - \eta(k)\nabla_{w_j} L_k(u_j(k), w_j^{pr}(k), \lambda(k)) =$$
$$= w_j^{pr} - \eta(k)u_j^{\beta}(k)(x(k+1) - w_j^{pr}(k)), \tag{5.22}$$

де $\eta(k)$ – параметр кроку навчання; $w_j^{pr}(k)$ – прототипи j-го кластера, що обчислюються на вибірці з k спостережень.

Процедура (5.21)-(5.22) схожа на процедуру навчання Чанга-Лі [113] і для $\beta = p = 2$ збігається з градієнтною процедурою кластеризації Парка–Деггера [114]:

$$u_j^{pr}(k) = \frac{\| x(k) - w_j(k) \|^{-2}}{\sum_{l=1}^{m} \| x(k) - w_l(k) \|^{-2}}, \tag{5.23}$$

$$w_j^{pr}(k+1) = w_j^{pr}(k) + \eta(k)u_j^2(k)(x(k+1) - w_j^{pr}(k)). \tag{5.24}$$

В межах можливісного підходу локальний критерій набуває вигляду

$$E_k(u_j(k), w_j) = \sum_{j=1}^{m} u_j^{\beta}(k) D^2(x(k), w_j) + \sum_{j=1}^{m} \mu_j (1 - u_j(k))^{\beta}, \tag{5.25}$$

а результат його оптимізації записується як

$$u_j^{pos}(k) = \left(1 + \left(\frac{D^2(x(k), w_j(k))}{\mu_j(k)} \right)^{\frac{1}{\beta-1}} \right)^{-1}, \tag{5.26}$$

$$w_j^{pos}(k+1) = w_j^{pos}(k) - \eta(k)u_j^{\beta}(k)(x(k+1) - w_j^{pos}(k)), \tag{5.27}$$

$$\mu_j(k+1) = \frac{\sum_{p=1}^{k} u_j^{\beta}(p) D^2(x(p), w_j(k+1))}{\sum_{p=1}^{k} u_j^{\beta}(p)}. \tag{5.28}$$

В квадратичному випадку (при $\beta = 2$) процедура (5.26)–(5.28) перетворюється в достатньо просту конструкцію

$$u_j^{pos}(k) = \frac{\mu_j(k)}{\mu_j(k) + \| x(k) - w_j(k) \|^2}, \tag{5.29}$$

$$w_j^{pos}(k+1) = w_j^{pos}(k) + \eta(k)u_j^{2}(k)(x(k+1) - w_j^{pos}(k)), \tag{5.30}$$

$$\mu_j(k+1) = \frac{\sum_{p=1}^{k} u_j^{2}(p) \| x(p) - w_j(k+1) \|^2}{\sum_{p=1}^{k} u_j^{2}(p)}. \tag{5.31}$$

Паралельне застосування адаптивних імовірнісної і можливісної процедур призводить до об’єднаної процедури нечіткої кластеризації [115] [116]

$$\begin{cases} w_j^{pr}(k) = w_j^{pos}(k-1) - \eta(k)\,\mathrm{u}_j^{pos\,\beta}(k-1)(x(k+1) - w_j^{pos}(k)), \\ u_j^{pr}(k) = (D^2(x(k), \mathrm{w}_j^{pr}(k)))^{\frac{1}{1-\beta}} \left(\sum_{l=1}^{m} (D^2(x(k), \mathrm{w}_j^{pr}(k)))^{\frac{1}{1-\beta}} \right)^{-1}, \\ w_j^{pos}(k) = w_j^{pr}(k-1) - \eta(k)\,\mathrm{u}_j^{pr\,\beta}(k)(x(k+1) - w_j^{pr}(k)), \\ \mu_j(k) = \left(\sum_{p=1}^{k} u_j^{pr\,\beta}(p) D^2(x(k), \mathrm{w}_j^{pos}(k)) \right) \left(\sum_{p=1}^{k} u_j^{pr\,\beta} \right)^{-1}, \\ u_j^{pos}(k) = \left(1 + \left(\frac{D^2(x(k), \mathrm{w}_j^{pos}(k))}{\mu_j(k)} \right) \right)^{-1}, \quad j = 1, 2, \ldots, m. \end{cases} \tag{5.32}$$

Ознакою правильного знаходження прототипів (а отже і коректної кластеризації), використовуючи процедуру (5.32), є виконання нерівності

$$\sum_{l=1}^{m} D^2(w_l^{pr}(k), w_l^{pos}(k)) \le \varepsilon, \tag{5.33}$$

де ε визначає прийнятну точність кластеризації.

Для евклідової метрики значення параметра $\mu_j(k)$ може обчислюватися відповідно до рекурентного співвідношення, яке безпосередньо випливає з (5.31):

$$\begin{cases} \beta_q(k) &= \beta_q(k-1) + w_j^{pr\,2}(k) s_q(x(k)), \quad q = 0, 1, 2, \\ \mu_j(k) &= \dfrac{\beta_2(k-1) - 2w_j^{pos\,T}(k)\beta_1(k-1) + \| \mathrm{w}_j^{pos}(k) \|^2 \beta_0(k-1)}{\beta_0(k-1)}, \end{cases} \tag{5.34}$$

де

$$s_q(x(k)) = \begin{cases} 1, & якщо\ q = 0, \\ x(k), & якщо\ q = 1, \\ \| x(k) \|^2, & якщо\ q = 2. \end{cases} \quad (5.35)$$

Початкові значення параметра $\beta_q(k)$ обираються як

$$\beta_q(N) = \sum_{p=1}^{N} (u_j^{pr}(p))^2 s_q(x(p)), \quad q = 0,1,2. \quad (5.36)$$

Таким чином, адаптивна процедура (5.32) може працювати як в пакетному режимі для ітеративної обробки заданої вибірки, так і в режимі реального часу, де кількість спостережень визначається дискретним часом $k = 1,2,\ldots,N,N+1,\ldots$ В останньому випадку за допомогою цієї процедури послідовно оброблюються спостереження, що надходять на опрацювання. Отже, у випадку нестаціонарних даних рівні належності та прототипи кластерів перестроюються відповідно до нових даних [67].

5.2 Архітектура ансамблю нечітких карт Т. Кохонена

Для вирішення завдання кластеризації в умовах, коли кількість кластерів невідома, пропонується використовувати ансамбль кластерувальних нейро-фаззі мереж Кохонена. Даний ансамбль містить $(M-1)$ $FSOM^{[m]}$, де індекс $[m]$ означає кількість кластерів, на яку ця мережа розбиває оброблювану вибірку – тобто кількість нейронів в шарі Кохонена KL, а p – індекс конкретного фаззіфікатора, що приймає q значень. Всі $FSOM^{[m]}$ навчаються за допомогою однотипних процедур (7), (8), які відрізняються один від одного тільки значеннями m та α.

Далі оцінюється якість кластеризації, що забезпечується конкретною FSOM, а вихідний шар ансамблю з $(M-1)$ результатів попередніх шарів виділяє

найкращий, тобто кількість кластерів m^* в оброблюваних даних, центроїди сформованих кластерів $w_1^*, w_2^*, ..., w_{m^*}^*$ і рівні належності кожного спостереження $u_1^*(k), u_2^*(k), ..., u_{m^*}^*(k)$ до відповідного кластеру.

Для оцінки якості кластеризації кожним з елементів ансамблю може бути використаний будь-який з індексів нечіткої кластеризації [50], де одним з найбільш популярних є індекс Ксі-Бені (Xie-Beni index) [102], який для FCM-процедури в разі m кластерів може бути записаний у формі

$$XB^{[m]} = \frac{\sum_{k=1}^{N}\sum_{j=1}^{m} u_j^2(k)\left\|x(k) - w_j\right\|^2}{\min\limits_{l \neq j}\left\|w_j - w_l\right\|^2} = \frac{NXB^{[m]}}{DXB^{[m]}}. \tag{5.37}$$

Для послідовної обробки можна ввести online версію XB-індексу у вигляді

$$XB^{[m]}(k) = \frac{NXB^{[m]}(k)}{DXB^{[m]}(k)} =$$

$$= \frac{NXB^{[m]}(k-1)}{\min\limits_{l \neq j}\left\|w_j(k) - w_l(k)\right\|^2} + \frac{\frac{1}{k}\left(\sum_{j=1}^{m} u_j^2(k)\left\|x(k) - w_j(k)\right\|^2 - NXB^{[m]}(k-1)\right)}{\min\limits_{l \neq j}\left\|w_j(k) - w_l(k)\right\|^2}. \tag{5.38}$$

Чим менше значення (5.37), (5.38), тим вище якість кластеризації.

В процесі обробки даних вихідний блок ансамблю знаходить $FSOM^{[m^*]}$ з найкращим значенням $MEXB^{[m^*]}$ і результати роботи саме цієї нейро-фаззі мережі визначають кінцевий результат кластеризації.

5.3Результати моделювання

Для вирішення проблеми визначення оптимального числа кластерів у наборах даних ми використали запропонований ансамбль для онлайнового нечіткого кластерування. Ми обрали набір даних Wine з UCI Machine Repository [82]. Ці дані є результатами хімічного аналізу вин, вирощених в одному регіоні Італії, але отриманих з трьох різних сортів. Аналіз визначив кількість 13 компонентів, що знаходяться в кожному з трьох видів вин. Вона містить 178 спостережень.

Атрибути:

- алкоголь;
- яблучна кислота;
- ясень;
- лужність;
- магній;
- всього фенолів;
- флавоноїди;
- нефлаваноїдні феноли;
- проантоціани;
- інтенсивність кольору;
- відтінок;
- вміст OD280 / OD315;
- пролін.

На рис. 5.1 та 5.2 представлена візуалізація з використанням аналізу головних компонент (PCA-аналіз, три головні компоненти). На рис. 5.1 представлені результати роботі імовірнісного ансамблю нейро-фаззі мереж, де можна побачити, що цей підхід не виявив присутність третього кластера у даній вибірці.

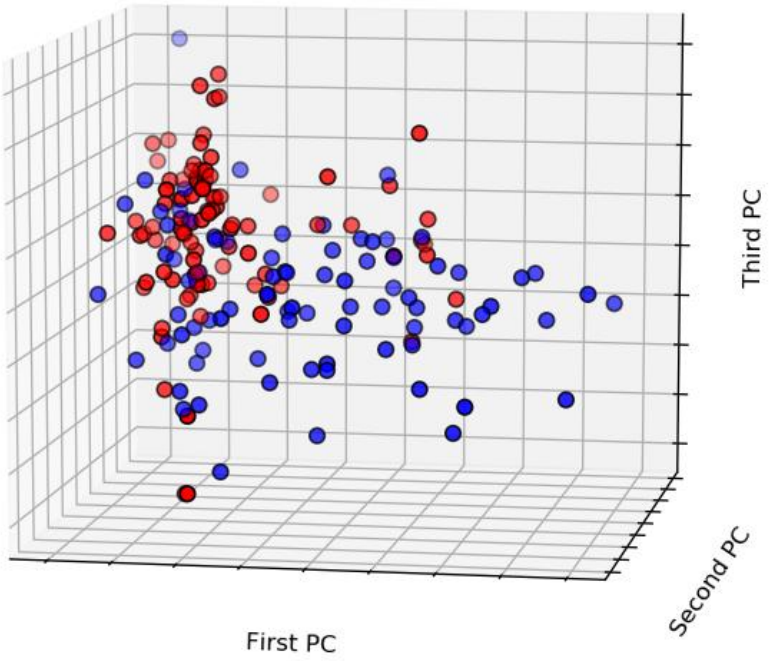

Рисунок 5.1 – Імовірнісний підхід для кластеризації даних

На рис. 5.2 представлені результати роботи можливісного підходу до кластеризації нейро-фаззі ансамблю мереж. За допомогою цього підходу точність кластеризації була більш точною, завдяки чому система змогла розпізнати три кластера на відміну від імовірнісного підходу.

