Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Кваліфікаційна наукова
праця на правах рукопису

КОБИЛІН ІЛЛЯ ОЛЕГОВИЧ

УДК 004.032.26

ДИСЕРТАЦІЯ
НЕЧІТКА КЛАСТЕРИЗАЦІЯ ЧАСОВИХ РЯДІВ В ІНТЕЛЕКТУАЛЬНОМУ АНАЛІЗІ ПОТОКІВ ДАНИХ

05.13.23 – системи та засоби штучного інтелекту
технічні науки

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
_____ Кобилін І.О.

Науковий керівник:
Бодянський Євгеній Володимирович,
доктор технічних наук, професор

Цей примірник дисертаційної роботи ідентичний за змістом
з іншими, поданими до спеціалізованої вченої ради Д 64.052.01

Учений секретар спеціалізованої
вченої ради Д 64.052.01 Є. І. Литвинова

Харків – 2019

АНОТАЦІЯ

Кобилін Ілля Олегович. Нечітка кластеризація часових рядів в інтелектуальному аналізі потоків даних. – Кваліфікаційна наукова праця 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 і науки України, Харків, 2019.

У дисертації викладено нове розв'язання актуального наукового завдання нечіткої кластеризації часових рядів.

Робота складається зі вступу, п’яти розділів, висновків, списку використаних джерел та двох додатків. Представлена робота належить до області штучного інтелекту, зокрема кластеризації часових рядів.

Мета дослідження – вирішення задач послідовної нечіткої кластеризації нерівномірно квантованих асинхронних нестаціонарних часових рядів в умовах, коли інформація на обробку надходить в онлайн режимі, і покращення результатів у вирішенні практичних задач.

Задачі дослідження: 1) проведення аналізу існуючих методів та підходів до кластеризації часових рядів; 2) розробка методу передобробки часових рядів, що спотворені аномальними викидами та збуреннями; 3) розробка онлайн модифікації методу кластеризації коротких часових рядів; 4) розробка методу онлайн кластеризації багатовимірних часових рядів; 5) розробка методу онлайн кластеризації асинхронних часових рядів, несхильного до впливу ефекту концентрації норм; 6) проведення експериментів на основі тестових та реальних даних.

Об'єкт дослідження – процес інтелектуального аналізу потоку даних у формі часових рядів.

Предмет дослідження – методи інтелектуального аналізу для нечіткої онлайн кластеризації багатовимірних часових рядів з асинхронними тактами квантування, що призначені для аналізу потоків даних.

Науково-практична задача – розробка модуля у якому реалізовано методи нечіткої кластеризації для використання у задачах моніторингу медичних даних в онлайн режимі.

Сутність дослідження - розробка моделей та методів нечіткої кластеризації даних, які послідовно надходять на обробку з нерівномірними тактами квантування, що базуються на апараті гібридних систем обчислювального інтелекту, та дозволяють підвищити ефективність застосування сучасних моніторингових систем.

Наукова новизна результатів дослідження.

1. Вперше запропоновано метод кластеризації, який несхильний до ефекту концентрації норм, що дозволяє вирішувати задачу кластеризації в онлайн режимі за умов перетину класів та асинхронних нерівномірно квантованих часових рядів за рахунок використання спеціальної цільової функції нечіткої кластеризації.

2. Вперше запропоновано послідовний онлайн метод кластеризації багатовимірних часових рядів, що базується на апараті гібридних систем обчислювального інтелекту, який дозволив вирішувати задачу кластеризації даних, які послідовно надходять на обробку з нерівномірними тактами квантування.

3. Отримав подальший розвиток метод адаптивної кластеризації, що базується на методах ймовірнісної та можливісної кластеризації коротких часових рядів, які, у свою чергу, засновані на метриці спеціального вигляду, що дозволяє значно спростити чисельну реалізацію методу, за рахунок використання метрики на основі тангенсів кутів нахилу, що на відміну від

відомих методів вирішує задачу кластеризації нерівномірно квантованих часових рядів.

4. Отримав подальший розвиток метод робастної адаптивної ідентифікації нестаціонарних часових рядів в онлайн режимі надходження потоку даних, який характеризується простотою обчислювальної реалізації та вирішує задачу обробки даних, що збурені аномальними викидами, за рахунок використання введеної модифікації критерія Гемана-МакКлюра.

Практична значущість отриманих результатів:

Використання запропонованих моделей та методів дозволяє підвищити ефективність застосування сучасних моніторингових систем для вирішення задач кластеризації даних, які послідовно надходять на обробку з нерівномірними тактами квантування, що базуються на апараті гібридних систем обчислювального інтелекту.

Реалізований модуль із запропонованими методами підтвердив свою ефективність у задачах моніторингу медичних даних в онлайн режимі. У такому разі моніторинг медичних даних дозволяє ефективно виявляти аномалії у хворих у режимі реального часу. Результати досліджень впроваджені у ТОВ «Інфобуд», м. Харків (акт впровадження від 03.10.2018) та у ТОВ «Сайтосс», м. Харків (акт впровадження від 06.10.2018). Результати досліджень впроваджені у Харківському національному університеті радіоелектроніки на кафедрі штучного інтелекту в освітній процес з дисципліни «Нейромережеві методи обчислювального інтелекту».

Публікації: За тематикою дослідження опубліковано 12 наукових праць, з них 1 розділ у колективній монографії, що входить до наукометричної бази SCOPUS; 1 стаття за кордоном, що входить до наукометричної бази SCOPUS; 3 статті у виданнях, які зазначені в переліках фахових видань України з технічних наук; 7 публікацій у матеріалах конференцій (2 включено до наукометричної бази даних SCOPUS).

СПИСОК ОПУБЛІКОВАНИХ РОБІТ ЗА ТЕМОЮ ДИСЕРТАЦІЇ

в яких опубліковані основні наукові результати дисертації:

1. Setlak, G., Bodyanskiy, Y., Pliss, I., Vynokurova, O., Peleshko, D., & Kobylin, I. (2017). Adaptive Fuzzy Clustering of Multivariate Short Time Series with Unevenly Distributed Observations Based on Matrix Neuro-Fuzzy Self-Organizing Network. In *Advances in Fuzzy Logic and Technology 2017* (pp. 308-315). Springer, Cham. (Входить до міжнародної наукометричної бази SCOPUS).

2. Bodyanskiy, Y., Vynokurova, O., Kobylin, I., & Kobylin, O. (2016). Adaptive Fuzzy Clustering of Short Time Series with Unevenly Distributed Observations in Data Stream Mining Tasks. *Information Technology and Management Science*, *19*(1), 23-28. (Входить до наукометричної бази SCOPUS).

3. Бодянский, Е. В., Винокурова, Е. А., Кобылин, И. О., & Мулеса, П. П. (2016). Робастная адаптивная идентификация нестационарных временных рядов с помощью ансамбля обучаемых гибридных адаптивных моделей. *Управляющие системы и машины*, (5), 76-83.

4. Бодянський, Є., Винокурова, О., Кобилін, І., & Мулеса, П. (2017). Адаптивна матрична нейро-фаззі самоорганізовна мережа для кластеризації багатовимірних потоків даних. *Вісник Національного університету «Львівська політехніка». Серія: Комп'ютерні науки та інформаційні технології*, (864), 314-319.

5. Бодянский, Е.В., Винокурова, Е.А., Кобылин, И.О., Кобылин, О.А., & Пелешко, Д.Д. (2017) Нечёткая кластеризация временных рядов с неравномерными и асинхронными тактами квантования. *Системи обробки інформації, 5(151)*, 47-54.

які засвідчують апробацію матеріалів дисертації:

6. Bodyanskiy, Y., Vynokurova, O., Szymański, Z., Kobylin, I., & Kobylin, O. (2016, August). Adaptive Robust Models for Identification of

Nonstationary Systems in Data Stream Mining Tasks. In *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)* (pp. 263-268). IEEE. (Входить до наукометричної бази SCOPUS).

7. Bodyanskiy, Y., Kobylin, I., Rashkevych, Y., Vynokurova, O., & Peleshko, D. (2018, February). Hybrid Fuzzy-Clustering Algorithm of Unevenly and Asynchronously Spaced Time Series in Computer Engineering. In *2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET)* (pp. 930-935). IEEE. (Входить до міжнародної наукометричної бази SCOPUS).

8. Бодянский, Е. В., Дейнеко, А. А., Кобылин, И. О., & Плисс, И. П. (2016). Адаптивная нечеткая кластеризация коротких временных рядов в интеллектуальном анализе потоков данных. *Intellectual Systems For Decision Making and Problems of Computational Intelligence*, 255-257.

9. Бодянський, Є. В., Винокурова, О. А., Ізонін, І. В., Кобилін, І. О., & Мулеса, П. П. (2017) Кластеризація багатовимірних часових рядів на основі адаптивної матричної нейро-фаззі самоорганізовної мережі. *Intellectual Systems For Decision Making and Problems of Computational Intelligence*, 247-248.

10. Бодянський, Є. В., Винокурова, О. А., Кобилін, І. О., & Мулеса, П.П. (2016). Адаптивна нечітка кластеризація багатовимірних часових рядів з нерівномірним тактом квантування. *Праці VIII-Й Міжнародної школи семінару- «Теорія Прийняття Рішень»* 56-57.

11. Кобылин, И.О., (2015) Об одном методе кластеризации коротких временных рядов. *"Радиоэлектроника и молодежь в XXI веке"* 30-31.

12. Кобылин, И.О., (2016) Адаптивная кластеризация коротких временных рядов с неравномерным тактом квантования. *"Радиоэлектроника и молодежь в XXI веке"* 21-22.

ABSTRACT

Kobylin I.O. Fuzzy clustering time series in data stream mining. – Qualified scientific paper as manuscript.

Dissertation for obtaining the scientific degree of the candidate of technical sciences on the specialty 05.13.23 "Systems and tools of artificial intelligence". – Kharkiv National University of Radio Electronics, Ministry of Education and Science of Ukraine, Kharkiv, 2019.

In the dissertation the actual problem of developing methods for fuzzy clustering of time series have solved.

The work consists of an introduction, five sections, conclusions, a list of sources used and two applications. The presented work relates to the field of artificial intelligence, in particular clusterization of time series.

The purpose and objectives of the study: the purpose of this work is to develop online methods for fuzzy clustering of nonsteady quantized asynchronous non-stationary time series to improve efficiency and reduce processing time in data mining.

Tasks of the research: carrying out analysis of existing methods and clustering of time series; development of the pre–preprocessing of time series contaminated with anomalous outlier and disturbances; development of online modification of the clustering method of short time series; development of the method of online clustering of multidimensional time series; development of the method of online clustering of asynchronous time series, non-susceptible to the effect of the concentration of norms; conducting experiments based on test and real data.

The object of research – is the process of data mining for the case of processing time series.

The subject of research is the methods of intellectual analysis for fuzzy online clustering of multidimensional short time series with non–uniform and asynchronous quantization cycles, designed to analyze data streams.

The scientific and practical task development of the module with methods of fuzzy clustering for use in the tasks of monitoring medical data in online mode.

The essence of the study the development of models and methods of fuzzy clustering of data that consistently arrive at processing with uneven quantization cycles, based on the device of hybrid systems of computing intelligence, and allow to increase the efficiency of application of modern monitoring systems.

Scientific novelty of research results:

1. For the first time, a clustering method was proposed, which allows solving the clustering problem under the conditions of overlapping of classes and in the online mode of asynchronous non–uniformly quantized time series, subjected to the norm concentration effect, due to the use of a special goal function of fuzzy clustering.

2. For the first time, a sequential online clustering method for multidimensional time series based on a hybrid computational intelligence system was proposed, which made it possible to solve the problem of clustering data, which is subsequently received for processing with non–uniform quantization cycles.

3. The method of adaptive clustering has been further developed, based on possibilistic and probabilistic clustering of short time series, which, in turn, is based on a special form metric which allows to significantly simplify the numerical implementation of the method by using a metric with tangent angles of the time series, which, unlike the known methods, solves the clustering problem of unevenly quantized time series.

4. The method of robust adaptive identification of non-stationary time series in the data stream mining has been further developed, which is characterized by simplicity of computational implementation and solves the problem of processing

data that is perturbed by anomalous emissions as a result of using the introduced Geman-McClure criteria modification.

The results of the dissertation are implemented: The use of the proposed models and methods allows to increase the efficiency of application of modern monitoring systems for solving data clustering problems that consistently arrive at processing with uneven quantization cycles based on the device of hybrid systems of computing intelligence.

The implemented module with the proposed methods has confirmed its effectiveness in the tasks of monitoring medical data in online mode. In this case, the monitoring of medical data allows you to effectively detect abnormalities in patients in real time. The research results were implemented at «InfoStroy», Kharkiv (implementation act dated 10.3.2018) and in «Sytoss», Kharkiv (implementation act dated 06.10.2018). Results of research implementation of the Kharkiv National University of Radio Electronics at the Department of Artificial Intelligence in the educational process of the course "Neural network methods of computational intelligence".

Publications: The dissertation materials are fully enunciated in 12 scientific papers: Section – in a joint monograph, included in the SCOPUS Scientometric Data Base; 1 article – included in the SCOPUS Scientometric Data Base; 3 articles – in editions indicated in the list of professional editions of Ukraine in technical sciences, included in the National Scientometric Data Bases; 7 abstracts – in the collections of research papers of international scientific and technical conferences, 2 are included in the SCOPUS Scientometric Data Base.

Key words: time series, fuzzy clustering of time series, asynchronous quantization, online fuzzy clustering procedure, adaptive learning procedures, data mining, robust goal functions, rheonomous nonlinear time series, statistically distributed clustering.

LIST OF PUBLICATIONS

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

1. Setlak, G., Bodyanskiy, Y., Pliss, I., Vynokurova, O., Peleshko, D., & Kobylin, I. (2017). Adaptive Fuzzy Clustering of Multivariate Short Time Series with Unevenly Distributed Observations Based on Matrix Neuro-Fuzzy Self-Organizing Network. In *Advances In Fuzzy Logic And Technology 2017* (pp. 308-315). Springer, Cham. (Входить до міжнародної наукометричної бази SCOPUS).

2. Bodyanskiy, Y., Vynokurova, O., Kobylin, I., & Kobylin, O. (2016). Adaptive Fuzzy Clustering of Short Time Series with Unevenly Distributed Observations in Data Stream Mining Tasks. *Information Technology and Management Science*, *19*(1), 23-28. (Входить до наукометричної бази SCOPUS).

3. Бодянский, Е. В., Винокурова, Е. А., Кобылин, И. О., & Мулеса, П. П. (2016). Робастная адаптивная идентификация нестационарных временных рядов с помощью ансамбля обучаемых гибридных адаптивных моделей. *Управляющие системы и машины*, (5), 76-83.

4. Бодянський, Є., Винокурова, О., Кобилін, І., & Мулеса, П. (2017). Адаптивна матрична нейро-фаззі самоорганізовна мережа для кластеризації багатовимірних потоків даних. *Вісник Національного університету «Львівська політехніка». Серія: Комп'ютерні науки та інформаційні технології*, (864), 314-319.

5. Бодянский, Е.В., Винокурова, Е.А., Кобылин, И.О., Кобылин, О.А., & Пелешко, Д.Д. (2017) Нечёткая кластеризация временных рядов с неравномерными и асинхронными тактами квантования. *Системи обробки інформації, 5(151)*, 47-54.

Results that confirm the approbation of the thesis:

6. Bodyanskiy, Y., Vynokurova, O., Szymański, Z., Kobylin, I., & Kobylin, O. (2016, August). Adaptive Robust Models for Identification of

Nonstationary Systems in Data Stream Mining Tasks. In *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)* (pp. 263-268). IEEE. (Входить до наукометричної бази SCOPUS).

7. Bodyanskiy, Y., Kobylin, I., Rashkevych, Y., Vynokurova, O., & Peleshko, D. (2018, February). Hybrid Fuzzy-Clustering Algorithm of Unevenly and Asynchronously Spaced Time Series in Computer Engineering. In *2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET)* (pp. 930-935). IEEE. (Входить до міжнародної наукометричної бази SCOPUS).

8. Бодянский, Е. В., Дейнеко, А. А., Кобылин, И. О., & Плисс, И. П. (2016). Адаптивная нечеткая кластеризация коротких временных рядов в интеллектуальном анализе потоков данных. *Intellectual Systems For Decision Making And Problems of Computational Intelligence*, 255-257.

9. Бодянський, Є. В., Винокурова, О. А., Ізонін, І. В., Кобилін, І. О., & Мулеса, П. П. (2017) Кластеризація багатовимірних часових рядів на основі адаптивної матричної нейро-фаззі самоорганізовної мережі. *Intellectual Systems For Decision Making And Problems of Computational Intelligence*, 247-248.

10. Бодянський, Є. В., Винокурова, О. А., Кобилін, І. О., & Мулеса, П.П. (2016). Адаптивна нечітка кластеризація багатовимірних часових рядів з нерівномірним тактом квантування. *Праці VIII-Й Міжнародної школи семінару- «Теорія Прийняття Рішень»* 56-57.

11. Кобылин, И.О., (2015) Об одном методе кластеризации коротких временных рядов. *"Радиоэлектроника и молодежь в XXI веке"* 30-31.

12. Кобылин, И.О., (2016) Адаптивная кластеризация коротких временных рядов с неравномерным тактом квантования. *"Радиоэлектроника и молодежь в XXI веке"* 21-22.

ЗМІСТ

Вступ 14
1 Огляд стану проблеми і постановка завдання дослідження 19
1.1 Види часових рядів та їх властивості 20
1.1.1 Передобробка часових рядів 22
1.1.2 Типи забруднень часових рядів 22
1.2 Методи формування векторів-ознак для часових рядів на основі статистичних досліджень 23
1.3 Проблеми кластеризації та класифікації часових рядів 27
1.4 Класифікація методів кластеризації часових рядів 30
1.5 Методи кластеризації на основі аналізу кореляцій в пакетному та онлайн режимі 32
1.6 Методи ієрархічної кластеризації 33
1.7 Алгоритм кластеризації k-середніх 37
1.8 Нечіткий алгоритм кластеризації c-середніх для часових рядів 39
Висновки до розділу 42
2 Адаптивні методи фільтрації та ідентифікації часових рядів 43
2.1 Адаптивна модель нестаціонарного часового ряду на базі квадратичного критерія якості 43
2.2 Робастні адаптивні моделі часових рядів 47
Висновки до розділу 56
3 Адаптивна нечітка кластеризація одновимірних часових рядів з нерівномірним тактом квантування в інтелектуальному аналізі потоків даних 57
3.1 Формування векторів ознак для одновимірних часових рядів 57
3.2 Пакетний метод нечіткої кластеризації часових рядів 60
3.3 Адаптивна можливісна нечітка кластеризація часових рядів 66
Висновки до розділу 68

4 Адаптивна нечітка кластеризація багатовимірних потоків даних з нерівномірним та асинхроними тактами квантування 70
4.1 Формування векторів ознак для багатовимірних часових рядів 70
4.2 Нечітка ймовірнісна кластеризація багатовимірних часових рядів 71
4.3 Оцінка відстані між реалізаціями часового ряду з нерівномірними асинхронними тактами квантування 75
4.4 Нечітка кластеризація часових рядів з нерівномірними асинхронними тактами квантування 79
Висновки до розділу 86
5 Імітаційне моделювання та розв’язання практичних задач 88
5.1 Імітаційне моделювання робастних адаптивних моделей часових рядів 88
5.2 Імітаційне моделювання методів адаптивної можливісної нечіткої кластеризації коротких часових рядів 94
5.3 Імітаційне моделювання послідовної онлайн нечіткої кластеризації багатовимірних рядів на базі модифікованої нейро-фаззі мережі Т.Кохонена .. 99
5.4 Імітаційне моделювання методу нечіткої кластеризації часових рядів з нерівномірними асинхронними тактами квантування та експериментальні дослідження 102
5.4 Застосування методів нечіткої кластеризації часових рядів у моніторингових системах 121
Висновки до розділу 138
Висновки 139
Перелік посилань 142
Додаток А Акти впровадження 153
Додаток Б Список опублікованих праць за темою дисертації 157

ВСТУП

Актуальність теми. На сьогоднішній час тенденція обробки великих обсягів інформації та їх аналізу за допомогою кластеризації дає змогу зрозуміти різноманіття процесів для її подальшого використання у сферах життєдіяльності, що супроводжують людину.

Значну частину інформації, пов’язану з обробкою великих обсягів даних, містять часові ряди. Однак, однією з типових проблем обробки часових рядів є їх нерівномірне квантування та багатовимірність.

Так, наприклад, при проведенні медичних досліджень (вимірювання артеріального тиску, термометрія) не вирішено задачу, пов’язану з аналізом часових рядів, які можуть бути несинхронізовані.

Дослідження часових рядів та методи їх обробки описані у працях Бездека, Густафсонна, Келлера, Клавонна, Кохонена, Хьоппнера та інших вчених, але в цих роботах не розглядались питання аналізу несинхронізованих даних, кластеризації даних, при якій виникає ефект концентрації норм, «прокльон розмірності», ряди, які не мають стохастичної природи або зашумлені викидами, нестаціонарність характеристик рядів, часовий ряд обробляється як вибірка загалом, а не окремими спостереженнями, тобто самі спостереження об’єднані у формі пакету і саме у такому вигляді подаються на обробку, а не в онлайн режимі, що значно зменшує час обробки ряду для отримання нових результатів. Також, до недоліків можна віднести неможливість обробки невеликих проміжків спостережень та невміння розпізнавати можливі похибки.

Для вирішення задач, враховуючи сучасні потреби ефективного аналізу нечіткої обробки коротких часових рядів з нерівномірно розподіленими спостереженнями, які не дозволяють використовувати стандартні методи для

обробки в онлайн режимі, доцільною є розробка нових методів у рамках концепції інтелектуального аналізу даних.

Кластеризація одновимірних та багатовимірних часових рядів з нерівномірними тактами квантування ускладнена тим, що для кожного з рядів ці такти можуть бути несинхронізовані, тобто нерівномірні та асинхронно квантовані. Таким чином, актуальною проблемою є аналіз коротких часових рядів за допомогою кластеризації.

Зв'язок роботи з науковими програмами, планами, темами. Дослідження, результати яких викладені у дисертаційній роботі, проводилися відповідно до держбюджетних тем НДР, що виконувались у Харківському національному університеті радіоелектроніки:

– «Нейро-фаззі системи для поточної кластеризації та класифікації послідовностей даних в умовах їх спотворення відсутніми і аномальними спостереженнями» (№ДР 0113U000361);

– «Динамічний інтелектуальний аналіз послідовностей нечіткої інформації в умовах суттєвої невизначеності на основі гібридних систем обчислювального інтелекту» (№ДР 0116U002539).

В рамках зазначених НДР здобувачем розроблено методи нечіткої кластеризації, що призначені для обробки даних в онлайн режимі, коли дані надходять на обробку послідовно, одні за одними, а кластери можуть перетинатися. Автор брав участь у виконанні цих робіт, як виконавець, щодо розроблення адаптивних методів навчання еволюційних нейро-фаззі систем для вирішення задач кластеризації в онлайн режимі.

Мета дослідження – є розробка онлайн методів нечіткої кластеризації нерівномірно квантованих асинхронних нестаціонарних часових рядів для підвищення ефективності та скорочення часу обробки в інтелектуальному аналізі потоків даних.

Об'єкт дослідження – процес інтелектуального аналізу потоку даних у формі часових рядів.

Предмет дослідження – методи інтелектуального аналізу для нечіткої онлайн кластеризації багатовимірних часових рядів з асинхронними тактами квантування, що призначені для аналізу потоків даних.

Задачі дослідження:

– проведення аналізу існуючих методів та підходів до кластеризації часових рядів;

– розробка методу передобробки часових рядів, що спотворені аномальними викидами та збуреннями;

– розробка онлайн модифікації методу кластеризації коротких часових рядів;

– розробка методу онлайн кластеризації багатовимірних часових рядів;

– розробка методу онлайн кластеризації асинхронних часових рядів, несхильного до впливу ефекту концентрації норм;

– проведення експериментів на основі тестових та реальних даних.

Практична значущість отриманих результатів. Використання запропонованих моделей та методів дозволяє підвищити ефективність застосування сучасних моніторингових систем для вирішення задач кластеризації даних, які послідовно надходять на обробку з нерівномірними тактами квантування, що базуються на апараті гібридних систем обчислювального інтелекту.

Реалізовано модуль із запропонованими методами, що підтвердив свою ефективність у задачах моніторингу медичних даних в онлайн режимі. У такому разі моніторинг медичних даних дозволяє ефективно виявляти аномалії у хворих в онлайн режимі.

Результати досліджень впроваджені у ТОВ «Інфобуд», м. Харків (акт впровадження від 03.10.2018 р.) та ТОВ «Сайтосс», м. Харків (акт впровадження від 06.10.2018 р.). Результати досліджень впроваджені у Харківському національному університеті радіоелектроніки на кафедрі

штучного інтелекту в освітній процес з дисципліни «Нейромережеві методи обчислювального інтелекту».

Особистий внесок здобувача. Усі положення, що виносяться на захист, основні результати теоретичних та експериментальних досліджень отримані здобувачем особисто (додаток Б). У публікаціях, написаних у співавторстві, автору належать: [1] – запропонований метод нечіткої кластеризації часових рядів з асинхронними тактами квантування з використанням поліноміального фаззіфікатора; [3] – запропонована адаптивна ідентифікація нестаціонарних часових рядів на основі модифікованого робастного критерія Гемана- МакКлюра та ансамбль адаптивних моделей на його основі; [4] – запропонована самоорганізовна нейро-фаззі мережа для нечіткої кластеризації багатовимірних часових рядів; [5] – запропонований метод адаптивної нейро-фаззі ймовірнісної кластеризації багатовимірних потоків даних; [2] – запропонований метод адаптивної нечіткої кластеризації коротких часових рядів; [7] – запропонована цільова функція нечіткої кластеризації на основі сферичної норми; [6] – запропонована модифікована функція Гемана-МакКлюра; [8] – запропонована оцінка відстані між рядами на основі тангенсів кутів нахилу; [9] – запропоновано використання WTM - правила самонавчання для нечіткої кластеризації часових рядів; [10] – запропоновано зважування координат у просторі ознак.

Апробація результатів дисертації. Основні результати дисертаційної роботи доповідалися й обговорювалися на конференціях: First International Conference on Data Stream Mining & Processing (23-27 August 2016, Lviv); XIV Міжнародній науково-технічній конференції «Advanced Trends in Radioelectronics, Telecommunications and Computer Engineering» (20-24 лютого 2018 р., м. Славське); Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту-ISDMCI-2016 (24-28 травня 2016 р., Залізний порт, м. Херсон); VIII Міжнародній школі-семінарі «Теорія

прийняття рішень» (26 вересня - 1 жовтня 2016 р., м. Ужгород); Інтелектуальні системи прийняття рішень та проблеми обчислювального інтелекту - ISDMCI-2017 (22–26 травня 2017 р., Залізний порт, м. Херсон); XIX Міжнародному молодіжному форумі «Радіоелектроніка та молодь у XXI столітті» (20-22 квітня 2015 р., м. Харків); XX Ювілейному міжнародному молодіжному форумі «Радіоелектроніка та молодь у XXI столітті» (19-21 квітня 2016 р., м. Харків).

Публікації. За тематикою дослідження опубліковано 12 наукових праць (додаток Б), з них 1 розділ у колективній монографії, що входить до наукометричної бази SCOPUS; 1 стаття за кордоном, що входить до наукометричної бази SCOPUS; 3 статті у виданнях, які зазначені в переліках фахових видань України з технічних наук, 7 публікацій у матеріалах конференцій (2 включено до наукометричної бази даних SCOPUS) .

1 ОГЛЯД СТАНУ ПРОБЛЕМИ І ПОСТАНОВКА ЗАВДАННЯ ДОСЛІДЖЕННЯ

В існуючих умовах динамічної обробки інформації та у зв’язку з необхідністю аналізу часових рядів в інтелектуальному аналізі потоків даних, необхідно проаналізувати стан проблеми кластеризації часових рядів та існуючі підходи до її вирішення. Пропонується аналіз існуючих публікацій з використанням стандартних метрик кластеризації часових рядів, їх обробки за умов недостатньої кількості даних у ряді, аномалій, а також при неможливості побудови моделей або при використані методів на основі статистики. Ставляться науково-практичні завдання дослідження в рамках інтелектуального аналізу даних.

Необхідно проаналізувати стан проблеми нечіткої кластеризації часових рядів та існуючі підходи до її вирішення. Розглянути різноманітні метрики, які застосовуються у методах нечіткої кластеризації часових рядів, на основі існуючих методів та метрик.

Мета – постановка задач для дослідження на основі огляду джерел, у яких розглядаються існуючі моделі, методи і алгоритми нечіткої кластеризації часових рядів, за допомогою об'єднання апаратів нейронних мереж та нечіткої логіки.

Завдання: 1) провести огляд існуючих методів, проаналізувати стан проблеми кластеризації часових рядів та підходи до її вирішення; 2) розглянути основні принципи нечіткої логіки та систем нечіткого розбиття; 3) провести аналіз існуючих методів кластеризації, методів їх навчання і самонавчання, що використовуються для вирішення завдань нечіткої кластеризації даних; 4) з’ясувати, як об'єднання апаратів нейронних мереж і нечіткої логіки може працювати в задачах нечіткої кластеризації часових рядів; 5) сформулювати задачу дослідження.

1.1 Види часових рядів та їх властивості

Часовий ряд – це множина спостережень, які генеруються послідовно у часі. Якщо час неперервний, часовий ряд називається процесом, а якщо час змінюється дискретно – дискретним часовим рядом. Спостереження дискретного часового ряду, зроблені в моменти часу $t_1, t_2, ...t_N$, можуть бути позначені як $x(t_1), x(t_2), ..., x(t_N)$. Часовий ряд, який підлягає аналізу, може розглядатися як одна часткова реалізація системи, що навчається та генерується прихованим ймовірнісним або іншим механізмом [1].

Розрізняють інтервальні, моментні та довільні часові ряди.

1. Інтервальним рядом називається такий ряд, кожен рівень якого характеризує явище за певний відрізок часу, де рівні характеризують значення показника у ряді за певні періоди часу, тому особливістю інтервальних часових рядів є можливість сумування їх рівнів.

2. Моментним рядом називається такий ряд, кожен рівень якого характеризує виникненням на певний момент часу. На відміну від інтервального ряду рівні моментного не піддаються перетворенням: їх не можна дробити або підсумовувати для утворення довільних рядів. У моментних часових рядах рівні характеризують значення показника станом на певні моменти часу. Окремі ж рівні даного ряду абсолютних величин містять елементи повторних відліків.

3. До часових (довільних) рядів належать ряди середніх показників і ряди відносних величин. Такі ряди утворюються розрахунковим шляхом на основі інтервальних і моментних рядів [1].

Часові ряди класифікуються:

а) за формою подання рівнів:

– ряди абсолютних показників;

– відносних показників;

– середніх величин;

б) за кількістю показників, для яких визначаються рівні в кожен момент часу:

– одновимірні часові ряди;

– багатовимірні часові ряди;

в) за характером часового параметра:

– моментні часові ряди;

– інтервальні часові ряди.

За відстанню між датами та інтервалами часу виділяють рівновіддалені – коли дати реєстрації або закінчення періодів слідують один за одним з рівними інтервалами та неповні (нерівномірні) – коли принцип рівних інтервалів не дотримується [2].

За наявністю пропущених значень виділяють повні і неповні часові ряди. Часові ряди бувають детермінованими і випадковими. Перші отримують на основі значень деякої невипадкової функції (ряд послідовних даних про кількість днів у місяцях), другі є результатом реалізації деякої випадкової величини.

Залежно від наявності основної тенденції виділяють стаціонарні ряди, в яких середнє значення і дисперсія постійні, і нестаціонарні, що містять основну тенденцію розвитку.

За кількістю тактів квантування у ряді можлива класифікація рядів на рівномірно та нерівномірно квантовані.

Рівномірне (однорідне) квантування базується на принципі розбиття діапазону значень тактів сигналу на відрізки однакової довжини і заміни цих значень на рівень квантування.

Таке квантування застосовується для збільшення точності квантування у ситуаціях, коли розподіл значень є нерівномірним [3].

1.1.1 Передобробка часових рядів

Основою для попередньої обробки є процедури швидких дискретних ортогональних перетворень, які реалізуються в різних функціональних базисах; процедури лінійної і нелінійної фільтрації, лінійної алгебри.

Основним завданням обробки є усунення аномальних викидів і шумів. Вирішити дане завдання повною мірою можна тільки в тому випадку, якщо ряд надходить з точно визначеними параметрами, тобто чим більше надходить корисних даних і чим менше перешкод, тим більша ймовірність виконати завдання якісно.

Усі завдання, які вирішуються методами аналізу даних, можна умовно розбити на шість класів: класифікація, регресія, кластеризація, асоціація, послідовні шаблони, аналіз відхилень. Проблеми аналізу формуються по-іншому, але рішення більшості з них зводиться до тієї чи іншої задачі аналізу даних або до їх комбінації. Для вирішення вищеописаних завдань використовуються різні методи і алгоритми аналізу даних. З огляду на те, що аналіз даних розвивається на стику таких дисциплін, як статистика, теорія інформації, машинне навчання та теорія баз даних, цілком закономірно, що більшість алгоритмів і методів аналізу були розроблені на основі різних методів з цих дисциплін [1].

1.1.2 Типи забруднень часових рядів

Гауссовський «білий» шум це нормальний розподіл ймовірностей, для одновимірного випадку. Функція Гаусса залежить від двох параметрів:

1) математичного сподівання (μ);

2) стандартного відхилення (σ).

У разі стандартного нормального розподілу перший параметр дорівнює нулю, другий – одиниці.

До адитивного належить спосіб, при якому білий шум і корисний сигнал підсумовуються, іноді гауссівський шум помилково вважають обов'язково білим, але ці два поняття не є еквівалентними. Для гауссівського шуму характерний нормальний розподіл значень сигналу. Термін «білий шум» стосується кореляції сигналу в два різних моменти часу, але кореляція від розподілу амплітуди шуму не залежить, тому «білий» шум може мати будь-який розподіл: Гаусса, Пуассона, Коші та ін. Назву «білий» шум отримано від білого світла, яке містить електромагнітні хвилі частот усього видимого діапазону електромагнітного випромінювання.

Ідеальний «білий шум», а саме шум з однаковою спектральною потужністю на всіх частотах, у природі і техніці не зустрічається. Інакше сигнал мав би нескінченну потужність, а під категорію білих потрапляють будь-які шуми з однаковою (або слабкою відмінністю) спектральною щільністю в певному частотному діапазоні [2].

1.2 Методи формування векторів-ознак для часових рядів на основі статистичних досліджень

Для малих, за обсягом, статистичних досліджень моделювання стає необхідною умовою забезпечення цілісності дослідження. Важливо, щоб зібрані у статистичному дослідженні дані були опрацьовані та відредаговані перш, ніж до них будуть застосовані основні статистичні методи. Перевірка даних необхідна для виявлення грубих помилок у дослідженні, а також помилок, допущених при кодуванні і перетворенні даних; виявлення

можливих викидів або аномальних спостережень. Існує декілька основних методів формування ознак, в основі яких лежать статистичні дослідження:

1. Метод максимальної правдоподібності, який дає можливість оцінювання невідомого параметра шляхом максимізації функції правдоподібності: $L(X_1,...,X_n;\hat{\theta}_{МП}) = \max L(X_1,...,X_n;\hat{\theta})$, де $L(X_1,\mathrm{X}_2,...,X_n\theta) = = f(X_1;\theta)\times f(X_2;\theta)\times...\times f(X_n;\theta)$ –функція правдоподібності, а $f(X_1;\theta)$ – закон розподілу вірогідності. Але для забезпечення працездатності методу максимальної правдоподібності необхідне точне визначення типу аналізованого закону розподілу $f(X_1;\theta)$, також даний метод є неефективним при малих обсягах вибірок [4–6].

2. Метод моментів, який полягає в прирівнюванні певної кількості вибіркових моментів до відповідних теоретичних моментів досліджуваної випадкової величини, причому останні є функціями від невідомих параметрів $\theta^{(1)},...,\theta^{(\mathrm{k})}$.

Розглядаючи кількість моментів, яка дорівнює k, що підлягають оцінці параметрів, і вирішуючи отримані рівняння щодо цих параметрів, отримуємо оцінки. Таким чином, оцінки $\hat{\theta}^{(1)}_{ММ},...,\hat{\theta}^{(\mathrm{k})}_{ММ}$ за методом моментів невідомих параметрів $\theta^{(1)},...,\theta^{(\mathrm{k})}$ є рішеннями системи рівнянь:

$$\begin{cases} \int x^{(t)} \times f(X;q)dX = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i^{(l)}, l = 1,2,..., p; \\ \int x^{(t)} \times x^{(m)} \times f(X;q)dX = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i^{(l)} x_i^{(m)}, \\ l,m = 1,2,..., p. \end{cases} \qquad (1.1)$$

До переваг методу моментів слід віднести його порівняно просту обчислювальну реалізацію, а також те, що оцінки, отримані як рішення

системи (1.1), є функціями від вибіркових моментів, що значно спрощує дослідження статистичних властивостей оцінок методу моментів. Тому результати, одержані за допомогою даного методу, приймаються за перше наближення [4–6].

3. Метод найменших квадратів (МНК) є одним із методів регресійного аналізу і призначений для оцінки невідомих величин за результатами вимірів, що містять випадкові похибки.

На рисунку 1.1 зображено графічну інтерпретацію причин, що обумовлюють необхідність використання МНК:

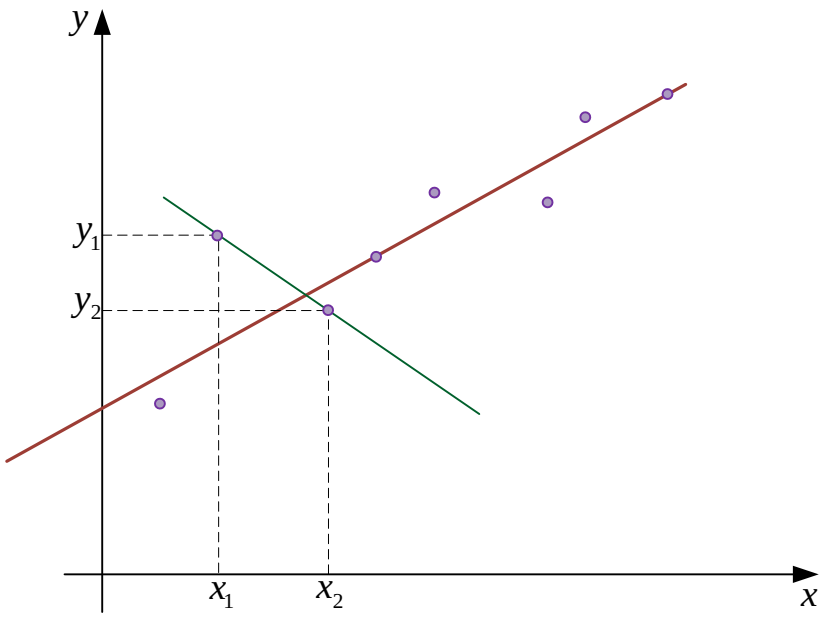


Рисунок 1.1 – Графічна інтерпретація МНК

Нехай відомо, що вихідний параметр процесу y лінійно залежить від вхідного параметра x (червона лінія на рис. 1.1). Статистична характеристика цього процесу може бути подана у вигляді $y = ax + b$, де a і b – коефіцієнти. Для визначення числових значень необхідно задати два значення x_1, x_2 вхідній величині x і заміряти відповідні їм значення y_1, y_2 вихідної величини y, оскільки лише за виконання цих умов можна скласти систему двох алгебраїчних рівнянь із двома невідомими a і b:

$$\begin{cases} y_1 = ax_1 + b, \\ y_2 = ax_2 + b. \end{cases} \quad (1.2)$$

Гаусс запропонував для визначення коефіцієнтів a і b моделі $y = ax + b$ сформувати суму квадратів різниць $\sum_{i=1}^{n}$ між теоретично заданими значеннями вихідної координати y_1 при значеннях аргументу x_i, $i = 1...\mathrm{N}$ та її експериментальними значеннями y_i:

$$\sum_{i=1}^{n} = \min \sum_{i=1}^{N} (y(x_i) - y_i)^2 , \quad (1.3)$$

а потім знайти значення коефіцієнтів a, b рівняння $y = ax + b$, для мінімізації виразу (1.3).