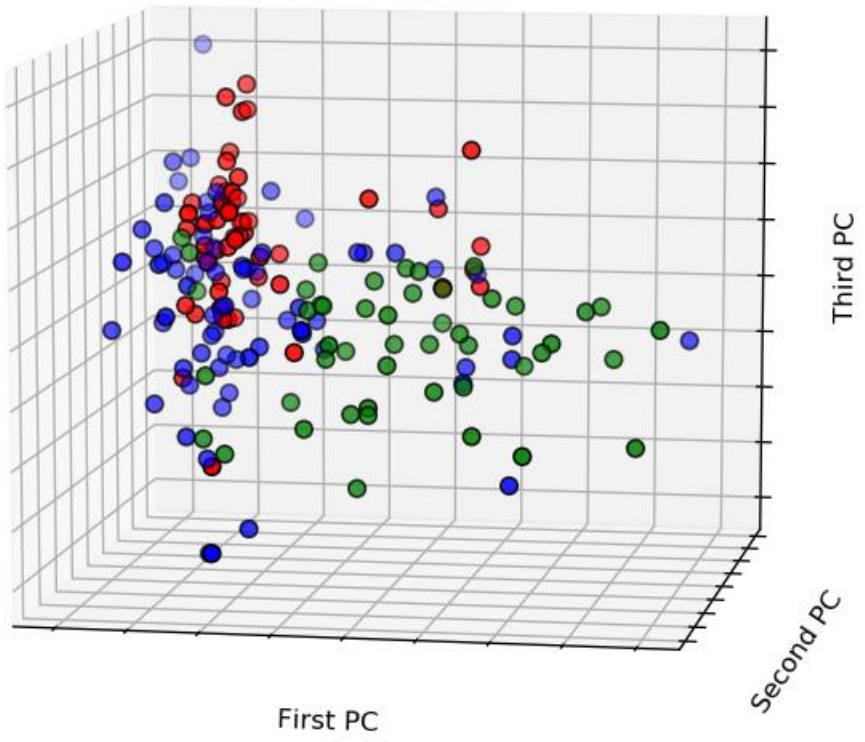

Рисунок 5.2 – Можливісний підхід для кластеризації даних

5.4 Висновки за розділом

1. Запропоновані архітектура та алгоритм самонавчання нейро-фаззі системи, призначеної для вирішення завдання online кластеризації потоку даних в умовах, коли кластери які формуються, перекриваються та їх число заздалегідь невідомо. Запропонована система є ансамблем нейро-фаззі самоорганізовних карт Т. Кохонена, кожна з яких відрізняється від інших кількістю нейронів. Налаштування кожного з членів ансамблю відбувається за допомогою модифікованого WTM правила самонавчання, при цьому в процесі налаштування проводиться автоматичне зважування усіх компонент оброблюваних векторів.

2. Запропонований підхід є узагальненням низки відомих процедур нечіткої імовірнісної та можливісної кластеризації та може бути використаний для вирішення задач аналізу потоків даних.

6 ІМІТАЦІЙНЕ МОДЕЛЮВАННЯ ТА ВИРІШЕННЯ ПРАКТИЧНИХ ЗАВДАНЬ З ВИКОРИСТАННЯМ АНСАМБЛЕВОГО ПІДХОДУ

У даному розділі наведені результати моделювання розроблених методів нейро-фаззі кластеризації даних високої розмірності. Імітаційне моделювання було виконано як на тестових вибірках так і з використанням реальних даних. Результати моделювання розроблених методів було порівняне з існуючими методами кластеризації даних.

Моделювання систем було виконано з використанням мови програмування Python 3.7. Python – це високорівнева мова програмування загального призначення, орієнтована на підвищення продуктивності розробника та читання коду. Синтаксис ядра Python мінімалістичний. Python підтримує кілька парадигм програмування: структурний, об'єктно-орієнтоване, функціональне, імперативне та аспектно-орієнтоване. У мові присутня динамічна типізація, автоматичне керування пам'яттю, повна інтроспекція, механізм обробки виключень, підтримка багатопоточних обчислень та зручні високорівневі структури даних. Програмний код на Python організовується у функції та класи, які можуть об'єднуватися в модулі, а вони в свою чергу можуть бути об'єднані в пакети. Мова Python зазвичай використовується як інтерпретуєма, але може бути скомпільована в байт-код Java і в MSIL (в рамках платформ. NET). За продуктивністю інтерпретуєма мова Python схожа на всі інші подібні мови, але можливість компіляції в байт-код дозволяє домогтися більшої продуктивності. У той же час стандартна бібліотека включає великий обсяг корисних функцій. Ця мова програмування насамперед має багато інструментів для моделювання та обробки даних, інтелектуального аналізу даних, візуалізації та іншого. Завдяки цьому стає все більш популярною у сфері штучного інтелекту, поступово витісняючи MATLAB, як універсальну платформу.

Всі розрахунки були зроблені в середовищі Spyder 3.3.2 – це вільна і кросплатформена інтерактивна IDE для наукових розрахунків на мові Python, що

забезпечує простоту використання функціональних можливостей і легковажність програмної частини.

Spyder є частиною модуля spyderlib для Python, заснованого на PyQt4, pyflakes, rope і Sphinx, що надає потужні віджети на PyQt4, такі як редактор коду, консоль Python (вбудована в додатку), графічний редактор змінних (в тому числі списків, словників і масивів) .

6.1 Імітаційне моделювання ансамблю кластерувальних мереж Т. Кохонена

Для підтвердження роботи розробленого ансамблю кластерувальних мереж на основі карт Т. Кохонена для потоків даних було використано декілька вибірок даних.

Были взяты наборы данных:

– Тестова вибірка «Іриси Фішера» із UCI-репозиторія [82]. Вибірка складається з даних о 150 екземплярах ірисів, по 50 екземплярів трьох видів – Iris setosa, Iris virginica та Iris versicolor. Для кожного екземпляра вимірувалися чотири характеристики:

1. Довжина чашолистка (англ. sepal length);
2. Ширина чашолистка (англ. sepal width);
3. Довжина пелюстки (англ. petal length);
4. Ширина пелюстки (англ. petal width).

– Набір даних морфометрії, який був отриманий у рамках науково-дослідної роботи, що фінансувалася Міністерством охорони здоров’я України «Встановити механізми адаптації до сполученої дії хімічних та фізичних чинників навколишнього середовища» (акт впровадження представлений у Додатку А). Загальна кількість щурів, що були досліджено складає 24. Кожен щур характеризується 20 ознаками (стан надниркового шару, стан селезінки, стан серця, печінки, легень, щитовидної залози та інші). Весь набір був розділений по 6 щурів на 4 групи, тобто, одна група була контроль, друга піддавалась електромагнітному

впливу, третя піддавалась холодовому впливу, та остання комбінованому впливу (електромагнітний та холодовий одночасно).

Для перевірки точності кластеризації ансамблю нейронних мереж Т. Кохонена використовувався критерій кластеризації Цаліньського-Харабаша. На основі цього критерію обирається та нейронна мережа Кохонена у якої значення цього критерію найвище. Середній результат для двох наборів даних наведено у таблиці 6.1. Ансамбль був налаштований на кількість кластерів від двох до восьми.

Таблиця 6.1 – Порівняння точності кластеризації на обраних наборах даних

Кількість кластерів	Набір даних «Іриси Фішера»	Набір даних «Морфометрія»
2	123,6	7,9
3	163,4	12,1
4	92,7	18,7
5	88,2	11,4
6	75,2	9,8
7	59,6	8,3
8	49,7	6,9

За результатами які наведено у таблиці 6.1 можна побачити, що для набору даних «Іриси Фішера» найкращий результат кластеризації дає карта Кохонена яка була налаштована на 3 кластера, цей показник найвищій. А для набору даних «Морфометрія» найкращій показник дає карта Кохонена, яка налаштована на 4 кластера. Залежність цих показників наведена на рисунку 6.1, де можна побачити як зменшується цей індекс в залежності від якості кластеризації.

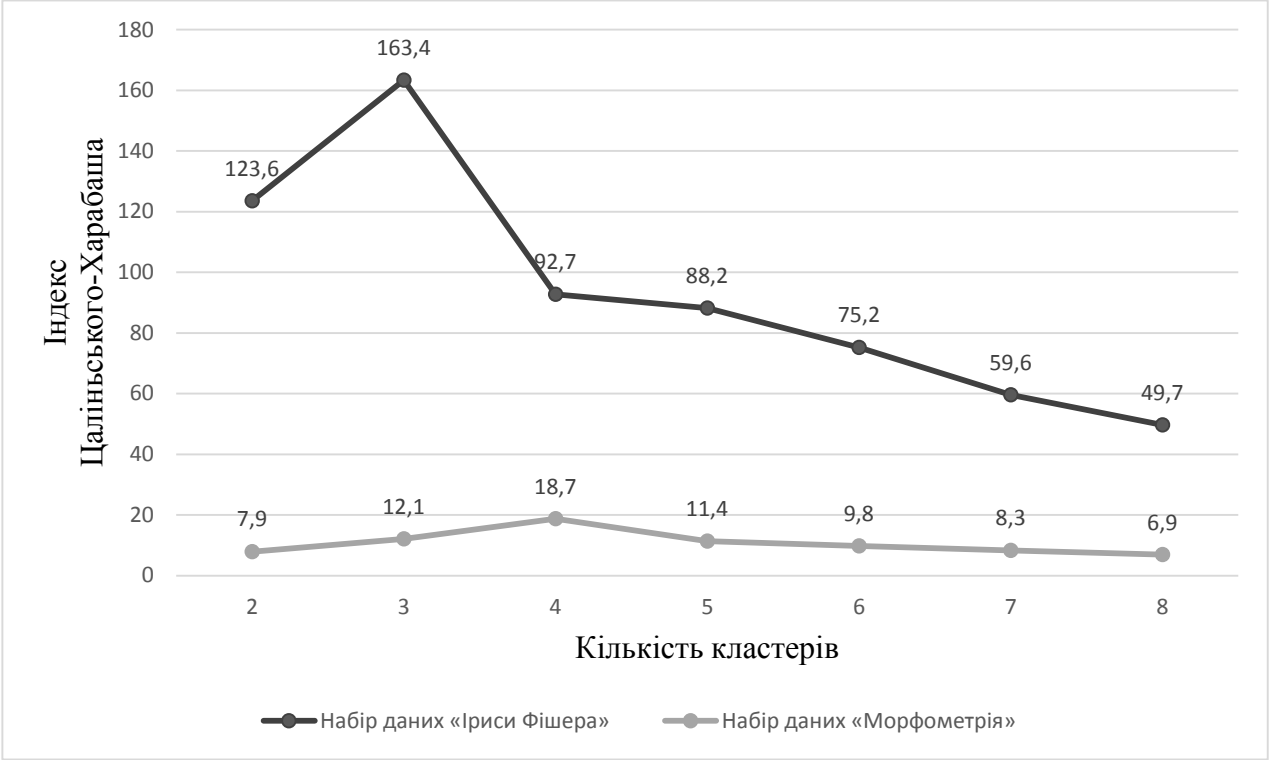

Рисунок 6.1 – Залежність індексу Цалінського-Харабаша від якості кластеризації

Візуалізація результату кластеризації набору даних «Іриси Фішера» карти Т. Кохонена, яка була налаштована на 3 кластери и була обрана за результатами індексу валідації у останньому шари ансамблю. Та візуалізація кластеризації набору даних «Морфометрія» наведена на рисунку 6.2 та 6.3 відповідно.