Методу найменших квадратів притаманні наступні особливості:

– простота і ефективність для вибраної структури моделі функціональної залежності та оптимальні значення її параметрів у межах заданого діапазону значень та його функції;

– вибір виду та структури моделі є прерогативою дослідника;

– клас заданих структур МНК дає змогу отримати оптимальні значення коефіцієнтів і оптимальну структуру моделі.

За допомогою обчислювальних алгоритмів МНК можна розв’язувати задачі оптимізації відновлення сигналів та задачі оптимального синтезу динамічної характеристики системи, як недолік можна відзначити неможливість синтезу основної моделі прогнозу [4–6].

1.3 Проблеми кластеризації та класифікації часових рядів

На сьогоднішній час використання, для обробки часових рядів, методів кластеризації має ряд істотних недоліків, наприклад, чутливість до змін у часовому ряді, надлишкова інформація, висока вартість обробки даних.

Метою кластеризації часових рядів є необхідність знайти невідомі закономірності у ряді, тому що ряди є не періодичними, а значення, що знаходяться у ряді, не можуть бути розпізнані з причини надлишкової або недостатньої інформації[92,98]. Таким чином, одним з основних завдань в області інтелектуального аналізу даних є завдання кластеризації, а саме завдання розбиття вихідних даних на однорідні групи в режимі навчання без вчителя (самонавчання) [17].

На рисунку 1.2 зображено графічну інтерпретацію, що обумовлюють необхідність кластеризації часових рядів, а на рисунках 1.3, 1.4 та 1.5 зображено можливі результати кластеризації.

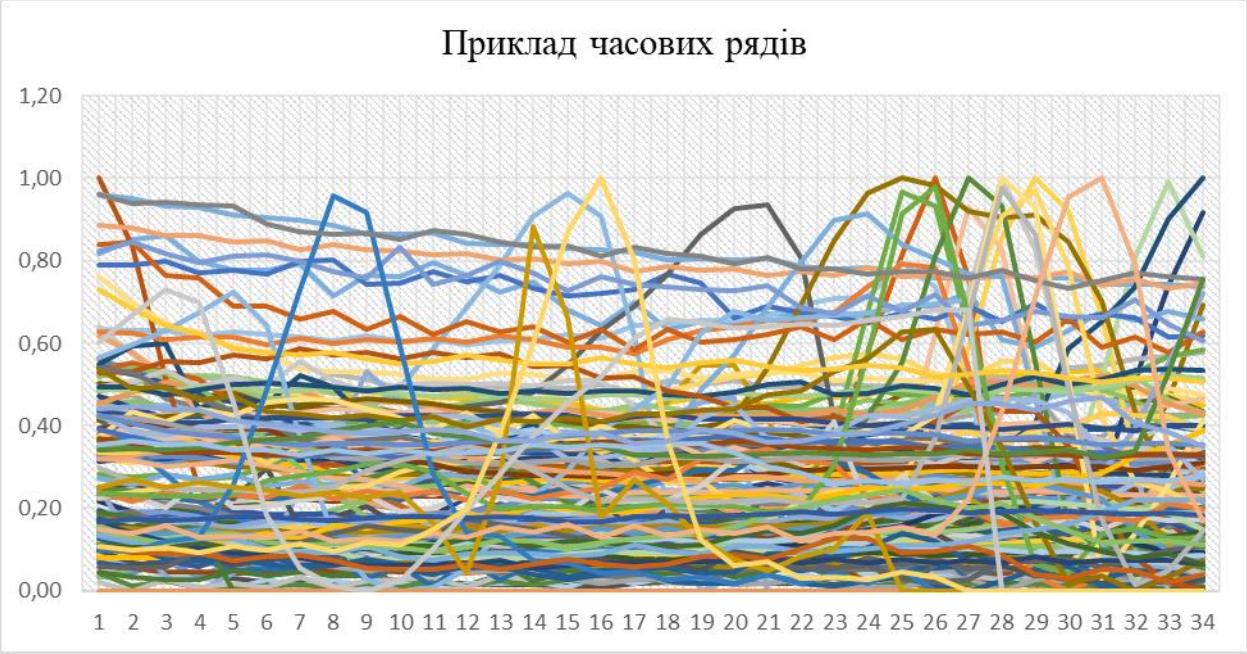


Рисунок 1.2 – Графічна інтерпретація часових рядів, що подаються на обробку

Рисунок 1.3 – Графічна інтерпретація результату кластеризації часових рядів по кластеру 1

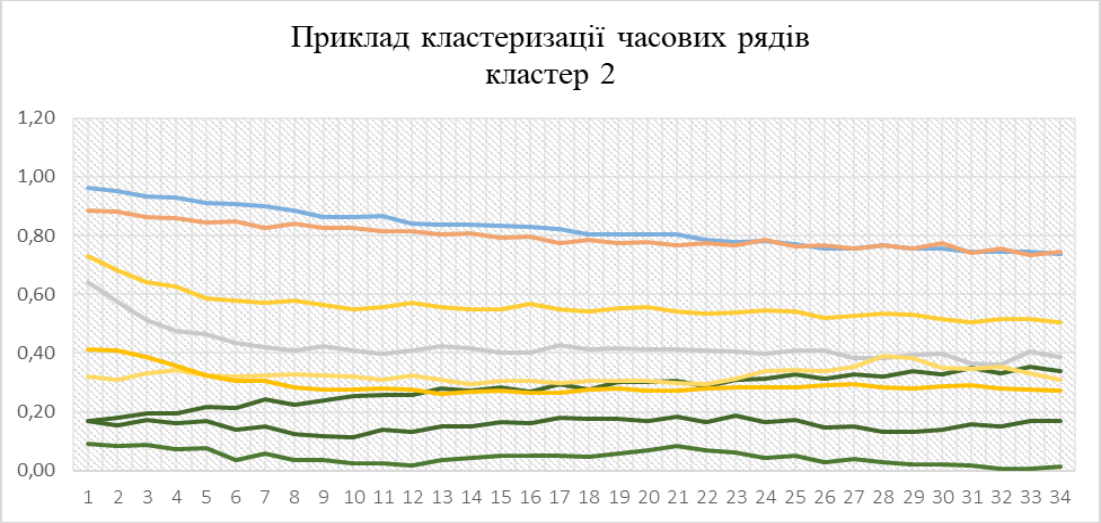


Рисунок 1.4 – Графічна інтерпретація результату кластеризації часових рядів по кластеру 2

Рисунок 1.5 – Графічна інтерпретація результату кластеризації часових рядів по кластеру 3

Кластеризація – це одна з найбільш важливих проблем неконтрольованого навчання. Як і кожна проблема такого типу, вона має справу з виявленням прихованої структури у сукупності даних.

Об’єднання в кластери – це процес організації об’єктів у групи, члени яких подібні до певної міри, а кластер – це безліч об’єктів, близьких між собою у смислі деякої міри схожості. У просторі змінних кластери являють собою скупчення точок (об’єктів) різної форми (рис. 1.6).

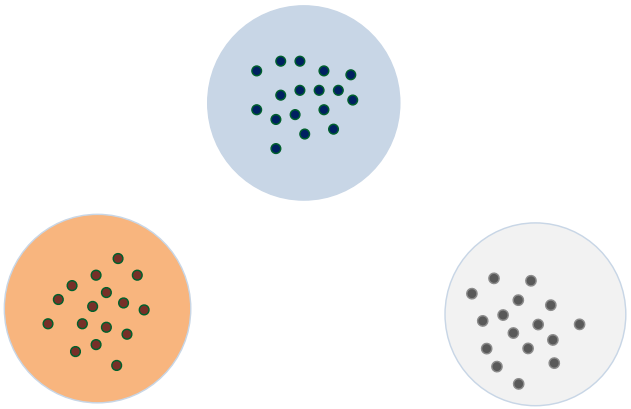


Рисунок 1.6 – Кластеризація

У випадку, представленому на рисунку 1.6, мірою подібності є відстань: два або більше об'єктів належать до одного кластеру, якщо вони «близькі», то, відповідно, це кластеризація на підставі відстані.

Інший тип кластеризації – нечітка кластеризація: де два або більше об'єктів належать до одного кластеру одночасно. Іншими словами, об'єкти групуються таким чином, щоб відповідати певному кластеру.

Кластерний аналіз займає одне з центральних місць серед методів аналізу даних і являє собою сукупність підходів, методів і алгоритмів, призначених для знаходження деякого розбиття досліджуваної сукупності об'єктів на підмножини щодо подібних, схожих між собою об'єктів [17]. При цьому для виділення таких підмножин, які отримали спеціальну назву «кластер», які так само іноді називають «таксонами» або просто класами,

служить лише неформальне припущення про те, що об'єкти, які належать до одного кластеру, повинні мати більшу схожість між собою, ніж з об'єктами з інших кластерів [16].

Істотна відмінність між кластерами даних часового ряду та кластеризацією об’єктів в евклідовому просторі полягає у тому, що кластер з часовими рядами може не мати однакової довжини. Якщо це не так, то всі часові ряди мають однакову довжину, і стандартні методи кластеризації можуть бути застосовані шляхом подання кожного часового ряду у вигляді вектора та з використанням традиційної відстані L_p – норми. Такий підхід можна використовувати тільки при схожості у часі, а подібності, як за формою так і в зміні значення не враховуються [7].

1.4 Класифікація методів кластеризації часових рядів

Інтуїтивно проблема вибору ознак тісно пов'язана з проблемою визначення кластерної тенденції набору функцій. Існують різні методи кластеризації (рис. 1.7). Методи вибору елементів визначають підмножини ознак, які максимізують основну тенденцію кластеризації. Існують два основних класи моделей для виконання вибору функції.

1. Моделі фільтрів. У цьому випадку оцінка пов'язана з кожною цільовою функцією з використанням критеріїв подібності. Даний критерій є фільтром, що забезпечує чітку умову для видалення функцій, які не відповідають необхідному результату. У деяких випадках ці моделі можуть кількісно визначати якість підмножини функцій як комбінацію, а не одну особливість. Такі моделі є більш потужними, оскільки вони неявно враховують зростаючий вплив додавання функції.

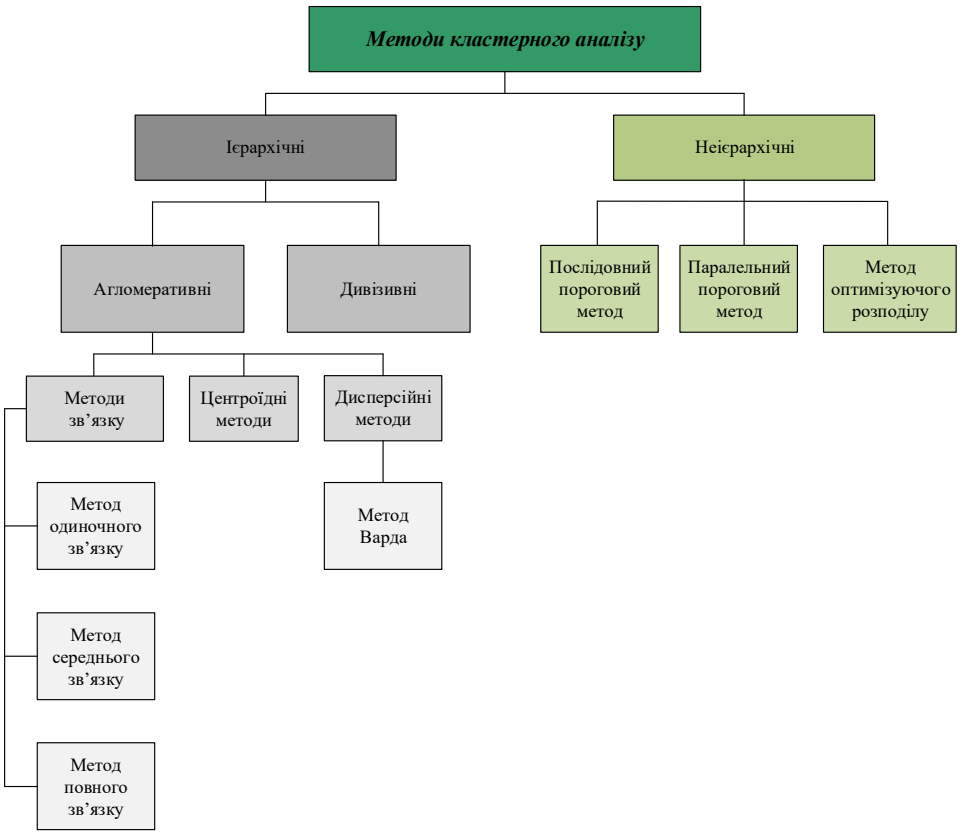


Рисунок 1.7 – Методи кластерного аналізу

2. Моделі згортки. У даному випадку алгоритм кластеризації використовується для оцінки якості підмножини ознак, що використовують для уточнення підмножини функцій, на яких виконується кластеризація. Це ітеративний підхід, при якому вибір функцій залежить від кластерів і навпаки. Обрані функції будуть залежати від конкретної методології, що використовується для кластеризації. Різні методи кластеризації можуть працювати краще з різними наборами функцій. Отже, дана методологія також може оптимізувати вибір функції для конкретної технології кластеризації. З іншого боку, внутрішня інформативність конкретних функцій може іноді не відбиватися в цьому підході через вплив конкретної кластеризації. Основна відмінність між фільтрами та згортковими моделями полягає у тому, що перша може бути виконана як фаза попередньої обробки, тоді як остання інтегрована безпосередньо у процес кластеризації.

Визначення кластерів часових рядів надзвичайно складне через труднощі у визначенні подібності в різних часових рядах, які можна масштабувати як за часовими, так і за поведінковими вимірами, тому концепція подібності важлива для кластеризації даних часового ряду. Коли міру подібності було визначено для часу, серію даних можна розглядати як абстрактний об'єкт, на якому використовуються методи, засновані на схожості, а саме спектральні методи або методи розбиття. Такі часові ряди дозволяють використовувати різноманітні формулювання для процесу кластеризації в залежності або від серій, що групуються на основі їх онлайн-кореляцій, або від тих, що групуються на основі їх форм.

Кластеризація часових рядів для всіх типів даних має на меті створення кластерів з високим ступенем внутрішньокластерної подібності і, одночасно, з низькою ідентичністю між різними кластерами. Зокрема, об'єкти, що належать до одного і того ж кластеру, повинні мати високий ступінь подібності між собою, тоді як об'єкти, що належать різним кластерам, повинні мати низьку схожість. Тому такі властивості як висока розмірність, наявність шуму і висока кореляція створюють унікальні завдання для проектування ефективних алгоритмів кластеризації [8].

1.5 Методи кластеризації на основі аналізу кореляцій в пакетному та онлайн режимі

Такі методи також можуть бути використані для того, щоб знайти час, який має подібні тенденції або, принаймні, корельовані тенденції, у випадках, пов'язаних з проблемами прогнозування і регресії часових рядів. Фактично деякі такі методи кластеризації використовують кореляційну онлайн-кластеризацію для вибору потоку та ефективного прогнозування.

Важливим аспектом є те, що їх часто необхідно виконувати у реальному часі, оскільки потоки еволюціонують з часом. Тому бажано виявляти істотні зміни в поведінці, що надає уявлення про те, як потік кореляційних тенденцій змінюється.

Офлайн кластеризація на основі форми: в цих випадках визначаються часові ряди аналогічних форм з даних. Основне завдання полягає в тому, щоб визначити відповідну схожість за формою функції. Залежно від області визначення часові ряди можуть бути масштабовані і деформовані. Тому функції подібності, такі як евклідова функція або динамічне деформування часу, використовуються для кластеризації часових рядів. Кластеризація більш результативна при багатовимірній кластеризації часових рядів, які вирішуються онлайн-методами на основі кореляції і тісно пов'язані з проблемою прогнозування.

Два потоки в одному кластері можуть бути позитивно корельовані. Фактично, такі два потоки з ідеальною негативною кореляцією можуть також належати до одного і того ж кластеру, якщо передбачуваність між різними потоками в пріоритеті, що частіше трапляється в багатьох реальних сценаріях, в яких деякі потоки можуть бути передбачені.

З іншого боку, в неконтрольованій кореляційній кластеризації бажано визначити найкращий набір представників, який може прогнозувати всі потоки даних [9].

1.6 Методи ієрархічної кластеризації

Існує два типи ієрархічної кластеризації – агломеративні та дивізивні, їх структуру зображено на рисунку 1.8. Перший полягає у визначенні меж кластерів як найбільш щільних ділянок у багатовимірному просторі вихідних

даних, тобто визначення кластера там, де є велике «згущення точок». Другий підхід полягає в мінімізації відмінності меж кластерів у багатовимірному просторі вихідних даних [10]:

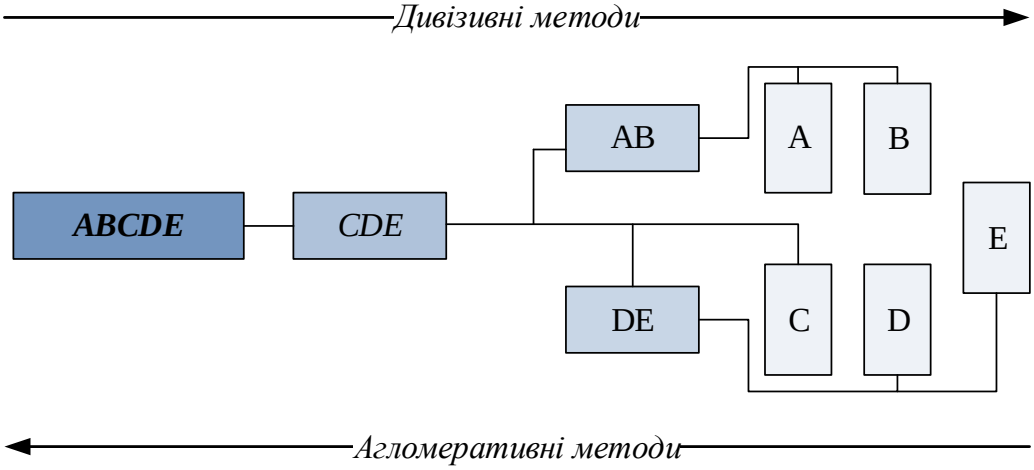


Рисунок 1.8 – Структура ієрархічних методів кластерного аналізу

Агломеративна ієрархічна кластеризація починається з розгляду кожного об'єкта даних як окремого кластера і продовжує пошук найбільш схожої пари кластерів. Тоді найбільш подібна пара об'єднується в один кластер, і процес триває до тих пір, поки не буде досягнуто бажану кількість кластерів [11].

Агломеративна ієрархічна кластеризація, загальна схема роботи зображена на рисунку 1.9, має безліч варіантів вибору двох кластерів, які є найбільш близькими один до одного і тому повинні бути об'єднані на поточному кроці, наприклад:

– єдиний зв'язок: при виборі одиночного зв'язку відстань між двома кластерами визначається як найкоротша відстань між усіма об'єктами.

Відстань у цьому випадку між кластерами C_i та C_j має вигляд:

$$D_{SL}\left(C_i, C_j\right) = \min \in C_i, y \in C_j \left(dist\left(x, y\right)\right), \tag{1.5}$$

де *dist* – обрана міра відстані;

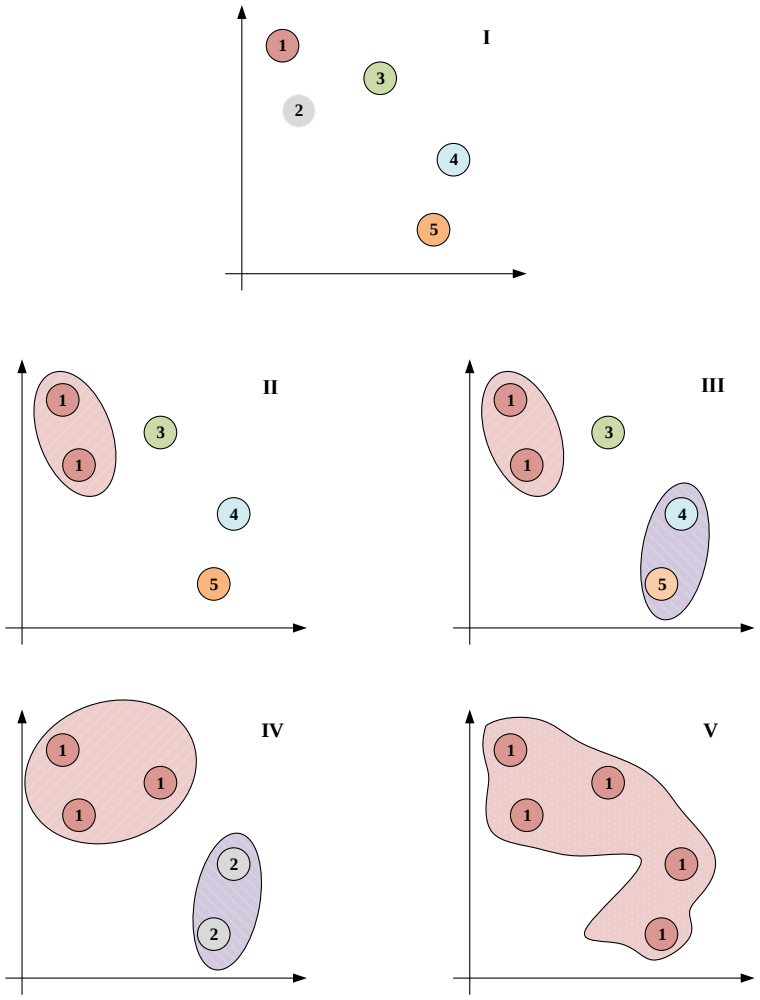


Рисунок 1.9 – Ієрархічні агломеративні методи кластерного аналізу

– повна прив'язка: при повному виборі зв'язку відстань між двома кластерами дорівнює відстані між усіма об'єктами-членами.

Відстань у цьому випадку між кластерами C_i та C_j, має вигляд:

$$D_{CL}\left(C_i, C_j\right) = \max_{x \in C_i, y \in C_j} \left(dist\left(x, y\right)\right), \tag{1.6}$$

де *dist* – обрана міра відстані;

– середній зв'язок: при виборі середньої прив'язки відстань між двома кластерами визначається як середня відстань між усіма об'єктами-членами. Відстань у цьому випадку середньої лінії між кластерами C_i и C_j, має вигляд:

$$D_{AV}\left(C_i, C_j\right) = avgx \in C_i, y \in C_j \left(dist\left(x, y\right)\right). \quad (1.7)$$

Ієрархічні дивізивні (розподільні) методи є методами розділення великого макрокластеру, що містить всі елементи та розділяється на дві групи, кожна з яких, у свою чергу, також розділяється на дві групи і так далі, схема роботи зображена на рисунку 1.10. Таким чином генерується ієрархія кластерів «зверху вниз».

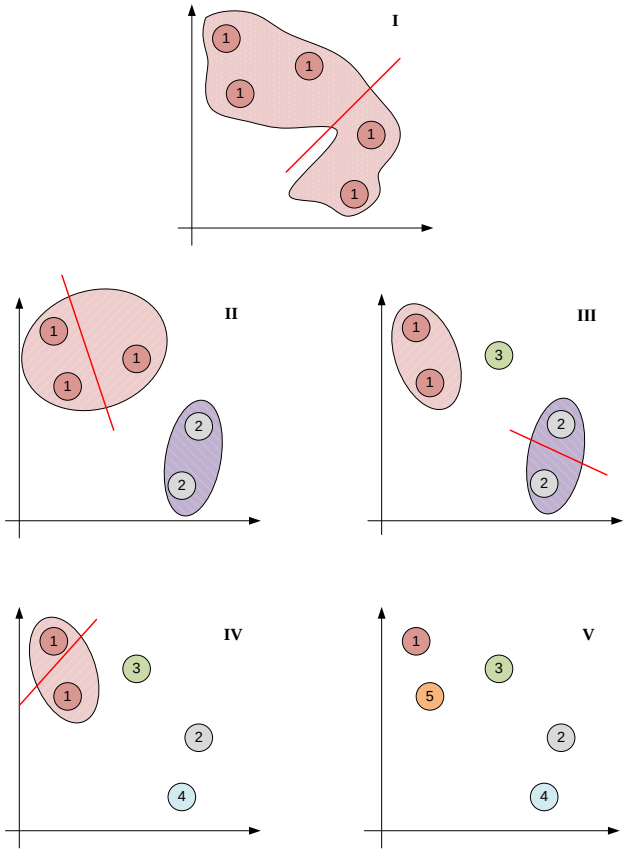


Рисунок 1.10 – Загальна схема роботи ієрархічних дивізивних методів

Але за умов обробки великої кількості спостережень ієрархічні методи кластерного аналізу не можуть працювати ефективно [12].

1.7 Алгоритм кластеризації *k*-середніх

Алгоритм *k*-середніх є найбільш широко використовуваним методом кластеризації, оскільки він групує аналогічні об'єкти в одному кластері і використовує метод ітераційної обробки при мінімізації функції помилки. Принцип дії алгоритму складається з початкового пошуку *k* початкових центрів кластерів та вибору *k* довільних об'єктів, потім призначається об'єкт ідентичний кластеру.

Ідентичним кластером є кластер з найближчого центру, відповідно до деякої функції відстані, наприклад, евклідової або динамічної деформації часу.

Після цього перераховуються кластерні центри *k* шляхом усереднення всіх призначених об'єктів для кожного кластера. Потім кластерні центри більше не переміщуються, поки функція помилки, яка є сумою квадратів помилок кожного центру кластера та його призначені об'єкти не будуть зведені до мінімуму.

Особливістю методу *k*-середніх є те, що як метрика використовується евклідова відстань, число кластерів наперед не відоме і вибирається заздалегідь, а якість кластеризації залежить від початкового розбиття. На рисунку 1.11 наведено приклад роботи алгоритму *k*-середніх:

Конструктивно алгоритм є ітераційною процедурою наступного виду:

Крок 1. Проініціалізувати початкове розбиття $U^{(-1)}$ (наприклад, випадковим чином), обрати точність δ (використовується у критерії завершення алгоритму), проініціалізувати номер ітерації $l = 0$.

Крок 2. Обрахувати центри кластерів за наступною формулою:

$$c_i^{(l)} = \frac{\sum_{j=1}^{|X|} u_{ij}^{(l-1)} \cdot x_j}{\sum_{j=1}^{|X|} u_{ij}^{(l-1)}}, \tag{1.8}$$

де $1 \le i \le c$.

Рисунок 1.11 – Приклад роботи алгоритму *k*-середніх

Крок 3. Оновити матрицю розбиття з тим, щоб мінімізувати квадрати помилок, використовуючи наступний вираз:

$$u_{ij}^{(l)} = \begin{cases} 1, \ при \ \ d\left(x_j, c_l^{(l)}\right) = \min\limits_{l \le k \le |C|} d\left(x_j, c_k^{(l)}\right), \\ 0, \ в \ інших \ випадках. \end{cases} \tag{1.9}$$

Крок 4. Перевірити умову $\left\|U^{(l)} - U^{(l-1)}\right\| < \delta$, якщо умова виконується – завершити процес, якщо ні – переходимо до кроку 2 та з номером ітерації $l = l + 1$.

Основний недолік полягає у дискретній множині значень елементів матриці належності, що часто невиправдано загрубляє рішення.

Складністю алгоритму *k*-середніх є $O(k \cdot N \cdot r \cdot D)$, де k – кількість очікуваних кластерів, N – кількість об'єктів, що підлягають кластеризації (що дорівнює розміру набору даних), r – число ітерацій для досягнення збіжності, D – розмірність простору об'єктів.

Неправильний вибір початкового числа кластерів *k* може привести до некоректних результатів. Саме тому при використанні методу *k*-середніх важливо спочатку провести перевірку відповідного числа кластерів для даного набору даних [15].

1.8 Нечіткий алгоритм кластеризації *c*-середніх для часових рядів

Взаємозв'язок між кластерним аналізом і теорією нечітких множин являє собою приклад структуризації складних систем, де більшість об'єктів виявляються неточними за своєю природою. Неточність полягає в тому, що перехід від належності до неналежності елементів в даних класах швидше поступовий.

Для будь-якої міри схожості величина належності спостереження кластеру залежить від схожості об'єкта і прототипу цього кластера. У разі, якщо мірою подібності є відстань, величина належності об'єкта обернено пропорційна його відстані до центроїда кластера.

Сума належностей спостережень кластерам у будь-який момент часу повинна дорівнювати 1.

Основні ідеї алгоритму для вирішення завдання нечіткої кластеризації були запропоновані Дж. К. Даному в 1973 р. [14]. Надалі алгоритм був розвинений Дж. Бездеком [13] і отримав назву нечітких *c*-середніх (FCM).

Як результат виконання даного алгоритму визначається локально-оптимальне нечітке розбиття, яке описується сукупністю функцій належності. На рисунку 1.12 зображено алгоритм обчислення нечіткої кластеризації [13].

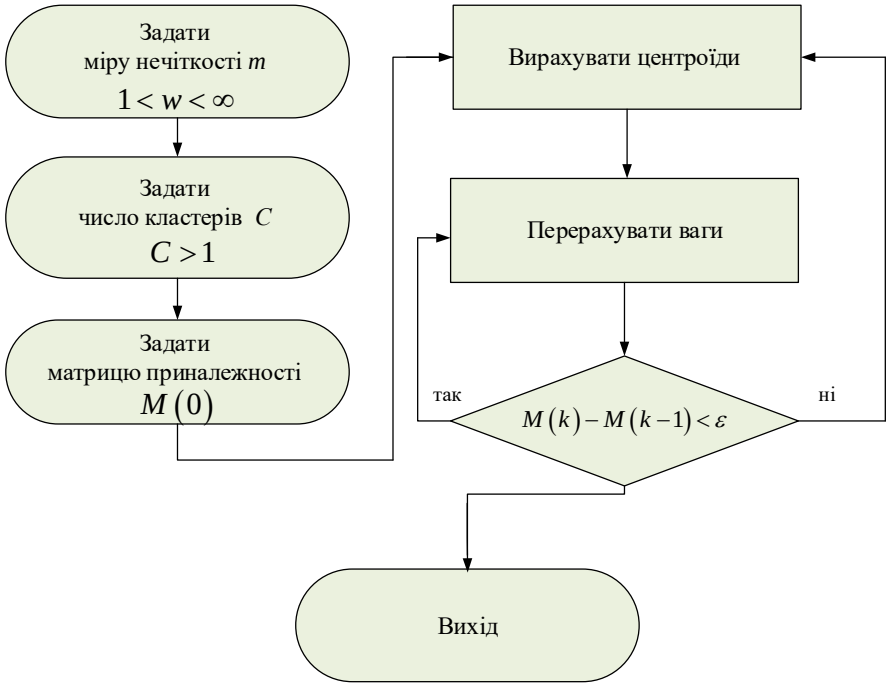


Рисунок 1.12 – Алгоритм обчислення нечіткої кластеризації

Поряд з традиційним ймовірнісним підходом до нечіткої кластеризації, коли кожен об'єкт з певною ймовірністю належить до кожного з кластерів, існує можливісний підхід до кластерного аналізу.

Можливісна кластеризація також розглядає нечіткі кластери і відповідні їм функції належності, що приймають значення з інтервалу [0,1]. Різниця полягає в тому, що ймовірнісна кластеризація передбачає наявність

строгого обмеження, а саме–сума належності об'єкта до всіх кластерів дорівнює 1, а можливісний кластерний аналіз не має подібного обмеження.

Алгоритм FCM, наведений на рисунку 1.13, має ітеративний характер послідовного поліпшення деякого нечіткого розбиття, яке задається користувачем або формується автоматично за деяким евристичним правилом [13]. Формально алгоритм FCM визначається у формі ітеративного виконання деякої послідовності кроків.

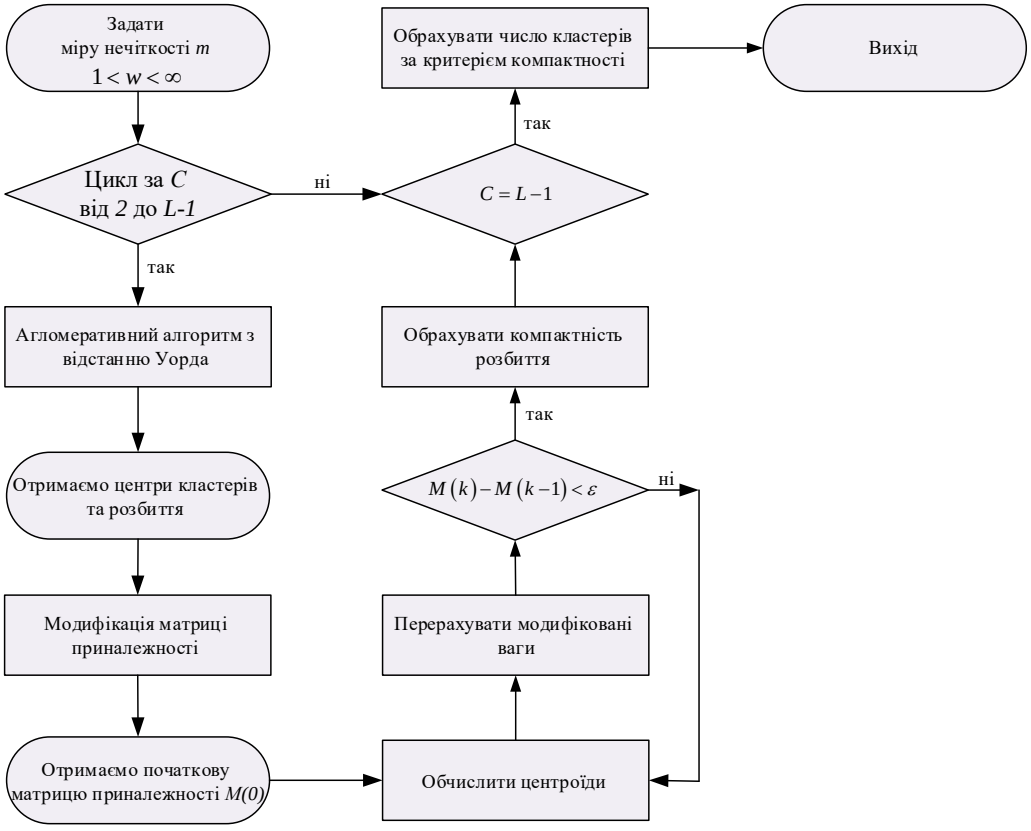


Рисунок 1.13 – Блок–схема FCM

На кожній з ітерацій рекуррентно перераховуються значення функцій належності об'єктів нечітким кластерам та їх типові представники (центроїди).

Алгоритм закінчить роботу у разі, коли виконання заданого апріорі деякого кінцевого числа ітерацій закінчиться, або коли мінімальна абсолютна

різниця між значеннями функцій належності (або центроїдами кластерів) на двох послідовних ітераціях не стане меншою за деяке апріорі задане значення [14]. Алгоритм FCM за своїм характером належить до наближених алгоритмів пошуку екстремуму цільової функції за наявності обмежень.

Висновки до розділу

1. Проаналізовано стан проблеми кластеризації даних і сформульовано існуючі підходи до її вирішення.
2. Розглянуто основні принципи нечіткої логіки та систем нечіткого розбиття.
3. Проведено аналіз існуючих методів кластеризації; методів їх навчання і самонавчання, що використовуються для вирішення завдань нечіткої кластеризації даних.
4. Показано, що об'єднання апаратів нейронних мереж і нечіткої логіки може ефективно вирішувати складні завдання, долаючи недоліки кожної з цих технологій в задачах нечіткої кластеризації коротких часових рядів.
5. Сформульовано задачу дослідження.

Список використаних джерел у цьому розділі наведено у повному списку використаних джерел під номерами [1–17, 92, 98].

2 АДАПТИВНІ МЕТОДИ ФІЛЬТРАЦІЇ ТА ІДЕНТИФІКАЦІЇ ЧАСОВИХ РЯДІВ

Мета розділу – розробка методів обробки даних, що можуть працювати в умовах зашумленості даних в онлайн режимі, при цьому властивості рядів можуть змінюватися у часі нестаціонарним чином. У цій ситуації на перший план виходить адаптивний підхід, пов'язаний з використанням адаптивних моделей, налаштування яких проводиться в реальному часі за допомогою тих чи інших процедур. Таким чином, актуальним є синтез адаптивних моделей, які б могли обробляти нестаціонарні часові потоки даних, у тому числі в ситуаціях, коли аналізований часовий ряд забруднений аномальними викидами.

Завдання: 1) розробити робастні методи фільтрації часових рядів, що дозволяють в онлайн режимі обробляти спотворені дані; 2) розглянути методи прості в чисельної реалізації; 3) розробити нейросистему, що характеризується високою швидкодією і простотою чисельної реалізації.

2.1 Адаптивна модель нестаціонарного часового ряду на базі квадратичного критерія якості

Зазвичай передбачається, що оброблювані послідовності задані заздалегідь, у формі пакету спостережень, чиї властивості не змінюються з плином часу [18, 19, 33, 34]. Тому для оцінювання параметрів стохастичних об'єктів в умовах невизначеності, які можуть бути описані за допомогою рівняння псевдолінійної регресії, в рамках рівняння (2.1) були описані популярні моделі часових рядів, такі як авторегресійна модель (AR), модель ковзного середнього (MA), авторегресійна модель ковзного середнього

(ARMA), авторегресійна інтегрована розширена модель ковзного середнього (ARIMAX), які пов'язані з концепцією фільтра Бокса-Дженкінса [22].

$$y(k) = w^T x(k) + \zeta(k), \tag{2.1}$$

де $y(k)$ – скалярний вихід об’єкту (відгук) у дискретний момент часу;

$k = 1, 2, ..., N$; $w = (w_0, w_1, ..., w_n)^T$ – $((n+1)\times 1)$-вектор невідомих параметрів, що підлягають визначенню;

$x(k) \in R^n$ – вектор вхідних змінних (факторів, регресорів);

$\zeta(k)$ – випадкове обурення (перешкода) з нульовим математичним сподівання і невідомою функцією щільності розподілу.