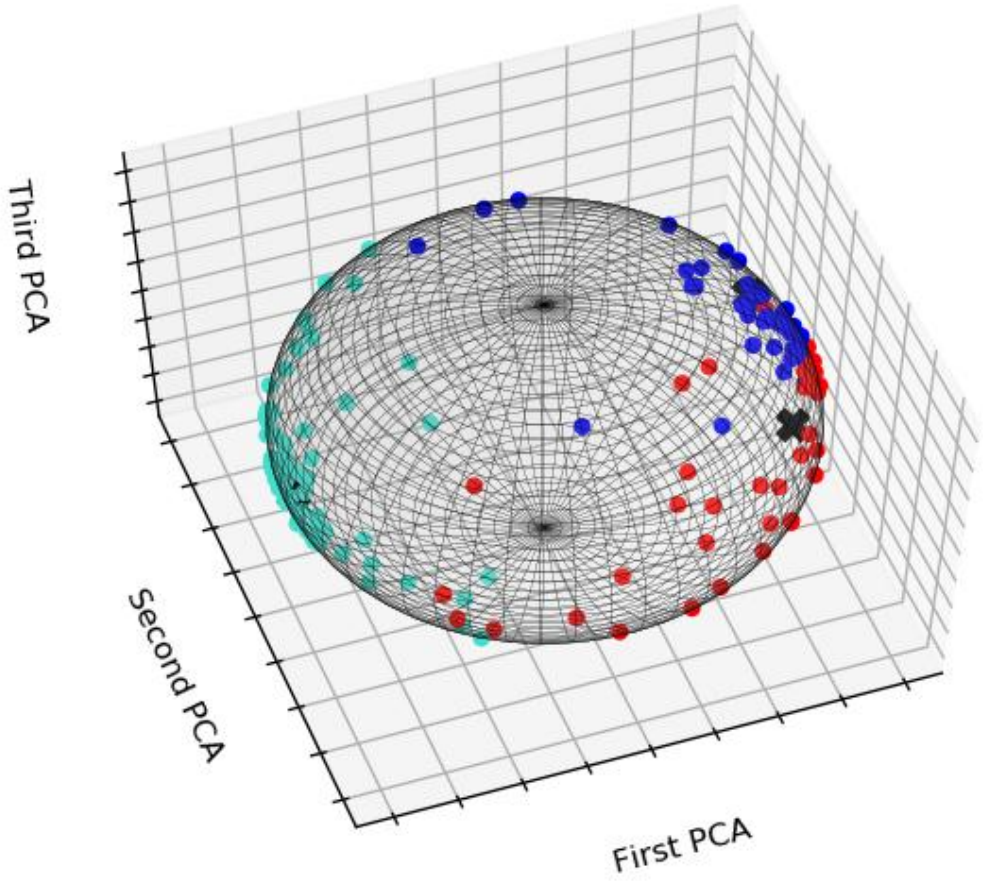

Рисунок 6.2 – Кластеризація набору даних «Іриси Фішера» на три кластера

За допомогою запропонованого підходу який базується на використанні ансамблю самоорганізовних карт Т. Кохонена для кластеризації даних коли кількість кластерів заздалегідь невідома, використавши індекс Цаліньського-Харабаша можна отримати найкращій результат кластеризації.

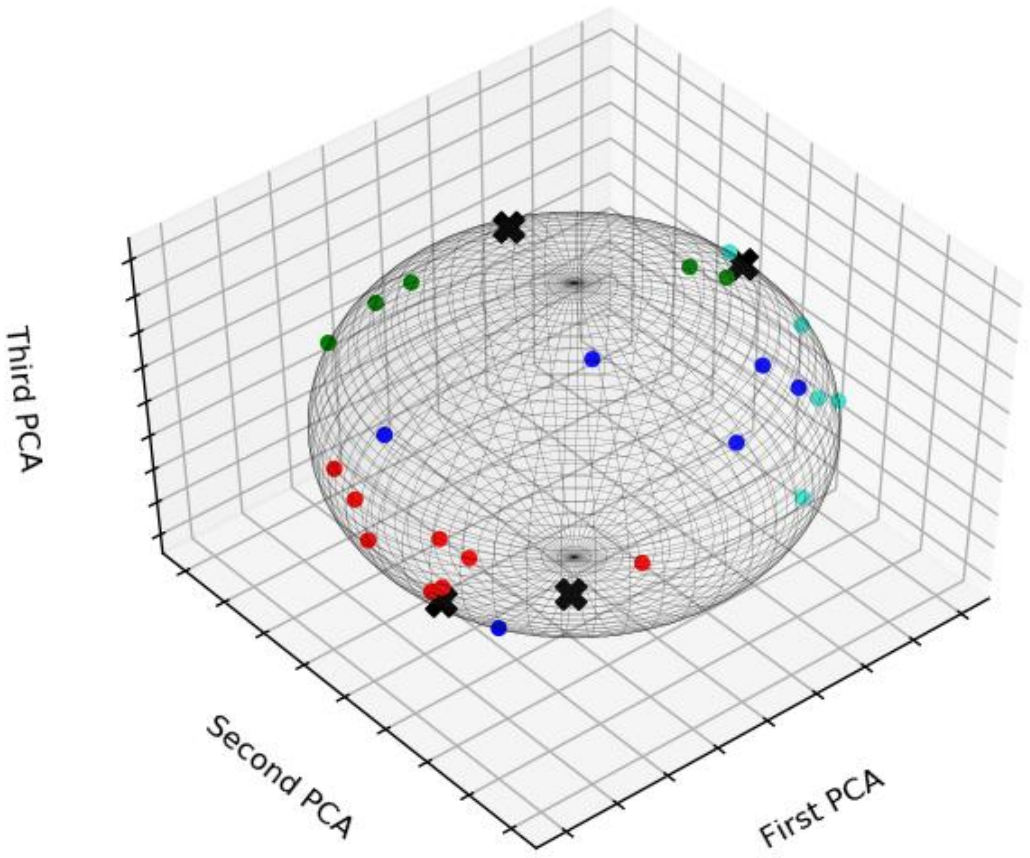

Рисунок 6.3 – Кластеризація набору даних «Морфометрія» на 4 кластера

6.2 Імітаційне моделювання ансамблю ядерних самоорганізовних карт Т. Кохонена для кластеризації потоків даних

Для підтвердження працездатності даного підходу, який заснований на ансамблевому підході з використанням ядерних функцій у першому шарі ансамблю, в якості яких були використані гаусіани для простоти реалізації з точки зору математичних розрахунків. Завдяки цьому шару вирішується проблема кластеризації даних коли класи є лінійно не роздільними та перетинаються у просторі ознак.

Ефективність запропонованого методу була досліджена з використанням набору даних, який було взято із UCI-репозиторію Pima Indians Diabetes [82].

Піма - група корінних американців, які живуть в Арізоні. Генетична схильність давала змогу цій групі нормально виживати на дієті, яка була бідна

вуглеводами протягом багатьох років. В останні роки, через різкий перехід від традиційних сільськогосподарських культур до оброблених харчових продуктів, разом із зниженням фізичної активності, вони набули найбільшу поширеність діабету 2 типу і з цієї причини вони були предметом багатьох досліджень.

Набір даних містить дані від 768 жінок з 8 характеристиками, зокрема:

– скільки разів були вагітні;

– концентрація глюкози в плазмі через 2 години в пероральному тесті на толерантність до глюкози;

– діастолічний артеріальний тиск (мм рт.ст.);

– товщина складки шкіри трицепса (мм);

– 2-годинний сироватковий інсулін (мкО / мл);

– індекс маси тіла (вага в кг / (висота в м)^2)

– функція племінного діабету;

– вік (років).

Кожну мережу Т. Кохонена було налаштована на свою кількість кластерів від 2 до 5, та за допомогою критерія валідації, а саме індексу Девіса-Булдена [79] обиралась та мережа у якої цей показник був найменший. Проміжні дані наведену у таблиці 6.2 у який наведено показник цього індексу для кожної мережі запропонованого ансамблю.

Таблиця 6.2 – Значення індексу Девіса-Булдена для набора даних Pima Indians Diabetes

Кількість кластерів	Значення індексу Девіса-Булдена
2	1,148
3	1,296
4	2,192
5	2,507

Використовуючи дані які наведено у таблиці 6.2 можна побудувати графік залежності зростання індексу Девіса-Булдена у ситуації помилкової кластеризації. Можна побачити, що мережа Т. Кохонена яка налаштована на 2 кластери дає найменший індекс валідації, що відповідає правильній кластеризації вибраного набору даних. Цей графік наведено на рисунку 6.4.

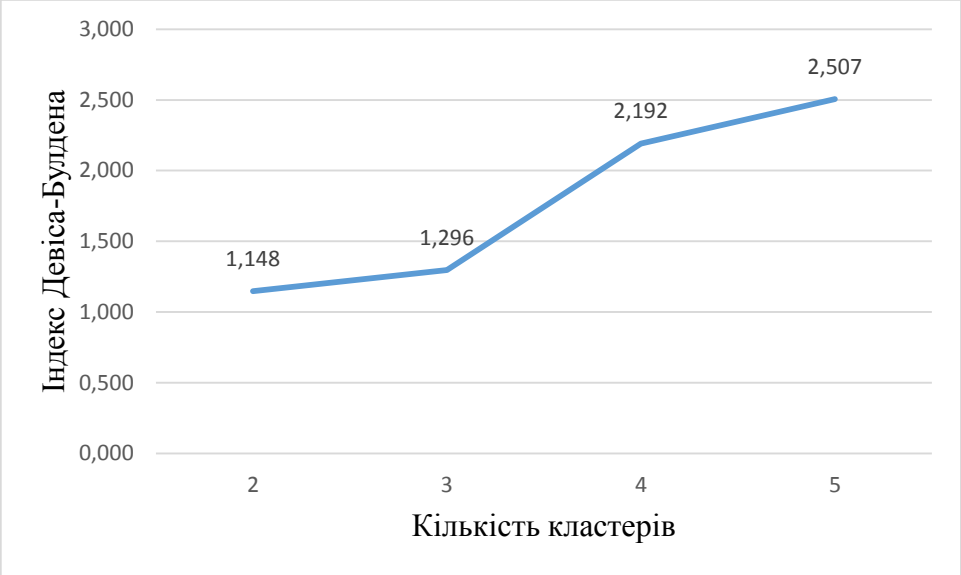

Рисунок 6.4 – Залежність індексу валідації від кількості кластерів з використанням набору даних Pima Indians Diabetes

На рисунку 6.5 представлено візуалізацію набору даних Pima Indians Diabetes на 2 класи, що було визначено завдяки індексу Девіса-Булдена.

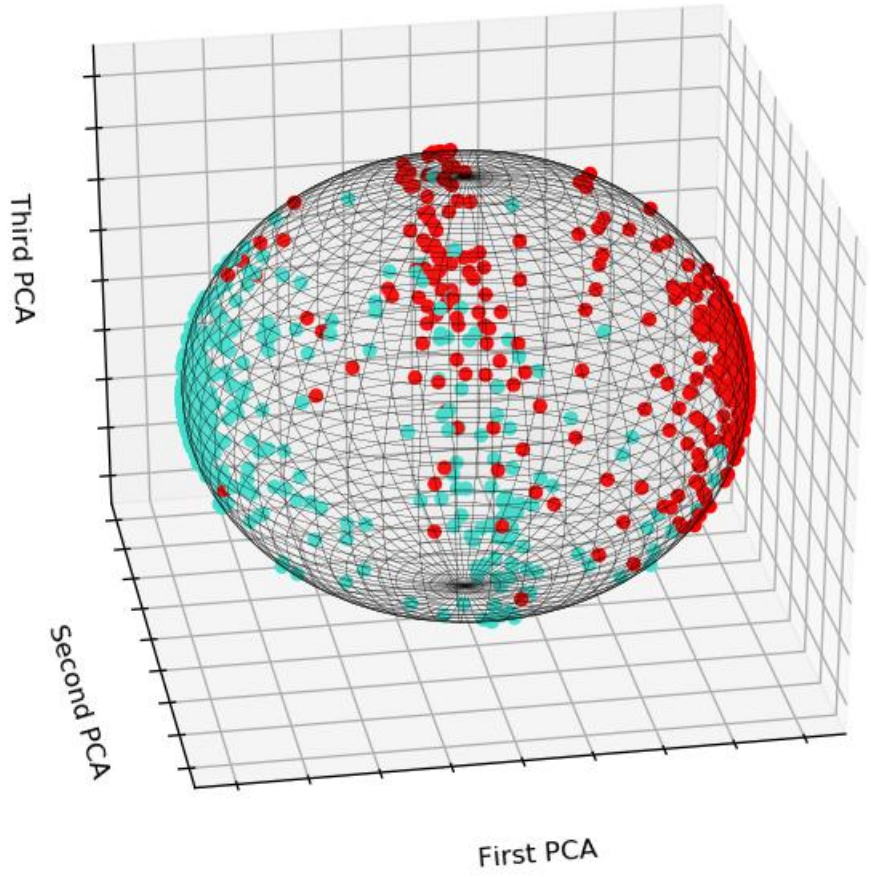

Рисунок 6.5 – Візуалізація кластеризації набору даних Pima Indians Diabetes на 2 класи

6.3 Імітаційне моделювання ансамблю нейро-фаззі самоорганізовних карт Т. Кохонена для кластеризації потоку даних

Ефективність запропонованого методу, який заснований на ансамблевому підході до кластеризації потоків даних, була досліджена на наборах даних, які було отримано у ході деяких досліджень, а саме:

1. Медичний набір даних був отриманий у КЗОЗ «Харківська міська клінічна лікарня №13» м. Харкова шляхом обробки форми 027/о – форма надходження пацієнту до стаціонару лікарні. Загальна кількість пацієнтів складає 132 людини. Кожен пацієнт характеризується 104 ознаками (стать, вік, скарги пацієнта – 24 ознаки, анамнез – 14 ознак, об'єктивний опис стану пацієнта – 26

ознак, клінічний аналіз крові – 10 ознак, біохімічний аналіз крові – 8 ознак, клінічний аналіз сечі – 10 ознаки, рентгенографія грудної клітки – 6 ознак, електрокардіограма (ЕКГ) – 8 ознак, спірометрія – 2 ознаки). Увесь набір даних був поділяється на три групи-діагнози: 46 пацієнтів мають хронічне обструктивне захворювання легень (ХОЗЛ), 53 хворих – бронхіальну астму та 33 пацієнти хворіють на пневмонію.

2. Набір даних морфометрії, який був отриманий у рамках НДР бюджетного фінансування «Встановити механізми адаптації до сполученої дії хімічних та фізичних чинників навколишнього середовища».

Результати кластеризації запропонованого ансамблю нейро-фаззі самоорганізовних карт Т. Кохонена були оцінені за допомогою критерія валідації, а саме за допомогою індексу Ксі-Бені. Налаштування кожної карти Т. Кохонена було в діапазоні від 2 до 7 кластерів. Результати кластеризації запропонованого ансамблю були порівняні з стандартним методом кластеризаціїї К-середніх та наведено у таблицях 6.3 та 6.4.

Таблиця 6.3 – Порівняння роботи методів кластеризації для набору даних «Пульмонологія»

Кількість кластерів	Алгоритм	Індекс Ксі-Бені
2	MHDFCM	23,26
	К-середні	2741,05
3	MHDFCM	11,59
	К-середні	26193,92
4	MHDFCM	163,60
	К-середні	6374357,80
5	MHDFCM	232,34
	К-середні	593006,54
6	MHDFCM	29359,38
	К-середні	15226854,27

Продовження таблиці 6.3

7	MHDFCM	1189,79
	К-середні	65445664,69

За даними таблиці був побудований графік залежності індексу Ксі-Бені від кількості кластерів, на яку налаштована кожна з карт Т. Кохонена, можна спостерігати зростання цього індексу при зростанні помилки кластеризації даних. Цей результат наведений на рисунку 6.6.

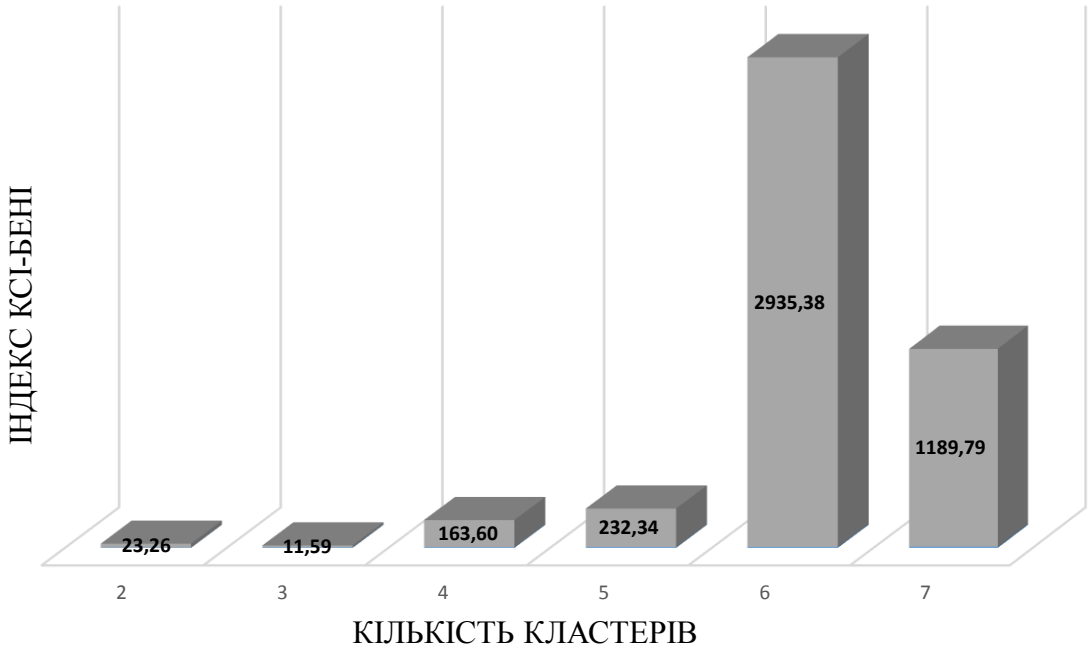

Рисунок 6.6 – Показник індексу Ксі-Бені в залежності від кількості кластерів для набору даних «Пульмонологія»

Таблиця 6.4 – Порівняння роботи методів кластеризації для набору даних «Морфометрія»

Кількість кластерів	Алгоритм	Індекс Ксі-Бені
2	MHDFCM	0,00015
	К-середні	303,12752

Продовження таблиці 6.4

3	MHDFCM	0,00124
	К-середні	713,51465
4	MHDFCM	0,00012
	К-середні	3129,59935
5	MHDFCM	0,00032
	К-середні	73510,69741
6	MHDFCM	0,00900
	К-середні	648536,96584
7	MHDFCM	0,00260
	К-середні	28954,99470

За допомогою даних які наведені у таблиці 6.3 був побудований графік на якому можна побачити залежність індексу Ксі-Бені від якості кластеризації. Як видно, що чим менший показник критерію валідації, тим краща якість роботи карти Т. Кохонена для кластеризації потоків даних. Ця залежність наведена на рисунку 6.7

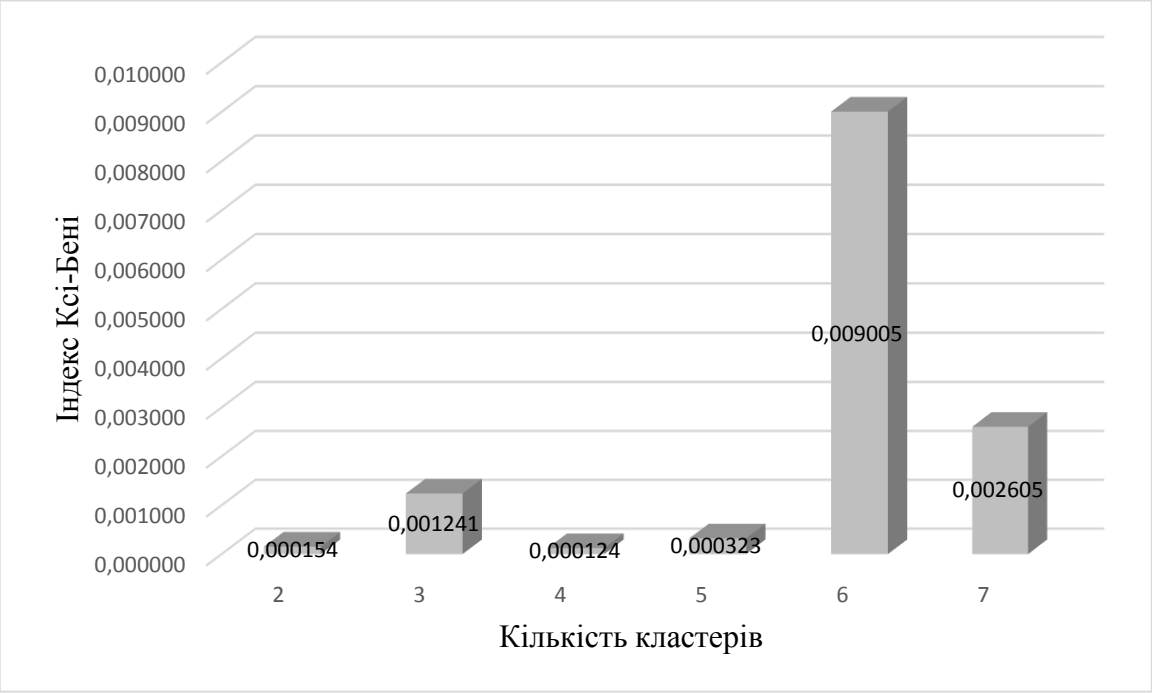

Рисунок 6.7 – Якість кластеризації набору даних «Морфометрія» в залежності від кількості кластерів

Як можна побачити у таблиці 6.3, що найменший індекс Ксі-Бені в наборі даних «Пульмонологія» був результатом роботи карти Т. Кохонена, яка була налаштована на 3 кластери, згідно з цим можна зробити висновок, що саме ця мережа дає вірний результат кластеризації. Візуалізація роботи саме цієї мережі підтверджує цей результат и представлений на рисунку 6.8.

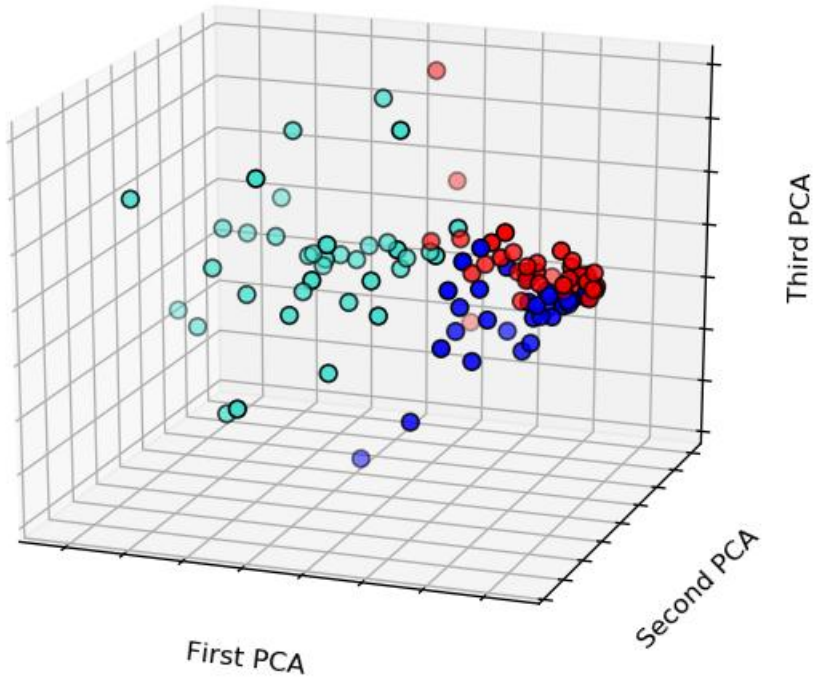

Рисунок 6.8 – Кластеризація набору даних «Пульмонологія» на три кластери

Також, аналізуючи дані, які були отримані у ході роботи ансамблю нейро-фаззі самоорганізовних карт Т. Кохонена з набором даних «Морфометрія», можно побачити, що найкращій результат кластеризації у мережі з налаштуванням на 4 кластери, що підтверджує результати, які були отримані у ході дослідження щурів у рамках науково-дослідної роботи, що фінансувалася Міністерством охорони здоров’я України «Встановити механізми адаптації до сполученої дії хімічних та фізичних чинників навколишнього середовища». Це підтверджує індекс Ксі-Бені який дав найменший результат кластеризації на 4 класи. Це також підтверджує візуалізація яка представлена на рисунку 6.8.