В основі значного числа процедур та алгоритмів синтезу таких моделей лежить гіпотеза про нормальний розподіл перешкод – $\zeta(k) \sim N(0, \sigma^2)$, що призвело до використання критерію мінімуму суми квадратів помилок оцінювання і пов'язаного з ним метода найменших квадратів в різних модифікаціях [20, 21]. У разі, якщо дані на обробку надходять послідовно або обсяг вибірки N не фіксований, перевагу слід віддати рекурентному МНК, який, однак, при великих значеннях може бути чисельно нестійкий, громіздкий і, в підсумку, призводить до «вибуху параметрів» коваріаційної матриці.

Ефективною альтернативою рекурентному МНК є спеціалізована процедура стохастичної апроксимації Гудвіна-Ремеджа-Кейнеса для ідентифікації об'єктів управління, що була запропонована у [24].

Вводячи у розгляд рівняння налаштовної моделі:

$$\hat{y}(k) = \hat{w}^T(k) x(k), \tag{2.2}$$

де $\hat{w}(k) - ((n+1)\times 1)$ –вектор налаштовних параметрів,

помилку ідентифікації:

$$e(k) = y(k) - \hat{y}(k) = y(k) - \hat{w}^T(k-1)x(k); \tag{2.3}$$

функцію втрат:

$$\rho(e(k)) = \frac{1}{2}e^2(k), \tag{2.4}$$

і заснований на функції втрат (2.4) критерій ідентифікації:

$$E(e(k)) = \sum_k \rho(e(k)) = \frac{1}{2}\sum_k e^2(k), \tag{2.5}$$

можна записати алгоритм адаптивної ідентифікації у вигляді системи рекурентних співвідношень:

$$\begin{cases} \hat{w}(k) = \hat{w}(k-1) + \dfrac{\left(y(k) - \hat{w}^T(k-1)x(k)\right)x(k)}{r(k)}, \\ r(k) = r(k-1) + \left\|x(k)\right\|^2, \; r(0) = 1. \end{cases} \tag{2.6}$$

У разі необхідності ідентифікації сигналу, параметри якого змінюються у часі, може бути використана модифікація процедури (2.6), заснована на критерії:

$$E(e(k)) = \sum_k \alpha^{k-1}\rho(e(k)) = \frac{1}{2}\sum_k \alpha^{k-1}e^2(k), \tag{2.7}$$

де $0 \le \alpha \le 1$– параметр забування інформації.

Тоді модифікований алгоритм має вигляд:

$$\begin{cases} \hat{w}(k) = \hat{w}(k-1) + \dfrac{\left(y(k) - \hat{w}^T(k-1)x(k)\right)x(k)}{r(k)}, \\ r(k) = \alpha r(k-1) + \left\|x(k)\right\|^2, \ \ 0 \le \alpha \le 1. \end{cases} \tag{2.8}$$

Процедура (2.8), на відміну від експоненціально зваженого рекурентного методу найменших квадратів, стійка при будь-яких значеннях параметра забування α та збігається при $\alpha = 1$ з алгоритмом Гудвіна-Ремеджа-Кейнеса, а при $\alpha = 0$ приймає форму популярного в теорії навчання штучних нейронних мереж алгоритму Качмажа-Уїдроу-Хоффа.

На рисунку 2.1 наведено структурну схему адаптивної моделі, що налаштовується за допомогою процедури (2.8).

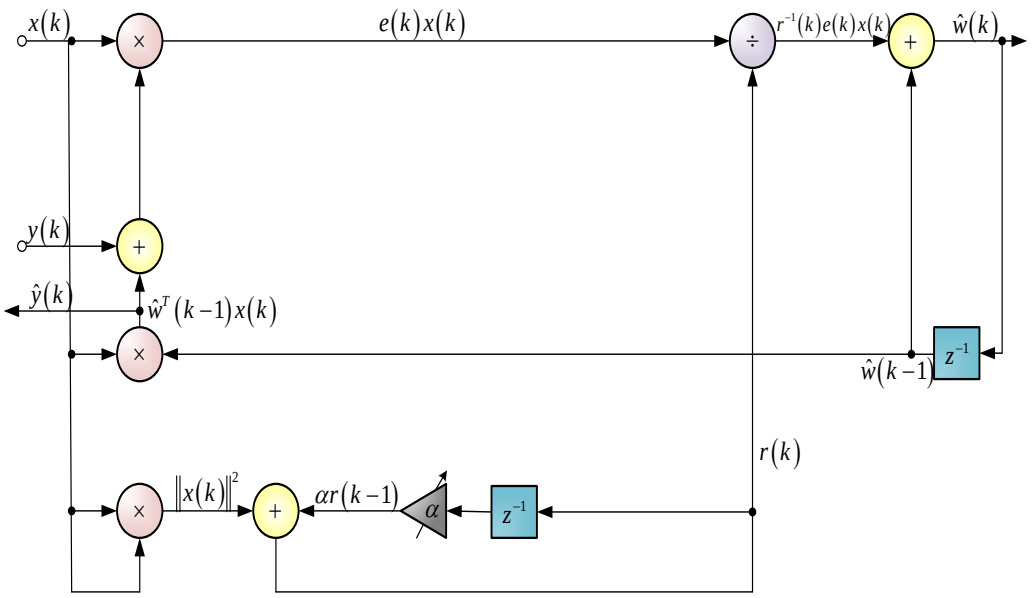


Рисунок 2.1 – Структурна схема адаптивної моделі, що налаштовується за допомогою процедури типу Гудвіна-Ремеджа-Кейнеса

Процес оцінювання реалізується за допомогою елементарних арифметичних операцій та операцій зсуву назад z^{-1}.

2.2 Робастні адаптивні моделі часових рядів

Завдання обробки інформації свідчать про те, що розподіл даних описується гаусівським законом, хоча більшість стандартних статистичних методів базується на явних або неявних представленнях про нормальність закону розподілу аналізованих рядів.

Як вже зазначалося, для вирішення завдання нечіткої кластеризації даних, що містять викиди, можуть бути використані цільові функції спеціального виду [25–27], тим або іншим способом ці аномалії пригнічують, а сама задача пов'язана з мінімізацією цих функцій.

У той же час відомо, що методи ідентифікації, засновані на критерії найменших квадратів, є надзвичайно чутливими до відхилень фактичного закону розподілу від нормального. В умовах різного роду збоїв, викидів, грубих помилок, негауссівських перешкод з «важкими хвостами» метод найменших квадратів втрачає свою ефективність.

Даний факт призвів до створення широкого класу методів робастного оцінювання, які засновані на мінімізації критеріїв, відмінних від квадратичного і призводять до необхідності вирішення задачі нелінійної оптимізації. Однак, при роботі в онлайн режимі рішення цього завдання ускладнене в силу чисельної громіздкості.

До теперішнього часу відомо багато робастних функцій втрат, спільною рисою яких є «придушення» віддалених від точки екстремуму спостережень (викидів). Так, важливою є функція Гемана-МакКлюра, що має вигляд [25–27]:

$$\rho(e(k)) = \frac{1}{2} \times \frac{e^2(k)}{1 + e^2(k)},$$

та форму, наведену на рисунку 2.2.

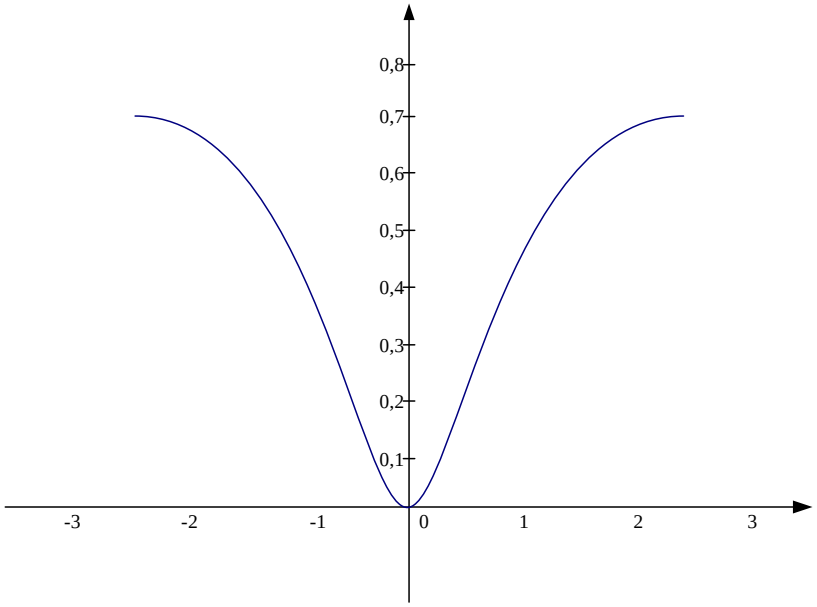


Рисунок 2.2 – Робастна функція втрат Гемана-МакКлюра

Функція Гемана-МакКлюра породжена функцією щільності розподілу Коші:

$$\rho_c(e(k)) = \frac{1}{1+e^2(k)}, \tag{2.9}$$

та являє собою «перевернутий» та зсунутий на одиницю кошіан:

$$\rho(e(k)) = 1 - \rho_c(e(k)). \tag{2.10}$$

Вводячи у (2.10) масштабуючий параметр σ^2, який визначає «ширину» кошіану:

$$\rho_c(e(k)) = \frac{1}{1+\frac{e^2(k)}{\sigma^2}} = \frac{\sigma^2}{\sigma^2 + e^2(k)}, \tag{2.11}$$

можна отримати модифіковану функцію Гемана-МакКлюра:

$$\rho(e(k)) = \frac{1}{2} \cdot \frac{e^2(k)}{\sigma^2 + e^2(k)}, \tag{2.12}$$

вид якої, при різних параметрах ширини σ^2, наведені на рисунку 2.3, 2.4, 2.5.

Представимо функцію впливу та вагову функцію для (2.5) у вигляді (2.13) і (2.14):

$$\psi(e(k)) = \rho'(e(k)) = \frac{e(k)}{(\sigma^2 + e^2(k))^2}, \tag{2.13}$$

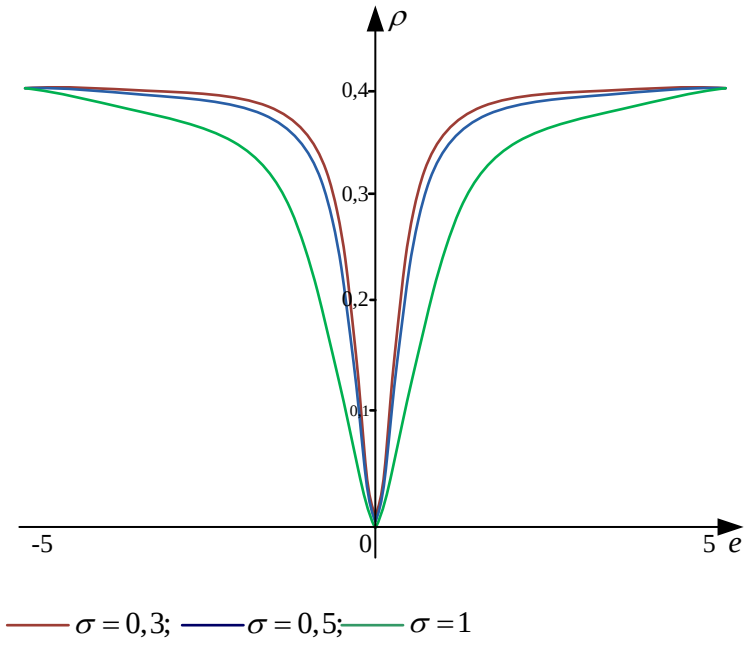


Рисунок 2.3 – Функція втрат, яка модифікована з використанням функції Гемана-МакКлюра

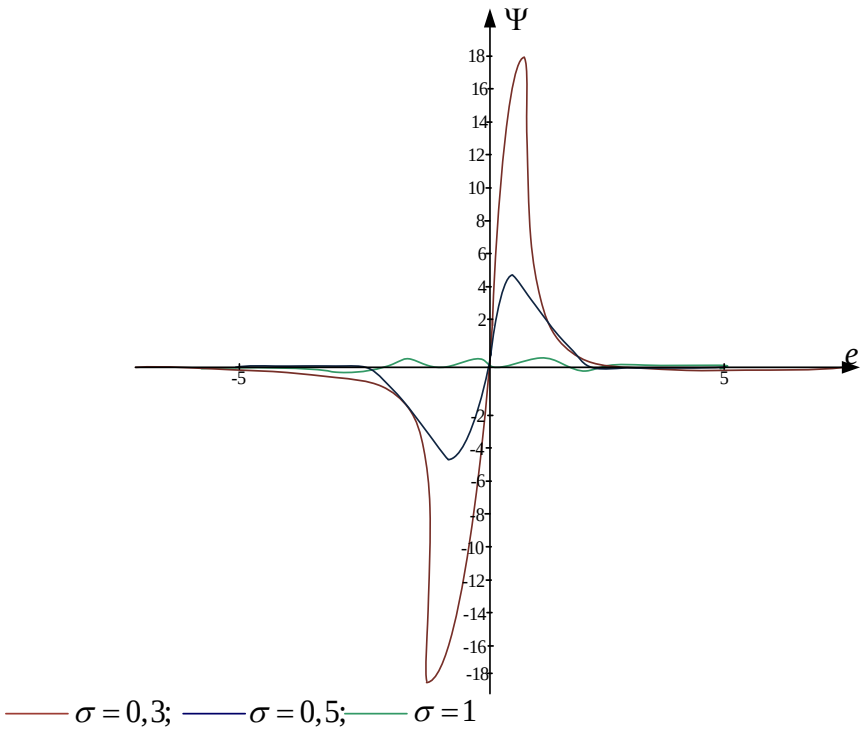


Рисунок 2.4 – Функція впливу, яка модифікована з використанням функції Гемана-МакКлюра

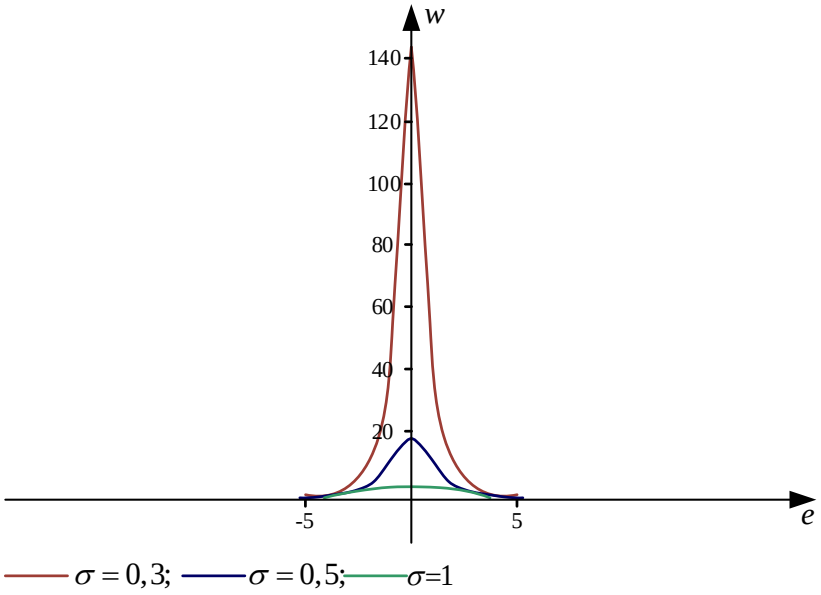


Рисунок 2.5 – Вагова функція, яка модифікована з використанням функції Гемана-МакКлюра

$$w(e(k)) = \frac{\psi(e(k))}{e(k)} = \frac{1}{(\sigma^2 + e^2(k))^2}, \tag{2.14}$$

та введемо цільову функцію:

$$E(e(k)) = \sum_k \rho(e(k)) = \frac{1}{2}\sum_k \frac{e^2(k)}{\sigma^2 + e^2(k)}. \tag{2.15}$$

Тоді, можна ввести градієнтний алгоритм її оптимізації у вигляді:

$$\begin{aligned} &\hat{w}(k) = \hat{w}(k-1) + \eta(k)\psi(e(k))x(k) = \hat{w}(k-1) + \eta(k)e(k)\frac{\sigma^2 x(k)}{(\sigma^2 + e^2(k))^2} = \\ &= \hat{w}(k-1) + \eta(k)e(k)J_G(k), \end{aligned} \tag{2.16}$$

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

Використовуючи техніку оптимізації алгоритмів навчання [23], можна ввести експоненціально зважену процедуру типу (2.8) у вигляді:

$$\begin{cases} \hat{w}(k) = \hat{w}(k-1) + \dfrac{\left(y(k) - \hat{w}^T(k-1)x(k)\right)J_G(k)}{r(k)}, \\ r(k) = \alpha r(k-1) + \left\|J_G(k)\right\|^2,\ 0 \le \alpha \le 1, \end{cases} \tag{2.17}$$

де $J_G(k) = \dfrac{\sigma^2 x(k)}{(\sigma^2 + e^2(k))^2}$.

На рисунку 2.6 наведено структурну схему налаштовної моделі, навченої за допомогою процедури (2.17).

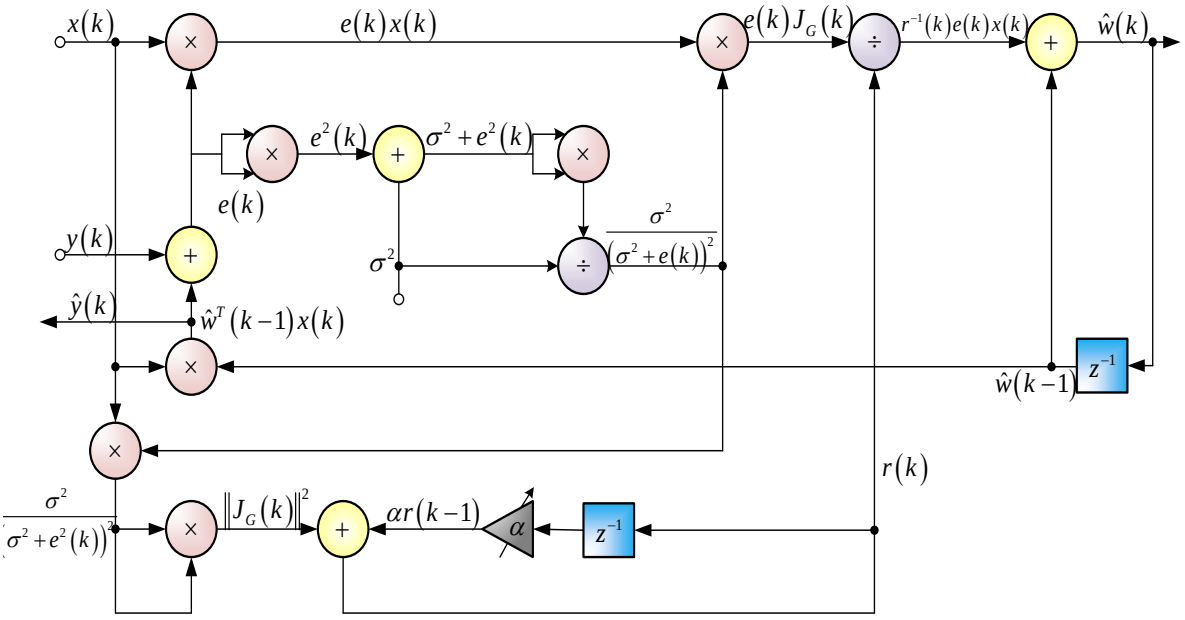


Рисунок 2.6 – Налаштовна модель, навчена за допомогою процедури Гемана-МакКлюра

Схема, наведена на рисунку 2.1, реалізується за допомогою елементарних операцій. Близькою за властивостями до (2.5) є робастна функція втрат Коші [21, 27, 28, 31]:

$$\rho(e(k)) = \frac{\sigma^2}{2}\ln\left(1+\frac{e^2(k)}{\sigma^2}\right), \tag{2.18}$$

вид якої, при різних значеннях σ^2, наведено на рисунках 2.7, 2.8 та 2.9

Функція впливу:

$$\psi(e(k)) = \frac{e(k)\sigma^2}{\sigma^2+e^2(k)} \tag{2.19}$$

та вагова функція:

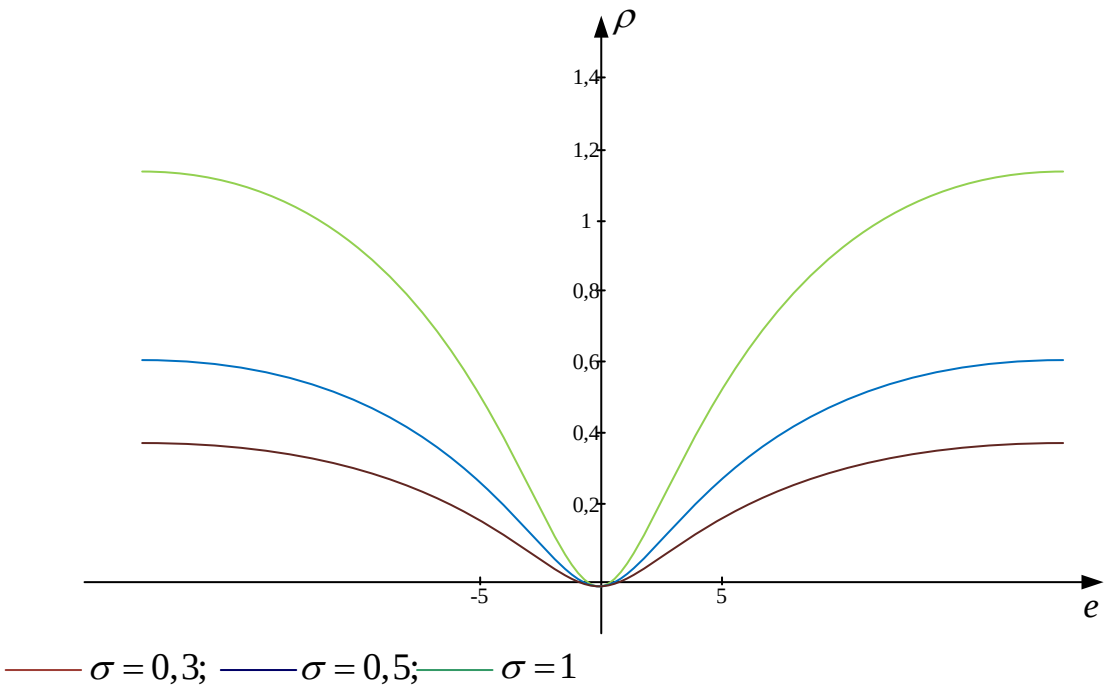


Рисунок 2.7– Функція втрат з використанням робастної функції Коші

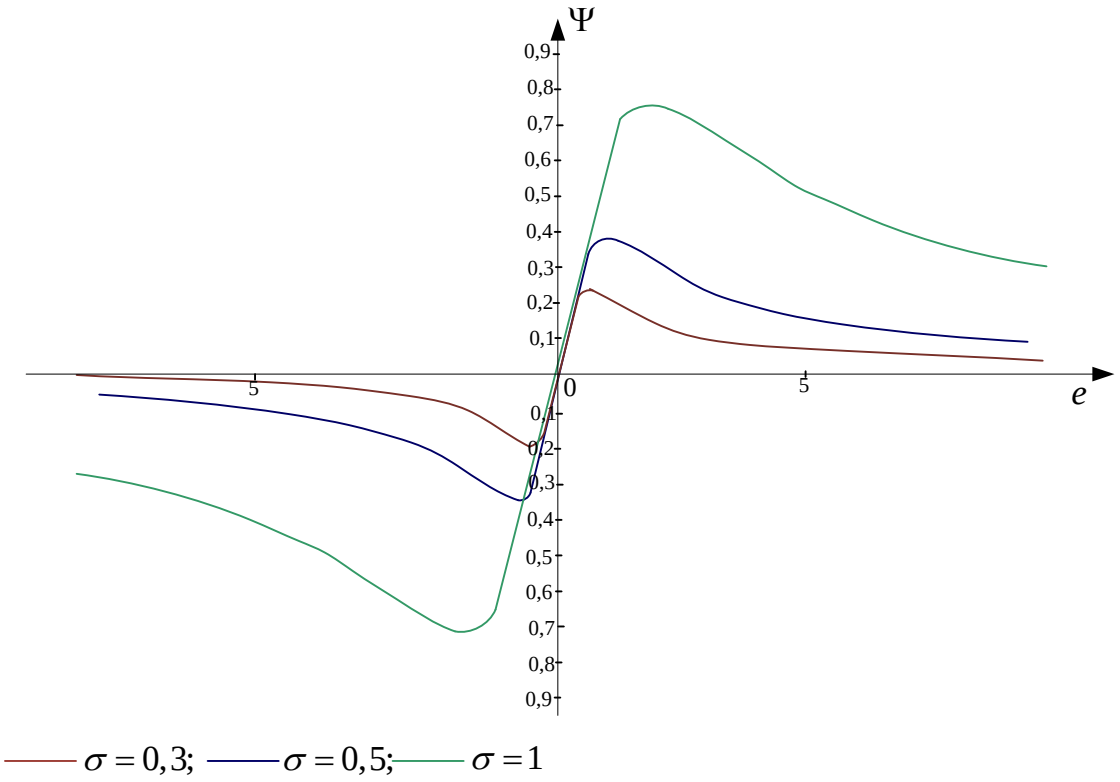


Рисунок 2.8– Функція впливу з використанням робастної функції Коші

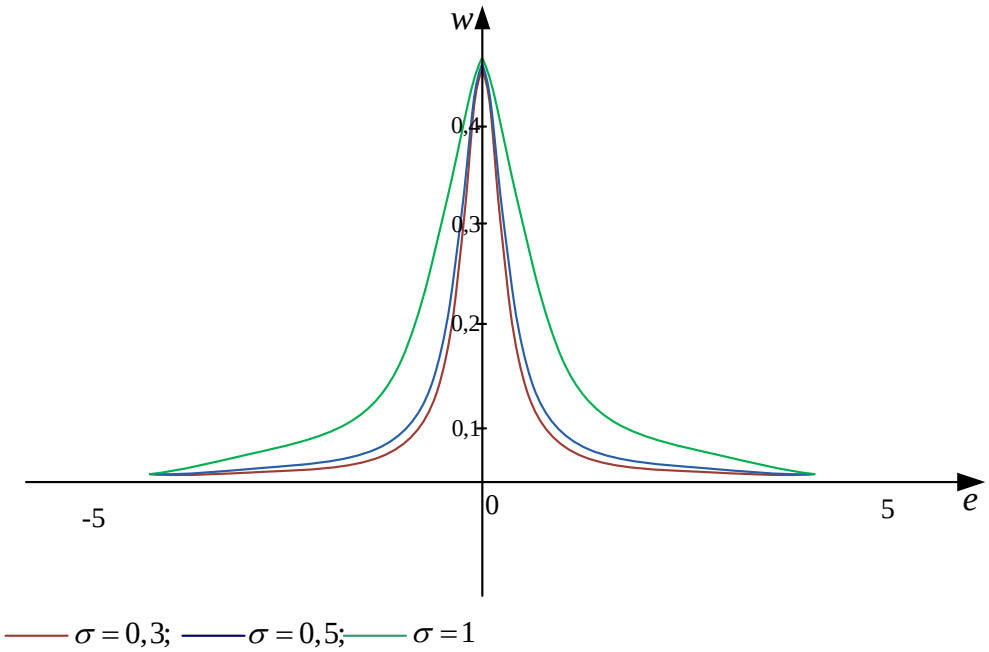


Рисунок 2.9– Вагова функція з використанням робастної функції Коші

$$w(e(k)) = \frac{\sigma^2}{\sigma^2 + e^2(k)}, \tag{2.20}$$

близькі $\rho_c(e(k))$ до розглянутих вище, тоді градієнтна процедура буде мати вигляд:

$$\hat{w}(k) = \hat{w}(k-1) + \eta(k)\psi(e(k))x(k) = \hat{w}(k-1) + \eta(k)e(k)\frac{\sigma^2}{\sigma^2 + e^2(k)} = \\ = \hat{w}(k-1) + \eta(k)e(k)J_C(k) \tag{2.21}$$

та аналог алгоритму (2.18) у вигляді:

$$\begin{cases} \hat{w}(k) = \hat{w}(k-1) + \dfrac{e(k)J_C(k)}{r(k)}, \\ r(k) = \alpha r(k-1) + \left\| J_C(k) \right\|^2, \end{cases} \tag{2.22}$$

де $J_C(k) = \dfrac{\sigma^2}{\sigma^2 + e^2(k)}$.

На рисунку 2.10 наведено структурну схему налаштовної моделі, навченої за допомогою процедури (2.15).

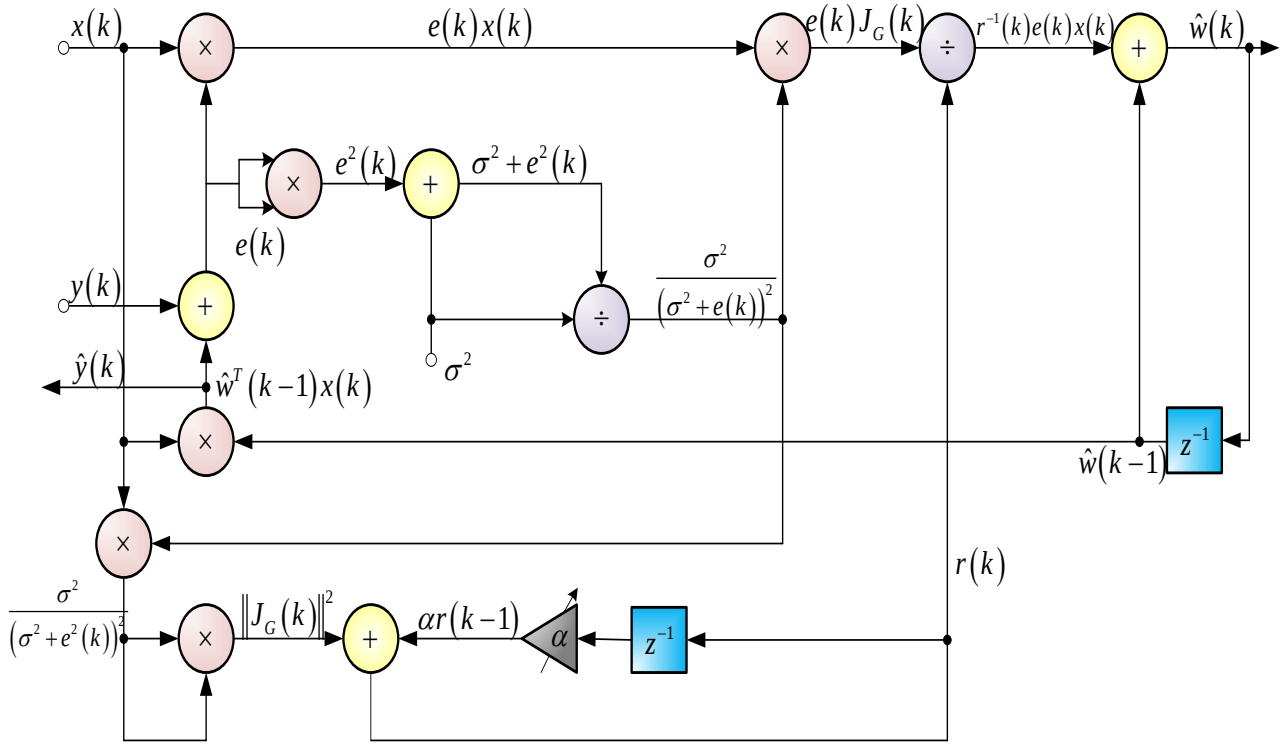


Рисунок 2.10 – Налаштовна модель, навчена за допомогою функції втрат Коші

Дана схема простіша за наведену на рисунку 2.4, обидві вони істотно простіші ніж нейромережева система робастної ідентифікації, описана в [28, 32], містять нелінійні перетворювачі та є розширенням на робастний випадок алгоритму навчання, введеного в [29, 30].

Висновки до розділу

1. Запропонована група адаптивних робастних методів фільтрації часових рядів, що дозволяють в онлайн режимі обробляти спотворені дані, які містять як пропуски так і аномальні викиди. Також введена спеціальна міра подібності, що дозволяє працювати з викривленою інформацією.

2. Розглянуті методи прості в чисельної реалізації і по суті є градієнтними процедурами оптимізації цільових функцій спеціального виду.

3. Введена нейросистема, яка утворена набором адаптивних лінійних асоціаторів, що характеризується високою швидкодією і простотою чисельної реалізації.

Список використаних у цьому розділі джерел наведено у повному списку використаних джерел під номерами [18–34].

3 АДАПТИВНА НЕЧІТКА КЛАСТЕРИЗАЦІЯ ОДНОВИМІРНИХ ЧАСОВИХ РЯДІВ З НЕРІВНОМІРНИМ ТАКТОМ КВАНТУВАННЯ В ІНТЕЛЕКТУАЛЬНОМУ АНАЛІЗІ ПОТОКІВ ДАНИХ

Кластеризація коротких часових рядів залежна як від кількості параметрів моделі, що підлягають оцінці, так і від кількості даних, що надійшли на обробку. Також важливою особливістю є те, що дані, які необхідно обробляти, подаються на обробку в онлайн режимі.

Мета розділу - розробка методу нечіткої кластеризації в основі якого алгоритм, що будує матрицю належностей, по якій відбувається розбиття на кластери.

Завдання: 1) розглянути методи нечіткої кластеризації коротких часових рядів з нерівномірним тактом квантування, які можуть бути представлені у формі пакету спостережень або послідовно надходити на обробку в онлайн режимі; 2) розглянути адаптивний варіант ймовірнісного та можливісного алгоритму нечіткої кластеризації коротких часових рядів.

3.1 Формування векторів ознак для одновимірних часових рядів

Основною метою кластеризації є визначення структури даних шляхом їх об'єктивної організації, в яких схожість внутрішньогрупового об'єкта мінімізується, а різниця між групами максимізується.

Кластеризація необхідна, коли дані не є доступними, незалежно від того, чи є вони двійковими, числовими, інтервальними, порядковими, реляційними, текстовими, просторовими, часовими, зображеннями, мультимедійними або сукупністю вищезазначених типів даних [36, 38, 39].

Припустимо, що вихідна інформація задана у формі набору вибірок $x_i(k)$ (де $i=1,2,...,n$ номер окремого спостереження у k-й реалізації, $k=1,2,...,N$), яка містить $N(N>n)$ часових послідовностей з нерівномірним тактом квантування, що підлягають кластеризації.

При цьому кожна така реалізація може бути представлена у формі $(n\times 1)$ вектора $x(k)=(x_1(k),x_2(k),...,x_n(k))^T$. Нерівномірність квантування означає що:

$$\Delta t_i = t_i - t_{i-1} \neq \Delta t_{i+1} = t_{i+1} - t_i,$$

тобто $\Delta t_i \neq const$.

У зв'язку з цим в [37] у якості подібності часових рядів було введено PS–відстань (Piecewise slope distance=PS–distance=STS–distance=short time series distance).

Для представлення цих рядів у вигляді кусочно-лінійних функцій:

$$x_t(k) = a_t(k) + b_t(k)t, \tag{3.1}$$

де $t_i \le t \le t_{i+1}$.

$$\begin{cases} a(k) = \dfrac{t_{i+1}x_i(k) - t_i x_{i+1}(k)}{t_{i+1} - t_i}, \\ b_t(k) = \dfrac{x_{i+1}(k) - x_i(k)}{t_{i+1} - t_i} \end{cases},$$

що оцінює відміну форм (нахилів) аналізованих вибірок.

Приклад однієї такої реалізації наведено на рисунку 3.1, де для оцінки відстані між такими вибірками не можуть бути використані ані традиційна евклідова метрика, ані класичні стохастичні критерії [50].

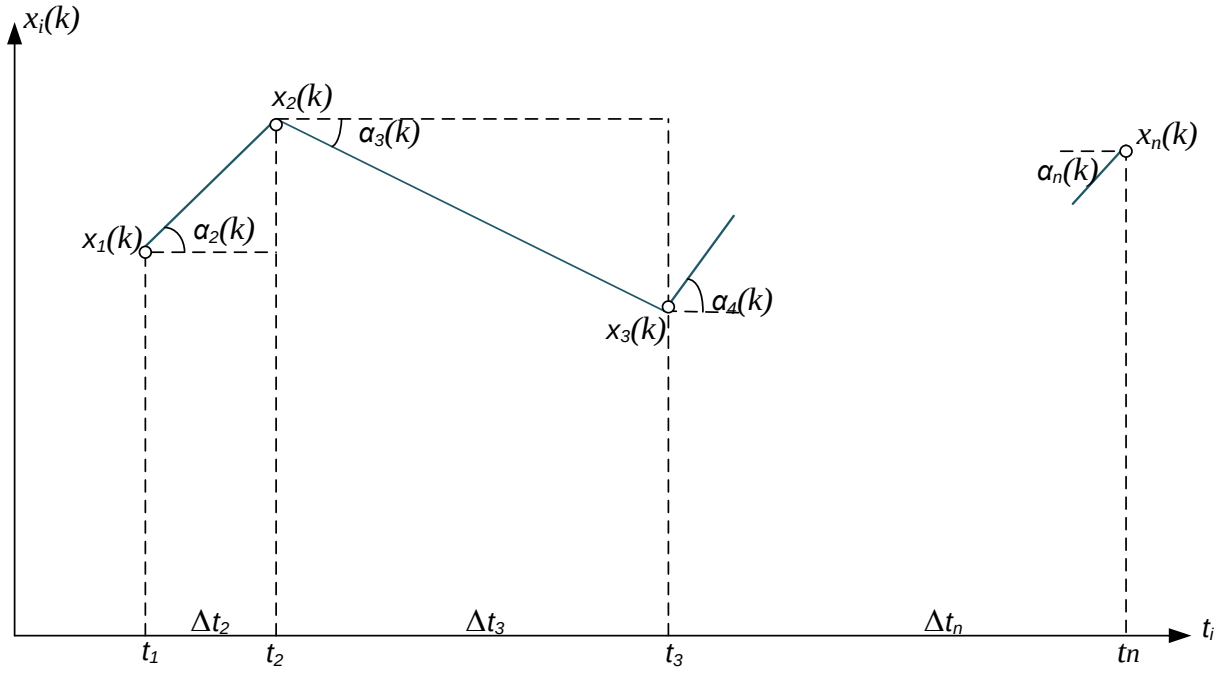


Рисунок 3.1–Часовий ряд з нерівномірним тактом квантування

Відстань між двома послідовностями $x(k)$ та $x(l)$ можна представити у вигляді виразу:

$$d^2_{STS}\left(x(k),x(l)\right)=\sum_{i=1}^{n-1}\left(\frac{x_{i+1}(k)-x_i(k)}{t_{i+1}-t_i}-\frac{x_{i+1}(l)-x_i(l)}{t_{i+1}-t_i}\right)^2= \\ =\sum_{i=1}^{n-1}\left(\frac{x_{i+1}(k)-x_i(k)}{\Delta t_{i+1}}-\frac{x_{i+1}(l)-x_i(l)}{\Delta t_{i+1}}\right)^2, \tag{3.2}$$

який задовольняє всім умовам, які визначають метрику.

3.2 Пакетний метод нечіткої кластеризації часових рядів

Під адаптивною кластеризацією розуміється така кластеризація, при якій параметри, що визначають результат, вибираються і коригуються в процесі онлайн виконання завдання, виходячи із заданих критеріїв та рекомендацій, для досягнення оптимального результату [35].