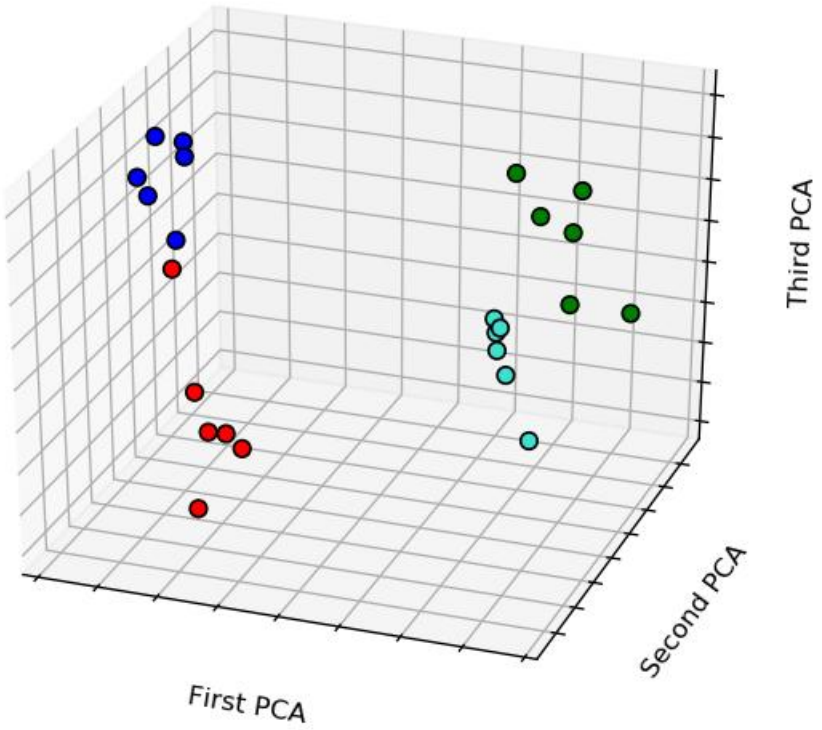

Рисунок 6.8 – Кластеризація набору даних «Морфометрія» на чотири кластери

6.4 Імітаційне моделювання нейро-фаззі мереж Т. Кохонена з використанням імовірнісно-можливісного підходу

Ефективність даного методу, який заснований на ансамблі нейро-фаззі карт Т. Кохонена з використанням імовірнісно-можливісного підходу, що одночасно дає змогу обробляти дані, які надходять на вхід системи, використовуючи одночасно ці два методи. Саме завдяки цьому методу кількість помилок зменшується, а точність кластеризації зростає. У таблиці 6.4 наведено порівняння роботи ансамблю нейро-фаззі самоорганізовних карт Т. Кохонена та ансамблю нейро-фаззі карт Т. Кохонена з використанням імовірнісно-можливісного підходу.

Для порівняння роботи цих ансамблів використовувався набір даних «Dermatology» із UCI-репозиторію [82].

Цей набір даних містить 34 атрибути, 33 з яких є лінійними та один з них номінальний.

Диференціальна діагностика еритемато-плоскоклітинних захворювань є реальною проблемою в дерматології. Всі вони поділяють клінічні особливості еритеми та масштабування, з дуже невеликими відмінностями. Захворювання в цій групі – псоріаз (112 пацієнтів), себорейний дерматит (61 пацієнт), червоний плоский лишай (72 пацієнти), рожевий лишай (49 пацієнтів), хронічний дерматит (52 пацієнти) та висівкоподібний волосяний лишай (20 пацієнтів). Зазвичай для діагностики необхідна біопсія, але, на жаль, ці захворювання також мають багато гістопатологічних особливостей. Інші труднощі для диференціальної діагностики полягає в тому, що захворювання може показати особливості іншого захворювання на початковому етапі та може мати характерні ознаки на наступних стадіях. Пацієнтів вперше оцінювали клінічно з 12 ознаками. Після цього брали зразки шкіри для оцінки 22 гістопатологічних ознак. Значення гістопатологічних ознак визначають шляхом аналізу зразків під мікроскопом.

У наборі даних, функція родинної історії має значення 1, якщо будь-яке з цих захворювань спостерігається у родині, та 0 в іншому випадку. Вікова функція просто відображає вік пацієнта. Кожної наступної ознаці (клінічному та гістопатологічному) було надано ступінь в діапазоні від 0 до 3. Тут 0 вказує, що ознака не була присутня, 3 вказує найбільшу можливу кількість, а 1, 2 – відносні проміжні значення.

У таблиці 6.5 наведено значення індексу валідації для двох запропонованих методів, які базуються на використанні ансамблю нейро-фаззі самоорганізовних карт Т. Кохонена. Modify high-dimension Fuzzu C-means (MHDFCM) використовує тільки імовірнісний підхід для кластеризації потоків даних. Modify high-dimension Fuzzu C-means probabilistic-possibilistic (MHDFCMpr-pos) базується на використанні одночасно двох підходів імовірнісного та можливісного, що дозволяє запобігати помилок при кластеризації, коли наступне спостереження знаходиться на однаковій відстані від усіх кластерів.

Таблиця 6.5 – Порівняння роботи MHDFCM та MHDFCMp-p

Метод кластеризації	Кількість кластерів	Індекс Ксі-Бені
MHDFCM	3	0,113
MHDFCMpr-pos		0,082
MHDFCM	4	0,375
MHDFCMpr-pos		0,166
MHDFCM	5	1,498
MHDFCMpr-pos		0,007
MHDFCM	6	0,178
MHDFCMpr-pos		0,005
MHDFCM	7	0,331
MHDFCMpr-pos		0,332
MHDFCM	8	1,385
MHDFCMpr-pos		0,280

Використовуючи дані, які знаходяться у таблиці 6.5, був побудований графік відношення критерію валідації та кількості кластерів з використанням обох підходів. Нескладно побачити, що робота ансамблю нейро-фаззі мереж з використанням імовірнісно-можливісного підходу дає кращій результат. Дані наведено на рисунку 6.10.

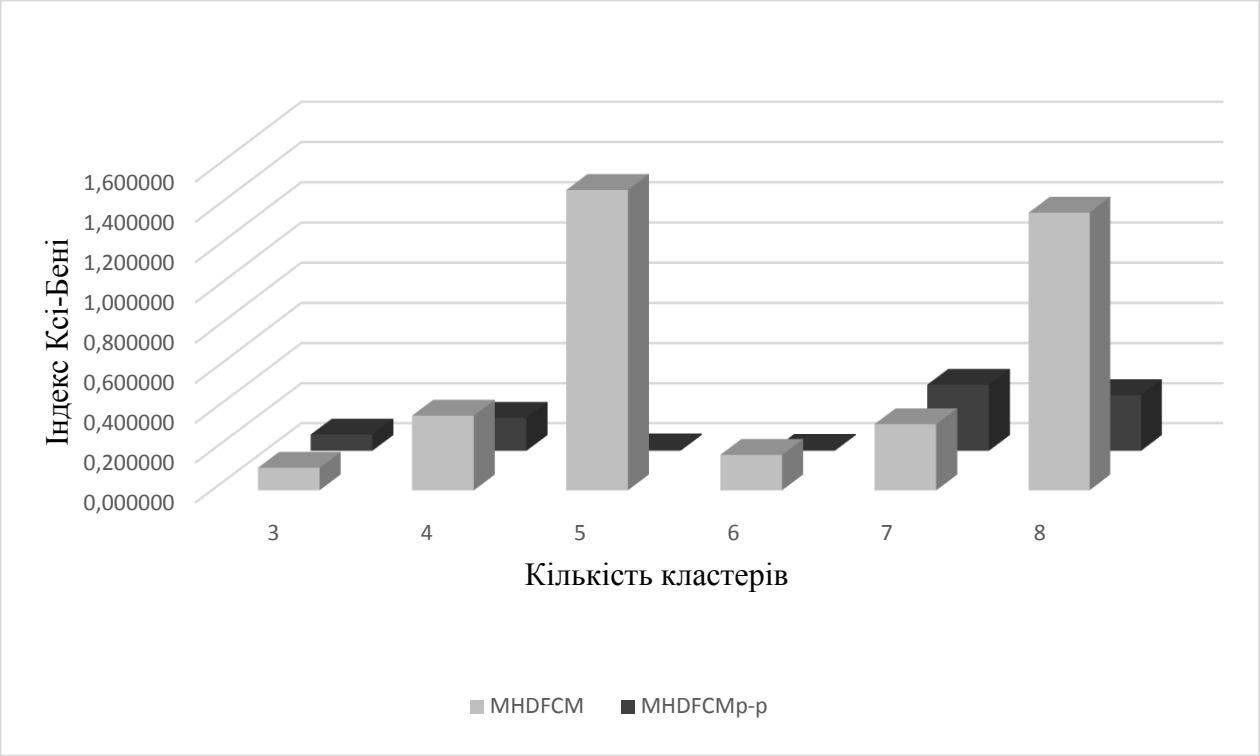

Рисунок 6.10 – Графік залежності індексу Ксі-Бені від кількості кластерів

Даний підхід не тільки дозволяє отримати більш точний індекс кластеризації, але й отримати більш точний результат кластеризації з меншою кількістю помилок. Це можна побачити на рисунках 6.11 та 6.12.

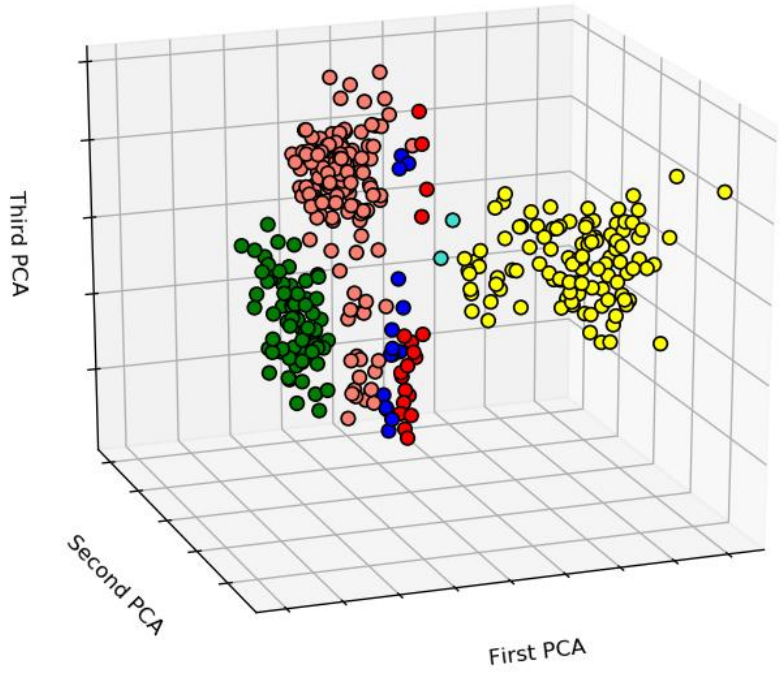

Рисунок 6.11 – Візуалізація набору даних «Дерматологія» на 6 кластерів за допомогою ансамблю нейро-фаззі самоорганізовних карт Т. Кохонена

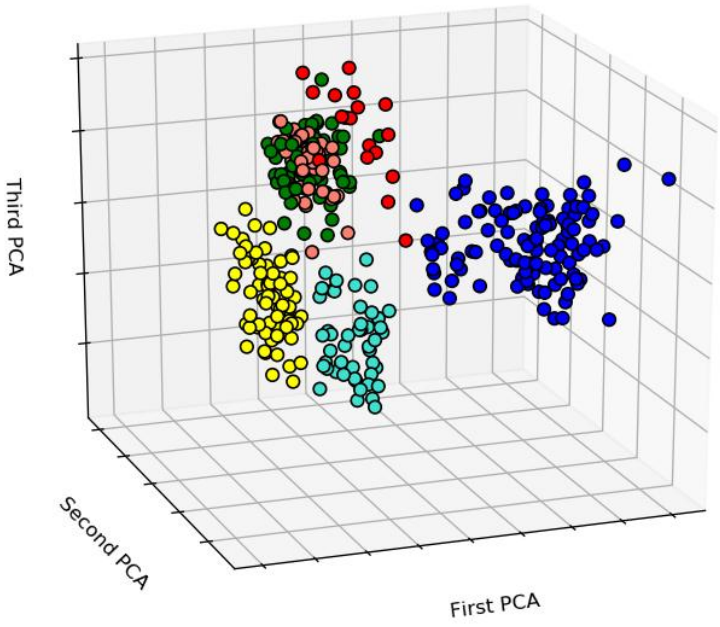

Рисунок 6.10 – Візуалізація набору даних «Дерматологія» на 6 кластерів за допомогою ансамблю нейро-фаззі мереж Т. Кохонена з використанням імовірнісно-можливісного підходу

6.5 Висновки за розділом

1. Була вирішена задача кластеризації на основі тестових вибірок з UCI-репозиторія за допомогою розробленого ансамблю самоорганізовних карт Т. Кохонена для кластеризації потоків даних коли кількість класів апріорно невідома.

2. Проведено імітаційне моделювання рішення задачі нечіткої кластеризації даних, на основі вибірки отриманої у рамках НДР бюджетного фінансування «Встановити механізми адаптації до сполученої дії хімічних та фізичних чинників навколишнього середовища», за допомогою запропонованого ансамблю самоорганізовних карт Т. Кохонена для кластеризації потоків даних.

3. Проведено імітаційне моделювання кластеризації даних на основі ансамблю нейро-фаззі самоорганізовних карт Т. Кохонена з використанням імовірнісно-можливісного підходу, що дозволяє обробляти дані коли апріорно невідома кількість та форма класів.

ВИСНОВКИ

У дисертаційній роботі представлені результати, які є відповідно до поставленої мети рішенням актуального завдання обробки багатовимірних масивів даних в умовах невизначеності за допомогою ансамблю нечітких методів ядерної кластеризації на основі ядерних функцій. Проведені дослідження дозволили зробити наступні висновки.

1. Розроблено ансамбль самоорганізовних карт Т. Кохонена для кластеризації даних за умов апріорі невідомої кількості класів з використанням онлайн модифікованого методу К-середніх;

2. Розроблено ансамбль нейро-фаззі самоорганізовних карт Т. Кохонена для кластеризації потоків даних за умов коли класи є лінійно нероздільними та довільним чином перетинаються у просторі ознак, що базується на використанні онлайн методу нечітких С-середніх за умов коли кількість кластерів апріорі невідома;

3. Розроблено ансамбль ядерних самоорганізовних карт Т. Кохонена для кластеризації потоків даних за умов коли кластери є лінійно нероздільними, що характеризується введенням додаткового ядерного шару для підвищення розмірності вхідного простору;

4. Розроблено ансамбль самоорганізовних нечітких карт Т. Кохонена, що одночасно реалізує процедури імовірнісної та можливісної кластеризації потоків даних;

5. Розроблено ансамблю нейро-фаззі мереж на основі імовірнісно-можливісного підходу для кластеризації потоків даних;

6. Проведено імітаційне моделювання розроблених методів та моделей та рішення практичних задач нечіткої кластеризації потоків даних високої розмірності використовуючи дані, які було взято з UCI-репозиторію.

7. Було проведено апробацію запропонованих методів на реальних даних для кластеризації даних високої розмірності, які були отримані у ході дослідження

щурів у рамках НДР бюджетного фінансування «Встановити механізми адаптації до сполученої дії хімічних та фізичних чинників навколишнього середовища».

8. Вирішено практичне завдання кластеризації даних. Порівняння з іншими методами кластеризації розроблені методи показали простоту реалізації з точки зору математичного апарату та підвищення точності кластеризації потоків даних.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] P. Zhernova, A. Deyneko, Z. Deyneko, I. Pliss and V. Ahafonov, "Data Stream Clustering in Conditions of an Unknown Amount of Classes," in *I Advances in Intelligent Systems and Computing*, Springer, Cham, 2019, pp. 410-419.

[2] Є. Бодянський, А. Дейнеко, П. Жернова, О. Золотухін та Я. Хаустова, «Послідовне ядерне нечітке кластерування великих масивів даних на основі гібридної системи обчислювального інтелекту,» *Вісник Національного університету "Львівська політехніка". Інформаційні системи та мережі,* № 829, pp. 20-24, 2017.

[3] Є. Бодянский, А. Дейнеко, П. Жернова та В. Репін, «Онлайн модифікація методу Х-середніх на основі ансамблю самоорганізовних мап Т. Когонена,» *Збірник наукових праць «Розвиток транспорту»,* № 1, pp. 96-107, 2017.

[4] П. Жернова та Є. Бодянський, «Ядерна нечітка кластеризація потоків даних на основі ансамблю нейронних мереж,» *Сучасний стан наукових досліджень та технологій в промисловості,* № 4(6), pp. 42-49, 2018.

[5] Y. Bodyanskiy, I. Perova and P. Zhernova, "Online fuzzy clustering of high - dimensional data based on ensembles in data stream mining tasks," *Innovative Technologles & Scientific Solutions for Industries,* no. 1(7), pp. 16-24, 2019.

[6] П. Жернова та Є. Бодянський, «Нечітка імовірнісно-можливісна послідовна кластеризація даних на основі ансамблевого підходу,» *Науково-технічний журнал «Прикладна радіоелектроніка»,* № 1,2, pp. 40-45, 2019.

[7] Е. Бодянский, А. Дейнеко, П. Жернова и В. Репин, «Адаптивная модификация метода Х-средних на основе ансамбля кластеризующих нейронных сетей Т. Кохонена,» в *Матеріали VI Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології»*, Одесса, 2017.

[8] Е. Бодянський, П. Жернова и А. Дейнеко, «Кластеризующий ансамбль нейронных сетей и его обучение в условиях неизвестного количества классов,» в *Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту»Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту»*, Залізний порт, Україна, 2018.

[9] А. Дейнеко, П. Жернова, І. Плісс та О. Чала, «Модифікована нечітка ймовірнісна нейронна мережа,» в *Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту»*, Залізний порт, Україна, 2018.

[10] P. Zhernova, A. Deyneko, Y. Bodyanskiy and V. Riepin, "Adaptive kernel data streams clustering based on neural networks ensembles in conditions of uncertainty about amount and shapes of clusters," in *IEEE Second International Conference on Data Stream Mining & Processing*, Lviv, Ukraine, 2018.

[11] A. Deineko, P. Zhernova, B. Gordon, O. Zayika, I. Pliss and N. Pabyrivska, "Data stream online clustering based on fuzzy expectation-maximization approaching formation on submission," in *IEEE Second International Conference on Data Stream Mining & Processing*, Lviv, Ukraine, 2018.

[12] П. Жернова та А. Лобинцев, «Кластеризація даних високої розмірності з використанням можливісного підходу,» в *Матеріали 23-го Міжнародного молодіжного форуму «Радіоелектроніка та молодь в 21 столітті»*, Харьков, 2019.

[13] П. Жернова та Є. Бодянський, «Нейро-фаззі мережа та її навчання для кластеризації потоків даних високої розмірності,» в *Матеріали V міжнародної науково-практичної конференції «Обчислювальний інтелект (результати, проблеми, перспективи)»*, Ужгород, 2019.

[14] J. Hwang, S. Lay, M. Maechler, R. Martin and J. Schimert, "Regression modeling in back-propagation and projection pursuit learning," in *IEEE Transactions on Neural Networks*, 1994.

[15] Ф. Розенблатт, «Модель памяти на нейронных сетях,» *Автоматика,* № 5(2), pp. 2-4, 1966.

[16] R. Roseline, G. Jenitha and J. Henri Amirhtaraj, "Analysis and Application of Clustering Techniques in Data Mining," *International Journal of Computing Algorithm*, 2014.

[17] S. Guha, R. Rastogi and R. Shim, "CURE: An Efficient Clustering Algorithm for Large Databases»," in *SIGMOD '98 Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, Seattle, Washington, 1998.

[18] S. Sarmah and D. Bhattacharyya, "An Effective Technique for Clustering Incremental Gene Expression data," *IJCSI International Journal of Computer Science Issues*, 2010.

[19] Suman and M. Mittal, "Comparison and Analysis of Various Clustering Methods in Data mining On Education data set using the weak tool," *International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)*, 2014.

[20] B. Chaudhari and M. Parikh, "A Comparative Study of clustering algorithms Using weka tools," *International Journal of Application or Innovation in Engineering & Management (IJAIEM)*, 2012.

[21] L. Rutkowski, Computational Intelligence. Methods and Techniques, Berlin-Heidelberg: Springer-Verlag, 2008.

[22] M. Verma, M. Srivastava, N. Chack, D. Kumar and G. Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining," *International Journal of Engineering Research and Applications (IJERA)*, 2012.

[23] A. Joshi and R. Kaur, "Comparative Study of Various Clustering Techniques in Data Mining," *International Journal of Advanced Research in Computer Science and Software Engineering*, 2013.

[24] И. А. Чубукова, Data Mining, Москва: Открытые системы, 2006, p. 382.

[25] R. Kruse, C. Borgelt, F. Klawonn, C. Moewes, M. Steinbrecher and P. Held, Computational Intelligence. A Methodological Introduction, Berlin: Springer, 2013.

[26] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data, Englewood Cliffs, N.J.: Prentice-Hall, 1988.

[27] O. Nelles, Nonlinear System Identification, Berlin: Springer, 2001.

[28] J. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, N.Y.: Plenum Press, 1981, p. 272.

[29] R. Babuska, Fuzzy Modeling and Identification. Delft, The Netherlands: Delft University of Technology, 1996.

[30] V. Patel and R. Mehta, "Impact of Outlier Removal and Normalization Approach in Modified K-Means Clustering Algorithm," *IJCSI International Journal of Computer Science Issues*, pp. 331-336, 2011.

[31] S. Chakraborty and N. Nagwani, "Analysis and Study of Incremental DBSCAN Clustering Algorithm," *International Journal of Enterprise Computing and Business Systems*, 2011.

[32] S. Mahmud, M. Rahman and A. N., "Improvement of K-means Clustering algorithm with better initial centroids based on weighted average," in *7th International Conference on Electrical and Computer Engineering*, 2012.

[33] M. Yedla, S. Pathakota and T. Srinivasa, "Enhancing K-means Clustering Algorithm with Improved Initial Center," *International Journal of Computer Science and Information Technologies(IJCSIT)*, pp. 121-125, 2010.

[34] S. Kushwah, K. Rawat and P. Gupta, "Analysis and Comparison of Efficient Techniques of Clustering Algorithms in Data Mining," *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 2012.

[35] J. A. Hartigan and M. A. Wong, "A k-means clustering algorithm," *Applied Statistics*, no. 28, pp. 100-108, 1979.

[36] D. Virmani, S. Taneja and G. Malhotra, "Normalization based K means Clustering Algorithm," *International Journal of Advanced Engineering Research and Science (IJAERS)*, 2015.

[37] B. Gu and V. Sheng, "Feasibility and Finite Convergence Analysis for Accurate On-Linev-Support Vector Machine," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1304 - 1315, 2013.

[38] M. Deng, Q. Liu, T. Cheng and Y. Shi, "An adaptive spatial clustering algorithm based on Delaunay triangulation," *Comput.Environ. UrbanSyst*, p. 320–332, 2011.

[39] M. Anderberg, Cluster Analysis for Applications, Academic Press, 1973.

[40] G. Golub and C. Van Loan, Matrix Computations, Johns Hopkins University Press, 1996, p. 728.

[41] K.-L. Du and M. N. S. Swamy, Neural Networks and Statistical Learning, London: Springer-Verlag, 2014.

[42] F.-H. Rhee, "Uncertain fuzzy clustering: insights and recommendations," *IEEE. Comp. Intel. Mag.*, no. 2(1), pp. 4- 56, 2007.

[43] F. Höppner, F. Klawonn and R. Kruze, Fuzzy Clusteranalyse, Braunschweig: Vieweg, 1999, p. 280.

[44] T. Kohonen, Self-Organizing Maps, Berlin: Springer-Verlag, 1995, p. 362.

[45] W. Sarle, "Measurement theory: Frequently asked questions," 1996.

[46] S. Gallant, Neural Networks Learning and Expert Systems, Cambridge: MIT Press, 1987.

[47] H. Ritter, T. Martinetz and K. Schulten, Neuronale Netze, Bonn: Addison-Wesley, 1991.

[48] R. M. Golden, Mathematical Methods for Neural Network Analysis and Design, Cambridge, Massachussetts: The MIT Press, 1996.

[49] C. C. Aggarwal and C. K. Reddy, Data Clustering. Algorithms and Application, Boca Raton: CRC Press, 2014.

[50] R. Xu and D. Wunsch, Clustering, Hoboken, NJ: John Wiley & Sons, Inc., 2009, p. 370.

[51] P. Vuorimaa, "Fuzzy self-organizing maps," *Fuzzy Sets and Systems*, no. 66, pp. 223-231, 1994.