Для визначення кількості кластерів в рамках адаптивного пошуку розв‘язання задачі кластеризації вводиться формальна схема або процедура, що зображена на рисунку 3.2, в якій центральною ланкою є оцінка якості:

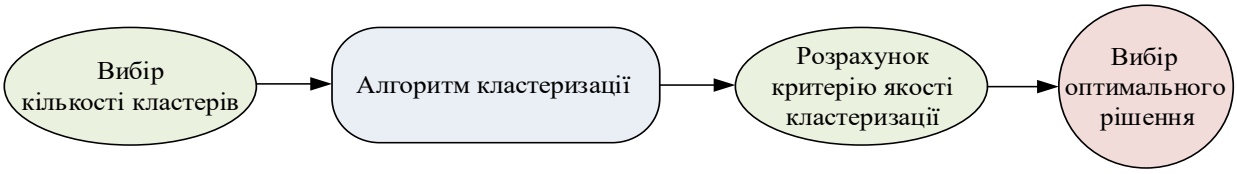


Рисунок 3.2– Узагальнена схема адаптивної кластеризації

Виходячи з цього, можна запропонувати дві стратегії знаходження розв`язку задачі кластеризації:

1. Формальний вибір, коли вибирається одне або кілька рішень у чіткій відповідності до екстремумів заданого критерію;

2. Комбінований вибір, при якому рішення вибирається фахівцем з невеликої множини рішень, що складається за результатами виявлення екстремумів заданого критерію.

Застосування формального вибору на практиці обмежене, оскільки може давати незадовільні результати, це може бути пов'язане як з недоліками використовуваних критеріїв оцінки якості, так і з некоректною вихідною множиною. Хоча, з іншого боку, аналізуючи результати, отримані за

допомогою формального вибору рішення, як правило, можна визначити коректне або некоректне рішення було отримано.

Представлена на рисунку 3.2 адаптивна схема пошуку найкращого рішення є ітераційною, пошук не є спрямованим, оскільки при його реалізації не робиться жодних припущень щодо характеру залежності значення критерію від кількості кластерів. Таким чином, перед вибором найкращого рішення необхідно провести кластеризацію для кожного значення кількості кластерів і з встановленого діапазону.

Відмінності можуть бути незалежно від того, чи є дані дискретно-значущими або дійсними, рівномірно або нерівномірно дискретизованими, одновимірними або багатовимірними, а також чи є ряди даних однакової або нерівної довжини.

Нерівномірно вибрані дані повинні бути перетворені в уніфіковані до того, як операції кластеризації можуть бути виконані, це може бути досягнуто за допомогою методів – від простої вибірки, на основі інтервалу вибірки, до складного методу моделювання та оцінки. Не розглядаючи їх відмінності, можна сказати, що в основі вони намагаються змінити існуючі алгоритми кластеризації статичних даних.

Таким чином, щоб дані часових рядів можна було обробляти або перетворювати в вигляді статичних даних, для того щоб існуючий алгоритм для кластеризації статичних даних був використаний, саме такі задачі часто зустрічаються в інтелектуальному аналізі даних [36].

Для їх розв`язання запропоновано безліч методів, включаючи алгоритми онлайн обробки інформації, що послідовно надходить на обробку [35].

Разом з тим, у багатьох практичних застосуваннях виникають ситуації, коли класичні підходи до аналізу часових рядів виявляються неефективними.

Методи кластеризації, засновані на цільових функціях [37], призначені для вирішення завдання кластеризації шляхом оптимізації деякого наперед заданого критерія якості кластеризації і є найбільш коректними з математичної точки зору.

Тому на підставі метрики (3.2), авторами [40] була введена пакетна (офлайн) процедура нечіткої кластеризації, яка є модифікацією алгоритму нечітких *c*-середніх (FCM) на випадок обробки часових рядів з нерівновіддаленими спостереженнями [49].

Компоненти виразу (3.2) - це перші різниці дискретного сигналу $x_i(k)$ або тангенси кутів нахилу лінійних функцій (3.1), тобто

$$\Delta x_{i+1}(k) = \frac{x_{i+1}(k) - x_i(k)}{\Delta t_{i+1}} = tg\alpha_{i+1}(k),$$

проте ряд, утворений першими різницями, містить на одну точку менше ніж вихідна вибірка, тобто $(n-1)$ спостережень $\Delta x_2(k), \Delta x_3(k), ..., \Delta x_n(k)$, або, що те ж саме, $tg\alpha_2(k), tg\alpha_3(k), ..., tg\alpha_n(k)$. Оскільки в результаті взяття різниць з ряду видаляється його середнє значення, для відновлення вихідної вибірки за її різницями, необхідно доповнити набір цих різниць будь-яким із спостережень вихідної послідовності, наприклад, $x_n(k)$. Тоді, маючи послідовність різниць $\Delta x_i(k)$, відновимо вихідний ряд за допомогою простих співвідношень:

$$\begin{cases} x_{n-1}(k) = x_n(k) - \Delta x_n(k)\Delta t_n, \\ x_{n-2}(k) = x_n(k) - \Delta x_{n-1}(k)\Delta t_{n-1}, \\ \vdots \\ x_1(k) = x_2(k) - \Delta x_2(k)\Delta t_2. \end{cases} \tag{3.3}$$

Вводячи далі у розгляд $(n\times 1)$-вектор ознак $\tilde{x}(k)=(\Delta x_2(k),\Delta x_3(k),...,\Delta x_n(k),x_n(k))^T$, перепишемо метрику (3.3) у традиційній формі:

$$d^2_{STS}\left(x(k),x(l)\right)=\left\|\tilde{x}(k)-\tilde{x}(l)\right\|^2, \tag{3.4}$$

тобто, фактично повернемося до стандартної евклідової відстані між різницями вихідних рядів.

Далі, використовуючи (3.3) та методику стандартного нечіткого ймовірнісного кластерного аналізу, шляхом знаходження сідлової точки функції Лагранжа:

$$L\left(u_j(k),\tilde{c}_j,\lambda(k)\right)=$$
$$=\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^{\beta}(k)\lambda(k)\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2+\sum_{k=1}^{N}\lambda(k)\left(\sum_{j=1}^{m}u_j(k)-1\right), \tag{3.5}$$

де $u_j(k)$– рівень належності вектора;

$\tilde{x}(k)$– j–му кластеру з прототипом – центроїдом $\tilde{c}_j$; j=1,2,...,m;

M – число кластерів, які встановлюються апріорно;

$\lambda(k)$–невизначений множник Лагранжа;

$\beta>1$ – параметр фаззіфікації (fuzzifier), який визначає «розмитість» границь між кластерами,

приходимо до стандартної процедури нечіткої ймовірнісної кластеризації:

$$\begin{cases} u_j(k) = \dfrac{\left(\left\|\tilde{x}(k) - \tilde{c}_j\right\|^2\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(\left\|\tilde{x}(k) - \tilde{c}_j\right\|^2\right)^{\frac{1}{1-\beta}}}, \\ \tilde{c}_j = \dfrac{\sum\limits_{k=1}^{N} u_j^{\beta}(k)\tilde{x}(k)}{\sum\limits_{k=1}^{N} u_j^{\beta}(k)}, \end{cases} \tag{3.6}$$

що збігається при $\beta = 2$ з популярним FCM–алгоритмом Дж.Бездека:

$$\begin{cases} u_j(k) = \dfrac{\left\|\tilde{x}(k) - \tilde{c}_j\right\|^{-2}}{\sum\limits_{l=1}^{m}\left\|\tilde{x}(k) - \tilde{c}_l\right\|^{-2}}, \\ \tilde{c}_j = \dfrac{\sum\limits_{k=1}^{N} u_j^{2}(k)\tilde{x}(k)}{\sum\limits_{k=1}^{N} u_j^{2}(k)}. \end{cases} \tag{3.7}$$

Оскільки вектори $\tilde{c}_j$, $j = 1,2,...,m$, є центроїдами кластерів, утворених рядами різниць, для відновлення прототипів вихідних даних $\tilde{c}_j$ можна скористатися співвідношеннями (3.3).

Процедури кластеризації (3.3) та (3.4) синтезовані у припущенні, що вся вихідна інформація задана у вигляді фіксованого масиву даних $x(1), x(2),...,x(N)$ і не змінюється в процесі обробки. Якщо ж вибірки $x(k)$ надходять на обробку послідовно у формі потоку даних, можна скористатися підходами, які використані в концепціях інтелектуального аналізу даних і, перш за все, в адаптивних методах.

Скориставшись для пошуку сідлової точки лагранжіана (3.5) рекурентним алгоритмом нелінійного програмування Ерроу-Гурвіца-Удзави, отримуємо адаптивну градієнтну процедуру нечіткої кластеризації [44-48]:

$$\begin{cases} u_j(k+1) = \dfrac{\left(\left\|\tilde{x}(k+1) - \tilde{c}_j(k)\right\|^2\right)^{\frac{1}{1-\beta}}}{\sum\limits_{l=1}^{m}\left(\left\|\tilde{x}(k) - \tilde{c}_j\right\|^2\right)^{\frac{1}{1-\beta}}}, \\ \tilde{c}_j(k+1) = \tilde{c}_j(k) + \eta(k) u_j^{\beta}(k+1)\left(\tilde{x}(k+1) - \tilde{c}_j(k)\right), \end{cases} \tag{3.8}$$

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

З позиції навчання самоорганізовних мап Кохонена [43], друге рекурентне співвідношення (3.6) є правилом самонавчання на основі принципу «Переможець отримує більше» (WTM).

Множник $u_j^{\beta}(k+1)$ відповідає функції сусідства, яка має форму кошіану замість традиційного гауссіана [41-42].

При $\beta = 0$ приходимо до стандартного принципу «Переможець отримує все» (WTA) – правила самонавчання $\tilde{c}_j(k+1) = \tilde{c}_j(k) + \eta(k)$, $(x(k+1) - \tilde{c}_j(k))$, який мінімізує цільову функцію:

$$E\left(\tilde{c}_j\right) = \sum_k \left\|\tilde{x}(k) - c_j\right\|^2, \tag{3.9}$$

при цьому при $\eta(k) = (k+1)^{-1}$ отримуємо процедуру стохастичної апроксимації:

$$\tilde{c}_j(k+1) = \tilde{c}_j(k) + \frac{1}{k+1}\left(\tilde{x}(k+1) - \tilde{c}_j(k)\right), \tag{3.10}$$

що приведе до стандартної оцінки середнього арифметичного як центроїда. Таким чином, для вирішення в онлайн режимі завдання нечіткої кластеризації коротких часових рядів з нерівномірним тактом квантування можна використовувати чисельно простий адаптивний алгоритм (3.4), який є розширенням WTM-правила самонавчання Кохонена на розглянуту задачу [43].

3.3 Адаптивна можливісна нечітка кластеризація часових рядів

Незважаючи на своє значне поширення, алгоритми, пов'язані з оптимізацією лагранжіана (3.5), мають істотний недолік, пов'язаний з необхідністю виконання обмеження:

$$\sum_{j=1}^{m} u_j(k) = 1. \tag{3.11}$$

Саме завдяки обмеженню (3.11) такі процедури отримали назву ймовірнісних, сам же недолік, що випливає з (3.11), полягає у тому, що вектор спостережень $\tilde{x}(k)$, рівноналежний всім кластерам, має ті ж рівні належності, що і вектор, що не належить жодному з класів, але рівновіддалений від усіх центроїдів.

Таким чином, аномальний викид буде віднесений до всіх наявних кластерів.

Альтернативою ймовірнісним алгоритмам кластеризації є можливісні методи [13], пов'язані з мінімізацією цільової функції:

$$E\left(u_j,\tilde{c}_j,\mu_j\right)=\sum_{j=1}^{m}\mu_j\sum_{k=1}^{N}\left(1-u_j\left(k\right)\right)^{\beta}+\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^{\beta}\left(k\right)\left\|\tilde{x}\left(k\right)-\tilde{c}_j\right\|^2, \tag{3.12}$$

де $\mu_j>0$ визначає відстань від $\tilde{x}(k)$ до $\tilde{c}_j$, на якій рівень належності приймає значення $0{,}5$, тобто $u_j\left(k\right)=0{,}5$ при

$$\left\|\tilde{x}\left(k\right)-\tilde{c}_j\right\|^2=\mu_j. \tag{3.13}$$

Оптимізація (3.12) за $u_j(k)$, $\tilde{c}_j$ та μ_j веде до результату у вигляді [40]:

$$\begin{cases} u_j\left(k\right)=\left(1+\left(\dfrac{\left\|\tilde{x}\left(k\right)-\tilde{c}_j\right\|^2}{\mu_j}\right)^{\frac{1}{\beta-1}}\right)^{-1}, \\ \mu_j\left(k\right)=\dfrac{\sum\limits_{k=1}^{N}u_j^{\beta}\left(k\right)\left\|\tilde{x}-\tilde{c}_j\right\|^2}{\sum\limits_{k=1}^{N}u_j^{\beta}\left(k\right)}, \\ \tilde{c}_j=\dfrac{\sum\limits_{k=1}^{N}u_j^{\beta}\left(k\right)\tilde{x}\left(k\right)}{\sum\limits_{k=1}^{N}u_j^{\beta}\left(k\right)}, \end{cases} \tag{3.14}$$

при цьому співвідношення для обчислення центроїдів (3.4) та (3.14) збігаються.

Адаптивний варіант можливісного алгоритму (3.4) може бути отриманий внаслідок градієнтної оптимізації цільової функції (3.3) у вигляді [12]:

$$\begin{cases} u_j(k) = \left(1 + \left(\dfrac{\left\| \tilde{x}(k) - \tilde{c}_j(k) \right\|^2}{\mu_j(k)} \right)^{\frac{1}{\beta-1}} \right)^{-1}, \\ \mu_j(k+1) = \dfrac{\sum\limits_{k=1}^{N} u_j^{\beta}(p) \left\| \tilde{x}(p) - \tilde{c}_j(k) \right\|^2}{\sum\limits_{p=1}^{k+1} u_j^{\beta}(p)}, \\ \tilde{c}_j(k+1) = \tilde{c}_j(k) + \eta(k) u_j^{\beta}(k+1) \left(x(k+1) - \tilde{c}_j(k) \right), \end{cases} \quad (3.15)$$

при цьому третє співвідношення також є WTM-правилом самонавчання, однак відрізняється від аналогічного рекурентного виразу (3.15) видом використаної функції сусідства.

Якщо у процесі обробки інформації виявляється, що деякий вектор спостережень має малі рівні належності до будь-якого з кластерів, це свідчить про те, що дані спостереження або є аномальним викидом або це сигнал про виникнення нового, відмінного від уже наявних кластера [50].

Висновки до розділу

1. Розглянута задача нечіткої кластеризації коротких часових рядів з нерівномірним тактом квантування, які можуть бути представлені у формі пакету спостережень або послідовно надходити на обробку в онлайн режимі.

2. У процесі дослідження введено адаптивний варіант ймовірнісного та можливісного алгоритму нечіткої кластеризації коротких часових рядів.

3. Запроваджена процедура може бути корисна при вирішенні завдань, що виникають в рамках аналізу даних, коли вихідні дані – короткі часові ряди, які не можуть оброблятися стандартними методами.

Список використаних у даному розділі джерел наведено у повному списку використаних джерел під номерами [35–50].

4 АДАПТИВНА НЕЧІТКА КЛАСТЕРИЗАЦІЯ БАГАТОВИМІРНИХ ПОТОКІВ ДАНИХ З НЕРІВНОМІРНИМ ТА АСИНХРОНИМИ ТАКТАМИ КВАНТУВАННЯ

Мета розділу-реалізація нечіткої кластеризації часових рядів з нерівномірними спостереженнями за ситуацією, коли дані надходять на обробку в онлайн режимі у формі багатовимірного потоку інформації в рамках концепції інтелектуального аналізу даних.

Завдання: 1) розглянути матричну модифікацію нейро-фаззі мережі Т. Кохонена, що навчається на основі правила «Переможець отримує більше»; 2) розглянути метод для адаптивної кластеризації, що заснований на використанні критеріїв оцінки якості рішення і дозволяє повністю формалізувати розв’язання задачі нечіткої кластеризації багатовимірних часових рядів, за допомогою оцінки якості кожного розбиття і вибір найкращого з них; 3) розглянути процедуру нечіткої кластеризації, що не схильна до ефекту «концентрації норм» і є узагальненням ряду відомих алгоритмів ймовірнісної нечіткої кластеризації; 4) розглянути процедуру, що обробляє вихідні дані в рамках інтелектуального аналізу потоків даних, коли дані мають велику розмірність.

4.1 Формування векторів ознак для багатовимірних часових рядів

Для забезпечення формування векторів ознак для багатовимірних часових рядів та їхньої подальшої кластеризації необхідно ввести квантування всередині кожного ряду. Особливістю цих завдань є те, що об'єктом кластеризації є не окремі спостереження, а вибірки загалом, самі спостереження фіксуються через нерівновіддалені моменти часу, а

сформовані кластери не перетинаються тобто, кожна вибірка може належати відразу до декількох класів [51–52]. При цьому також передбачається, що вся оброблювана інформація задана у формі фіксованого масиву даних, обсяг якого не змінюється.

При цьому стандартні методи не мають можливості обробляти такі дані з причини недостатньої кількості досліджень у часових рядах та подальшої неможливості побудувати корелограми зв’язку або застосувати існуючі метрики [98].

Ситуація ще більше ускладняється, якщо вихідна інформація задана у формі багатовимірних часових рядів, тобто двовимірних полів спостережень. Прикладом таких двовимірних полів можуть бути електромагнітні та оптичні поля, області забруднення повітря і води, біомедичні масиви спостережень та сигнали цифрового відео, які формують дискретні двовимірні поля [53–56].

У зв'язку з цим представляється доцільність реалізації нечіткої кластеризації коротких часових рядів з нерівномірними спостереженнями [57] на випадок, коли дані надходять на обробку в онлайн режимі у формі багатовимірного потоку інформації в рамках концепції інтелектуального аналізу даних.

4.2 Нечітка ймовірнісна кластеризація багатовимірних часових рядів

Розглянута у розділі 3 метрика 3.2 повністю задовольняє початковому аналізу інформації при обробці багатовимірних часових рядів.

Нескладно помітити, що компоненти відстані (3.2) є за суттю першими різницями дискретних сигналів $x_p(k)$ і $x_p(l)$, тобто тангенсами кутів нахилу кусочно-лінійних функцій на випадок обробки багатовимірного часового ряду (рис. 4.1):

$$\Delta x_{i+1,p}(k) = \frac{x_{i+1,p}(k) - x_{ip}(k)}{\Delta t_{i+1}} = tg\alpha_{i+1,p}(k),$$

$$\Delta x_{i+1,p}(l) = \frac{x_{i+1,p}(l) - x_{ip}(l)}{\Delta t_{i+1}} = tg\alpha_{i+1,p}(l).$$

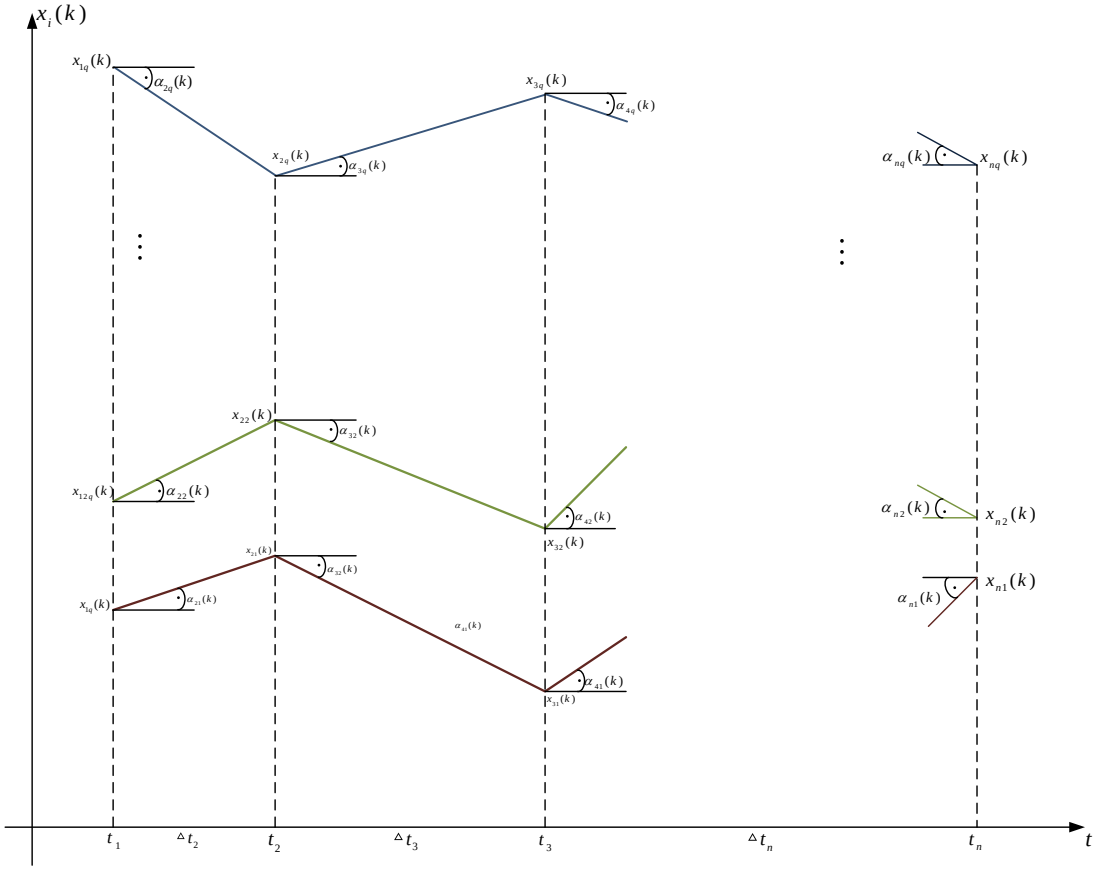


Рисунок 4.1 – Багатовимірний часовий ряд з нерівномірним тактом квантування

Для того, щоб скористатися ідеєю оцінки відстаней між рядами за їх першими різницями, введемо до розгляду $(q \times n)$ – матрицю

$$\tilde{X}(k) = \begin{pmatrix} \Delta x_{11}(k) & \Delta x_{21}(k) & \cdots & \Delta x_{nl}(k) & x_{nl}(k) \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \vdots & \cdots & \Delta x_{ip}(k) & \cdots & \vdots \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ \Delta x_{1q}(k) & \Delta x_{2q}(k) & \cdots & \Delta x_{nq}(k) & x_{nq}(k) \end{pmatrix},$$

і замість евклідової відстані – сферичну норму:

$$D_{PS}^2(X(k),X(l)) = Tr(\tilde{X}(k)-\tilde{X}(l))(\tilde{X}(k)-\tilde{X}(l))^T, \quad (4.1)$$

що є узагальненням (3.4) на матричний випадок.

На базі відстані (4.1) може бути проведена нечітка кластеризація $\tilde{X}(1),\tilde{X}(2),\ldots,\tilde{X}(N)$.

Далі, використовуючи методику нечіткого ймовірнісного кластерного аналізу, введемо до розгляду цільову функцію:

$$E(u_j(k),\tilde{C}_j) = \sum_{k=1}^{N}\sum_{j=1}^{m} u_j^{\beta}(k) D_{PS}^2(\tilde{X}(k),\tilde{C}_j) =$$

$$= \sum_{k=1}^{N}\sum_{j=1}^{m} u_j^{\beta}(k) Tr(\tilde{X}(k)-\tilde{C}_j)(\tilde{X}(k)-\tilde{C}_j)^T$$

за наявності стандартних обмежень

$$\sum_{j=1}^{m} u_j(k) = 1, \text{ чи } \sum_{j=1}^{m} u_j(k) - 1 = 0,$$

де $k = 1,2,\ldots N$, $0 < \sum_{j=1}^{m} u_j(k) < N$, $j = 1,2,\ldots,m$;

$u_j(k)$ – рівень належності матриці $\tilde{X}(k)$ j-му кластеру з матричним центроїдом $\tilde{C}_j$;

m – кількість кластерів, що задається апріорно.

Результатом кластеризації є $(N \times m)$–матриця $U = \{u_j(k)\}$, що має назву матриці нечіткого розбиття, та m матриць-центроїдів $\tilde{C}_j$, $j = 1,2,\ldots,m$.

Записавши матричну функцію Лагранжа:

$$L(u_j(k),\tilde{C}_j,\lambda(k)) = \sum_{k=1}^{N}\sum_{j=1}^{m} u_j^{\beta}(k) Tr(\tilde{X}(k)-\tilde{C}_j)(\tilde{X}(k)-\tilde{C}_j)^T + \\ + \sum_{k=1}^{N}\lambda(k)\left(\sum_{j=1}^{m} u_j(k) - 1\right), \tag{4.2}$$

де $\lambda(k)$– невизначені множники Лагранжа), та розв'язавши систему рівнянь Каруша-Куна-Таккера:

$$\begin{cases} \partial L(u_j(k),\tilde{C}_j,\lambda(k))/\partial u_j(k)) = \beta u_j^{\beta-1}(k) Tr(\tilde{X}(k)-\tilde{C}_j)(\tilde{X}(k)-\tilde{C}_j)^T + \\ \qquad + \lambda(k) = 0, \\ \partial L(u_j(k),\tilde{C}_j,\lambda(k))/\partial\lambda(k) = \sum_{j=1}^{m} u_j(k) - 1 = 0, \\ \{\partial L(u_j(k),\tilde{C}_j,\lambda(k))/\partial \tilde{C}_{jip}\} = -2\sum_{k=1}^{N} u_j^{\beta}(k)(\tilde{X}(k)-\tilde{C}_j) = \Theta, \end{cases} \tag{4.3}$$

де $\{\partial L(\mathrm{u}_j(k),\tilde{C}_j,\lambda(k))/\partial \tilde{C}_{jip}\}$–$(q\times n)$ – матриця, що сформована частковими похідними;

Θ – матриця тієї ж розмірності, що утворена нулями, приходимо до результату у вигляді [54]:

$$\begin{cases} u_j(k) = \dfrac{(Tr(\tilde{X}(k)-\tilde{C}_j)(\tilde{X}(k)-\tilde{C}_j)^T)^{\frac{1}{1-\beta}}}{\sum_{g=1}^{m}(Tr(\tilde{X}(k)-\tilde{C}_g)(\tilde{X}(k)-\tilde{C}_g)^T)^{\frac{1}{1-\beta}}}, \\ \lambda(k) = -\left(\sum_{g=1}^{m}\left(\beta Tr(\tilde{X}(k)-\tilde{C}_g)(\tilde{X}(k)-\tilde{C}_g)\right)^{\frac{1}{1-\beta}}\right)^{1-\beta}, \\ \tilde{C}_j = \dfrac{\sum_{k=1}^{N} u_j^{\beta}(k)\tilde{X}(k)}{\sum_{k=1}^{N} u_j^{\beta}(k)}, \end{cases} \tag{4.4}$$

структурно близького при $\beta = 2$ до алгоритму Дж. Бездека [54,58–65], та що є його узагальненням на матричний випадок:

$$\begin{cases} u_j(k) = \dfrac{(Tr(\tilde{X}(k) - \tilde{C}_j)(\tilde{X}(k) - \tilde{C}_j)^T)^{-1}}{\sum\limits_{g=1}^{N}(Tr(\tilde{X}(k) - \tilde{C}_g)(\tilde{X}(k) - \tilde{C}_g)^T)^{-1}}, \\ \tilde{C}_j = \dfrac{\sum\limits_{k=1}^{N} u_j^2(k)\tilde{X}(k)}{\sum\limits_{k=1}^{N} u_j^2(k)}. \end{cases} \quad (4.4)$$

Оскільки матриці $\tilde{C}_j$, $j = 1,2,\ldots m$ є центроїдами кластерів, що утворено рядами різниць, для відновлення центроїдів вихідних даних $\tilde{C}_j$ необхідно скористатися співвідношеннями (4.2) [66-68].

4.3 Оцінка відстані між реалізаціями часового ряду з нерівномірними асинхронними тактами квантування

Ситуація ускладнюється, якщо квантування виконується не тільки нерівномірно, але й відрізняється для кожної з реалізацій контрольованого сигналу. Саме в цьому випадку на перший план виходить ефект «концентрації норм», ігнорування якого не дозволить отримати прийнятне рішення. У зв'язку з цим видається доцільним поширення підходу, введеного в [70], на ситуацію, коли інформація на обробку подається у формі вибірок з нерівномірними і асинхронними тактами квантування. Тому для реалізації оцінки відстані припустимо, що вихідна інформація задана у формі набору вибірок $x_{i(k)}(k)$ (де $i(k) = 1,2,\ldots,n(k)$ – номер певного спостереження у *k*-й

реалізації, при цьому кожна нова реалізація може містити різну кількість спостережень $n(k), k = 1,2,\ldots,N)$, що містить N послідовностей з нерівномірним тактом квантування, які підлягають нечіткій кластеризації, при цьому кожна вибірка може бути представлена в формі $(n(k)\times 1)$-вектора $x(k) = \left(x_1(k), x_2(k), ..., x_{n(k)}(k)\right)^T$. Кожне спостереження $x_{i(k)}(k)$ спостерігається в момент часу $0 \le t_{i(k)}(k) \le T$. Зрозуміло, що два вектора-вибірки $x_{i(k)} \in R^{n(k)}$ та $x(l) \in R^{n(l)}$ при $n(k) \ne n(l)$ в принципі непорівнянні.

Нерівномірність ж квантування означає те, що $\Delta t_{i(k)} = t_{i(k)} - t_{i-1(k)} \ne \Delta t_{i+1(k)} = t_{i+1(k)} - t_{i(k)}$, тобто $\Delta t_{i(k)} \ne const$. Крім того, у загальному випадку $t_{i(k)} \ne \Delta t_{i(l)}$.

Інтервал спостереження всього набору даних може бути заданий у вигляді $\left[t_1 = t_{1\min} = \min\left\{t_1(k)\right\} - t_n = T = \max\left\{t_{n(k)}(k)\right\}\right]$.

Для оцінки відстані між вибірками використаємо модифіковану PS-відстань:

$$x_t(k) = a_t(k) + b_t(k)t, \tag{4.5}$$

де $t_i(k) \le t \le t_{i+1}(k)$,

$$\begin{cases} a_t(k) = \dfrac{t_{i+1}(k)x_{i(k)}(k) - t_i(k)x_{i+1(k)}(k)}{t_{i+1}(k) - t_i(k)}, \\ b_t(k) = \dfrac{x_{i+1}(k) - x_i(k)}{t_{i+1}(k) - t_i(k)} \end{cases} \tag{4.6}$$

і фактично оцінює відміну форм аналізованих вибірок співвідношення (4.5) та (4.6), були використані для оцінки відстані між вибірками [69-74].

На рисунку 4.2 наведено приклад, де показані реалізації $x(1)=\left(x_1(1),x_2(1),...,x_{n(1)}(1)\right)^T$, $x(k)=x_1(k),x_2(k),...,x_{n(k)}(k))^T$ та $x(N)=x_1(N),x_2(N),...,x_{n(N)}(N))^T$, при цьому у загальному випадку $t_1(1)\neq t_1(k)\neq t_1(N)$ та $t_{n(1)}(1)\neq t_{n(k)}(k)\neq t_{n(N)}(N)$.

У разі асинхронних вибірок у розгляд можна ввести квазіспостереження, отримані за допомогою виразів (4.5) та (4.6).

Тому, повертаючись до рисунку 4.2, можна записати квазіспостереження послідовностей $x(1)$ та $x(N)$ в момент часу $t_2(k)$ у вигляді:

$$\begin{cases} \hat{x}_{t_2(k)}(1)=a_{t_2(k)}(1)+b_{t_2(k)}(1)t_2(k), \\ a_{t_2(k)}(1)=\dfrac{t_2(1)x_2(1)-t_1(1)x_1(1)}{t_2(1)-t_1(1)}, \\ b_{t_2(k)}(1)=\dfrac{x_2(1)-x_1(1)}{t_2(1)-t_1(1)}, \end{cases} \tag{4.7}$$

та

$$\begin{cases} \hat{x}_{t_2(k)}(N)=a_{t_2(k)}(N)+b_{t_2(k)}(N)t_2(k), \\ a_{t_2(k)}(N)=\dfrac{t_2(N)x_2(N)-t_1(N)x_1(N)}{t_2(N)-t_1(N)}, \\ b_{t_2(k)}(N)=\dfrac{x_2(N)-x_1(N)}{t_2(N)-t_1(N)}. \end{cases} \tag{4.8}$$

Цілком аналогічно можна ввести квазіспостереження $\hat{x}(k)$ на підставі реальних спостережень рядів $x(1),...,x(k-1),x(k+1),...,x(N)$ [71–74].

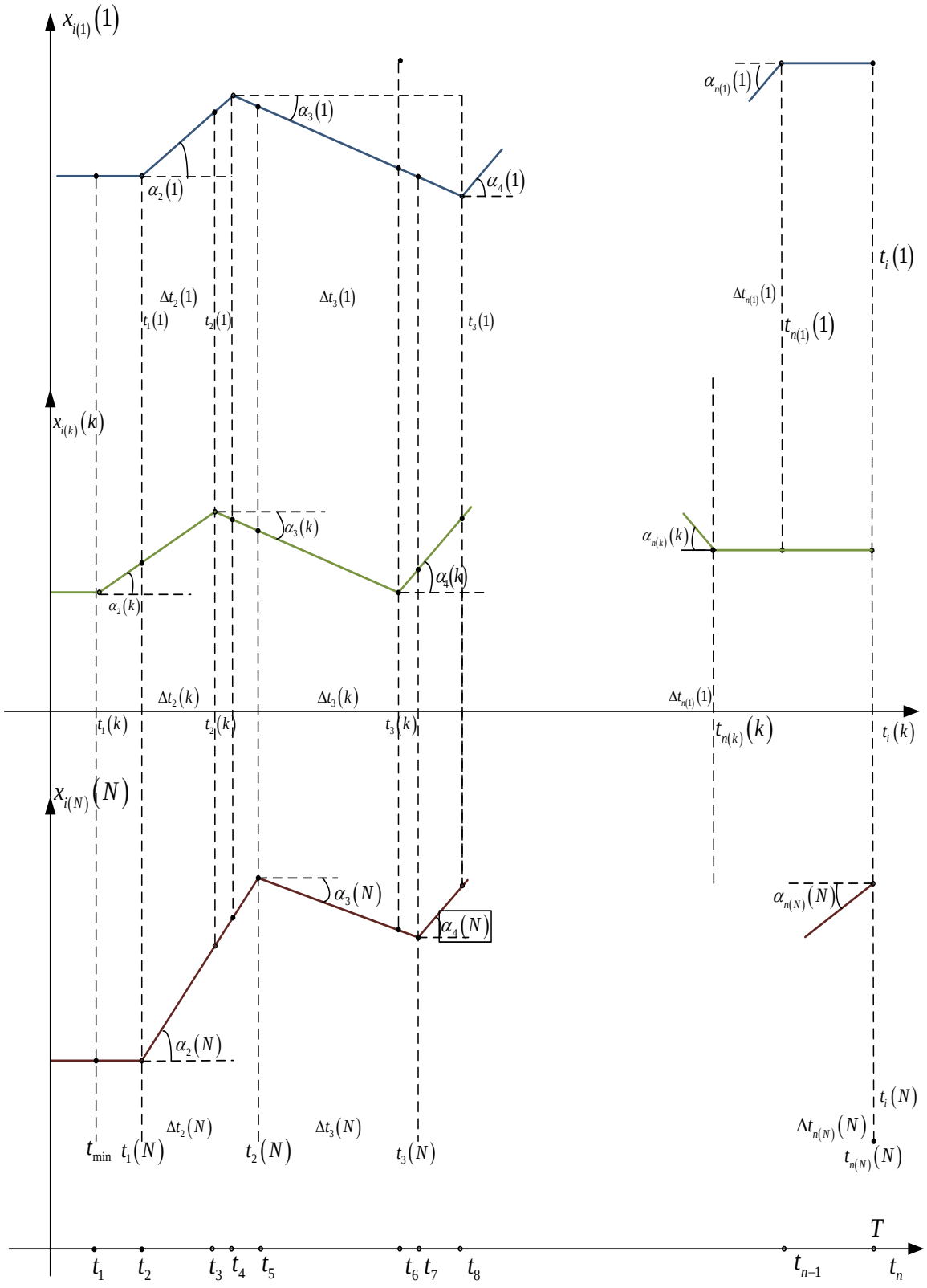


Рисунок 4.2 – Часові ряди з нерівномірними асинхронними тактами квантування

Як результат формується загальна часова шкала (нижня вісь, рис. 4.2), яка містить $n(1)=...=n(k)=...=n(N)$ моментів у випадку повністю синхронізованих вибірок та $n=\sum_{k=1}^{N} n(k)$ точок, якщо моменти фіксації даних у всіх рядах повністю не збігаються. На основі часової шкали може бути сформований набір з N векторів-вибірок:

$$\hat{x}(1)=\left(\hat{x}_1(1),\hat{x}_2(1),...,\hat{x}_i(1),...,\hat{x}_n(1)\right)^T,...,\hat{x}(k)=$$
$$=\left(\hat{x}_1(k),...,\ \hat{x}_i(k)...,\hat{x}_n(k)\right)^T,...,\hat{x}(N)=\left(\hat{x}_1(N),...,\hat{x}_i(N),...,\hat{x}_n(N)\right)^T,$$

які мають однакову розмірність $(n\times 1)$, при цьому компонентами цих векторів $\hat{x}_i(k)$ можуть бути як реальні спостереження, так і квазіспостереження типу (4.7) та (4.8).

4.4 Нечітка кластеризація часових рядів з нерівномірними асинхронними тактами квантування

На підставі метрики (4.9) авторами [75–82] була введена процедура нечіткої кластеризації, що є модифікацією алгоритму нечітких *c*-середніх (FCM) для обробки часових рядів з нерівновіддаленими спостереженнями, а у [80] – її онлайн версія.

Завдання кластеризації в рамках стандартного методу нечітких *c*-середніх зводиться до мінімізації цільової функції:

$$E^{FC}\left(u_j(k),\tilde{c}_j\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^{\beta}(k)\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2 \tag{4.9}$$

при наявності обмежень:

$$\sum_{j=1}^{m}u_j(k)=1,\quad 0<\sum_{k=1}^{N}u_j(k)\le N\,. \tag{4.10}$$

Ефективність підходу до кластеризації коротких часових рядів на основі критерія (4.9) була підтверджена в [83], проте, коли загальна довжина оброблених вибірок n (в межі $n=\sum_{k=1}^{N}n(k)$) може бути досить велика, починає проявлятися ефект «концентрації норм», що призводить до неефективності використання цього критерія [91]. Для подолання цього негативного ефекту був запропонований цілий ряд альтернативних (4.9) критеріїв, серед яких слід, перш за все, відзначити цільову функцію [84–85]:

$$E^{KH}\left(u_j(k),\tilde{c}_j\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2(k)+(1-\alpha)u_j(k)\right)\left\|\tilde{x}(k)-\tilde{c}_j^2\right\|^2, \tag{4.11}$$

де $0<\alpha\le 1$– налаштовний параметр, який встановлює компроміс між FCM і чітким методом *k*-середніх, з обмеженнями (5.2).