[52] G. Gan, C. Ma and J. Wu, Data Clustering: Theory, Algorithms and Application, Philadelphia: SIAM, 2007.

[53] J. Abonyi and B. Feil, Cluster Analysis for Data Mining and System Identification, Basel: Birkhauser, 2007.

[54] D. L. Olson and D. Dursun, Advanced Data Mining Techniques, Berlin: Springer, 2008.

[55] Y. Jin and B. Hammer, "Computational Intelligence in Big Data," *IEEE Computational Intelligence Magazine*, pp. 12-15, August 2014.

[56] D. Pelleg and A. Moor, "X-means: extending K-means with efficient estimation of the number of clusters," in *Proc. 17th Int. Conf. on Machine Learning*, San Francisco, 2000.

[57] T. Ishioka, "An expansion of X-means for automatically determining the optimal number of clusters," in *Proc. 4th IASTED Int. Conf. Computational Intelligence*, Calgary, Alberta, 2005.

[58] A. Bifet, Adaptive Stream Mining: Pattern Learning and Mining from Evolving Data Streams, Amsterdam: IOS Press, 2010, p. 224.

[59] J. Kacprzyk and W. Pedrycz, Springer Handbook of Computational Intelligence, Berlin Heidelberg: Springer – Verlag, 2015.

[60] H. Braun, Neuronale Netze. Optimierung durch Lernen und Evolution, Berlin: Springer-Verlag, 1997.

[61] A. Strehl and J. Ghosh, "Cluster Ensembles – A knowledge reuse framework for combining multiple partitions," *Journal of Machine Learning Research*, pp. 583-617, 2002.

[62] A. Topchy, A. Jain and W. Punch, "Clustering ensembles: models of consensus and weak partitions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 27, pp. 1866-1881, 2005.

[63] H. Alizadeh, B. Minaei-Bidgoli and H. Parvin, "To improve the quality of cluster ensembles by selecting a subset of base clusters," *Journal of Experimental & Theoretical Artificial Intelligence*, no. 26, pp. 127-150, 2013.

[64] M. Charkhabi, T. Dhot and S. Mojarad, "Cluster ensembles, majority vote, voter eligibility and privileged voters," *Int. Journal of Machine Learning and Computing*, no. 4, pp. 275-278, 2014.

[65] Y. Bodyanskiy, "Computational intelligence techniques for data analysis," in *Lecture Notes in Informatics*, Bonn, GI, 2005, pp. 15-36.

[66] Е. Бодянский и О. Руденко, Искусственные нейронные сети: архитектуры, обучение, применения, Харьков: ТЕЛЕТЕХ, 2004, p. 372.

[67] Є. В. Бодянський, Д. Д. Пелешко, О. А. Винокурова, С. В. Машталір та Ю. С. Іванов, Аналіз та обробка потоків даних засобами обчислювального інтелекту, Львів: Вид-во Львів. політехніки, 2016, p. 235.

[68] R. Rojas, Neural Networks. A Systematic Introduction, Berlin: Springer-Verlag, 1996.

[69] R. Schalkoff, Artificial Neural Networks, N.Y.: The McGraw–Hill Comp., Inc., 1997.

[70] L. H. Tsoukalas and R. E. Uhrig, Fuzzy and Neural Approaches in Engineering., N.Y.: John Wiley & Sons, Inc., 1997.

[71] S. Haykin, Neural Networks. A Comprehensive Foundation. - Upper Saddle River, N.J.: Prentice Hall, Inc., 1999.

[72] C. J. Harris, Ed., Advances in Intelligent Control, London: Taylor and Francis, 1994.

[73] D. Zahirniak, R. Chapman, S. Rogers, B. Suter, M. Kabritsky and V. Piati, "Pattern recognition using radial basis function network," in *Proc 6th Ann. Aerospace Application of Artificial Intelligence Conf.*, Dayton, OH, 1990.

[74] Z.-P. Lo, Y. Yu and B. Bavarian, "Analysis of the convergence properties of topology preserving neural networks," *IEEE Trans. on Neural Networks,* no. 4, pp. 207-220, 1993.

[75] Y. Bodyanskiy, O. Chaplanov, V. Kolodyazhniy and P. Otto, "Adaptive quadratic radial basis function network for time series forecasting," in *Proc. East West Fuzzy Coll*, Zittau, 2002.

[76] M. Cottrel and J. Fort, "A stochastic model of retinotopy: a self-organizing process," *Biological Cybernetics,* no. 54, p. 234 – 249, 1986.

[77] H. Ritter and K. Schulten, "On stationary state of the Kohonen self-organizing sensory mapping," *Biological Cybernetics,* no. 54, p. 234 – 249, 1986.

[78] H. Ritter and K. Schulten, "Convergence properties of Kohonen's topology conserving maps: fluctuation, stability, and dimension selection," *Biological Cybernetics,* no. 60, pp. 59-71, 1986.

[79] D. Davies and D. Bouldin, "A Cluster Separation Measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* no. 4, pp. 224-227, 1979.

[80] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions Pattern Analysis Machine Intelligence,* no. 24(12), p. 1650 – 1654, 2002.

[81] S. Saitta, B. Raphael and I. F. C. Smith, "A bounded index for Cluster validity," in *Proceedings of the 5th International Conference on Machine Learning and Data Mining in Pattern Recognition*, 2007.

[82] D. Dua and C. Graff, "UCI Machine Learning Repository," CA: University of California, School of Information and Computer Science, Irvine, 2019.

[83] J. C. Bezdek, J. Keller, R. Krishnapuram and N. Pal, Fuzzy Models and Algorithms for Pattern Recognition and Image Processing. The Handbook of Fuzzy Sets, vol. 4, Kluwer, Dordrecht, Netherlands: Springer, 1999.

[84] J. Friedman, T. Hastie and R. Tibshirani, he Elements of Statistical Learning. Data Mining, Inference, and Prediction, Berlin: Springer, 2003.

[85] T. M. Cover, "Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition," *IEEE Trans. on Electronic Computers*, no. 14, pp. 326-334, 1965.

[86] M. Girolami, "Mercer kernel-based clustering in feature space," *IEEE Trans. on Neural Networks*, vol. 13, no. 3, pp. 780-784, 2002.

[87] D. MacDonald and C. Fyfe, "Clustering in data space and feature space," in *ESANN'2002 Proc. European Symp. on Artificial Neural Networks*, Bruges (Belgium), 2002.

[88] F. Camastra and A. Verri, "A novel kernel method for clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, no. 5, pp. 801-805, 2005.

[89] J. Park and I. Sandberg, "Universal approximation using radial-basis-function networks," *Neural Computation*, no. 3, pp. 246-257, 1991.

[90] D.-Q. Zhang and S.-C. Chen, "Kernel based fuzzy and possibilistic c-means clustering," in *Proc. Int. Conf. Artificial Neural Networks ICANN*, Turkey, 2003.

[91] Y. Bodyanskiy, A. Deineko and Y. Kutsenko, "On-line kernel clustering based on the general regression neural network and T. Kohonen's self-organizing map," *Automatic Control and Computer Sciences*, pp. 55-62, 2017.

[92] S. Kung, Kernel Methods and Machine Learning, Cambridge: University Press, 2014.

[93] C. Mumford and L. Jain, Computational Intelligence. Collaboration, Fuzzy and Emergence, Berlin: Springer-Vergal, 2009.

[94] Y. Gorshkov, V. Kolodyazhniy and Y. Bodyanskiy, "New recursive learning algoritms for fuzzy Kohonen clustering network," in *In Proc. 17th Int. Workshop on Nonlinear Dynamics of Electronic Systems*, Rapperwil, Switzerland, 2009.

[95] F. Höppner, F. Klawonn and R. Kruse, Fuzzy-Clusteranalyse. Verfahren für die Bilderkennung, Klassifikation und Datenanalyse, Braunschweig: Vieweg, 1996, p. 292.

[96] K. J. Cios and W. Pedrycz, "Neuro-fuzzy algorithms," in *Handbook of Neural Computation*, Oxford, IOP Publishing Ltd and Oxford University Press, 1997, pp. 97-111.

[97] N. Shakhovska, M. Medykovsky and P. Stakhiv, "Application of algorithms of classification for uncertainty reduction," *Przeglad Elektrotechniczny,* no. 4, pp. 284-286, 2013.

[98] F. Klawonn, F. Höppner and B. Jayaram, "What are clusters in high dimensions and are they difficult to find?," *Lecture Nodes in Computer Science,* vol. 7627, pp. 14-33, 2015.

[99] B. Kolchygin and Y. Bodyanskiy, "Adaptive fuzzy clustering with a variable fuzzifier," *Cybernetics and Systems Analysis,* vol. 49, no. 3, pp. 366-374, 2013.

[100] A. Keller and K. F., "Fuzzy Clustering with weighting of data variables," *Uncertainty, Fuzziness and Knowledge Based Systems,* no. 8, pp. 735-746, 2000.

[101] Y. Bodyanskiy, B. Kolchygin and I. Pliss, "Adaptive neuro-fuzzy Kohonen network with variable fuzzifier," *Inform. Theories and Appl,* vol. 18, no. 3, pp. 215-223, 2011.

[102] X. Xie and G. A. Beni, "Validity Measure for Fuzzy Clustering," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* no. 13, pp. 841-847, 1991.

[103] Y. Bodyanskiy, O. Tyshchenko and D. Kopaliani, "An Evolving Connectionist System for Data Stream Fuzzy Clustering and Its Online Learning," *Neurocomputing,* no. 262, p. 41 – 56, 2017.

[104] F. Klawonn, R. Kruse and H. Timm, "Fuzzy shell cluster analysis," in *Learning, Networks and Statistics*, Wien, Springer-Verlag, 1997, pp. 105-120.

[105] L. Pau, Failure diagnosis and performance monitoring, N.Y.: Marcel Dekker, 1981.

[106] E. Gustafson and W. Kessel, "Fuzzy clustering with a fuzzy covariance matrix," in *Proc. IEEE CDC*, San Diego, California, 1979.

[107] I. Gath and A. Geva, "Unsupervised optimal fuzzy clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 11, p. 773–781, 1989.

[108] R. Krishnapuram and J. Keller, "A possibilistic approach to clustering," *IEEE Trans. on Fuzzy Systems,* no. 1, pp. 98-110, 1993.

[109] R. Krishnapuram and J. Keller, "Fuzzy and possibilistic clustering methods for computer vision," *Neural Fuzzy Systems.,* no. 12, pp. 133-159, 1994.

[110] Y. Bodyanskiy, V. Kolodyazhniy and A. Stephan, "Recursive fuzzy clustering algorithms," in *Proc. 10-th East-West Fuzzy Coll.*, Zittau, 2002.

[111] F. Klawonn and R. Kruse, "Constructing a fuzzy controller from data," *Fuzzy Sets and Systems,* no. 85, pp. 117-193, 1997.

[112] K. Arrow, L. Hurwitz and H. Uzawa, Studies in Linear and Nonlinear Programming, Stanford: Stanford University Press, 1958, p. 242.

[113] F. Chung and T. Lee, "Fuzzy competitive learning," *Neural Networks,* vol. 7, no. 3, pp. 539-552, 1994.

[114] D. Park and I. Dagher, "Gradient based fuzzy c-means (GBFCM) algorithm," in *Proc. IEEE Int. Conf. on Neural Networks*, Orlando, FL, USA, 1994.

[115] Y. Bodyanskiy, Y. Gorshkov, I. Kokshenev and V. Kolodyazhniy, "Outlier resistant recursive fuzzy clustering algorithms," in *Computational Intelligence, Theory and Applications. Advances in Soft Computing*, Berlin; Heidelberg, Springer-Verlag, 2006, p. 647–652.

[116] П. Жернова, «Вероятностно-возможностный подход для кластеризации потоков данных на основе ансамблей нейронных сетей,» в *Материалы международной научно-практической конференции «Информационные технологии и системы»*, Харьков, 2019.

[117] B. Ripley, Pattern Recognition and Neural Networks, Cambridge University Press., 2008.

ДОДАТОК А

СПИСОК ПУБЛІКАЦІЙ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

Список публікацій здобувача, в яких опубліковані основні наукові результати дисертації:

1. P. Zhernova, A. Deyneko, Z. Deyneko, I. Pliss and V. Ahafonov, "Data Stream Clustering in Conditions of an Unknown Amount of Classes," In: Hu Z., Petoukhov S., Dychka I., He M. (eds) *Advances in Computer Science for Engineering and Education. ICCSEEA 2018. Advances in Intelligent Systems and Computing*, vol 754. Springer, Cham, pp. 410-419, 2019. (Входить до міжнародних наукометричних баз SCOPUS).

2. Є. Бодянський, А. Дейнеко, П. Жернова, О. Золотухін та Я. Хаустова, «Послідовне ядерне нечітке кластерування великих масивів даних на основі гібридної системи обчислювального інтелекту,» *Вісник Національного університету "Львівська політехніка". Інформаційні системи та мережі,* № 829, pp. 20-24, 2017. (Входить до міжнародної наукометричної бази Google Scholar).

3. Є. Бодянський, А. Дейнеко, П. Жернова та В. Репін, «Онлайн модифікація методу Х-середніх на основі ансамблю самоорганізовних мап Т. Когонена,» *Збірник наукових праць «Розвиток транспорту»,* № 1, pp. 96-107, 2017.

4. П. Жернова та Є. Бодянський, «Ядерна нечітка кластеризація потоків даних на основі ансамблю нейронних мереж,» *Сучасний стан наукових досліджень та технологій в промисловості,* № 4(6), pp. 42-49, 2018. (Входить до міжнародної наукометричної бази Index Copernicus International).

5. Y. Bodyanskiy, I. Perova and P. Zhernova, "Online fuzzy clustering of high - dimensional data based on ensembles in data stream mining tasks," *Innovative Technologles & Scientific Solutions for Industries*, no. 1(7), pp. 16-24, 2019.

6. П. Жернова та Є. Бодянський, «Нечітка імовірнісно-можливісна послідовна кластеризація даних на основі ансамблевого підходу,» *Науково-*

технічний журнал «Прикладна радіоелектроніка», № 1,2, pp. 40-45, 2019. (Входить до міжнародної наукометричної бази Index Copernicus International).

Результати, які засвідчують апробацію матеріалів дисертації:

7. Е. Бодянский, А. Дейнеко, П. Жернова и В. Репин, «Адаптивная модификация метода X-средних на основе ансамбля кластеризующих нейронных сетей Т. Кохонена,» в *Матеріали VI Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології»*, Одесса, 2017.

8. Е. Бодянский, П. Жернова и А. Дейнеко, «Кластеризующий ансамбль нейронных сетей и его обучение в условиях неизвестного количества классов,» в *Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту»*, Залізний порт, Україна, 2018.

9. А. Дейнеко, П. Жернова, І. Плісс та О. Чала, «Модифікована нечітка ймовірнісна нейронна мережа,» в *Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту»*, Залізний порт, Україна, 2018.

10. P. Zhernova, A. Deyneko, Y. Bodyanskiy and V. Riepin, "Adaptive kernel data streams clustering based on neural networks ensembles in conditions of uncertainty about amount and shapes of clusters," in *IEEE Second International Conference on Data Stream Mining & Processing*, Lviv, Ukraine, 2018. (Входить до міжнародної науково-метричної бази SCOPUS).

11. Deineko, P. Zhernova, B. Gordon, O. Zayika, I. Pliss and N. Pabyrivska, "Data stream online clustering based on fuzzy expectation-maximization approaching formation on submission," in *IEEE Second International Conference on Data Stream Mining & Processing*, Lviv, Ukraine, 2018. (Входить до міжнародної науково-метричної бази SCOPUS).

12. П. Жернова, «Вероятностно-возможностный подход для кластеризации потоков данных на основе ансамблей нейронных сетей,» в *Материалы*

международной научно-практической конференции «Информационные технологии и системы», Харьков, 2019.

13. П. Жернова та А. Лобинцев, «Кластеризація даних високої розмірності з використанням можливісного підходу,» в *Матеріали 23-го Міжнародного молодіжного форуму «Радіоелектроніка та молодь в 21 столітті»*, Харьков, 2019.

14. П. Жернова та Є. Бодянський, «Нейро-фаззі мережа та її навчання для кластеризації потоків даних високої розмірності,» в *Матеріали V міжнародної науково-практичної конференції «Обчислювальний інтелект (результати, проблеми, перспективи)»*, Ужгород, 2019.

ДОДАТОК Б

ВІДОМОСТІ ПРО АПРОБАЦІЮ РЕЗУЛЬТАТІВ ДИСЕРТАЦІЇ

1. Е. Бодянский, А. Дейнеко, П. Жернова и В. Репин, «Адаптивная модификация метода Х-средних на основе ансамбля кластеризующих нейронных сетей Т. Кохонена,» в *Матеріали VI Міжнародної науково-практичної конференції «Інформаційні управляючі системи та технології»*, Одесса, 2017. – *очна участь з доповіддю*

2. Е. Бодянский, П. Жернова и А. Дейнеко, «Кластеризующий ансамбль нейронных сетей и его обучение в условиях неизвестного количества классов,» в *Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту»*, Залізний порт, Україна, 2018. – *очна участь з доповіддю*

3. А. Дейнеко, П. Жернова, І. Плісс та О. Чала, «Модифікована нечітка ймовірнісна нейронна мережа,» в *Матеріали міжнародної наукової конференції «Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту»*, Залізний порт, Україна, 2018. – *очна участь з доповіддю*

4. P. Zhernova, A. Deyneko, Y. Bodyanskiy and V. Riepin, "Adaptive kernel data streams clustering based on neural networks ensembles in conditions of uncertainty about amount and shapes of clusters," in *IEEE Second International Conference on Data Stream Mining & Processing*, Lviv, Ukraine, 2018. (Входить до міжнародної науково-метричної бази SCOPUS). – *очна участь з доповіддю*

5. Deineko, P. Zhernova, B. Gordon, O. Zayika, I. Pliss and N. Pabyrivska, "Data stream online clustering based on fuzzy expectation-maximization approaching formation on submission," in *IEEE Second International Conference on Data Stream Mining & Processing*, Lviv, Ukraine, 2018. (Входить до міжнародної науково-метричної бази SCOPUS). – *очна участь з доповіддю*

6. П. Жернова, «Вероятностно-возможностный подход для кластеризации потоков данных на основе ансамблей нейронных сетей,» в *Материалы*

международной научно-практической конференции «Информационные технологии и системы», Харьков, 2019. – *очна участь з доповіддю*

7. П. Жернова та А. Лобинцев, «Кластеризація даних високої розмірності з використанням можливісного підходу,» в *Матеріали 23-го Міжнародного молодіжного форуму «Радіоелектроніка та молодь в 21 столітті»*, Харьков, 2019. – за*очна участь з доповіддю*

8. П. Жернова та Є. Бодянський, «Нейро-фаззі мережа та її навчання для кластеризації потоків даних високої розмірності,» в *Матеріали V міжнародної науково-практичної конференції «Обчислювальний інтелект (результати, проблеми, перспективи)»*, Ужгород, 2019. – *очна участь з доповіддю*

ДОДАТОК В

ДОКУМЕНТИ, ЩО ПІДТВЕРДЖУЮТЬ ВПРОВАДЖЕННЯ

«ЗАТВЕРДЖУЮ»
Проректор з наукової роботи
Харківського національного
медичного університету
проф. В.В. М'ясоєдов
«___» _____ 20__ р.

АКТ ПРО ВПРОВАДЖЕННЯ

1. Найменування пропозиції (метод профілактики, діагностики, лікування, пристрій, форма організаційної роботи та ін.)
   Методи та моделі ансамблю нейро-фаззі систем, а саме: ансамбль нейро-фаззі самоорганізовних карт Т.Кохонена, метод нечіткої кластеризації даних великої розмірності; критерій валідації оцінювання якості кластеризації медичних показників для інтелектуальної обробки даних медико-біологічних досліджень.
2. Запропонований молодшим науковим співробітником, асистентом кафедри системотехніки Харківського національного університету радіоелектроніки Жерновою Поліною Євгеніївною
3. Джерело інформації (методичні рекомендації, інформаційний лист, звіт про НДР, дисертація, монографія, з'їзди, конференції, семінари та ін.)

3.1 Стаття P. Zhernova, A. Deineko, Zh. Deineko, I. Pliss, V. Ahafonov "Data Stream Clustering in Conditions of Classes Unknown Amount" Advances in Computer Science for Engineering and Education, 2018, v.754, pp.410-418. DOI: 10.1007/978-3-319-91008-6_41

3.2 Тези доповіді P. Zhernova, A. Deineko, Ye. Bodyanskiy, V. Riepin. Adaptive Kernel Data Streams Clustering Based on Neural Networks Ensembles in Conditions of Uncertainty About Amount and Shapes of Clusters // Proc. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), August 21-25, 2018, Lviv, Ukraine, pp. 7-12

3.3 Стаття Bodyanskiy Ye., Perova I., Zhernova P. Online fuzzy clustering of high-dimensional data based on ensembles in data stream mining tasks. Innovative Technologles & Scientific Solutions for Industries, 2019, № 5(7).

4. Впроваджено на кафедрі гігієни та екології № 2
5. Результати застосування методу за період з 2018 по 2019 рр.
   Встановлення закономірностей формування відповідної реакції організму на сполучений вплив екологічних чинників; використання методичних підходів, щодо визначення гігієнічної значущості біологічних ефектів сполученої дії електромагнітного випромінювання та позитивних низьких температур при аналізі результатів НДР бюджетного фінансування «Встановити механізми адаптації до сполученої дії хімічних та фізичних чинників навколишнього середовища»
6. Ефективність впровадження за критеріями, висловленими в джерелі

інформації (п.3)

Підтверджується працездатність розробленого ансамблю нейро-фаззі самоорганізовних карт Т. Кохонена, метода нечіткої кластеризації даних великої розмірності та критерію валідації оцінювання якості кластеризації медичних показників для обробки результатів клініко-лабораторних досліджень при визначенні біологічних ефектів під час дії електромагнітного випромінювання при ізольованій дії або у сполученні з позитивними низькими температурами на біологічний об'єкт та визначення найбільш інформативних показників у багатовимірних часових рядах. Для аналізу були взяті такі відомості про біологічний об'єкт (щури-самці лінії *Wistar*), як результати біохімії сечі та сироватки крові, імунологічних показників, морфометричного аналізу органів, репродуктивної функції, сформовані у таблицю «об'єкт-властивість» і обробляється послідовно у онлайн-режимі. В результаті роботи ансамблю нейро-фаззі самоорганізовних карт Т. Кохонена були отримані нечіткі функції належності кожного об'єкта до кожної з контрольних груп, що дозволило визначити кількість кластерів за допомогою індексу валідації.

6. Зауваження, пропозиції: немає

Відповідальний за впровадження:
в.о. завідувача кафедри
гігієни та екології № 2 ХНМУ,
к. мед.н., доцент М.О. Сидоренко

Затверджую

Ректор Харківського
національного університету
радіоелектроніки

_____ Семенець В.В.

_____ 2019 р.

Акт

щодо впровадження у навчальний процес кафедри системотехніки Харківського національного університету радіоелектроніки результатів дисертаційної роботи Жернової Поліни Євгеніївни на тему «Нечітка кластеризація потоків даних за умов невідомої кількості кластерів».

Комісія у складі:

Голови комісії: завідувача кафедри системотехніки, д.т.н., проф. Гребеннік І.В.,

Членів комісії: професора кафедри системотехніки, д.т.н., доц. Нечипоренко А.С.;
професор кафедри системотехніки, к.т.н., доц. Іванов В.Г.;
доцента кафедри системотехніки, к.т.н., с.н.с. Решетнік В.М.

розглянули результати наукових досліджень Жернової П.Є. та прийняли рішення щодо впровадження їх у навчальний процес кафедри системотехніки ХНУРЕ при викладанні дисципліни «Нейросистеми та генетичні алгоритми».

Отримані в дисертаційній роботі нові наукові результати, що стосуються розробки методів кластеризації потоків даних в умовах апріорної невизначеності кількості та форми кластерів на основі ансамблевого підходу, дозволять підвищити науковий рівень вказаної дисципліни та доповнити її прикладами практичного застосування.

Голова комісії: _____ І.В. Гребеннік

Члени комісії: _____ А.С. Нечипоренко

_____ В.Г. Іванов

_____ В.М. Решетнік