Авторами [86–87] було введено пакетний варіант мінімізації (4.7), а в [82] його онлайн версію. У [86] було введено процедуру нечіткої кластеризації із зважуванням компонентів оброблених даних. При цьому використовується цільова функція:

$$E^{KK}\left(u_j(k),\tilde{c}_j\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^{\beta}(k)\sum_{i=1}^{n}\gamma_{ji}^{t}\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)^2, \tag{4.12}$$

з додатковим обмеженням (крім (4.10)):

$$\sum_{i=1}^{n}\gamma_{ji}=1, \tag{4.13}$$

де γ_{ji} – параметр зважування для кожного компоненту вектора $\tilde{x}(k)$;

$t>0$ – параметр, який має смисл аналогічний фаззіфікатору та обирається емпірично.

У 2013 р. Ф. Клавонн запропонував як цільову функцію використати гібрид (4.11) та (4.12) виду:

$$\begin{gathered}E^{FK}\left(u_j(k),\tilde{c}_j\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}u_j^{\beta}(k)\times\\ \times\left(\alpha\sum_{i=1}^{n}\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)^2+(1-\alpha)\sum_{i=1}^{n}\gamma_{ji}^{t}\left(\tilde{x}(k)-\tilde{c}_{ji}\right)^2\right),\end{gathered} \tag{4.14}$$

що забезпечує компроміс між стандартним FCM та процедурою кластеризації із зважуванням даних. У розглянутому процесі пропонується використовувати цільову функцію:

$$\begin{gathered}E^{KB}\left(u_j(k),\tilde{c}_j\right)=\\ =\sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2(k)\sum_{i=1}^{n}\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)^2+(1-\alpha)u_j(k)\sum_{i=1}^{n}\gamma_{ji}^{t}\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)^2\right)=\\ =\sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2(k)\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2+(1-\alpha)u_j(k)\left\|\tilde{x}(k)-\tilde{c}_j\right\|_{\Gamma_j^t}^2\right),\end{gathered} \tag{4.15}$$

де:

$$\Gamma_j=diag\left(\gamma_{j1},\gamma_{j2},\ldots,\gamma_{jn}\right),\mathrm{Tr}\Gamma_j=1. \tag{4.16}$$

Для мінімізації (4.15) з урахуванням (4.16) введемо у розгляд функцію Лагранжа:

$$L\left(u_j(k),\tilde{c}_j,\rho_j\right)=\sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2(k)\sum_{i=1}^{n}\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)^2+(1-\alpha)u_j(k)\times\right.$$
$$\left.\times\sum_{i=1}^{n}\gamma_{ji}^t\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)^2\right)-\sum_{j=1}^{m}\rho_j\left(\sum_{i=1}^{n}\gamma_{ji}-1\right),$$

де ρ_j – невизначені множники Лагранжа,

та систему рівнянь Каруша-Куна-Таккера:

$$\begin{cases}\dfrac{\partial L\left(u_j(k),\tilde{c}_j,\rho_j\right)}{\partial\gamma_{ji}}=\sum\limits_{k=1}^{N}(1-\alpha)u_j(k)t\gamma_{ji}^{t-1}\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)^2-\rho_j=0,\\ \dfrac{\partial L\left(u_j(k),\tilde{c}_j,\rho_j\right)}{\partial\rho_j}=\sum\limits_{k=1}^{N}\gamma_{ji}-1=0.\end{cases}\qquad(4.17)$$

Проводячи ланцюжок очевидних перетворень:

$$\rho_j=(1-\alpha)t\gamma_{ji}^{t-1}\sum_{k=1}^{N}u_j(k)\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)^2,$$

$$\gamma_{ji}^{t-1}=\left(\frac{\rho_j}{(1-\alpha)t\sum\limits_{k=1}^{N}u_j(k)\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)^2}\right)^{\frac{1}{t-1}},$$

$$1=\left(\frac{\rho_j}{(1-\alpha)t}\right)^{\frac{1}{t-1}}\left(\frac{1}{\sum\limits_{k=1}^{N}u_j(k)\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)^2}\right)^{\frac{1}{t-1}},$$

$$\rho_j = \frac{(1-\alpha)t}{\left(\sum_{i=1}^{n}\left(\frac{1}{\sum_{k=1}^{N} u_j(k)\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)^2}\right)^{\frac{1}{t-1}}\right)^{t-1}},$$

$$\gamma_{ji}^{t-1} = \frac{\rho_j}{(1-\alpha)t\sum_{k=1}^{N} u_j(k)\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)^2},$$

Як результат отримуємо:

$$\gamma_{ji} = \frac{1}{\sum_{l=1}^{n}\left(\frac{\sum_{k=1}^{N} u_j(k)\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)^2}{\sum_{k=1}^{N} u_j(k)\left(\tilde{x}_l(k)-\tilde{c}_{jl}\right)^2}\right)^{\frac{1}{t-1}}}, \tag{4.18}$$

звідки випливає, що параметри γ_{ji} не залежать від α, тобто зважування окремих компонентів кластеризованих векторів проводиться по типу [88].

Для знаходження центроїдів кластерів запишемо рівняння:

$$\frac{\partial E^{KB}\left(u_j(k),\tilde{c}_j\right)}{\partial \tilde{c}_{ji}} = -2\sum_{k=1}^{N}\left(\alpha u_j^2(k)\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)+ \right.$$
$$\left. +(1-\alpha)u_j(k)\gamma_{ji}^t\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)\right)=0, \tag{4.19}$$

звідки

$$\tilde{c}_{ji} = \frac{\sum_{k=1}^{N}\left(\alpha u_j^2(k) + (1-\alpha)u_j(k)\gamma_{ji}^t\right)\tilde{x}_i(k)}{\sum_{k=1}^{N}\left(\alpha u_j^2(k) + (1-\alpha)u_j(k)\gamma_{ji}^t\right)}. \qquad (4.20)$$

При $\alpha = 1$ приходимо до стандартного FCM-алгоритму, а при $\gamma_{ji} = 1$ – до алгоритму, запропонованого у [89–90].

Для обліку обмежень (4.10) введемо ще одну функцію Лагранжа:

$$L\left(u_j(k), \tilde{c}_j, \lambda(k)\right) = \sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2(k)\sum_{i=1}^{n}\left(\tilde{x}_i(k) - \tilde{c}_{ji}\right)^2 + (1-\alpha)u_j(k)\times\right.$$

$$\left.\times\sum_{i=1}^{n}\gamma_{ji}^t\left(\tilde{x}_i(k) - \tilde{c}_{ji}\right)^2\right) - \sum_{k=1}^{N}\lambda(k)\left(\sum_{j=1}^{m}u_j(k) - 1\right) = \qquad (4.21)$$

$$= \sum_{k=1}^{N}\sum_{j=1}^{m}\left(\alpha u_j^2(k)\left\|\tilde{x}(k) - \tilde{c}_j\right\|^2 + (1-\alpha)u_j(k)\left\|\tilde{x}(k) - \tilde{c}_j\right\|_{\Gamma_j^t}^2 - \sum_{j=1}^{m}u_j(k) - 1,\right.$$

де $\lambda(k) - N$ невизначених множників Лагранжа,

та систему Каруша-Куна-Таккера:

$$\begin{cases} \dfrac{\partial L\left(u_j(k), \tilde{c}_j, \lambda(k)\right)}{\partial u_j(k)} = 2\alpha u_j(k)\left\|\tilde{x}(k) - \tilde{c}_j\right\|^2 + \\ \qquad + (1-\alpha)\left\|\tilde{x}(k) - \tilde{c}_j\right\|_{\Gamma_j^t}^2 - \lambda(k) = 0, \\ \dfrac{\partial L\left(u_j(k), \tilde{c}_j, \lambda(k)\right)}{\partial \lambda(k)} = \sum_{j=1}^{m}u_j(k) - 1 = 0. \end{cases} \qquad (4.22)$$

Здійснюючи аналогічно попереднім низку перетворень:

$$2\alpha u_j(k)\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2=\lambda(k)-(1-\alpha)\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2_{\Gamma_j^t},$$

$$u_j(k)=\frac{\lambda(k)-(1-\alpha)\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2_{\Gamma_j^t}}{2\alpha\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2},$$

$$\sum_{j=1}^{m}\frac{\lambda(k)-(1-\alpha)\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2_{\Gamma_j^t}}{2\alpha\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2}=1,$$

$$\lambda(k)\frac{1}{2\alpha}\sum_{j=1}^{m}\left\|\tilde{x}(k)-\tilde{c}_j\right\|^{-2}=1+\frac{(1-\alpha)}{2\alpha}\sum_{j=1}^{m}\frac{\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2_{\Gamma_j^t}}{\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2},$$

$$\lambda(k)=\frac{1+\frac{(1-\alpha)}{2\alpha}\sum_{j=1}^{m}\frac{\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2_{\Gamma_j^t}}{\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2}}{\frac{1}{2\alpha}\sum_{j=1}^{m}\left\|\tilde{x}(k)-\tilde{c}_j\right\|^{-2}},$$

$$2\alpha u_j(k)\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2+(1-\alpha)\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2_{\Gamma_j^t}-\frac{1+\frac{(1-\alpha)}{2\alpha}\sum_{j=1}^{m}\frac{\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2_{\Gamma_j^t}}{\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2}}{\frac{1}{2\alpha}\sum_{j=1}^{m}\left\|\tilde{x}(k)-\tilde{c}_j\right\|^{-2}}=0,$$

остаточно отримуємо процедуру нечіткої кластеризації на основі критерію (4.15):

$$\begin{cases} u_j(k) = \dfrac{1 + \dfrac{(1-\alpha)}{2\alpha}\sum\limits_{j=1}^{m}\dfrac{\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2_{\Gamma_j^t}}{\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2} - (1-\alpha)\left\|\tilde{x}(k)-\tilde{c}_j\right\|^2_{\Gamma_j^t}}{\dfrac{1}{2\alpha}\sum\limits_{k=1}^{m}\left\|\tilde{x}(k)-\tilde{c}_j\right\|^{-2}}, \\ \gamma_{ji} = \left(\sum\limits_{l=1}^{n}\left(\dfrac{\sum\limits_{k=1}^{N}u_j(k)\left(\tilde{x}_i(k)-\tilde{c}_{ji}\right)^2}{\sum\limits_{l=1}^{N}u_j(k)\left(\tilde{x}_l(k)-\tilde{c}_{jl}\right)^2}\right)^{\frac{1}{t-1}}\right)^{-1}, \\ \tilde{c}_{ji} = \dfrac{\sum\limits_{k=1}^{N}\left(\alpha u_j^2(k)+(1-\alpha)u_j(k)\gamma_{ji}^t\right)\tilde{x}_i(k)}{\sum\limits_{k=1}^{N}\left(\alpha u_j^2(k)+(1-\alpha)u_j(k)\gamma_{ji}^t\right)}. \end{cases} \quad (4.23)$$

Тому співвідношення (4.23) є узагальненням алгоритмів нечіткої кластеризації, заснованих на цільових функціях (4.9), (4.11), (4.12), (4.14), тобто при відповідному виборі вільних параметрів, що перетворюються у відомі процедури.

Висновки до розділу

1. Розглянуто задачу нечіткої кластеризації багатовимірних коротких часових рядів з нерівномірним тактом квантування, які можуть бути представлені у формі пакету спостережень або послідовно надходити на обробку в онлайн режимі.

2. Розглянуто матричну модифікацію нейро-фаззі мережі Т. Кохонена, що навчається на основі правила «Переможець отримує більше».

3. Запропоновано метод для адаптивної кластеризації, що заснований на використанні критеріїв оцінки якості рішення і дозволяє повністю формалізувати розв’язання задачі нечіткої кластеризації багатовимірних часових рядів. Відмінною особливістю методики є оцінка якості кожного розбиття і вибір найкращого з них.

4. Запропоновано процедуру нечіткої кластеризації, що не схильна до ефекту «концентрації норм» і є узагальненням ряду відомих алгоритмів ймовірнісної нечіткої кластеризації.

5. Запроваджена процедура може бути корисна при вирішенні завдань, що виникають в рамках інтелектуального аналізу потоків даних, коли вихідні дані мають високу розмірність.

Список використаних у цьому розділі джерел наведено у повному списку використаних джерел під номерами [50–90], [92].

5 ІМІТАЦІЙНЕ МОДЕЛЮВАННЯ ТА РОЗВ'ЯЗАННЯ ПРАКТИЧНИХ ЗАДАЧ

Мета розділу - провести експерименти на основі тестових та реальних даних. Розробити програмний модуль, що може бути застосований для моніторингу ряду практичних впроваджень. Зокрема реалізувати модуль, що підтвердив свою ефективність у задачах моніторингу медичних даних в онлайн режимі.

Завдання: 1) провести імітаційне моделювання методі навчання робастних адаптивних моделей часових рядів; 2) провести імітаційне моделювання методів адаптивної можливісної нечіткої кластеризації часових рядів; 3) провести імітаційне моделювання послідовної онлайн нечіткої кластеризації багатовимірних рядів на базі модифікованої нейро-фаззі мережі Т. Кохонена; 4) розглянути процедуру фаззі-кластеризації з асинхронними тактами квантування, що не схильна до «концентрації норм» і розв'язання практичних задач в рамках концепції інтелектуального аналізу даних.

5.1 Імітаційне моделювання робастних адаптивних моделей часових рядів

Для виключення емпіричного вибору тієї чи іншої моделі ідентифікації нестаціонарного нелінійного сигналу, який забруднений викидами з невідомими законами розподілу, доцільно побудувати ансамбль адаптивних моделей, який би навчався в процесі обробки потоку даних.

Таким чином, вхідний сигнал $X(k) = \left(x_1(k), x_2(k), \ldots, x_n(k)\right)^T$ паралельно обробляється P адаптивними моделями ідентифікації, що формують на своїх

виходах скалярні сигнали $y_i(k)$, які можуть бути об'єднані в $(P \times 1)$ – вектор виходів $Y(k) = \left(y_1(k), y_2(k), \ldots, y_P(k) \right)^T$. Далі цей сигнал надходить на вхід оптимізаційного блоку, де перетворюється в найкращий в сенсі прийнятого зовнішнього критерію скалярний сигнал $\overline{y}(k)$. На рисунку 5.1 представлено архітектуру ансамблю адаптивних моделей ідентифікації.

У більшості випадків для визначення якості вирішення завдань ідентифікації та прогнозування прийнято використовувати критерій виду:

$$MAPE_i = \frac{100}{N} \sum_{k=1}^{N} \left| \frac{y(k) - \hat{y}_i(k)}{y(k)} \right|, \qquad (5.1)$$

де N – кількість спостережень у вибірці;

y – фактичне значення сигналу;

$\hat{y}_i$ – значення, отримане на виході i –моделі;

$MAPE_i$ (Mean Absolute Percentage Error) – середня абсолютна процентна помилка i –ї моделі.

Отже, в блоці селекції ансамблю адаптивних гібридних моделей в кожен момент дискретного часу буде проводитися вибір найкращого вихідного сигналу в рамках прийнятого критерія (5.1) у такий спосіб:

$$MAPE(k) = \min_i [MAPE_i(k)], \qquad (5.2)$$

де $\hat{y}_{ans}(k) = y_i(k)$.

Чисельне моделювання розробленого ансамблю гібридних адаптивних моделей ідентифікації проводилося на основі сигналу, багатого на перешкоди інтенсивними викидами.

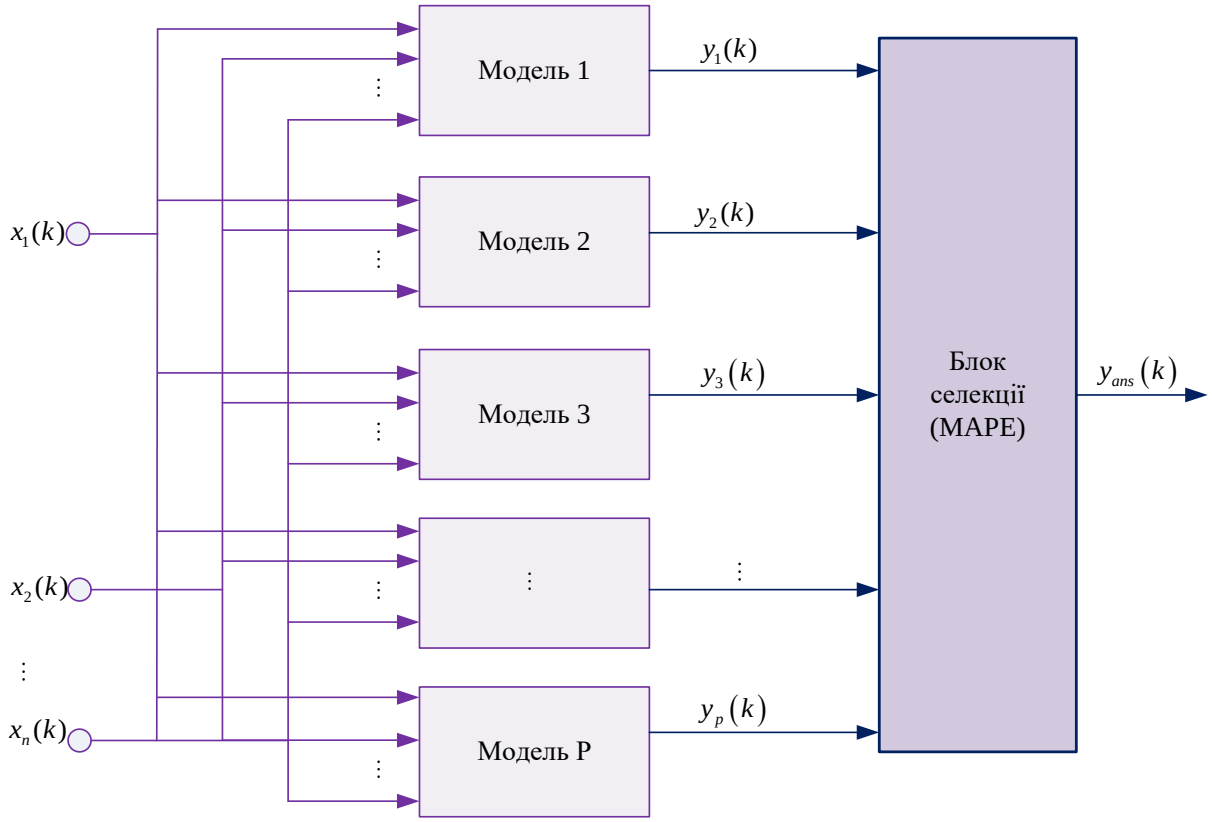


Рисунок 5.1 – Ансамбль адаптивних моделей ідентифікації

Сигнали, отримані на основі ряду Мекі-Гласса [94-96], які були згенеровані за допомогою рівняння:

$$\dot{x} = \frac{0.2x(t-\tau)}{1+x^{10}(t-\tau)} - 0.1x(t). \qquad (5.3)$$

Значення часового ряду для кожної точки були отримані за допомогою методу Рунге-Кутта четвертого порядку, який використовує часовий крок для даного методу 0.1, а початкові умови $x(0)=1.2$, затримка $\tau=17$ та $x(t)$ були отримані для $t=0\ldots51000$ і накладеного на них випадкового шуму, згенерованого за розподілом Коші.

На рисунку 5.2 представлений загальний вид оброблюваного сигналу з зашумленям з розподілу Коші. Значення $x(t-18), x(t-12), x(t-6), x(t)$ були використані для ідентифікації $x(t+6)$. Початкові значення синаптичних ваг були згенеровані випадковим способом в діапазоні від –0.1 до 0.1.

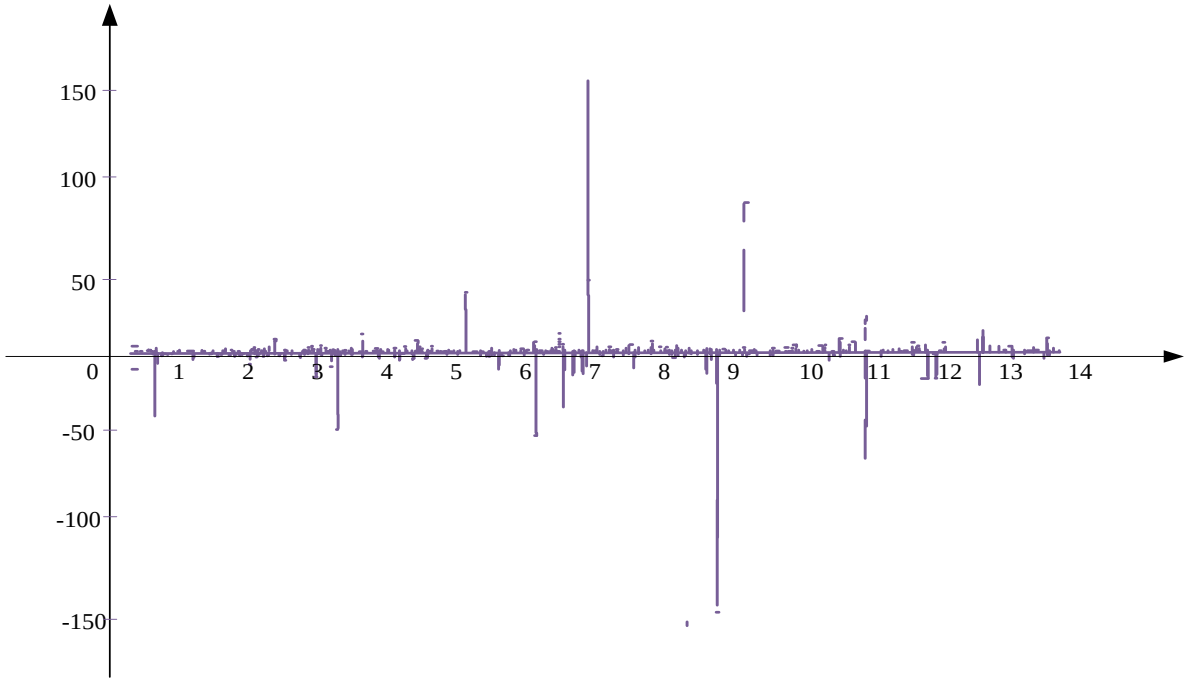
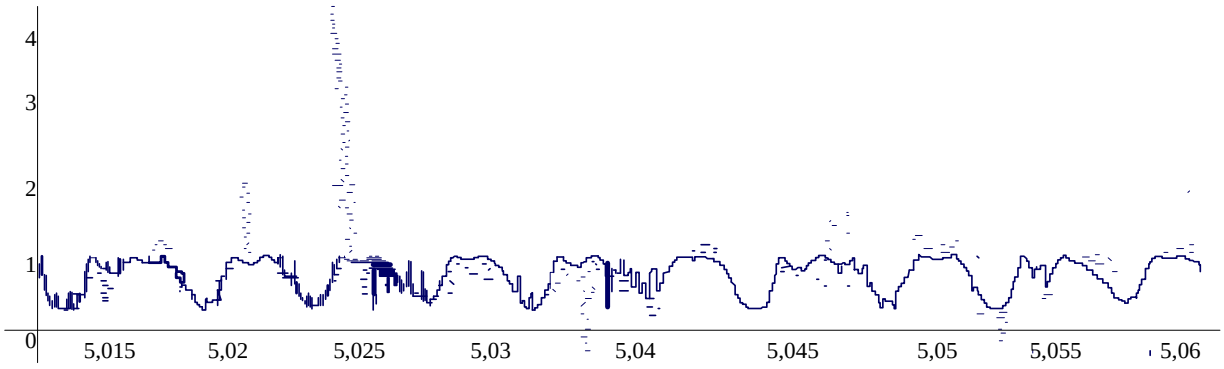


Рисунок 5.2 – Часовий ряд Мекі-Гласса з шумом, що має розподіл Коші

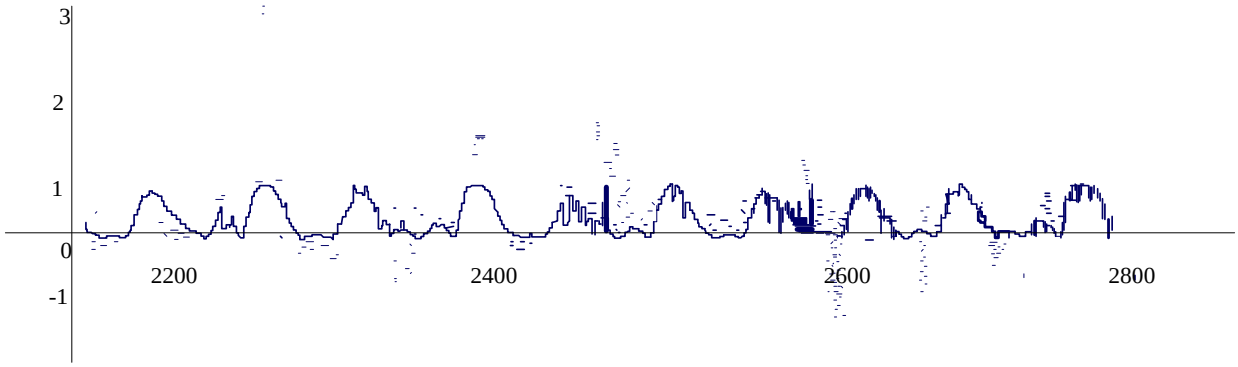
Як критерій якості була використана середньоквадратична помилка (RMSE). На рисунку 5.3 (а) представлено результати ідентифікації зашумленного часового ряду (реальні значення (пунктирна лінія) і значення вихідної системи ідентифікації (суцільна лінія)) [93]. На рисунку 5.3 (б) представлено сегмент процесу навчання. Як видно, ряд викидів з великою амплітудою, присутніх на початку вибірки, не зробив суттєвого впливу на метод навчання.

Порівняння результатів ідентифікації на основі робастного методу навчання проводилося з результатами ідентифікації на основі методів стохастичної апроксимації та алгоритму на основі рекурентного методу

найменших квадратів, де структура моделі і кількість параметрів налаштувань були однаковими. На рисунку 5.3 (а) представлено результати ідентифікації сигналу на основі методу стохастичної апроксимації (фактичні значення (пунктирна лінія) і значення вихідної системи ідентифікації (суцільна лінія)). На рисунку 5.4 (б) представлено сегмент процесу навчання. Як видно, перший же викид на початку вибірки сильно вплинув на процес навчання і як результат – велика помилка ідентифікації.

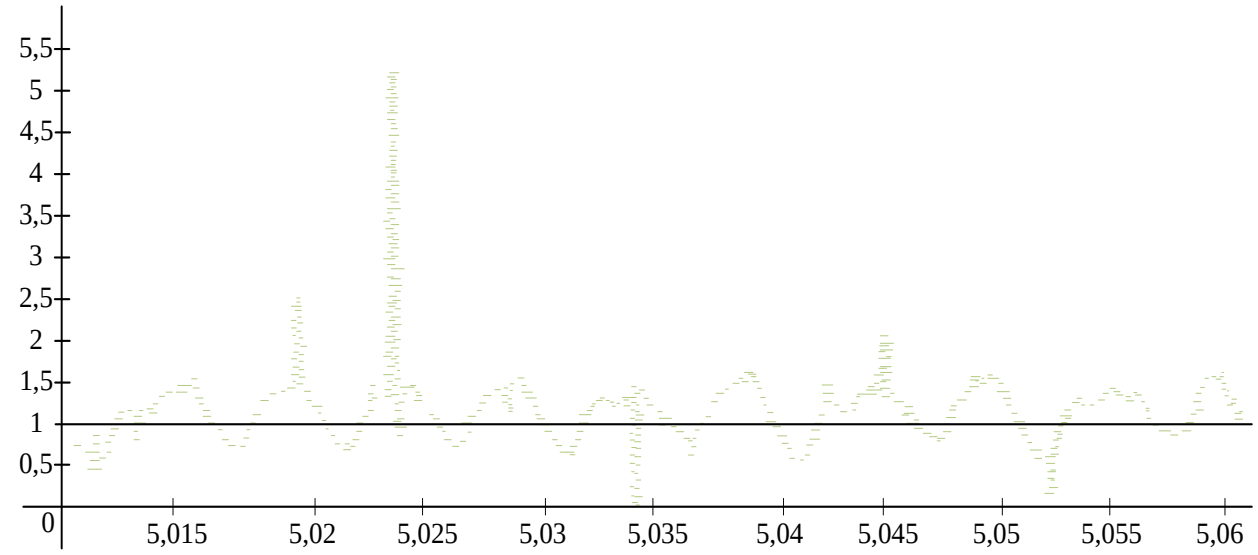


а)

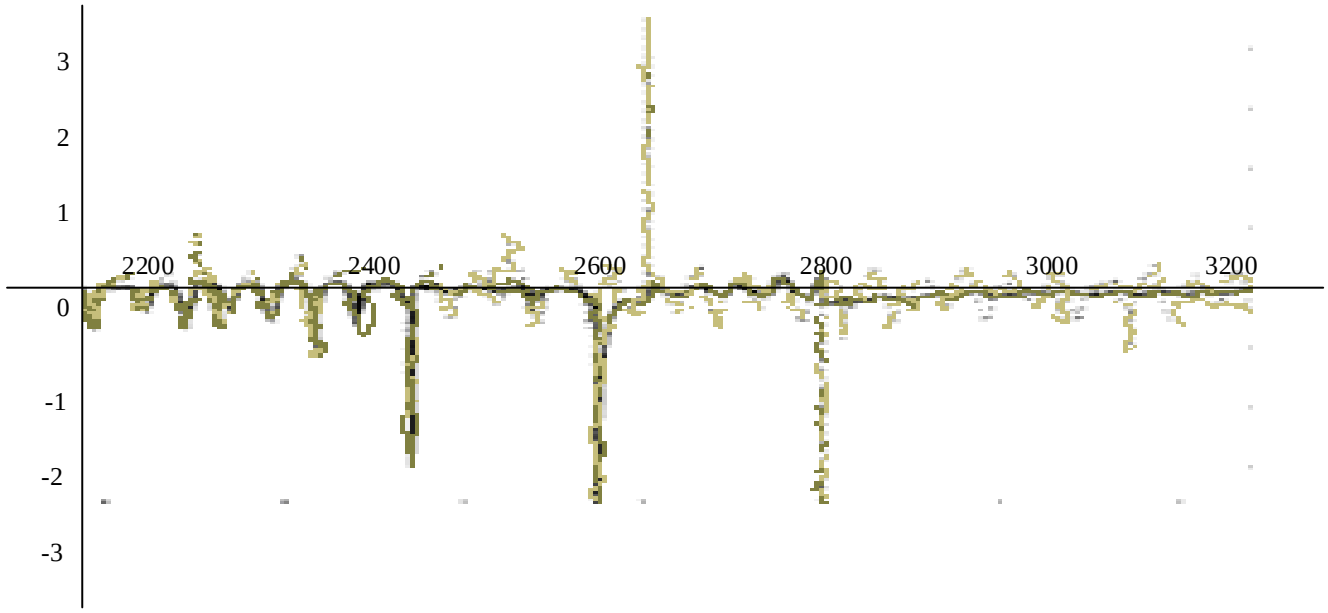


б)

Рисунок 5.3– Результати обробки сигналу за допомогою робастного методу навчання



а)



б)

Рисунок 5.4 – Результати обробки сигналу за допомогою алгоритмів стохастичної апроксимації

При навчанні моделі рекурентним методом найменших квадратів при першому викиді відбувається «вибух параметрів» коваріаційної матриці та як

результат – неможливість ідентифікації сигналів, зашумлених аномальними викидами [97–107, 111]. У таблиці 5.1 наведено результати порівняння значення вихідної системи ідентифікації сигналу на основі різних підходів.

Таблиця 5.1 – Порівняння результатів ідентифікації зашумленних сигналів

Модель та метод навчання	RMSE
Налаштовна модель, що навчається за допомогою алгоритму Гудвіна–Ремеджа–Кейнеса	0.3203
Налаштовна модель, навчена на основі квадратичної функції втрат	0.1932
Налаштовна модель, навчена на основі функції втрат Коши	0,0723
Налаштовна модель, навчена на основі реккурентного методу найменших квадратів	∞
Ансамбль гібридних адаптивних моделей ідентифікації	0.0735

Таким чином, видно, що запропонований підхід дозволяє обробляти сигнали в умовах істотного забруднення викидами з невідомим законом розподілу.

5.2 Імітаційне моделювання методів адаптивної можливісної нечіткої кластеризації коротких часових рядів

Для підтвердження ефективності запропонованого підходу до кластеризації-сегментації коротких часових рядів з нерівномірно

розподіленими спостереженнями була розглянута задача кластеризації-сегментації часових рядів погодинного споживання енергії.

Результати запропонованого підходу дозволяють підвищити якість аналізу і прогнозування часових рядів.

Часовий ряд складається з 2400 спостережень, для кластеризації даний часовий ряд був розділений на сегменти з 8 спостереженнями. Для отримання нерівномірно розподілених спостережень в кожному сегменті 3-є і 5-є спостереження були видалені з набору даних.

Отже, згідно з (4.3) отримуємо набір даних у вигляді таблиці «об'єкта - властивості» з 300 спостереженнями і 6 властивостями. Ряд кластерів $m = 3$ (ранковий, денний і вечірній сегменти енергоспоживання).

Всі алгоритми кластеризації були перевірені одним і тим же набором даних. Критерієм якості результатів кластеризації була прийнята середня помилка середнього класу (ПСК).

У першому експерименті порівняно продуктивність алгоритмів кластеризації в завданні класифікації, коли в наборі даних, що використовуються для кластеризації, були присутні екземпляри всіх доступних класів, тобто число класів було відоме апріорі і дорівнює 3. Набори даних були розділені на набори для навчання і тестування з 70% та 30% даних відповідно.

Для підвищення продуктивності алгоритмів рекурсивної кластеризації набори даних були випадковим способом перемішані. Навчальні набори використовувалися для ініціалізації класифікатора за допомогою нечіткої кластеризації, а тестові набори використовувалися для порівняння точності класифікації.

Використовуючи швидкість навчання $\eta = 0.01$ в рекурсивних процедурах, де параметр фаззіфікатора був узятий $\beta = 1.1$. Обидві можливі

процедури (пакетна і рекурсивна) були ініційовані за результатами ймовірнісної кластеризації за допомогою алгоритму нечітких *c*-середнього.

Було виконано 10 ітерацій для процедур пакетної кластеризації та 10 прогонів по навчальних даних для процедур рекурсивної кластеризації. Експеримент повторювали 50 разів, а потім розраховували середні результати.

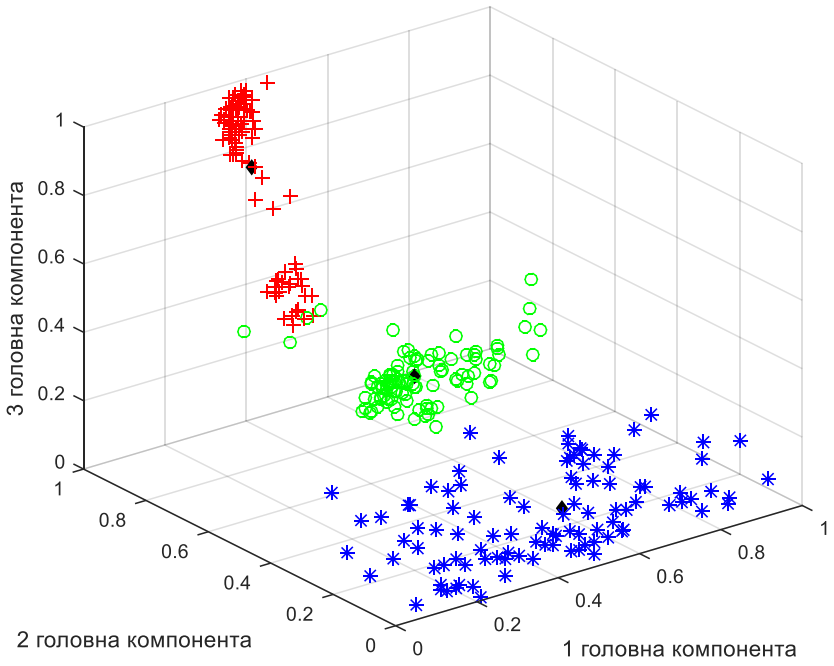
Результати наведені в таблиці 5.2 і представляють відсоток неправильно класифікованих об'єктів з набору даних тестування.

На рисунках 5.5 і 5.6 наведені результати кластеризації, отримані за допомогою запропонованих підходів.

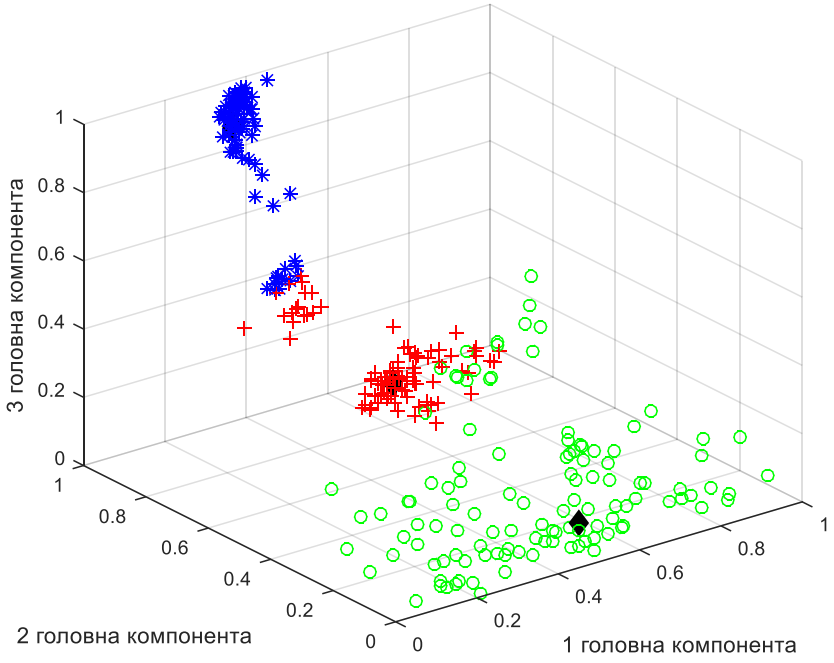
Як видно з отриманих результатів, алгоритми нечіткої ймовірнісної кластеризації мають найкращу якість кластеризації (як пакетного, так і адаптивного режимів) [103]. Також видно, що результати адаптивних режимів алгоритмів кластеризації мають кращу якість, ніж в пакетному режимі.

Таблиця 5.2 – Результати кластер-сегментації часової серії

Процедури кластеризації	M{ПСК}
Алгоритм нечіткої ймовірнісної кластеризації	1.6 % (5)
Алгоритм адаптивної нечіткої ймовірнісної кластеризації	1.3 % (4)
Ймовірнісний алгоритм кластеризації	11.1 % (33)
Адаптивний алгоритм можливісної кластеризації	6.3 % (19)

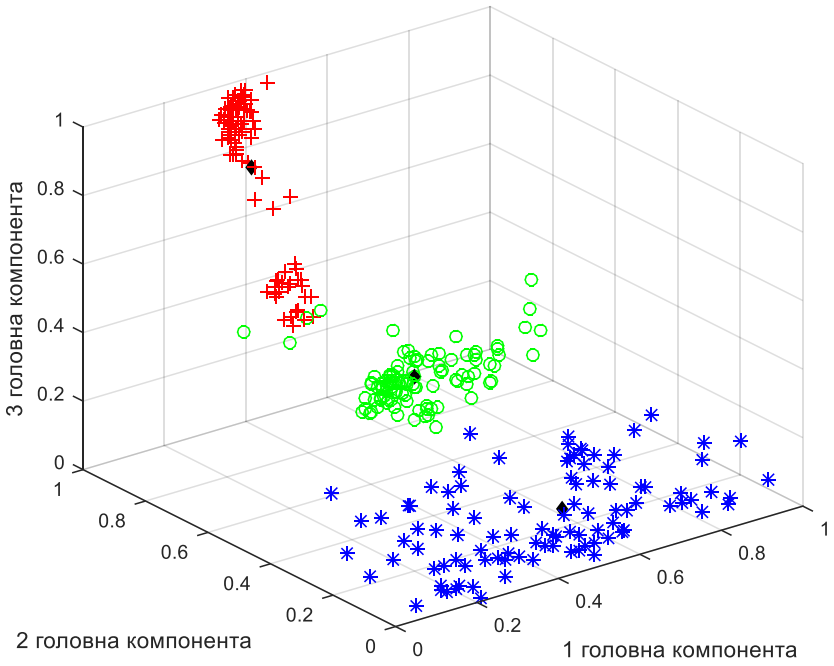


а)

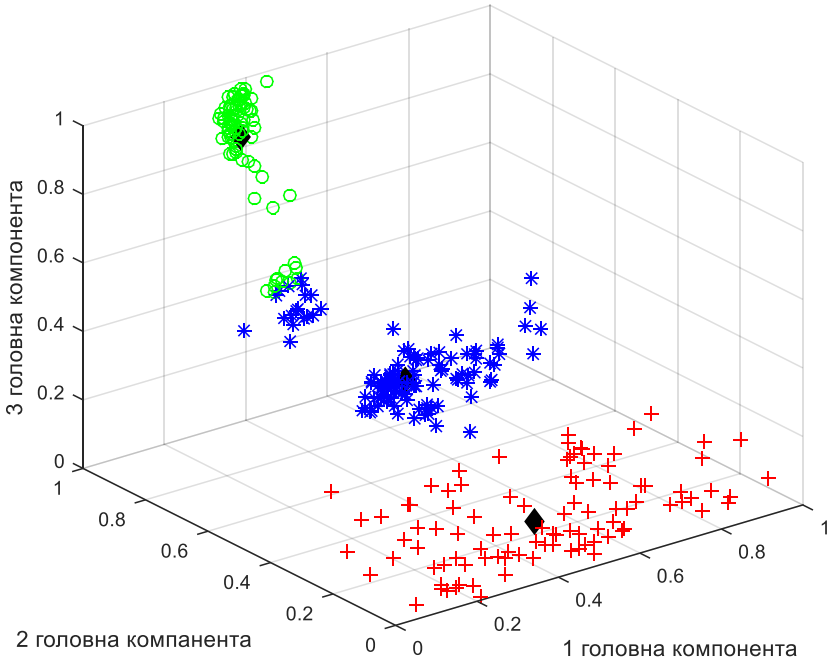


б)

Рисунок 5.5 – Проекція набору даних і прототипів кластерів на головні компоненти для алгоритмів нечіткої ймовірнісної кластеризації:
а) – пакетний режим, б) – адаптивний режим



а)



б)

Рисунок 5.6 – Проекція набору даних і прототипів кластерів на основні компоненти для алгоритмів можливісної кластеризації

а) – пакетний режим, б)– адаптивний режим

5.3 Імітаційне моделювання послідовної онлайн нечіткої кластеризації багатовимірних рядів на базі модифікованої нейро-фаззі мережі Т.Кохонена

Однією з особливостей мап Кохонена є наявність етапу в процесі самоорганізації, коли нейрон–переможець визначає локальну область топологічного сусідства, в якій збуджується не тільки він сам, але і його найближче оточення, при цьому більш близькі до переможця нейрони збуджуються сильніше, ніж віддалені [99].

Самоорганізовна мапа має дуже просту архітектуру з прямою передачею інформації. Крім нульового (рецепторного) шару, вона містить єдиний шар нейронів, який дуже часто називають шаром Кохонена [100].

Саме завдяки такій організації кожен нейрон мережі отримує всю інформацію про аналізований образ і генерує на своєму виході відповідний відгук. Після цього у шарі Кохонена виникає режим конкуренції, в результаті якої визначається єдиний нейрон-переможець з максимальним вихідним сигналом. Даний сигнал по латеральним зв'язкам забезпечує збудження найближчих «сусідів» переможця і придушення реакції далеко віддалених вузлів. На рисунку 5.7 представлена 1D-мапа Кохонена:

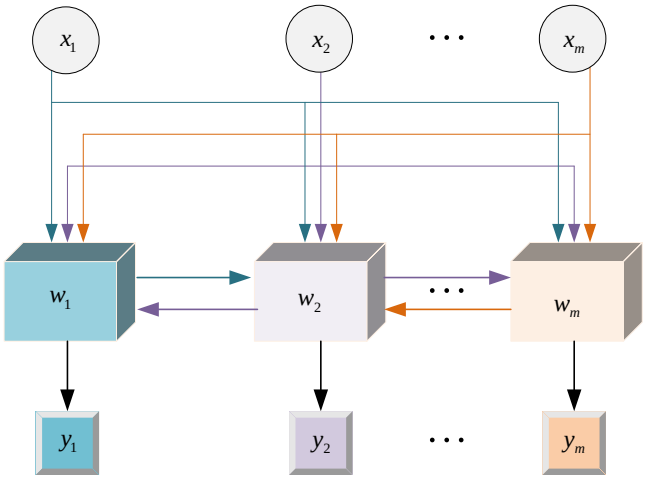


Рисунок 5.7 – 1D-мапа Кохонена

У процесі навчання сусідні нейрони впливають один на одного сильніше, ніж ті, які розташовані далі. Саме латеральні зв'язки в мережі забезпечують збудження одних нейронів і гальмування інших. Самоорганізовні карти можуть мати різну топологію, однак найбільш часто рецептори і нейрони розташовуються у вузлах одновимірної або двовимірної решітки.

Тому процедури кластеризації (4.3), (4.4) були введені в припущенні, що вся інформація задана у вигляді фіксованого масиву даних $X(1), X(2), \ldots, X(N)$ і не змінюється з плином часу. Якщо ж дискретні поля $X(k)$ надходять на обробку послідовно у формі потоку даних, можна скористатися підходами, що використовуються в динамічному інтелектуальному аналізі даних і, насамперед, адаптивними методами [101].

Для послідовної обробки даних найкраще пристосовані кластерні нейронні мережі – самоорганізовні мапи Т. Кохонена [102]–[104], що дозволяють в онлайн режимі самонавчання провести чітке розбиття потоку векторних спостережень.

За умов, коли вихідна інформація надходить у формі $(q \times n)$ – матричних спостережень класів, що перетинаються, можна скористатися матричною нейро-фаззі кластерувальною мережею [105].

Скориставшись для пошуку сідлової точки лагранжіана (4.5) рекурентним алгоритмом нелінійного програмування Ерроу-Гурвіца-Удзави, можна записати адаптивні процедури кластеризації багатовимірних коротких часових рядів з нерівномірним тактом квантування у вигляді:

$$\begin{cases} u_j(k) = \dfrac{(Tr(\tilde{X}(k) - \tilde{C}_j(k-1))(\tilde{X}(k) - \tilde{C}_j(k-1))^T)^{\frac{1}{1-\beta}}}{\sum\limits_{g=1}^{N}(Tr(\tilde{X}(k) - \tilde{C}_g(k-1))(\tilde{X}(k) - \tilde{C}_g(k-1))^T)^{\frac{1}{1-\beta}}}, \\ \tilde{C}_j(k) = \tilde{C}_j(k-1) - \eta(k)\{\partial L(\mathrm{u}_j(k), \tilde{C}_j, \lambda(k)) / \partial \tilde{C}_{jip}\} = \\ = \tilde{C}_j(k-1) + \eta(k)\,\mathrm{u}_j^{\beta}(k)(\tilde{X}(k) - \tilde{C}_j(k-1)), \end{cases} \tag{5.4}$$

для довільного значення фаззіфікатора β (тут $\eta(k)$ – параметр кроку навчання) і

$$\begin{cases} u_j(k) = \dfrac{(Tr(\tilde{X}(k) - \tilde{C}_j(k-1))(\tilde{X}(k) - \tilde{C}_j(k-1))^T)^{-2}}{\sum\limits_{g=1}^{m}(Tr(\tilde{X}(k) - \tilde{C}_g(k-1))(\tilde{X}(k) - \tilde{C}_g(k-1))^T)^{-2}}, \\ \tilde{C}_j(k) = \tilde{C}_j(k-1) + \eta(k)u_j^2(k)(\tilde{X}(k) - \tilde{C}_j(k-1)), \end{cases} \tag{5.5}$$

для β=2.

Нескладно помітити, що з позиції самонавчання кластерувальних мереж Т. Кохонена, інші рекурентні співвідношення (5.4) та (5.5) є модифікаціями на матричний випадок правил налаштування на базі принципу «Переможець отримує більше» [107], де множник $u_j^{\beta}(k)$ виконує роль функції сусідства.

На рисунку 5.8 представлена архітектура запропонованої адаптивної матричної нейро–фаззі самоорганізовної мережі.

Таким чином, для розв’язання задачі нечіткої кластеризації багатовимірних часових рядів можуть бути використані архітектури, що є за суттю самоорганізовними мапами з $(q \times n)$-матричним входом і m-матричними вузлами [106].

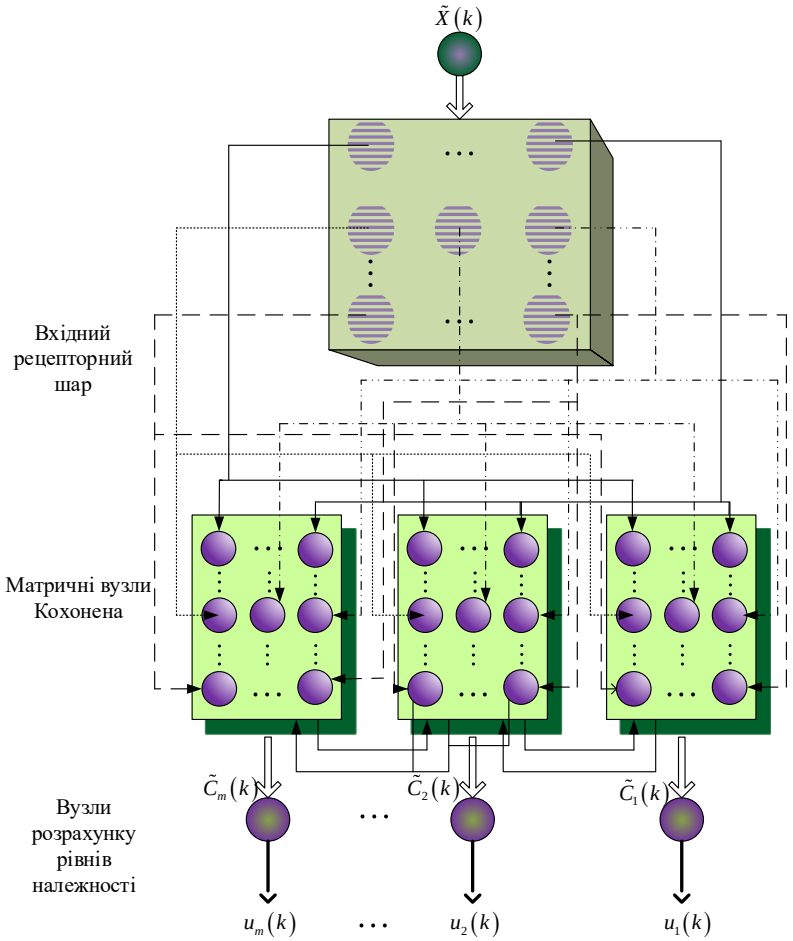


Рисунок 5.8 – Архітектура адаптивної матричної нейро-фаззі самоорганізовної мережі

5.4 Імітаційне моделювання методу нечіткої кластеризації часових рядів з нерівномірними асинхронними тактами квантування та експериментальні дослідження

У підрозділі розглянуто задачу нечіткої кластеризації часових рядів з нерівномірними і асинхронними тактами квантування, проведено моделювання. Запроваджена процедура може бути корисна при вирішенні завдань, що виникають в рамках інтелектуального аналізу, коли вихідні дані мають велику розмірність [107].

Актуальною медичною проблемою у світі, що займає лідируючі позиції є захворювання серцево-судинної системи (в Україні майже 57 % пацієнтів хворіють та помирають як серцево-судинні хворі). Однією з поширених патологій є ішемічна хвороба серця (ІХС), а стенокардія, у свою чергу, є найчастішою формою ішемічної хвороби серця. На сьогодні спостерігається позитивна тенденція в областях діагностики, профілактики, медикаментозного та хірургічного лікування серцево-судинних захворювань. Наприклад, операція коронарного шунтування (КШ) - один з найбільш ефективних методів хірургічного лікування хворих на ІХС. Також, відбувається розвиток реабілітології, відзначається тенденція до індивідуалізації та застосування нових відновлювальних методик, розширення показань для їх призначення. Це дозволяє ширше і ефективніше застосовувати можливості відновлювальної медицини та прискорювати процес реабілітації.

На сьогоднішній день хворі, які перенесли операцію на серці, проходять післяопераційні етапи, розбиті відповідно до стадій (I, II, III) захворювання стенокардією. Так, кожному періоду відновлювальної терапії відповідає певний набір значень який необхідно аналізувати і який залежить від тяжкості стану пацієнта, патології та індивідуальних особливостей, тому кожна група відповідної стадії стенокардії розбита на чотири групи. Дані збираються безпосередньо в день операції, потім в перший, третій, сьомий, чотирнадцятий та двадцять восьмий дні. Значення кожного хворого в кожній групі характеризуються багатовимірним часовим рядом не синхронізованим у часі.

Наприклад, проведемо аналіз артеріального тиску. Моніторинг артеріального тиску в період відновлювальної терапії після операції коронарного шунтування та підтримання його на оптимальному рівні є життєво важливим. Тому для експериментів обрані параметри артеріального тиску у післяопераційний період в день операції, в перший, третій, сьомий,

чотирнадцятий і двадцять восьмий дні. Значення артеріального тиску змінюються протягом доби.

Інформація у формі набору вибірок $x_{i(k)}(k)$ кожного пацієнта має кількість спостережень $N=6$ – відповідно це в післяопераційний період в день операції, в перший, третій, сьомий, чотирнадцятий і двадцять восьмий дні, $n(k), k=1,2,\ldots,N$. Кожна послідовність підлягає нечіткій кластеризації. На рисунку 5.9 наведено дані всіх пацієнтів, які перенесли операцію на серці та проходять післяопераційні етапи.

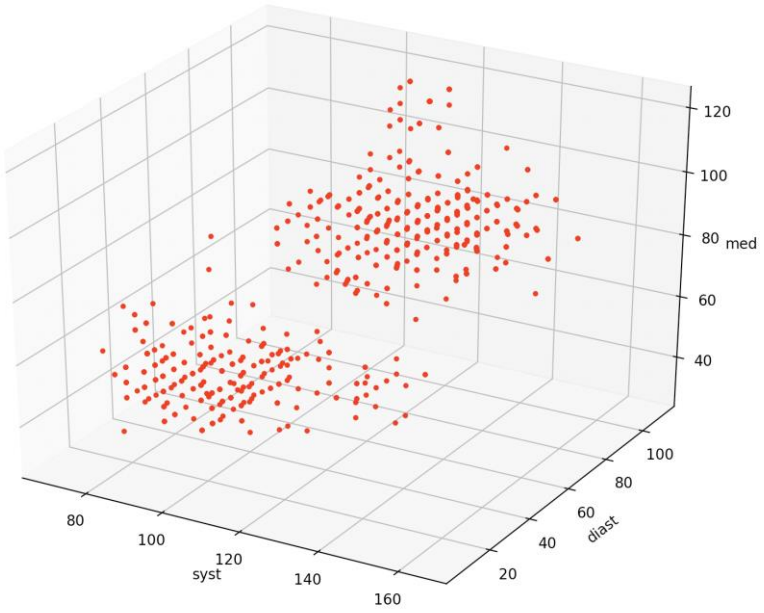


Рисунок 5.9 – Розподіл масиву даних вектору $x(k)=\left(x_1(k), x_2(k),..., x_{n(k)}(k)\right)^T$

Вибірка по кожному пацієнту може бути представлена у формі вектору $x(k)=\left(x_1(k), x_2(k),..., x_{n(k)}(k)\right)^T$. Значення $x_{i(k)}(k)$ спостерігається в момент часу $0 \le t_{i(k)}(k) \le T$. Вектор вибірки $x_{i(k)} \in R^{n(k)}$ та $x(l) \in R^{n(l)}$ при $n(k) \ne n(l)$ неможливо порівняти, тому $\Delta t_{i(k)} \ne const$ та $t_{i(k)} \ne \Delta t_{i(l)}$. Розподіл середнього

значення верхнього та нижнього рівня артеріального тиску хворих для 12 груп відображено на рисунках 5.10 та 5.11 відповідно.

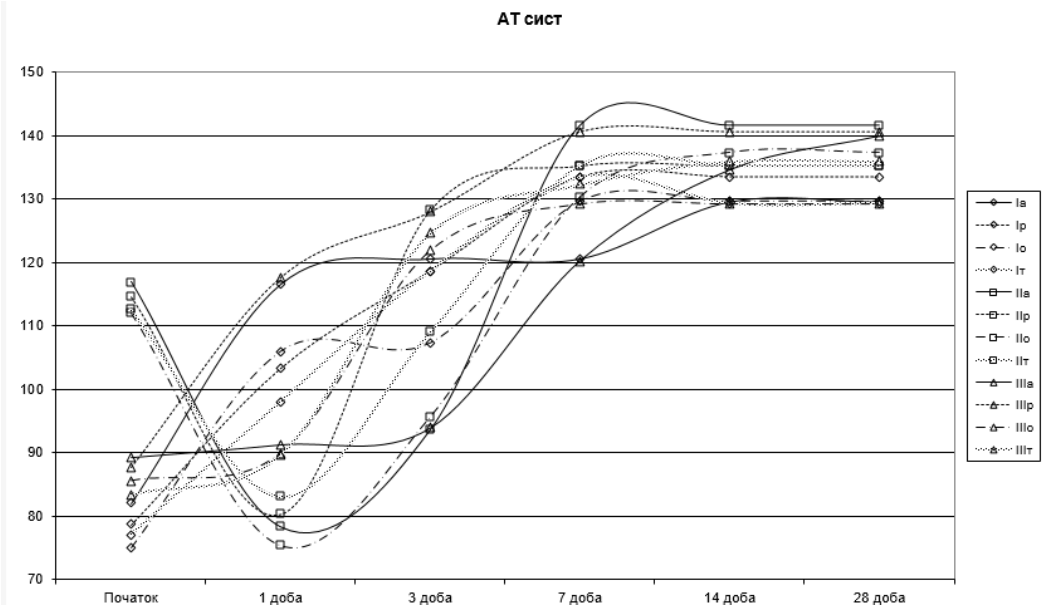


Рисунок 5.10 – Розподіл середнього систоличного значення артеріального тиску

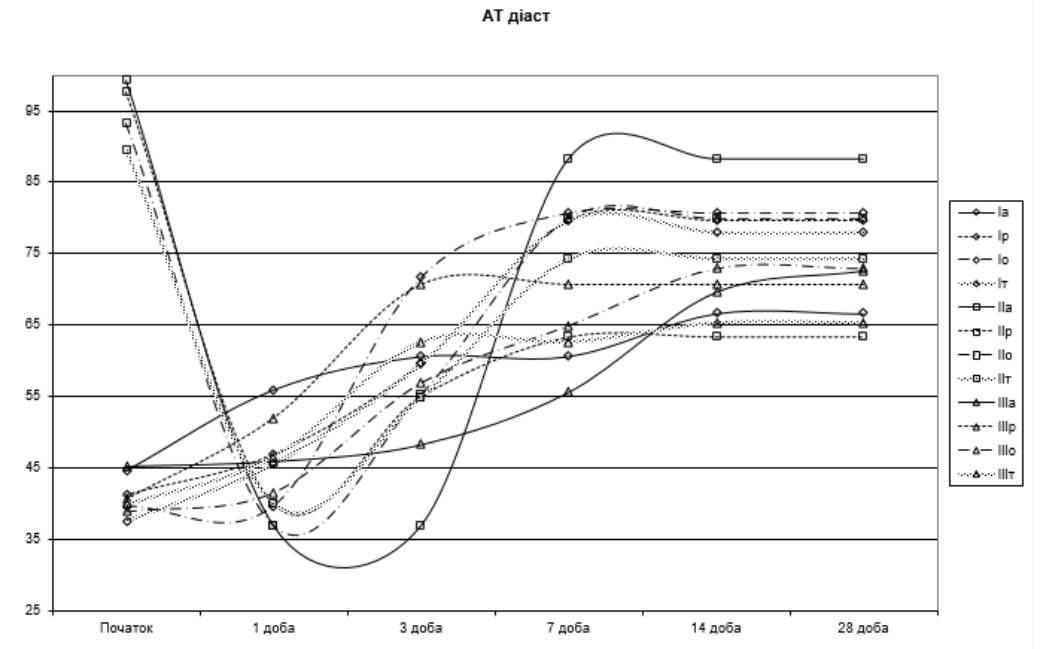


Рисунок 5.11 – Розподіл середнього дистоличного значення артеріального тиску

При цьому інтервал спостереження може бути представлений у вигляді $\left[t_1 = t_{1\min} = \min\left\{t_1(k)\right\} - t_n = T = \max\left\{t_{n(k)}(k)\right\}\right]$. Як результат формується загальна часова шкала у випадку повністю синхронізованих вибірок та $n = \sum_{k=1}^{N} n(k)$ точок, якщо моменти фіксації даних у всіх рядах повністю не збігаються.

Так, отримаємо набір з N векторів–вибірок:

$$\hat{x}(1) = \left(\hat{x}_1(1), \hat{x}_2(1), ..., \hat{x}_i(1), ..., \hat{x}_n(1)\right)^T, ..., \hat{x}(k) =$$
$$= \left(\hat{x}_1(k), ..., \hat{x}_i(k) ..., \hat{x}_n(k)\right)^T, ..., \hat{x}(N) = \left(\hat{x}_1(N), ..., \hat{x}_i(N), ..., \hat{x}_n(N)\right)^T,$$

що мають $n(1) = ... = n(k) = ... = n(6)$ моментів, які мають однакову розмірність$(n \times 1)$, при цьому компонентами цих векторів $\hat{x}_i(k)$ можуть бути як реальні спостереження, так і квазіспостереження типу (5.3) та (5.4).

Використовуючи метрику (3.2) [3], можливо виконати пакетну (офлайн) процедуру нечіткої кластеризації, яка є модифікацією алгоритму нечітких *c*–середніх (FCM). При виконанні виникають кластери сферичної форми, відображених на рисунку 5.12.

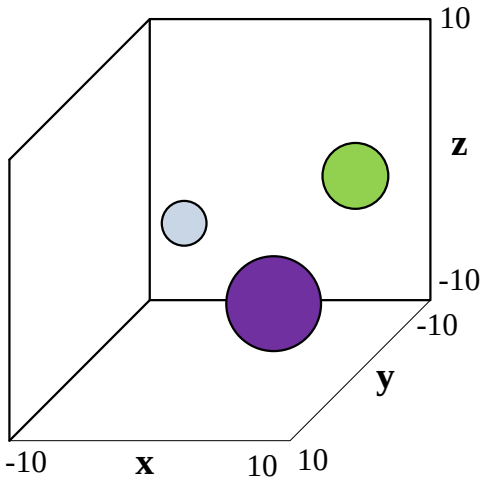


Рисунок 5.12– Форма кластерів при виконанні пакетної (офлайн) процедури нечіткої кластеризації

При реалізації кластеризації алгоритмом *k*-середніх кластери мають еліпсоїдну форму. Слід відзначити, що сферичне розподілення підходить далеко не для всіх задач кластерного аналізу.

Як результат реалізації пакетної (офлайн) процедури нечіткої кластеризації для стадій (I, II, III) захворювання стенокардією виникають сферичні кластери з дванадцяти груп, що наведені на рисунку 5.13.

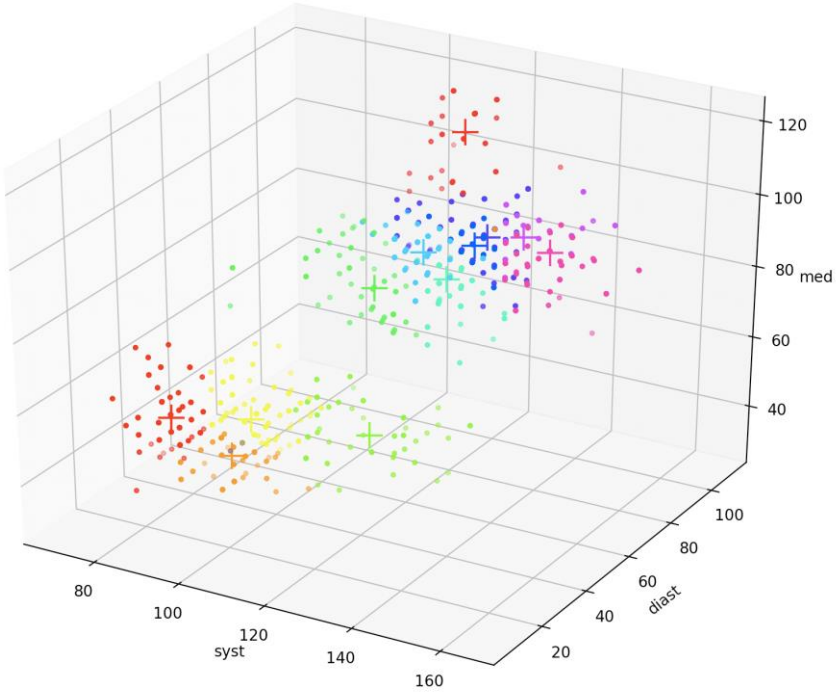


Рисунок 5.13 – Реалізації пакетної (оффлайн) процедури нечіткої кластеризації для стадій (I, II, III) захворювання стенокардією для 12 груп

Вибір оптимального рішення заснований на якості кластеризації, під даним терміном розуміється ступінь наближення результату кластеризації до ідеального рішення, а оскільки ідеальне розв’язок задачі кластеризації невідоме оцінити якість можна експертним або формальним способами.

Експертний вибір найкращого розв’язання задачі полягає в оцінці рішення фахівцями в даній предметній області, але експертна оцінка часто об'єктивно неможлива через великий обсяг та складність даних, тому важливу роль відіграють формальні критерії оцінки якості кластеризації.

Оцінка якості кластеризації, яка вимірює відповідність результатів кластеризації, розглядається як одна з суттєвих проблем, важливих для успіху додатків кластеризації. Основними параметрами даної оцінки є компактність та поділ. Під компактністю розуміється що відстань між елементами кластру повинна бути мінімальною. Дану властивість можна виразити через відстані між елементами кластера, щільність всередині кластера або ж обсяг, займаний кластером в багатовимірному просторі. І навпаки, під властивістю роздільності розуміється максимальна відстань між різними кластерами. Відстань між кластерами зазвичай вимірюється одним із наступних способів:

1) як відстань між найближчими елементами кластерів;

2) як відстань між найбільш віддаленими один від одного елементами кластерів;

3) як відстань між центрами кластерів.

За індекси для тестування були обрані такі: індекс силуета, індекс Девіса-Болдуїна, Calinski-Harabasz індекс.

Розглянемо індекс оцінки силуета (Silhouette index) [112]. Для визначення силуета кожного кластеру припустимо, що елемент x_j належить кластеру c_p, визначимо середню відстань від заданого об'єкта до інших об'єктів з того ж кластера c_p через a_{pj} та визначимо середню відстань від x_j до об'єктів з другого кластеру $c_q, q \neq p$ через a_{pj}. Припустимо, що $b_{pj} = \min_{q \neq p} d_{qj}$ - це значення визначає міру неідентичності окремих елементів з елементами максимально наближеного кластера, таким чином, силует кожного окремого елемента визначається як

$$S_{x_j} = \frac{b_{pj} - a_{pj}}{\max\left(a_{pj}, b_{pj}\right)}.$$

Знаменник введемо для нормалізації, тоді велике значення показника S_{x_j} характеризує покращену приналежність елемента x_j до кластера p. Тоді оцінка для всієї кластерної структури досягається усередненням показника за елементом

$$SWC = \frac{1}{N}\sum_{j=1}^{N} S_{x_j} .$$

Максимальне значення SWC визначає оптимальне розбиття, що досягається найменшою відстанню всередині кластера a_{pj}, а відстань між елементами сусідніх кластерів b_{pj} максимальна.

Далі розглянемо індекс Девіса-Болдуїна [17], так, нехай $S_i = \left\{ \frac{1}{n_{c_i}} \sum_{x \in c_i} \left\| x - v_i \right\|^q \right\}^{\frac{1}{q}}$ –це міра розкиду всередині кластеру c_i та $d_{ij} = \left\{ \sum_{k=1}^{d} \left(v_i^k - v_j^k \right)^p \right\}^{\frac{1}{p}}$ –міра відмінності між кластерами. Тоді мірою подібності між кластерами c_i та c_j може будь яка функція R_{ij}, що відповідає таким умовам:

1. $R_{ij} \geq 0$;
2. $R_{ij} = R_{ji}$;
3. при $S_i = 0, S_j = 0 \quad R_{ij} = 0$;
4. при $S_j > S_k \quad d_{ij} = d_{ik} \quad R_{ij} > R_{ik}$;
5. при $S_i = S_k \quad d_{ij} < d_{ik} \quad R_{ij} > R_{ik}$.

Індекс Девіса–Болдуїна буде обчислюватися за наступним виразом:

$$DB = \frac{1}{c}\sum_{i=1}^{c} R_i ,$$

де $R_i = \max_{i, j \in \{1 \ldots c\}, i \neq j} \left(R_{ij} \right)$.

Тобто індекс Девіса-Болдуїна визначає середню подібність між кластером c_i та найбільш близьким до нього кластером. Оскільки кластери у структурі значно відрізняються один від одного, оптимальною буде структура з мінімальним індексом.

І останнім розглянемо індекс Calinski-Harabasz [108]. Так, припустимо що середній квадрат відстані між елементами у кластеруємій множині – $\bar{d}^2$, а середній квадрат відстані між елементами у кластері c_i – $\bar{d}_{c_i}^2$, тоді сума відстаней всередині груп буде дорівнювати:

$$WGSS = \frac{1}{2}\sum_{i=1}^{c}\left(n_{ci}-1\right)\bar{d}_{c_i}^2 ,$$

де c – кількістькластерів;

n_{c_i} – кількість елементів у кластері;

c_i, d – розмірність множини X,

а сума відстаней між групами:

$$BGSS = \frac{1}{2}\left(\left(c-1\right)\bar{d}^2 + \left(N-c\right)A_c\right),$$

де $A_c = \frac{1}{N-c}\sum_{i=1}^{c}\left(n_{ci}-1\right)\left(\bar{d}^2 - \bar{d}_{c_i}^2\right)$–зважена середня різниця відстані між центрами кластерів та загальним центром множини.

Тоді індекс визначається як

$$VRC = \frac{\frac{BGSS}{c-1}}{\frac{WGSS}{N-c}} = \frac{\bar{d}^2 + \frac{N-c}{c-1}A_c}{\bar{d}^2 - A_c} = \frac{1+\frac{N-c}{c-1}a_c}{1-a_c},$$

де $a_c = \frac{A_c}{\bar{d}^2}$.

Якщо відстані між точками ідентичні, тоді $a_c = 0$ та $VRC = 1$. $a_c = 1$ винятково для безумовної кластеризації, і тоді у кластерах не буде відхилень та похибок. При нормальному розподілі даних a_c повільно але постійно збільшується при збільшенні c, але VRC спадає при постійному a_c та зростаючому c. При цьому зростання a_c балансує у випадку нормального розподілу, а максимальне значення індексу VRC відповідає оптимальній структурі кластерів.

У результаті кластеризації вектору ознак за допомогою *k*–середніх (рис.5.14) отримаємо 12 центрів які практично не формують сферичні кластери. Кластеризація методом Mini Batch *k*-means (рис. 5.15) та Agglomerative Clustering (рис. 5.16) візуально схожі між собою, але відрізняються за суттю.

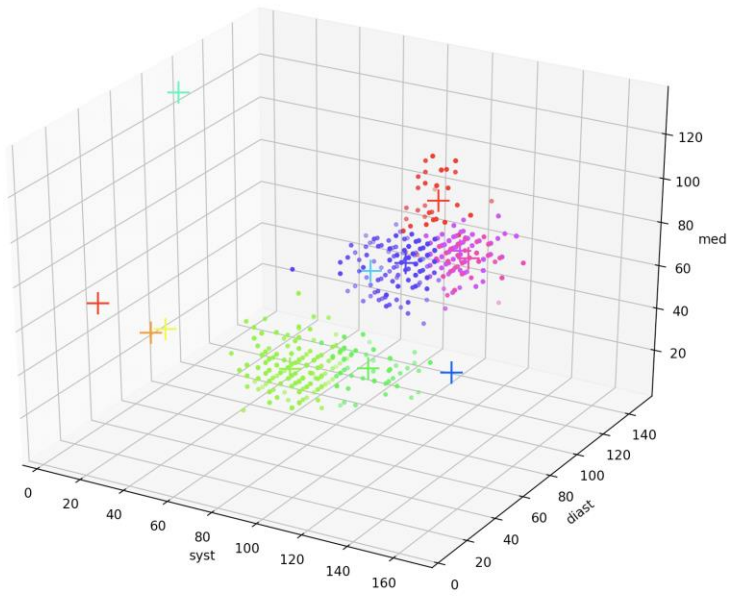


Рисунок 5.14 – Реалізація кластеризації за допомогою *k*–середніх

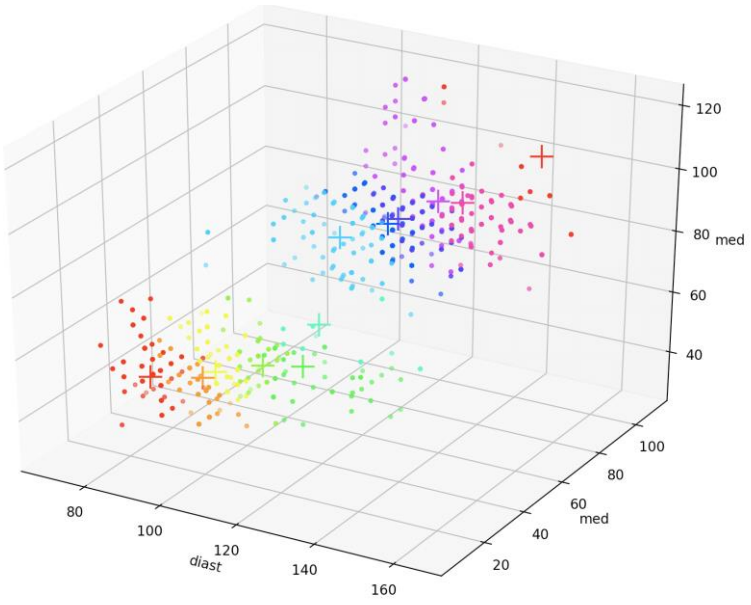


Рисунок 5.15 – Реалізація кластеризації за допомогою Mini Batch k-means

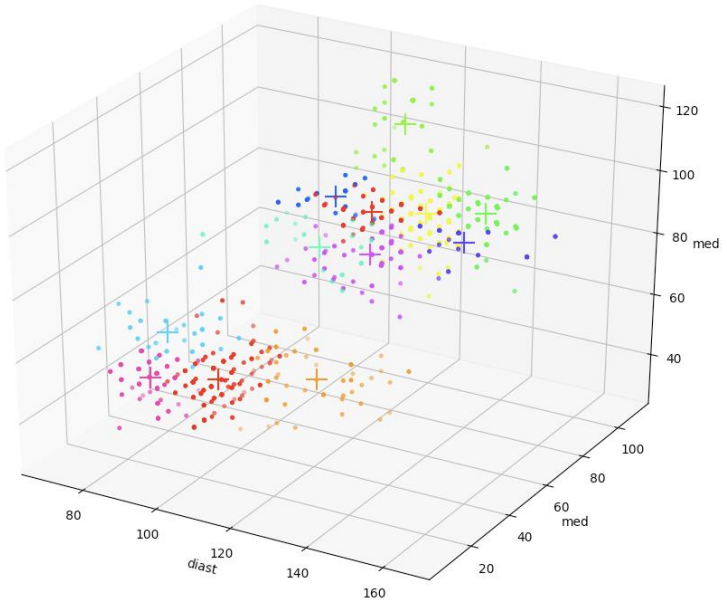


Рисунок 5.16 – Реалізація кластеризації за допомогою Agglomerative Clustering

Результати розрахунку порівнянь оцінок за допомогою індекса силуета, Calinski-Harabasz індекса, індекса Девіса-Болдуїна для алгоритму адаптивної нечіткої ймовірнісної кластеризації, FCM, *k*-середніх, Agglomerative Clustering, Mini Batch методів представлені у таблиці 5.3.

Таблиця 5.3 – Результати розрахунку значень для оцінки якості кластеризації

Індекси\Кластерізація	Індекс силуета	Calinski-Harabasz індекс	Індекс Девіса-Болдуїна
Алгоритм адаптивної нечіткої ймовірнісної кластеризації	0.2326	921.58	1.28
Алгоритм FCM	0.2354	986.39	1.23
Алгоритм *k*-середніх	0.3676	1419.28	1.09
Agglomerative Clustering	0.2790	1117.58	1.08
Mini Batch k-means	0.1904	836.87	1.50

Отже, індекс силуета показує, наскільки середня відстань до об'єктів свого кластера відрізняється від середньої відстані до об'єктів інших кластерів. Дана величина лежить в діапазоні [-1, 1]. Значення, близькі до -1, відповідають «поганим» (розрізненим) типам кластеризації. Значення, близькі до нуля, свідчить про те, що кластери перетинаються і накладаються один на одного. Значення, близькі до 1, відповідають «щільним» чітко виділеним кластерам. Таким чином, чим більше силует, тим чіткіше виділені кластери, і вони є компактними, щільно згрупованими хмарами точок. Як можна побачити з індексу силуета метод відновлення данних працює досить добре. Чим вище значення у індексі Calinski-Harabasz, тим кращим є рішення. У індексі Девіса-Болдуїна значення близькі до нуля вказують на кращий розділ, тобто, як можна побачити, майже при всіх втратах данних розподіл «гарний», отже метод добре відпрацював. Для моніторингу стану пацієнта важливо мати щоденні значення артеріального тиску, тому у кожного пацієнта у різні проміжки часу та діб проводились контрольні заміри тиску. Інформація у формі набору вибірок $x_{i(k)}(k)$ кожного пацієнта має кількість спостережень $N = 6$. Кожна послідовність формує вектор, якій підлягає нечіткій кластеризації. Реалізація нечіткої кластеризації вектору

$x(k)=\left(x_1(k),x_2(k),...,x_{n(k)}(k)\right)^T$ для 12 груп та результати розрахунку оцінки якості, яка містить різну кількість спостережень та відсутнє одне значення наведено на рисунку 5.17 та таблиці 5.4.

Таблиця 5.4 – Результати розрахунку значень для оцінки якості кластеризації, де відсутнє одне значення

Індекси\Кластерізація	Індекс силуета	Calinski-Harabasz індекс	Індекс Девіса-Болдуїна
Алгоритм адаптивної нечіткої ймовірнісної кластеризації	0.2326	921.58	1.28

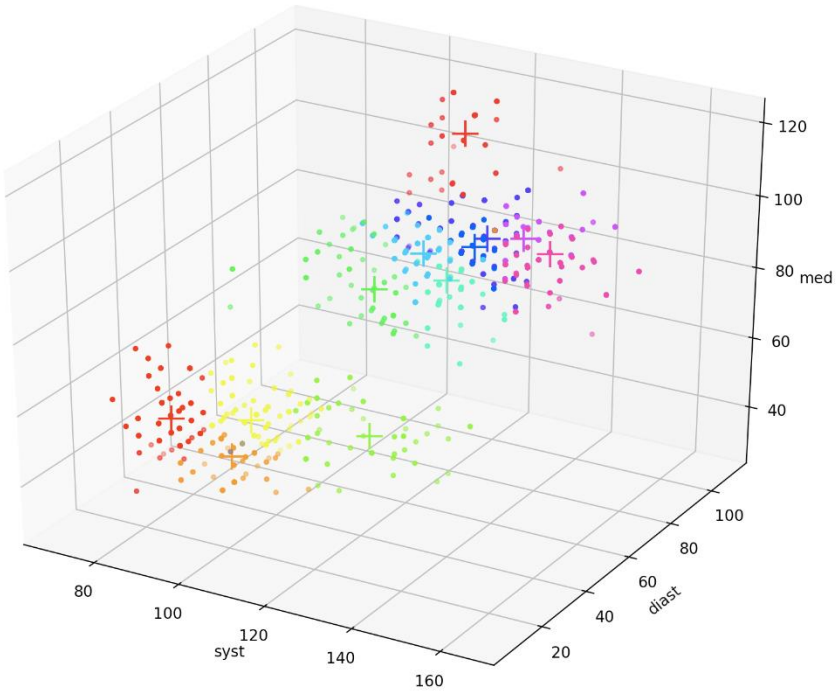


Рисунок 5.17 – Реалізація нечіткої кластеризації для 12 груп у якій відсутне одне значення

Для виявлення аномалій проведено місячний аналіз набору вибірок $x_{i(k)}(k)$ кожного пацієнта, що містить різну кількість спостережень $N=28$ $n(k), k=1,2,\ldots,N$. Реалізація нечіткої кластеризації вектору $x(k)=\left(x_1(k), x_2(k),..., x_{n(k)}(k)\right)^T$ для 12 груп та результати розрахунку оцінки якості, яка містить різну кількість спостережень та відсутнє не більш п'яти значень, наведено на рисунку 5.18 та у таблиці 5.5.

Таблиця 5.5 – Результати розрахунку значень для оцінки якості кластеризації з можливісними втратами не більше 5 значень

Індекси\Кластерізація	Індекс силуета	Calinski-Harabasz індекс	Індекс Девіса-Болдуїна
Алгоритм адаптивної нечіткої ймовірнісної кластеризації	0.2552	945.67	1.25

Однак, є можливість виникнення більшої кількості втрат, у цьому випадку інформація у формі набору вибірок $x_{i(k)}(k)$ кожного пацієнта має різну кількість спостережень $N=28$, $n(k), k=1,2,\ldots,N$. Реалізація нечіткої кластеризації вектору $x(k)=\left(x_1(k), x_2(k),..., x_{n(k)}(k)\right)^T$ для 12 груп та результати розрахунку оцінки якості, яка містить різну кількість спостережень та відсутнє не більш десяти значень, наведено на рисунку 5.17 та у таблиці 5.6.

Таблиця 5.6 – Результати розрахунку значень для оцінки якості кластеризації з можливісними втратами не більше 10 значень

Індекси\Кластерізація	Індекс силуета	Calinski-Harabasz індекс	Індекс Девіса-Болдуїна
Алгоритм адаптивної нечіткої ймовірнісної кластеризації	0.2206	846.46	1.22

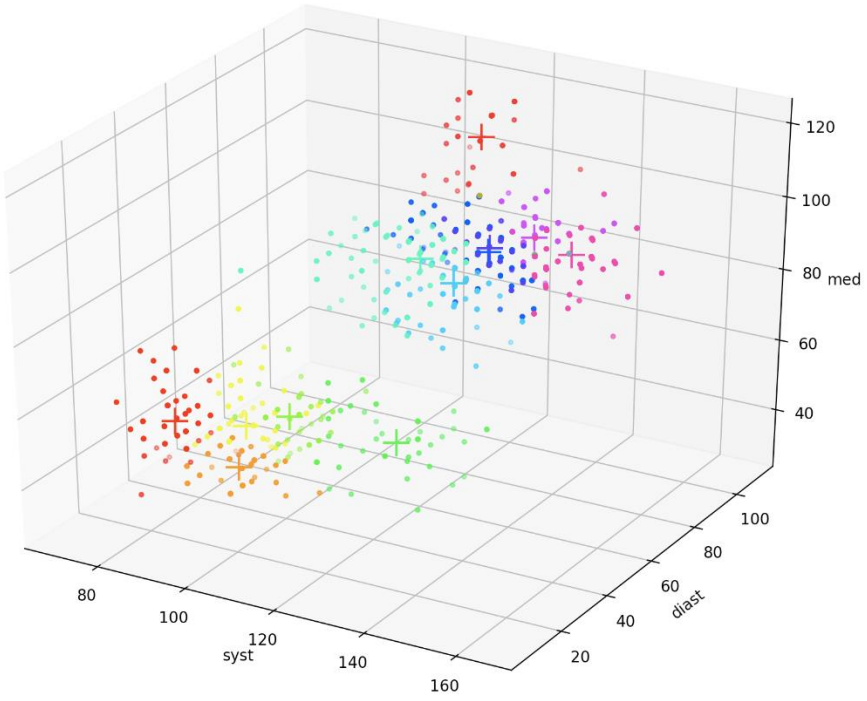


Рисунок 5.18 – Реалізація нечіткої кластеризації для 12 груп, у якій відсутнє більше десяти значень

Порівнюючи результати нечіткої кластеризації вектору $x(k)=\left(x_1(k),x_2(k),...,x_{n(k)}(k)\right)^T$ для 12 груп, де дані являють собою асинхронний багатовимірний ряд, що дорівнює 28 та пропусками значень не більше 5 та 10, можна наочно побачити, як виникають зміни у визначенні центрів кластерів, що пов’язано з формуванням ряду, у залежності від відновлюваних значень. Загалом реалізація нечіткої кластеризації часових рядів з нерівномірними асинхронними тактами квантування проведена вдало, з урахуванням введеної метрики.

Однак серії рядів можуть бути не однотипними–ряди можуть бути схожими за часом, формою та за структурою. У процедурі синхронізації необхідно також враховувати особливості кожної серії рядів. Так, наприклад, відновлення значень ряду з малим часом квантування не дозволяє коректно відновлювати значення та відповідно проводити кластеризацію

багатовимірних рядів. Наочно особливості обробки можна спостерігати на рисунку 5.19.

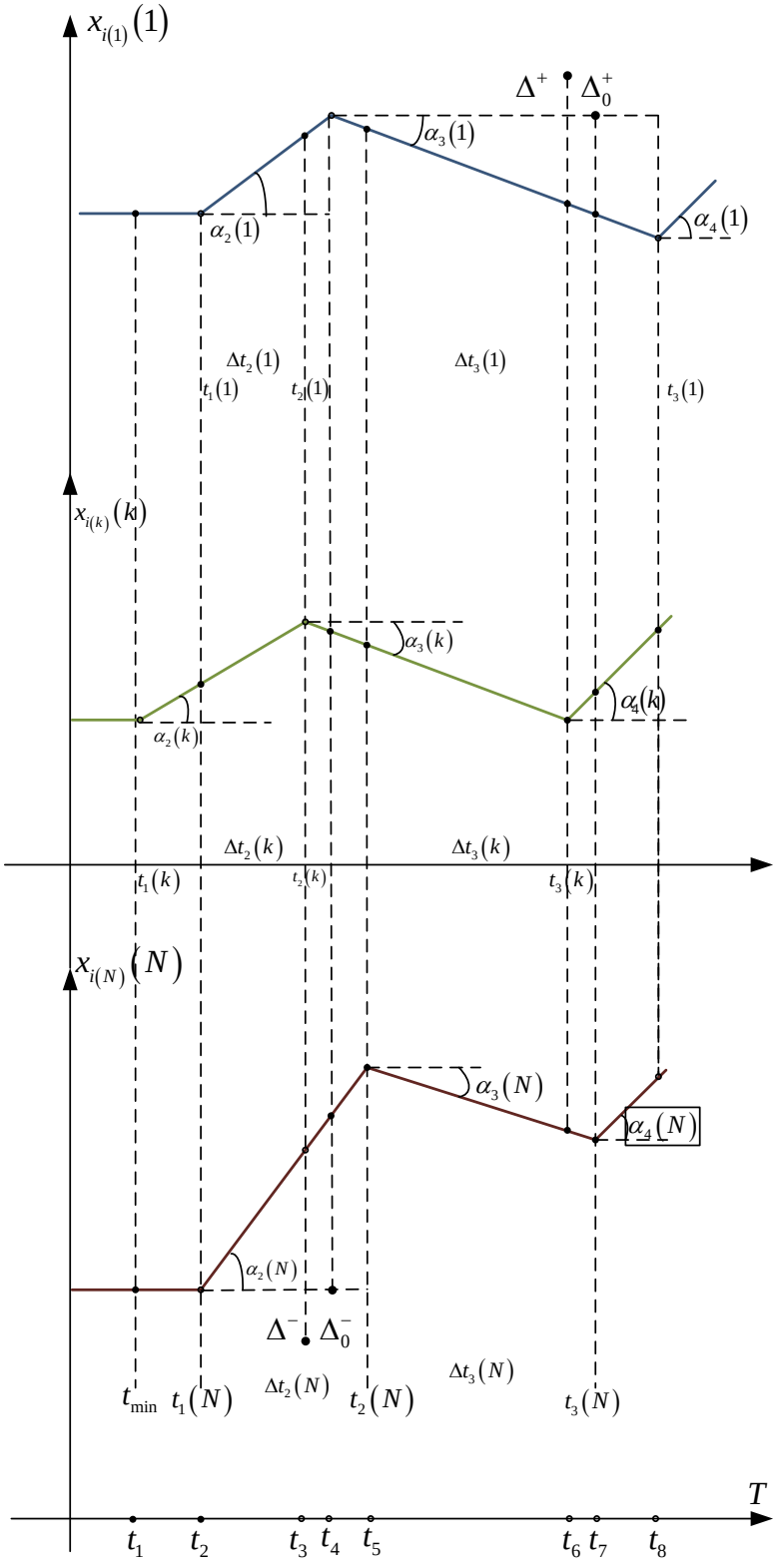


Рисунок 5.19 – Особливості часових рядів з нерівномірними асинхронними тактами квантування з урахуванням Δ^+ та Δ^-

Значення між t_3 и t_4 відносно мале, тому для виконання коректної синхронізації та розрахунку значень необхідно відкидати значення або вводити поріг для часової вісі [108].

Значення кута буде дорівнювати 0 в точці Δ_0^+ або Δ_0^-, це можливо у випадку повторення значень у процедурі синхронізації часових рядів. Тоді вираховуване значення у ряді буде повторюватися. Можлива помилка визначення значень часового асинхронного ряду при виконанні нечіткої кластеризації часових рядів з нерівномірними асинхронними тактами квантування. Тому попередньо для коректності обробки багатовимірних рядів бажано знати параметри та характеристики ряду.

У нашому випадку при аналізі багатовимірного часового ряду медичних даних функціонального стану серцево-судинної системи це практично неможливо [109]. Однак складність полягає в тому, що багатовимірний кластеризуємий часовий ряд може змінювати свої характеристики. Багатовимірний ряд може бути представлений періодично синхронним, аперіодично синхронним, періодично асинхронним та аперіодично асинхронним.

Моделювання виконання нечіткої кластеризації часових рядів з нерівномірними асинхронними тактами квантування при значеннях близьке до Δ_0^+ и Δ_0^- та з можливісними втратами значень не більше 10 наведене у таблицях 5.7 та 5.8 [110].

Як видно з результатів експериментального моделювання з можливісними втратами значень не більше 10 спостережень метод є стійким до пропусків.

У порівнянні із запропонованими методами метод кластеризації даних який містить різну кількість спостережень та з пропусками демонструє задовільні результати [113].

Таблиця 5.7 – Результати кластеризації розрахунку значень при Δ_0^+ для оцінки якості кластеризації з можливісними втратами значень не більше 10

Індекси\Кластерізація	Алгоритм адаптивної нечіткої ймовірнісної кластеризації	Алгоритм FCM	Алгоритм *k*–середніх
Індекс силуета	0.2528	0.2657	0.3255
Calinski-Harabasz індекс	946.1828	989.4909	1553.09
Індекс Девіса-Болдуїна	1.3281	1.3575	1.0771

Таблиця 5.8 – Результати кластеризації розрахунку значень при Δ_0^- для оцінки якості кластеризації з можливісними втратами значень не більше 10

Індекси\Кластерізація	Алгоритм адаптивної нечіткої ймовірнісної кластеризації	Алгоритм FCM	Алгоритм *k*–середніх
Індекс силуета	0.2506	0.2787	0.3344
Calinski-Harabasz індекс	946.1838	989.5209	1499.68
Індекс Девіса-Болдуїна	1.3381	1.3476	1.0931

Для наочності на рисунку 5.20 наведено графічне 2D представлення нечіткої кластеризації для 12 груп, у якій відсутні не більш десяти значень.

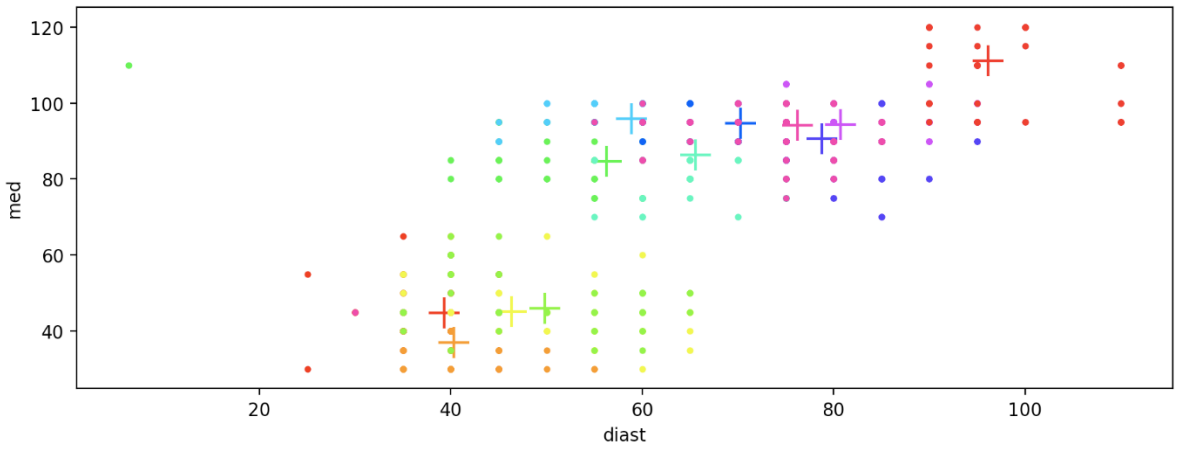


Рисунок 5.19 – Реалізація нечіткої кластеризації для 12 груп у якої відсутнє не більш десяти значень

Моделювання виконання нечіткої кластеризації часових рядів з нерівномірними асинхронними тактами квантування при значеннях близьким к Δ^+ и Δ^- наведено у таблицях 5.9 та 5.10.

Таблиця 5.9 – Результати розрахунку значень при Δ^+ для оцінки якості кластеризації с з можливісними втратами значень не більше 10

Індекси / значення Δ^+	Δ^+
Індекс силуета	0.2708
Calinski-Harabasz індекс	998.1828
Індекс Девіса-Болдуїна	1.4384

Таблиця 5.10 – Результати розрахунку значень при Δ^- для оцінки якості кластеризації с з можливісними втратами значень не більше 10

Індекси / значення Δ^-	Δ^-
Індекс силуета	0.2609
Calinski-Harabasz індекс	996.1828
Індекс Девіса-Болдуїна	1.4181

Як видно з результатів експериментального моделювання з можливісними втратами значень не більше 10 спостережень, при Δ^+ та Δ^- якість кластеризації погіршилась. Погіршення пояснюється тим, що відбувається зниження якості та втрата даних, що тягне за собою зниження якості кластеризації [114]. Графічно це подається достатньо наочно, а саме змінюється розташування центрів та приналежність до кластерів. Тобто, саме зміна якості даних має найбільш негативний вплив на кластеризацію та значення Δ^+ та Δ^- найбільш «небезпечними» для кластеризації. Також вищевказане стосується дослідження кутів нахилу від $30 < tg\alpha < 60$, результати наведено у таблиці 5.11, а $60 < tg\alpha < 90$ та таблиці 5.12.

Таблиця 5.11 – Результати розрахунку значень при Δ^+ для оцінки якості кластеризації з можливісними, без змін, значень не більше 10

Індекси / кут нахилу	$30 < tg\alpha < 60$
Індекс силуета	0.2608
Calinski-Harabasz індекс	996.1928
Індекс Девіса-Болдуїна	1.5181

Таблиця 5.12 – Результати розрахунку значень при Δ^- для оцінки якості кластеризації з можливісними, без змін, значень не більше 10

Індекси / кут нахилу	$60 < tg\alpha < 90$
Індекс силуета	0.2608
Calinski-Harabasz індекс	986.1528
Індекс Девіса-Болдуїна	1.5385

Запропонований метод кластерізації реалізовано на мові програмування Python. Засобами програмних модулів досліджувалися реальні медичні дані [115].

Для експерименту були розглянуті ситуації із втратою даних, адже для діагностики ішемічної хвороби серця важливо бачити повну картину даних. В результаті були отримані відкластеровані дані відповідно до стадій (I, II, III) захворювання стенокардією.

Можливо проводити кластеризацію багатовимірного масиву яким є медичні дані, з використанням різних відстаней і порівнювати результати.

5.4 Застосування методів нечіткої кластеризації часових рядів у моніторингових системах

У теперішній час у медицині існує потреба в інтелектуальному аналізі медичних даних. Сучасні медичні моніторингові системи повинні проводити централізований контроль стану пацієнтів. Моніторинг електрокардіограми (ЕКГ) пацієнтів є складним і не вирішеним завданням.

Сигнал ЕКГ (рис. 5.21) представляє собою часовий ряд, який є нестаціонарним та схильним до численних видів перешкод.

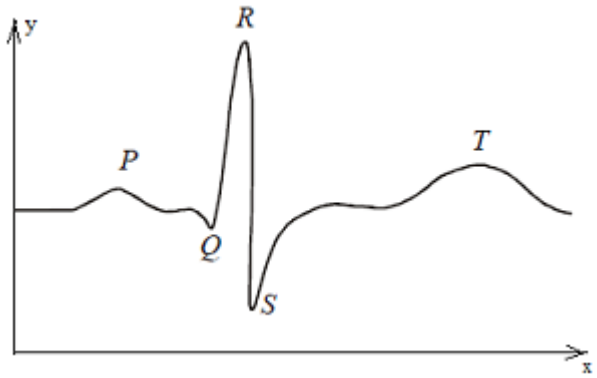


Рисунок 5.21 – Приклад ЕКГ: інтервали P, Q, R, S, T

Найпоширенішим методом тривалого реєстрування поверхневої ЕКГ є метод Холтера або холтерівське моніторування (ХМ). Даний метод дослідження дозволяє проводити безперервну реєстрацію динаміки серця на ЕКГ за допомогою портативного пристрою (холтера), відстежувати зміни в роботі серця і контролювати артеріальний тиск пацієнта протягом доби, в умовах його активності. Завдяки добовому моніторуванню серця можливо виявити або попередити розвиток таких захворювань серцево-судинної системи як аритмія (порушення серцевого ритму); стенокардія; гіпертонія або гіпотонія (підвищення або зниження артеріального тиску); ішемічна хвороба серця, тощо.

Сучасні технології реєстрації поверхневої ЕКГ дозволяють виконувати запис та зберігають сигнали різної тривалості та якості. Прилади що забезпечують даний вид моніторування, відносять до системи амбулаторного моніторування електрокардіограми (АМЕКГ) Такий прилад є портативним автоматичним і призначений для стаціонарного або амбулаторного застосування, забезпечуючи при цьому автоматичну реєстрацію активності серця у будь який час. Даний пристрій має два типи реєстрації:

– зовнішній реєстратор, а саме внутрішньогоспітальне ЕКГ (до 24 годин), стандартний метод холтерівського моніторування (до 48 годин), накладні ЕКГ (до 14 діб), та дистанційна телеметрія (від 30 діб);

– імплантуємий реєстратор, імплантуємі петелькові реєстратори (до 3 років).

Можливість довготривалої реєстрації ЕКГ сигналу дозволяє зберігати клінічні та електрокардіографічні дані як основних параметрів: кількість та тривалість епізодів, середня частота, навантаження, середнє значення, діапазон, тощо, так і додаткових: аналіз варіабельності та турбулентності ритму серця, оцінка довжини інтервалу QT, аналізи наявності пізніх потенціалів передсердя та пневмограми. Також, сучасні прилади мають опцію, що дозволяє пацієнту активізувати систему під час отримання ним будь яких симптомів, тобто фіксація періодів погіршення стану здоров’я дозволяє провести кореляцію між симптомами та фактом наявності або відсутності порушень функцій організму. Таким чином, верифікація даних дозволяє в повному обсязі розглянути доказову базу для постановки вірного діагнозу пацієнту.

Стандартний метод холтерівського моніторування електрокардіограми належить до неінвазивного, другорядного методу дослідження пацієнтів із захворюваннями серцево-судинної системи і полягає у реєстрації поверхневої ЕКГ в умовах вільної активності пацієнта, наприклад, дослідження поточного серцебиття, у нічний час, при фізичному навантаженні, тощо. Стандартна тривалість такого моніторування складає від 24 до 96 годин, по завершенні, сигнал вилучається та надходить на розшифровку, тобто сигнали, фільтровані з придушенням синфазної складової і посилені вхідними підсилювачами сучасних ХМ ЕКГ, надходять в аналого-цифровий перетворювач, де відбувається їх перетворення у цифрову форму. Витяг даних з реєстратора здійснюється або через спеціальний з'єднувальний кабель, або через флеш-карту, або за допомогою

бездротової системи передачі (Wi-Fi, Bluetooth). Комп'ютеризований програмний модуль забезпечує подальшу обробку сигналу з дешифруванням і перетворенням в ЕКГ сигнал, автоматичний аналіз даних з визначенням якісних і кількісних характеристик серцевого ритму, провідності, видачею статистичної звітності та попередніх висновків, акумулює записи пацієнта в базі даних, що дозволяє виконувати ретроспективний аналіз записів ХМ ЕКГ пацієнта. Всі існуючі різноманітні системи реєстрації ХМ ЕКГ дозволяють автоматичну інтерпретацію та візуалізацію отриманих даних.

Внутрішньогоспітальний ЕКГ моніторинг застосовується у випадку коли пацієнт знаходиться у стані загрози життю у відділенні інтенсивної терапії. Моніторинг проводиться у режимі реального часу та безперервно збирає та аналізує наступні дані: артеріальний тиск, температуру тіла, частоту серцевих скорочень та дихання, серцеві викиди, параметри гемодинаміки, рівень сатурації кисню, та дозволяє контролювати фотоплетизмограму, капнограму, насичення крові киснем, а при змінах граничних значень подає аудіовізуальний сигнали тривоги.

Тобто ЕКГ сигнал подається на обробку до центральної та портативної моніторингової станції де він аналізується та забезпечує автоматичну сигналізацію при виникненні порушень ритму, діагностичних змін сегменту та зберігає інформацію у базі даних.

Реєстратори подій (Event Recorders) поділяються на реєстратори з постійним та непостійним записом.

Подієвий ЕКГ реєстратор з постійним записом серцебиття безперервно працює протягом місяця реєструє показання. Також у нього є такі опції як можливість фіксувати раптовий симптом (активізує пацієнт) та фрагмент передсимптомного стану, автоматичної детекції події та безсимптомних епізодів. У сучасних приладах є можливість передавати дані в онлайн режимі до лікарні, що дозволяє співробітникам оперативно реагувати на стан пацієнта.

Подієвий ЕКГ реєстратор з непостійним записом це пристрій який задіється пацієнтом лише у разі виникнення симптому (запис триває не більше 90 секунд). Даний прилад зберігає невелику кількість ЕКГ записів загальною тривалістю 10 хвилин.

Імплантуємі петелькові реєстратори (Loop recorders) це портативний прилад, що на весь період спостережень, кріпиться безпосередньо до пацієнта. Імплантуємий реєстратор використовує власний алгоритм для аналізу аритмічних епізодів, а при їх виникненні подає сигнал та передає дані в онлайн режимі до лікарів спостережного центру.

Мобільні системи амбулаторного серцевого моніторингу (дистанційна телеметрія) передають інформацію у реальному часу до моніторингового центру. Завдяки створенню системи телемедицини з’явилась можливість постійно контролювати стан хворого, дані функції є як у зовнішніх, так і в імплантуємих приладах.

ЕКГ телеметрія забезпечує необхідний моніторинг протягом місяця, при використанні систем CardioNet, Mobile Cardiac Outpatient Telemetry System та HEARTLink II System, які мають вбудований модуль, необхідні дані передаються до кардіоцентру, системи містять реєструємий сигнал (який знаходиться на тілі пацієнта або імплантується у нього) та портативний пристрій, який приймає сигнал з реєструю чого приладу та знаходиться біля хворого. Програмне забезпечення безперервно проводить інтелектуальний аналіз даних (за власним алгоритмом) і у випадку виникнення проблем із серцево–судинною системою передає дані у режимі онлайн до моніторингового центру, у якому зберігається та аналізується вся інформація про стан пацієнта, що дозволяє діагностувати реальні аритмічні події та надати необхідну допомогу.

Підсумковою частиною дослідження є фінальний протокол, завданням якого є подання лікарю-кардіологу максимально інформативного висновку, з

обов’язковим відображенням тих параметрів ритму серця, що здатні вплинути на тактику лікування пацієнта, та прогнозів щодо одужання.

Важливою є фіксація та документування всіх оціночних параметрів дослідження, а саме: таблиць, трендів, зразків нормальної та аномальної ЕКГ, порушення ритму, графіків, цифрових показників що використовують додаткові опції, інтерпретація отриманих даних, порівняння із специфічними нормативними параметрами тощо.

У висновку дослідження представлено аргументовану думку щодо окремих положень протоколу, з відокремленням клінічно значущих параметрів, до яких відносяться наступні групи:

1. Аналіз ЕКГ:

– визначення базового ритму серця (синусовий, миготлива аритмія, ритм електрокардіостимулятору тощо);

– наявність другорядних ритмів, їх характеристика, тривалість, умови виникнення і припинення.

2.Тахікардія:

визначення типу (суправентрикулярная, шлуночкова, блокована, абберантна, вузлова);

– електрофізіологічний механізм;

– кількість та тривалість епізодів;

– частота серцевих скорочень (ЧСС) на піку;

– особливості початку і закінчення (ЧСС, активність, прийом препаратів);

3. Брадіаритмія:

– паузи ритму - можливий електрофізіологічний механізм (синоатріальна, АВ блокада);

– кількість та тривалість епізодів;

– тривалість та циркадність пауз;

– довжина максимальної паузи;

– особливості початку і закінчення (ЧСС, активність, прийом препаратів);

– характер активності і симптомів у момент реєстрації аритмії.

4. Частота серцевих скорочень за даними автоматичного аналізу:

– визначення середньодобової частоти серцевих скорочень;

– max та min частота, з вказанням часу їх виникнення;

– розрахунок циркадного індекса

5. Екстрасистолія:

– визначення типу, щільності та частоти (поодинокі, рідкісні, помірні, часті);

– визначення циркадності - нічний, денний, змішаний;

– визначення характеру - парні, групові, інтерпольовані, періоди бі- або трігеменії;

– визначення морфології - мономорфні, поліморфні.

6. Симптоматика.

До даної групи належать час і характер симптомів та зміни ЕКГ у період їх виникнення

7. Оцінка сегмента ST і зубця Т.

Також протокол може містити інші клінічно-фізіологічні інтерпретації додаткових опцій, якщо ці опції допоможуть встановити правильний діагноз та визначити тактику лікування.

Задача кластеризації ЕКГ у моніторингових системах даних допомагає в обробці та виявленні аномалій у людей для діагностики проблем серцево-судинної системи. Для оцінки сегмента ST і зубця Т було використано базу даних аритмії MIT-BIH Arrhythmia [117], яка використовується при виявленні випадків інфаркту міокарда та для фундаментальних досліджень динаміки серця. База даних аритмії MIT-BIH (рис. 5.22) має набір з більш ніж 4000 довгострокових записів, які були отримані у Бостонській лікарні Beth Israel (медичний центр Beth Israel Deaconess) в період з 1975 по 1979 рік. Дані

записи було отримано від змішаної популяції пацієнтів (близько 60%) та амбулаторних пацієнтів (близько 40%).

Рисунок 5.21 – Приклад ЕКГ бази даних аритмії MIT-BIH

Кластерізація ЕКГ дозволяє виділити чотири кластера:

– NOR – без патологій;

– LBBB (Left Bundle Branch Block) – блокада лівої ніжки пучка Гіса;

– RBBB (Right Bundle Branch Block) – блокада правої ніжки пучка Гіса;

– PVC (Premature Ventricular Contraction) – шлуночкова екстрасистола.

Вхідним файлом для аналізу бази даних аритмії MIT-BIH може бути файл CSV. Структура файлу приймається стандартною для медичних баз даних часових рядів.

Результати кластеризації візуалізуються у вигляді графіків часових рядів, що віднесені до відповідних кластерів, та у вигляді переліку номерів відповідних часових рядів у кожному кластері.

Для реалізації кластерного аналізу часових рядів використаємо медичний набір даних часових послідовностей електрокардіограм (ЕКГ) серцебиття. Даний набір даних складається з колекції сигналів серцебиття, отриманих з бази даних діагностичних ЕКГ MIT-BIH Arrhythmia.

Сигнали відповідають формам електрокардіограми (ЕКГ) серцевих скорочень для нормального випадку і випадків ураження різними аритміями та інфарктом міокарда. Ці сигнали попередньо оброблялися і сегментувалисяся, причому кожен сегмент відповідає одному серцевому удару [116].

Числові характеристики набору даних:

- кількість зразків (часових послідовностей): 87554;
- кількість категорій (класів), визначених при формуванні бази даних ЕКГ: 5;
- кількість часових відліків: 187;
- частота дискретизації: 125 Гц;
- джерело даних: MIT-BIH аритмічний набір даних;
- класи: [A: 0, B: 1, C: 2, D: 3, E: 4].

У клас E були включені всі часові послідовності, які неможливо віднести до класів кластеризації.

Оригінальні аналогові записи були зроблені і відтворювалися з використанням дев'яти двоканальних аналогових магнітофонів, які для зменшення перешкод живилися від постійного струму.

Аналогові сигнали з виходу блоку відтворення були відфільтровані і нормалізовані для обмеження насичення аналого-цифрового перетворювача (АЦП) і для згладжування з використанням фільтру і смугою пропускання від 0,1 до 100 Гц відносно реального часу, що значно виходить за межі найнижчих і найвищих частот записів.

Записи були оцифровані при використанні частоти квантування 360 відліків у секунду на канал з 11-бітною роздільною здатністю в діапазоні ±10 мВ.

Сигнали, пропущені через фільтр, були оцифровані при використанні частоти дискретизації 360 відліків в секунду (360 Гц) на канал з 11-бітною роздільною здатністю в діапазоні ±10 мВ.

Більшість медичних баз даних, зокрема PhysioBank, використовують коди приміток (анотацій), зазначені в таблиці 5.13.

Таблиця 5.13 – Коди анотацій (приклади) серцевих ударів

Код	Опис
N	Звичайний удар (відображається як " · " в базах даних PhysioBank, LightWAVE, pschart і psfd)
L	Білий блок гілок лівого пучка
R	Правий блок розгалужується
A	Передсердя передчасно б'ють
J	Нодальний (сполучний) передчасний удар
V	Передчасне скорочення шлуночків
r	R-на-T передчасне скорочення шлуночків
F	Злиття шлуночків і нормального биття
i	Втеча з передсердя
/	Темп збився
Q	Некласифікований удар
?	Удар не класифікується під час навчання
j	Nodal (junctional) перебіг удару
n	Надшлуночковий бічний потік (атріальний або вузловий)

Частота дискретизації була обрана для полегшення реалізації цифрових фільтрів на 60 Гц (частоти мережі в США) в детекторах аритмії.

Для уточнення інформації про окремі часові відліки використовуються додаткові коди анотацій і коди ритм-анотації, наведені у таблиці 5.14 та 5.15 відповідно.

Таблиця 5.14– Коди додаткових анотацій серцевих ударів

Код	Опис
[	Початок шлуночкового тріпотіння / фібриляції
!	Хвиля шлуночків
]	Кінець шлуночкового тріпотіння / фібриляції
x	Непроведена Р-хвиля (блокований АПК)
(	Формування сигналу
)	Кінець сигналу
p	Пік Р-хвилі
t	Пік Т-хвилі
u	Пік U-хвилі
`	Перехрестя PQ
'	J-точка
^	Артефакт кардіостимулятора
|	Ізольований артефакт типу QRS
~	Зміна якості сигналу
+	Зміна ритму
s	Зміна сегмента ST

Для опису рядів серцевих ударів був визначений стандартний набір кодів анотацій для ЕКГ [116].

Кожен екземпляр анотації може мати до шести атрибутів:

− time - час в межах запису (записується у файлі анотації як номер зразка вибірки, до якого вказує анотація);

− anntyp - числовий код анотації;

− subtyp, chan, num - три малих цілих числа (від -128 до 127), які визначають атрибути, залежні від контексту;

− aux - вільний текстовий рядок.

Анотації можна читати за допомогою додатків C, C++, Python, що використовують функцію getann , і можуть бути записані за допомогою функцій putann , визначених у бібліотеці WFDB - waveform-database package.

Таблиця 5.15 – Коди ритм-анотацій

Код	Опис
(AB	Передсердний бигемин
(AFIB	Миготлива аритмія
(AFL	Тріпотіння передсердь
(B	Шлуночкові великиміни
(BII	2 серцевий блок
(IVR	Ідіовентрикулярний ритм
(N	Нормальний синусовий ритм
(NOD	Нодальний (AV-сполучний) ритм
(P	Темп ритму
(PREX	Попереднє збудження (WPW)
(SBR	Синусова брадикардія
(SVTA	Надшлуночкова тахіаритмія
(T	Шлуночкові тригеміни
(VFL	Шлуночкові тріпотіння
(VT	Шлуночкова тахікардія

Програми Matlab і Octave можуть читати і писати анотації за допомогою m-файлів. Також анотації можна читати програмами мови сценаріїв, що використовують функцію rdann , і вони можуть бути записані з використанням програм wrann , що належать до пакету програм WFDB.

Поле aux анотації звичайно містить URL-посилання (уніфікований локатор ресурсу, у формі http://machine.name/some/data), придатний для переходу до веб-браузера. Якщо це можливо, текст посилання відображається підкресленим і синім кольором.

Анотації посилань можна використовувати для поєднання розширеного тексту, зображень або інших даних з файлом анотацій.

Набір з 87554 часових послідовностей розбивався на 4 кластери. Результати роботи кластеризації програми відображені на рисунку 5.23.

Центри кластерів змінюються у часі, тому лінія кластер-центру виділяється червоним кольором.

Результати кластеризації різними методами зведемо у таблицю 5.16. Результати були віднесені до класів [A: 0, B: 1, C: 2, D: 3]. Клас [E: 4] був вилучений з порівняння згідно його визначення в таблиці як «Некласифікований удар».

Таблиця 5.16 – Порівняння методів кластеризації

Метод	Кластер A	Кластер B	Кластер C	Кластер D	Загальний % співпадіння
FCM-adp	42,59%	36,70%	61,87%	54,25%	47,50%
soft-DTW	38,62%	36,85%	53,00%	53,22%	48,50%
Euclidean k-means	40,67%	32,58%	63,50%	57,97%	48,00%

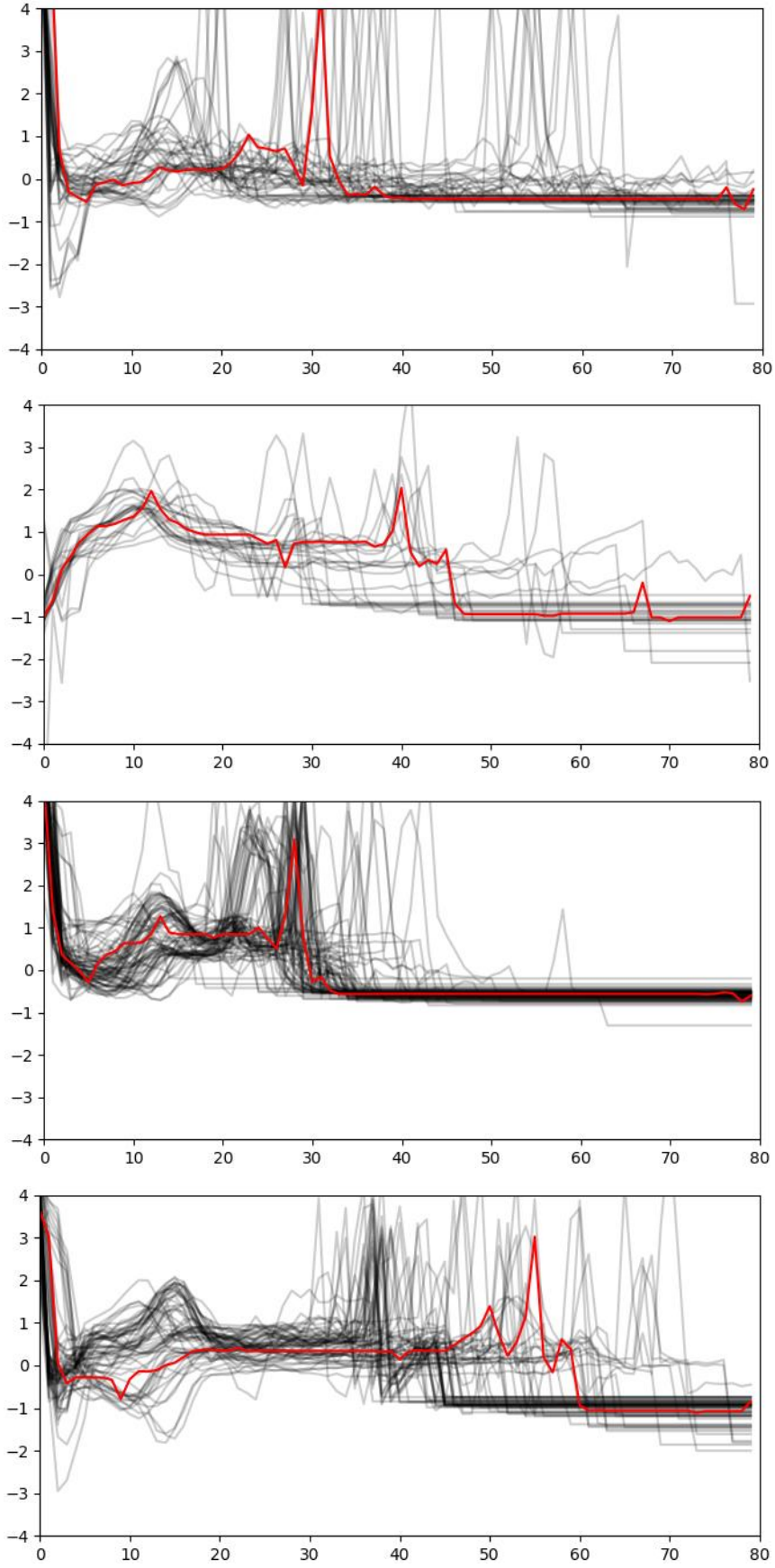


Рисунок 5.23 – Результати кластеризації бази даних аритмії MIT-BIH на 4 кластери

З таблиці видно, що методи кластеризації часових рядів для наданої виборки вхідних даних дають приблизно однаковий результат, тому що медична часова послідовність попередньо ділилася на сегменти серцевого удару. Результати роботи кластеризації програми методом soft-DTW відображено на рисунку 5.24.

Одним з ключових трендів у медицині є інтеграція сучасних методів машинного навчання у сучасні моніторингові системи для класифікації та кластеризації даних в онлайн режимі.

Для створення сучасних моніторингових систем можливо використовувати такі сервіси машинного навчання та інструменти для інтелектуального аналізу даних:

– платформа Google Cloud Machine Learning Engine;

– студія машинного навчання Microsoft Azure;

– Amazon SageMaker.

Платформа Google Cloud Machine Learning Engine дозволяє самостійно створювати та використовувати моделі та методи машинного навчання. Навчання моделей використовує великі обчислювальні ресурси. Для оптимізації обчислювальних ресурсів можливо використовувати дані з різних джерел: Google Cloud Dataflow, Google BigQuery, Google Cloud Dataproc, Google Cloud Storage і Google Cloud Datalab.

Існуючі навчені моделі включають API для створення сучасних хмарних сервісів, таких як Cloud Vision API, Google Translate API, Google Cloud Speech API. Платформа використовує систему Tensorflow з відкритим кодом. Система дозволяє істотно прискорити процес розробки моніторингових систем та розширити їх функціональні можливості для вирішення задач кластеризації даних.

Студія машинного навчання Microsoft Azure призначена для створення, тестування і розгортання рішень для прогнозного аналізу будь яких даних.

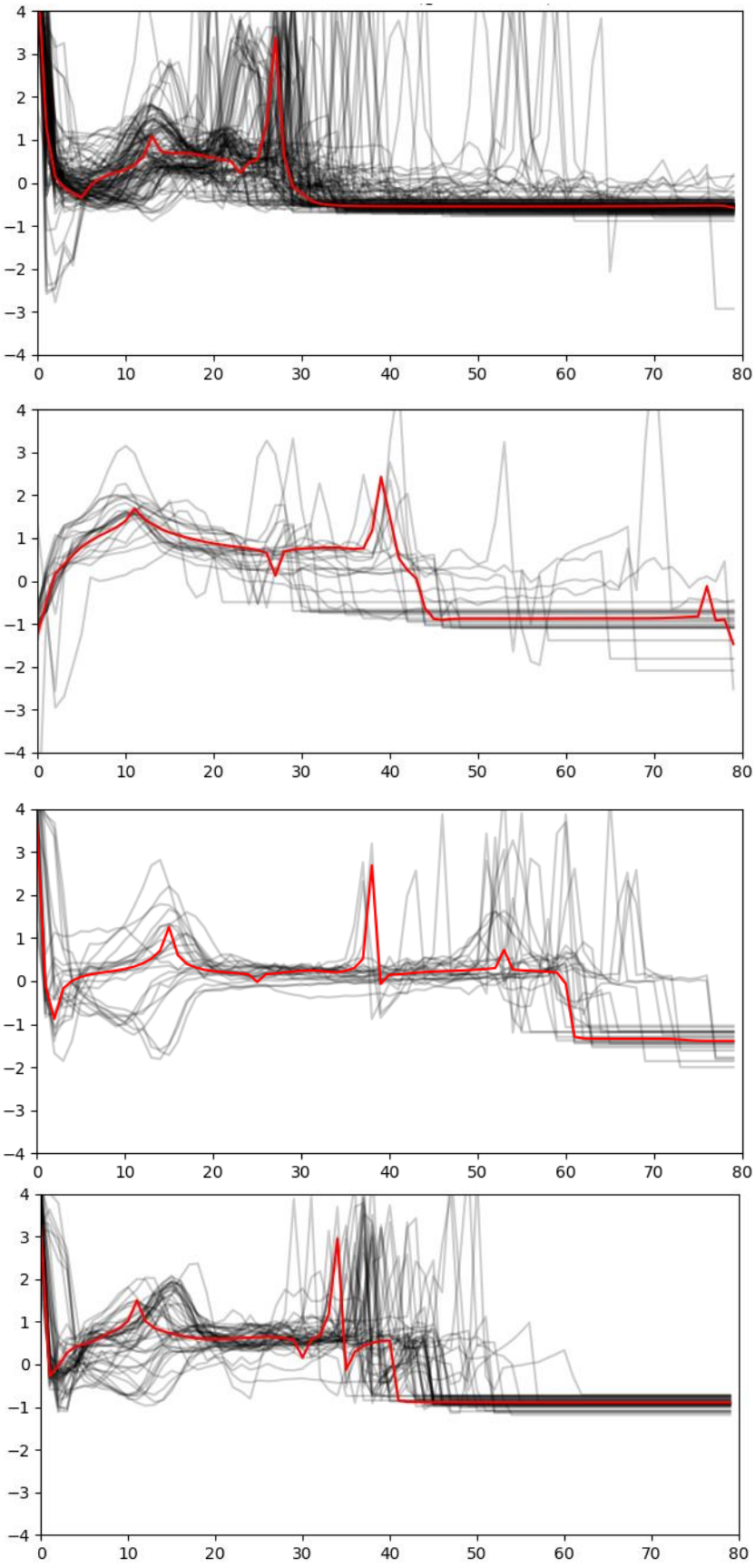


Рисунок 5.24 – Результати кластеризації бази даних аритмії MIT-BIH на 4 кластери методом soft-DTW

Студія машинного навчання Azure надає інтерактивний візуальний робочий простір, що спрощує аналіз даних. На інтерактивному полотні створюється експеримент який дозволяє пов’язати набір даних з модулями аналізу. Надалі експеримент можна опублікувати як веб-службу, щоб модель та методи стали доступні розробникам. Реалізовані моделі та методи як веб-служби можливо використовувати в моніторингових системах та вебзастосунках.

Також Azure дає можливість використовувати власний код на мові R та Python для експериментів. Середовище Python в Azure використовує Anaconda, яке включає в себе пакети Python, в тому числі NumPy, SciPyi scikit-learn.

Amazon SageMaker дозволяє створювати, навчати і розгортати моделі машинного навчання та включає в себе три модулі: модуль збірки, модуль навчання та модуль розгортання. Модуль збірки надає середовище розміщення для роботи з даними та візуалізує результати. Модуль навчання дозволяє навчати і налаштовувати моделі. Модуль розгортання надає керовану середовище, в якому є можливість розміщувати і тестувати моделі.

Amazon SageMaker надає керовані інстанси, які використовують блокноти Jupyter для дослідження і навчання. У ці блокноти завантажені драйвери, пакети Anaconda та бібліотеки для TensorFlow, Apache MXNet, PyTorch, Chainer. Amazon SageMaker автоматично конфігурує і оптимізує TensorFlow, Apache MXNet, Chainer, PyTorch, Scikit-learn і SparkML. Amazon SageMaker можна використовувати з будь-якою платформою для створення моніторингових систем. Для цього потрібно упакувати її в контейнер Docker і зберегти в Amazon EC2 Container Registry.

Висновки до розділу

1. Проведено імітаційне моделювання методів навчання робастних адаптивних моделей часових рядів.

2 Проведено імітаційне моделювання методів адаптивної можливісної нечіткої кластеризації часових рядів.

3. Проведено імітаційне моделювання послідовної онлайн нечіткої кластеризації багатовимірних рядів на базі модифікованої нейро-фаззі мережі Т. Кохонена.

4. Запропоновано процедуру фаззі-кластеризації з асинхронними тактами квантування, що не схильна до концентрації норм і розв’язання практичних задач в рамках концепції інтелектуального аналізу даних.

5. Розв’язано практичну задачу на базі розроблених методів кластеризації для медичних даних у сучасних моніторингових системах.

Список використаних у даному розділі джерел наведено у повному списку використаних джерел під номерами [91–115].

ВИСНОВКИ

У дисертаційній роботі представлені результати, які відповідно до поставленої мети є рішеннями актуального завдання розробки методів нечіткої кластеризації коротких часових рядів в інтелектуальному аналізі потоків даних, що можуть містити аномальні спостереження, які базуються на самонавчанних нейро-фаззі моделях і системах. Отримані результати мають важливе практичне значення для створення систем відновлення та кластеризації даних, які надходять на обробку послідовно, в реальному часі. В ході наукових досліджень отримані такі результати.

1. Проаналізовано стан проблеми кластеризації даних і сформульовано існуючі підходи до її вирішення; було розглянуто основні принципи нечіткої логіки та систем нечіткого розбиття; здійснено аналіз існуючих методів кластеризації; методів їх навчання і самонавчання, що використовуються для вирішення завдань нечіткої кластеризації даних. Показано і доведено, що об'єднання апаратів нейронних мереж і нечіткої логіки може ефективно вирішувати складні завдання, долаючи недоліки кожної з цих технологій в задачах нечіткої кластеризації коротких часових рядів.

2. Вперше запропоновано метод кластеризації, який несхильний до ефекту концентрації норм, що дозволяє вирішувати задачу кластеризації в онлайн режимі за умов перетину класів та асинхронних нерівномірно квантованих часових рядів за рахунок використання спеціальної цільової функції нечіткої кластеризації.

3. Вперше запропоновано послідовний онлайн метод кластеризації багатовимірних часових рядів, що базується на апараті гібридних систем обчислювального інтелекту, який дозволив вирішувати задачу кластеризації даних, які послідовно надходять на обробку з нерівномірними тактами квантування.

4. Отримала подальший розвиток процедура нечіткої кластеризації, що може бути використана для вирішення широкого класу задач, які пов’язані з невизначено структурованими даними великого обсягу, та для інтелектуального аналізу даних. Запропоновано використання WTM - правила самонавчання для нечіткої кластеризації часових рядів та запропоновано зважування координат у просторі ознак.

5. Отримав подальший розвиток метод адаптивної кластеризації, що базується на методах ймовірнісної та можливісної кластеризації коротких часових рядів, які, у свою чергу, засновані на метриці спеціального вигляду, що дозволяє значно спростити чисельну реалізацію методу, за рахунок використання метрики на основі тангенсів кутів нахилу, що на відміну від відомих методів вирішує задачу кластеризації нерівномірно квантованих часових рядів. Відмінною особливістю методики є оцінка якості кожного розбиття і вибір найкращого з них.

6. Отримав подальший розвиток метод робастної адаптивної ідентифікації нестаціонарних часових рядів в онлайн режимі надходження потоку даних, який характеризується простотою обчислювальної реалізації та вирішує задачу обробки даних, що збурені аномальними викидами, за рахунок використання введеної модифікації критерія Гемана-МакКлюра.

7. Отримав подальший розвиток ансамбль гібридних адаптивних моделей ідентифікації, на основі модифікованого робастного критерія Гемана-МакКлюра, який дозволяє уникнути емпіричного вибору тієї чи іншої моделі для нестаціонарного нелінійного сигналу, що забруднений викидами, з невідомим законом розподілу.

Запропоновані методи, орієнтовані на послідовну адаптивну обробку коротких часових рядів забезпечують істотне підвищення якості обробки інформації в умовах її забруднення і спотворення, оскільки всі відомі аналоги орієнтовані на обробку даних в пакетному режимі.

Проведені експериментальні дослідження довели, що запропоновані системи можуть бути успішно використані для вирішення прикладних задач. Так, метод нечіткої кластеризації нерівномірно квантованих асинхронних часових рядів застосовується для вирішення задач обробки медичних даних у сучасних моніторингових системах.

Запропоновані в роботі методи передобробки та обробки часових рядів можуть бути використані в різних областях, де дані представлені в числовій формі у вигляді таблиць «об’єкт-властивість» або часових послідовностей в онлайн режимі. Розглянуті архітектури і методи їх навчання довели свою ефективність при розв'язанні практичних задач моніторингу даних в онлайн режимі у ТОВ «Інфобуд», та під час моніторингу даних та їх кластеризації за допомогою запропонованого методу нечіткої кластеризації у ТОВ «Сайтосс». Усі впровадження підтверджено відповідними актами.

Результати досліджень впроваджено у Харківському національному університеті радіоелектроніки на кафедрі штучного інтелекту в навчальний процес з дисципліни «Нейромережеві методи обчислювального інтелекту».

Отримані теоретичні результати можуть бути використані для інтелектуального аналізу даних і обробки медико-біологічної, технічної, економічної інформації. Середовище проектування: Python, платформи, на яких проводилися дослідження: Microsoft Windows X і macOS Mojave.

ПЕРЕЛІК ПОСИЛАНЬ

1. Бокс, Д., Дженкинс, Г., & Левшин, А. Л. (1974). *Анализ временных рядов: Прогноз и управление. Вып. 2*. Мир.

2. Kay, S. M. (1993). *Fundamentals of statistical signal processing*. Prentice Hall PTR.

3. Hamilton, J. (1994). D. (1994), Time Series Analysis.

4. Айвазян, С. А., Енюков, И. С., & Мешалкин, Л. Д. (1983). *Прикладная статистика: Основы моделирования и первичная обработка данных*. Финансы и статистика.

5. Айвазян, С. А., Енюков, И. С., & Мешалкин, Л. Д. (1985). Прикладная статистика: Исследование зависимостей: Справ. изд. *М.: Финансы и статистика*, *487*.

6. Айвазян, С. А. (1989). *Прикладная статистика: Классификация и снижение размерности: Справочное издание* (Vol. 3). Финансы и статистика.

7. Haykin, S. (1994). *Neural networks: a comprehensive foundation*. Prentice Hall PTR.

8. Abonyi, J., & Feil, B. (2007). *Cluster analysis for data mining and system identification*. Springer Science & Business Media.

9. Borgelt, C. (2006). Prototype-based classification and clustering.

10. Klawonn, F., Kruse, R., & Timm, H. (1997). Fuzzy shell cluster analysis. In *Learning, networks and statistics* (pp. 105-119). Springer, Vienna.

11. Gath, I., & Geva, A. B. (1989). Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7), 773-780.

12. Gustafson, D. E., & Kessel, W. C. (1979, January). Fuzzy clustering with a fuzzy covariance matrix. In *1978 IEEE conference on decision and control including the 17th symposium on adaptive processes* (pp. 761-766). IEEE.

13. Hathaway, R. J., & Bezdek, J. C. (2001). Fuzzy c-means clustering of incomplete data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *31*(5), 735-744.

14. Miyamoto, S., Ichihashi, H., Honda, K., & Ichihashi, H. (2008). *Algorithms for fuzzy clustering* (pp. 1394-1399). Heidelberg: Springer.

15. Li, J., Song, S., Zhang, Y., & Zhou, Z. (2016). Robust k-median and k-means clustering algorithms for incomplete data. *Mathematical Problems in Engineering*, *2016*.

16. Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, *38*(11), 1857-1874.

17. Бодянский, Е. В., & Руденко, О. Г. (2004). Искусственные нейронные сети: архитектуры, обучение, применения. *Харьков: Телетех*, *369*.

18. Poljak, B. T., & Tsypkin, J. Z. (1980). Robust identification. *Automatica*, *16*(1), 53-63.

19. Huber, P. J. (2011). *Robust statistics* (pp. 1248-1251). Springer Berlin Heidelberg.

20. Rey, W. J. (2006). *Robust statistical methods* (Vol. 690). Springer.

21. Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions* (Vol. 196). John Wiley & Sons.

22. Gorshkov, Y., Kokshenev, I., Bodyanskiy, Y., Kolodyazhniy, V., & Shylo, O. (2006, September). Robust recursive fuzzy clustering-based segmentation of biological time series. In *2006 International Symposium on Evolving Fuzzy Systems* (pp. 101-105). IEEE.

23. Ljung, L. (2002). System Identification: Theory for the User Pers. *Peking: Tsinghua University Press and Prentice*.

24. Goodwin, G. C., Ramadge, P. J., & Caines, P. E. (1981). A globally convergent adaptive predictor. *Automatica*, *17*(1), 135-140.

25. Li, S. Z. (2009). *Markov random field modeling in image analysis*. Springer Science & Business Media.

26. Zhang, Z. (1997). Parameter estimation techniques: A tutorial with application to conic fitting. *Image and vision Computing*, *15*(1), 59-76.

27. Lee, C. C., Chiang, Y. C., Shih, C. Y., & Tsai, C. L. (2009). Noisy time series prediction using M-estimator based robust radial basis function neural networks with growing and pruning techniques. *Expert Systems with Applications*, *36*(3), 4717-4724.

28. Davé, R. N., & Krishnapuram, R. (1997). Robust clustering methods: a unified view. *IEEE Transactions on fuzzy systems*, *5*(2), 270-293.

29. Bodyanskiy, Y., Kolodyazhniy, V., & Stephan, A. (2001, October). An adaptive learning algorithm for a neuro-fuzzy network. In *International Conference on Computational Intelligence* (pp. 68-75). Springer, Berlin, Heidelberg.

30. Otto, P., Bodyanskiy, Y., & Kolodyazhniy, V. (2003). A new learning algorithm for a forecasting neuro-fuzzy network. *Integrated Computer-Aided Engineering*, *10*(4), 399-409.

31. Du, W., Inoue, K., & Urahama, K. (2005, August). Robust kernel fuzzy clustering. In *International Conference on Fuzzy Systems and Knowledge Discovery* (pp. 454-461). Springer, Berlin, Heidelberg.

32. Cochocki, A., & Unbehauen, R. (1993). *Neural networks for optimization and signal processing*. John Wiley & Sons, Inc.

33. Graupe, D. (2016). *Deep learning neural networks: Design and case studies*. World Scientific Publishing Company.

34. Цыпкин, Я. З. (1984). *Основы информационной теории идентификации*. Наука. Гл. ред. физ.-мат. лит.

35. Aggarwal, C. C., & Reddy, C. K. (2014). Data clustering. *Algorithms and Application, Boca Raton: CRC Press*.

36. Aggarwal, C. C. (2015). *Data mining: the textbook*. Springer.

37. Möller-Levet, C. S., Klawonn, F., Cho, K. H., & Wolkenhauer, O. (2003, August). Fuzzy clustering of short time-series and unevenly distributed sampling points. In *International Symposium on Intelligent Data Analysis* (pp. 330-340). Springer, Berlin, Heidelberg.

38. Cruz, L. P., Vieira, S. M., & Vinga, S. (2015, September). Fuzzy clustering for incomplete short time series data. In *Portuguese Conference on Artificial Intelligence* (pp. 353-359). Springer, Cham.

39. Evers, F. T., Höppner, F., Klawonn, F., Kruse, R., & Runkler, T. (1999). *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*. John Wiley & Sons.

40. Bezdek, J. C., Keller, J., Krisnapuram, R., & Pal, N. (1999). *Fuzzy models and algorithms for pattern recognition and image processing* (Vol. 4). Springer Science & Business Media.

41. Bifet, A. (2010, July). Adaptive stream mining: Pattern learning and mining from evolving data streams. In *Proceedings of the 2010 conference on adaptive stream mining: Pattern learning and mining from evolving data streams* (pp. 1-212). Ios Press.

42. Tsoukalas, L. H., & Uhrig, R. E. (1996). *Fuzzy and neural approaches in engineering*. John Wiley & Sons, Inc.

43. Kohonen, T. (2012). *Self-organization and associative memory*(Vol. 8). Springer Science & Business Media.

44. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436.

45. Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1, No. 10). New York: Springer series in statistics.

46. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, *61*, 85-117.

47. Havens, T. C., Bezdek, J. C., & Palaniswami, M. (2012). Incremental kernel fuzzy c-means. In *Computational Intelligence*(pp. 3-18). Springer, Berlin, Heidelberg.

48. Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

49. Arrow, K. J., & Hurwicz, L. (1958). Gradient method for concave programming, I: Local results. *Studies in Linear and Nonlinear Programming. Stanford University Press, Stanford, CA*, *31*, 322-338.

50. Chung, F. L., & Lee, T. (1994). Fuzzy competitive learning. *Neural Networks*, *7*(3), 539-551.

51. Zadeh, L. A. (2015). Fuzzy logic—a personal perspective. *Fuzzy sets and systems*, *281*, 4-20.

52. Zadeh, L. A. (1965). Fuzzy sets. *Information and control*, *8*(3), 338-353.

53. Chung, F., & Rhee, H. (2007). Uncertain fuzzy clustering: Insights and recommendations. *IEEE Computational Intelligence Magazine*, *2*(1), 44-56.

54. Bezdek, J. C. (2013). *Pattern recognition with fuzzy objective function algorithms*. Springer Science & Business Media.

55. Mendel, J. M. (2007). Type-2 fuzzy sets and systems: an overview. *IEEE computational intelligence magazine*, *2*(1), 20-29.

56. Zarandi, M. H. F., Zarinbal, M., & Türksen, I. B. (2009, July). Type-II Fuzzy Possibilistic C-Mean Clustering. In *IFSA/EUSFLAT Conf.*(pp. 30-35).

57. Bodyanskiy, Y., Kolodyazhniy, V., & Stephan, A. (2002, September). Recursive fuzzy clustering algorithms. In *Proc. 10th East West Fuzzy Colloquium* (pp. 276-283).

58. Klawonn, F., & Kruse, R. (1997). Constructing a fuzzy controller from data. *Fuzzy sets and systems*, *85*(2), 177-193.

59. Prieto, A., Prieto, B., Ortigosa, E. M., Ros, E., Pelayo, F., Ortega, J., & Rojas, I. (2016). Neural networks: An overview of early research, current frameworks and new challenges. *Neurocomputing*, *214*, 242-268.

60. Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1.

61. Schilling, R. J., Carroll, J. J., & Al-Ajlouni, A. F. (2001). Approximation of nonlinear systems with radial basis function neural networks. *IEEE Transactions on neural networks*, *12*(1), 1-15.

62. Nelles, O. (2013). *Nonlinear system identification: from classical approaches to neural networks and fuzzy models*. Springer Science & Business Media.

63. Specht, D. F. (1991). A general regression neural network. *IEEE transactions on neural networks*, *2*(6), 568-576.

64. Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford university press.

65. Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, *33*(3), 1065-1076.

66. Gan, G., Ma, C., & Wu, J. (2007). *Data clustering: theory, algorithms, and applications* (Vol. 20). Siam.

67. Du, K. L., & Swamy, M. N. (2013). *Neural networks and statistical learning*. Springer Science & Business Media.

68. Rosenblatt, F. (1962). Principles of Neurodynamics, Washington, D. *DC: Spartan Books*.

69. Bodyanskiy, Y., Pliss, I., & Vynokurova, O. (2007). A learning algorithm for forecasting adaptive wavelet-neuro-fuzzy network. In *Fifth International Conference INFORMATION RESEARCH AND APPLICATIONS* (p. 211).

70. Rutkowski, L. (2008). *Computational intelligence: methods and techniques*. Springer Science & Business Media.

71. Kruse, R., Borgelt, C., Braune, C., Mostaghim, S., & Steinbrecher, M. (2016). *Computational intelligence: a methodological introduction*. Springer.

72. Gorshkov, Y., Kolodyazhniy, V., & Bodyanskiy, Y. (2009, June). New recursive learning algorithms for fuzzy Kohonen clustering network. In *Proc. 17th Int. Workshop on Nonlinear Dynamics of Electronic Systems* (pp. 58-61).

73. Bodyanskiy, Y., Kolchygin, B., & Pliss, I. (2011). Adaptive neuro-fuzzy Kohonen network with variable fuzzifier. *Inform. Theories and Appl*, *18*(3), 215.

74. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning: prediction, inference and data mining. *Springer-Verlag, New York*.

75. Bodyanskiy, Y., Gorshkov, Y., Kokshenev, I., & Kolodyazhniy, V. (2010). Evolving Fuzzy Classification of Nonstationary Time Series. *Evolving Intelligent Systems*, 301.

76. Субботін, С. О., & Субботин, С. А. (2008). Подання й обробка знань у системах штучного інтелекту та підтримки прийняття рішень.

77. Bodyanskiy, Y., Skuratov, M., & Volkova, V. (2012, September). Adaptive matrix fuzzy c-means clustering. In *Proc. 19th East-West Fuzzy-Colloquium* (pp. 111-116).

78. Bodyanskiy, Y., Volkova, V., & Skuratov, M. (2011). Matrix neuro-fuzzy self-organizing clustering network. *Scientific Journal of Riga Technical University. Computer Sciences*, *45*(1), 54-58.

79. Bodyanskiy, Y. V., Boiko, O. O., & Pliss, I. P. (2015). Adaptive Method of Hybrid Learning for an Evolving Neuro-Fuzzy System. *Cybernetics and Systems Analysis*, *51*(4), 500-505.

80. Bodyanskiy, Y., Gorshkov, Y., Kokshenev, I., Kolodyazhniy, V., & Shilo, O. (2006). Recursive fuzzy clustering algorithm for segmentation of biomedical time series. In *East West Fuzzy Colloquium, Zittau-Görlitz: HS* (pp. 130-139).

81. Bodyanskiy, Y., Otto, P., Pliss, I., & Teslenko, N. (2007). Nonlinear process identification and modeling using general regression neuro-fuzzy network.

In *Proc. 52-nd Int. Sci. Colloquium "Computer Science Meets Automation"* (p. 27).

82. Kolchygin, B. V., & Bodyanskiy, Y. V. (2013). Adaptive fuzzy clustering with a variable fuzzifier. *Cybernetics and Systems Analysis*, *49*(3), 366-374.

83. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

84. Klawonn, F., & Höppner, F. (2003, August). What is fuzzy about fuzzy clustering? Understanding and improving the concept of the fuzzifier. In *International symposium on intelligent data analysis*(pp. 254-264). Springer, Berlin, Heidelberg.

85. Klawonn, F. (2013, November). What can Fuzzy cluster analysis contribute to clustering of high-dimensional data?. In *International Workshop on Fuzzy Logic and Applications* (pp. 1-14). Springer, Cham.

86. Keller, A., & Klawonn, F. (2000). Fuzzy clustering with weighting of data variables. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, *8*(06), 735-746.

87. Höppner, F., & Klawonn, F. (2013). *Fuzzy-Clusteranalyse: Verfahren für die Bilderkennung, Klassifizierung und Datenanalyse*. Springer-Verlag.

88. Bodyanskiy, Y. (2005). Computational Intelligence Techniques for Data Analysis. In *Leipziger Informatik-Tage* (pp. 15-36).

89. Song, Q., & Kasabov, N. (2001). ECM-A novel on-line, evolving clustering method and its applications. *Foundations of cognitive science*, 631-682.

90. Chen, X. W., & Lin, X. (2014). Big data deep learning: challenges and perspectives. *IEEE access*, *2*, 514-525.

91. Filippone, M., Camastra, F., Masulli, F., & Rovetta, S. (2008). A survey of kernel and spectral methods for clustering. *Pattern recognition*, *41*(1), 176-190.

92. Gorokhovatskiy, V. A., Gorokhovatskiy, A. V., & Peredrii, E. O. (2017). Vector Quantization, Learning and Recognition in the Space of Descriptors of

Structural Features of Images. *Telecommunications and Radio Engineering, 76*(19).

93. Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis* (Vol. 344). John Wiley & Sons.

94. Lughofer, E. (2011). *Evolving fuzzy systems-methodologies, advanced concepts and applications* (Vol. 53). Berlin: Springer.

95. Prechelt, L. (1997). Investigation of the CasCor family of learning algorithms. *Neural Networks*, *10*(5), 885-896.

96. Watts, M. J. (2009). A decade of Kasabov's evolving connectionist systems: a review. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *39*(3), 253-269.

97. Narendra, K. S., & Parthasarathy, K. (1990). Identification and control of dynamical systems using neural networks. *IEEE Transactions on neural networks*, *1*(1), 4-27.

98. Gorokhovatskiy, V. A., Gorokhovatskiy, A. V., & Berestovsky, A. Y. (2016). Intellectual Data Processing and Self-Organization of Structural Features at Recognition of Visual Objects. *Telecommunications and Radio Engineering*, *75*(2).

99. Lee, T. T., & Jeng, J. T. (1998). The Chebyshev-polynomials-based unified model neural networks for function approximation. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *28*(6), 925-935.

100. Olson, D. L., & Delen, D. (2008). *Advanced data mining techniques*. Springer Science & Business Media.

101. Veenman, C. J., & Reinders, M. J. (2005). The nearest subclass classifier: A compromise between the nearest mean and nearest neighbor classifier. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(9), 1417-1429.

102. Yin, Y., Kaku, I., Tang, J., & Zhu, J. (2011). *Data mining: Concepts, methods and applications in management and engineering design*. Springer Science & Business Media.

103. Гороховатський, О. В. (2016). Особливості розпізнавання зображень символів із використанням лінійних описів та корекції результатів. *Системи обробки інформації*, (4), 149-151.

104. De Oliveira, J. V., & Pedrycz, W. (Eds.). (2007). *Advances in fuzzy clustering and its applications*. John Wiley & Sons.

105. Suresh, S., Sundararajan, N., & Savitha, R. (2013). *Supervised learning with complex-valued neural networks* (pp. 125-132). Berlin: Springer.

106. Babinec, Š., & Pospíchal, J. (2008, November). Gating echo state neural networks for time series forecasting. In *International Conference on Neural Information Processing* (pp. 200-207). Springer, Berlin, Heidelberg.

107. Schalkoff, R. J. (1997). *Artificial neural networks* (Vol. 1). New York: McGraw-Hill.

108. Rojas, R. (2013). *Neural networks: a systematic introduction*. Springer Science & Business Media.

109. Moody, J., & Darken, C. J. (1989). Fast learning in networks of locally-tuned processing units. *Neural computation*, *1*(2), 281-294.

110. Zahirniak, D. R., Chapman, R., Rogers, S. K., Suter, B. W., Kabrisky, M., & Pyati, V. (1990). Pattern recognition using radial basis function network. In *Proc. 6th Ann. Aerospace Application of AI Conf., Dayton, OH* (pp. 249-260).

111. Nadaraya, E. A. (1965). On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, *10*(1), 186-190.

112. Takagi, T., & Sugeno, M. (1993). Fuzzy identification of systems and its applications to modeling and control. In *Readings in Fuzzy Sets for Intelligent Systems* (pp. 387-403). Morgan Kaufmann.

113. Пришляк, М. Ю., Субботин, С. А., & Олейник, А. А. (2018). Анализ глубоких моделей нейронных сетей на базе ограниченных машин больцмана.

114. Jang, J. S., & Sun, C. T. (1995). Neuro-fuzzy modeling and control. *Proceedings of the IEEE*, *83*(3), 378-406.

115. Mitsa, T. (2010). *Temporal data mining*. Chapman and Hall/CRC.

116. PhysioBank Annotations – PhysioNet. Updated Wednesday, 6 July 2016 at 13:41 EDT. https://www.physionet.org/physiobank/annotations.shtml

117. Moody GB, Mark RG. The impact of the MIT-BIH Arrhythmia Database. IEEE Eng in Med and Biol 20(3):45-50 (May-June 2001). (PMID: 11446209).

Додаток А Акти впровадження

ЗАТВЕРДЖУЮ
Директор ТОВ «Інфобуд»
_____Погорєлов І.М.
«__» _____ 2018 р.

Акт про впровадження результатів дисертаційної роботи на здобуття наукового ступеня кандидата технічних наук
КОБИЛІНА ІЛЛІ ОЛЕГОВИЧА

Комісія у складі:

Голова	Гаєвський А.О. - заступник директора з розробки програмного забезпечення;
Члени комісії	Матат О.О. - начальник відділу контролю якості програмного забезпечення;
	Кравченко С.Ю. - керівник проектів.

склала даний акт про те, що у ТОВ «Інфобуд» був застосований онлайн метод кластеризації багатовимірних часових рядів, що базується на апараті гібридних систем обчислювального інтелекту, які надходять на обробку в онлайн режимі. Реалізований модуль із запропонованим методом підтвердив свою ефективність у задачах моніторингу медичних показників, зокрема артеріального тиску.

Результати впровадження довели, що розроблені Кобиліним І.О. методи, які ґрунтуються на сучасних інтелектуальних технологіях, мають переваги над існуючими підходами.

Акт складений для пред'явлення до спеціалізованої вченої ради із захисту дисертацій і не є підставою для фінансових розрахунків.

Голова комісії: _____ А.О. Гаєвський

Члени комісії: _____ О.О. Матат

_____ С. Ю. Кравченко

ЗАТВЕРДЖУЮ
Директор ТОВ «САЙТОСС»
_____Луців В.В
«08» 10 2018 р.

Акт про впровадження результатів дисертаційної роботи на здобуття наукового ступеня кандидата технічних наук
КОБИЛІНА ІЛЛІ ОЛЕГОВИЧА

Комісія у складі:
Голова — Луців В.В. - директор;

Члени комісії — Харкевич О.М. - менеджер;
Матікайнен Т.О. - менеджер.

склала даний акт про те, що у ТОВ «САЙТОСС» був застосований модуль, у якому реалізовано метод нечіткої кластеризації, несхильний до ефекту концентрації норм за умов перетину класів в онлайн режимі. Запропонований Кобиліним І.О. модуль орієнтовано на онлайн процедуру нечіткої кластеризації і дає змогу вирішувати завдання аналізу несинхронізованих даних. Запропонований модуль дозволяє ефективно виявляти аномалії у хворих після аортокоронарного шунтування у режимі реального часу.

Результати впровадження довели, що розроблені Кобиліним І.О. методи, мають переваги над існуючими підходами в інтелектуальному аналізі потоків даних.

Голова комісії: _____ В.В Луців

Члени комісії: _____ О.М. Харкевич

_____ Т.О. Матікайнен

Затверджую
проректор з науково-методичної
роботи ХНУРЕ
_____ проф. Рубан І. В.
«04» ____06_____ 2018 р.

АКТ

про впровадження в навчальний процес результатів дисертаційної роботи на здобуття наукового ступеня кандидата технічних наук «Нечітка кластеризація часових рядів в інтелектуальному аналізі потоків даних»
аспіранта кафедри штучного інтелекту
Харківського національного університету радіоелектроніки
Кобиліна Іллі Олеговича

Комісія у складі декана факультету комп'ютерних наук, д. т. н., проф. Єрохіна А. Л., завідувача кафедри штучного інтелекту, д. т. н., проф. Філатова В. О., проф. каф. штучного інтелекту, к. т. н., доц. Рябової Н. В. підтверджує, що результати дисертаційної роботи Кобиліна І. О., що пов'язані із розробкою нечіткої кластеризації часових рядів в інтелектуальному аналізі потоків даних, впроваджені в навчальний процес на кафедрі штучного інтелекту в курсі «Нейромережеві методі обчислювального інтелекту».

Декан факультету КН, д.т.н., проф. _____А.Л. Єрохін

Завідувач кафедри ШІ, д.т.н., проф. _____В.О. Філатов

Професор кафедри ШІ, к.т.н., доц. _____Н.В. Рябова

Додаток Б Список опублікованих праць за темою дисертації

1. Setlak, G., Bodyanskiy, Y., Pliss, I., Vynokurova, O., Peleshko, D., & Kobylin, I. (2017). Adaptive Fuzzy Clustering of Multivariate Short Time Series with Unevenly Distributed Observations Based on Matrix Neuro-Fuzzy Self-Organizing Network. In *Advances in Fuzzy Logic and Technology 2017* (pp. 308-315). Springer, Cham. (Входить до міжнародної наукометричної бази SCOPUS).

2. Bodyanskiy, Y., Vynokurova, O., Kobylin, I., & Kobylin, O. (2016). Adaptive Fuzzy Clustering of Short Time Series with Unevenly Distributed Observations in Data Stream Mining Tasks. *Information Technology and Management Science*, *19*(1), 23-28. (Входить до наукометричної бази SCOPUS).

3. Бодянский, Е. В., Винокурова, Е. А., Кобылин, И. О., & Мулеса, П. П. (2016). Робастная адаптивная идентификация нестационарных временных рядов с помощью ансамбля обучаемых гибридных адаптивных моделей. *Управляющие системы и машины*, (5), 76-83.

4. Бодянський, Є., Винокурова, О., Кобилін, І., & Мулеса, П. (2017). Адаптивна матрична нейро-фаззі самоорганізовна мережа для кластеризації багатовимірних потоків даних. *Вісник Національного університету «Львівська політехніка». Серія: Комп'ютерні науки та інформаційні технології*, (864), 314-319.

5. Бодянский, Е.В., Винокурова, Е.А., Кобылин, И.О., Кобылин, О.А., & Пелешко, Д.Д. (2017) Нечёткая кластеризация временных рядов с неравномерными и асинхронными тактами квантования. *Системи обробки інформації, 5(151)*, 47-54.

6. Bodyanskiy, Y., Vynokurova, O., Szymański, Z., Kobylin, I., & Kobylin, O. (2016, August). Adaptive Robust Models for Identification of

Nonstationary Systems in Data Stream Mining Tasks. In *2016 IEEE First International Conference on Data Stream Mining & Processing (DSMP)* (pp. 263-268). IEEE. (Входить до наукометричної бази SCOPUS).

7. Bodyanskiy, Y., Kobylin, I., Rashkevych, Y., Vynokurova, O., & Peleshko, D. (2018, February). Hybrid Fuzzy-Clustering Algorithm of Unevenly and Asynchronously Spaced Time Series in Computer Engineering. In *2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET)* (pp. 930-935). IEEE. (Входить до міжнародної наукометричної бази SCOPUS).

8. Бодянский, Е. В., Дейнеко, А. А., Кобылин, И. О., & Плисс, И. П. (2016). Адаптивная нечеткая кластеризация коротких временных рядов в интеллектуальном анализе потоков данных. *Intellectual Systems For Decision Making and Problems of Computational Intelligence*, 255-257.

9. Бодянський, Є. В., Винокурова, О. А., Ізонін, І. В., Кобилін, І. О., & Мулеса, П. П. (2017) Кластеризація багатовимірних часових рядів на основі адаптивної матричної нейро-фаззі самоорганізовної мережі. *Intellectual Systems For Decision Making and Problems of Computational Intelligence*, 247-248.

10. Бодянський, Є. В., Винокурова, О. А., Кобилін, І. О., & Мулеса, П.П. (2016). Адаптивна нечітка кластеризація багатовимірних часових рядів з нерівномірним тактом квантування. *Праці VIII-Й Міжнародної школи семінару- «Теорія Прийняття Рішень»* 56-57.

11. Кобылин, И.О., (2015) Об одном методе кластеризации коротких временных рядов. *"Радиоэлектроника и молодежь в XXI веке"* 30-31.

12. Кобылин, И.О., (2016) Адаптивная кластеризация коротких временных рядов с неравномерным тактом квантования. *"Радиоэлектроника и молодежь в XXI веке"* 21-22.