DOI: https://doi.org/10.26565/2304-6201-2022-55-01

УДК (UDC) 519.254

Донець Володимир Віталійович
аспірант
Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 6, Харків, Україна, 61022
e-mail: vol.donets@gmail.com
https://orcid.org/0000-0002-5963-9998

Стрілець Вікторія Євгенівна
к.т.н., доцент кафедри теоретичної та прикладної системотехніки
Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 6, Харків, Україна, 61022
e-mail: viktoria.strilets@karazin.ua
https://orcid.org/0000-0002-2475-1496

Шевченко Дмитро Олександрович
аспірант
Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 6, Харків, Україна, 61022
e-mail: dimyich24@gmail.com
https://orcid.org/0000-0002-7897-250X

Шматков Сергій Ігорович
д.т.н., проф., завідувач кафедри теоретичної та прикладної системотехніки
Харківський національний університет імені В.Н. Каразіна, майдан Свободи, 6, Харків, Україна, 61022
e-mail: s.shmatkov@karazin.ua;
https://orcid.org/0000-0002-0298-7174

Агентно-орієнтований метод кластеризації даних оптового дистриб'ютора

Мета роботи полягає в підвищенні точності кластеризації даних, та визначення цільової кількості кластерів даних, генерованих динамічними економічними системами, за допомогою використання агентно-орієнтованого методу кластеризації з впровадженням методів попередньої обробки даних.
Методи дослідження: в ході виконання досліджень були використані методи обробки та підготовки даних, міри елементної відстані та методи кластеризації. Програмне забезпечення розроблено за допомогою мови Python, були використані бібліотеки scikit-learn, NumPy, SciPy, Pandas, PyTorch й інші.
У **результаті** роботи дані оптового дистриб'ютора було оброблено методами попередньої обробки даних, такими як: визначення пропущених значень, визначення асиметрії та перетворення Бокса-Кокса, проведена нормалізація даних з методом мін-макс нормалізації та проведене зменшення розмірності з методам PCA та t-SNE. Був застосований агентно-орієнтований метод кластеризації з різними метриками (Мангеттенська відстань, відстань Махаланобіса з оберненим значення функції приналежності, дивергенція Кульбака-Лейблера та крос-ентропія). Дивергенція Кульбака-Лейблера показала найкращі результати точності й була обрана для подальшого тестування. Також була протестована спроможність агентно-орієнтованого методу визначати кількість кластерів. Використання методів попередньої обробки даних показало явну присутність 3-х цільових кластерів, що було підтверджено методом.
Висновки: розроблений метод показав високі результати точності кластеризації за рахунок проведеної обробки даних, правильно обраної міри елементної відстані та використання агентно-орієнтованого підходу. Цей метод можна використовувати для покращення якості кластеризації даних динамічних економічних систем, але метод вимагає доопрацювання в збільшенні гнучкості щодо визначення розміру агентів-кластерів
Ключові слова: нечітка кластеризація, мультиагентний підхід, обробка даних, перетворення Бокса-Кокса, метод PCA, метод t-SNE, автокодувальник, дивергенція Кульбака-Лейблера, відстань Махаланобіса, Мангеттенська відстань.

Як цитувати: Донець В. В., Шевченко Д. О., Стрілець В. Є., Шматков С. І. Агентно-орієнтований метод кластеризації даних оптового дистриб'ютора. *Вісник Харківського національного університету імені В.Н. Каразіна, сер. «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління»*. 2022. вип. 55. С.6-18. https://doi.org/10.26565/2304-6201-2022-55-01

How to quote: V.V. Donets, D.O. Shevchenko, V.Y. Strilets, S.I. Shmatkov, "Agent-oriented method of clustering the wholesale distributor data" *Bulletin of V.N. Karazin Kharkiv National University, series "Mathematical modelling. Information technology. Automated control systems*, vol. 55, pp. 6-18, 2022. https://doi.org/10.26565/2304-6201-2022-55-01

1 Вступ

У сучасному світі значимість використання методів машинного навчання в різних сферах діяльності людини важко переоцінити. Однією з таких сфер є управління динамічним економічними системами, дослідження яких вимагає ретельного аналізу великих обсягів даних та виявлення складних взаємозв'язків. Використання методів машинного навчання для ефективного управління економічними системами може надати нові підходи та результати, що сприятимуть покращенню прогнозування економічних показників, виявленню тенденцій та залежностей.

Існують різні методи машинного навчання, які можуть допомогти з покращенням якості прогнозування економічних показників [1], але вони не передбачають використання методів попередньої обробки даних та не мають задовільної точності кластеризації. Крім того їм притаманна проблема невизначеності цільової кількості кластерів.

Мета роботи полягає в підвищенні точності кластеризації даних, та визначення цільової кількості кластерів даних, генерованих динамічними економічними системами, за допомогою використання агентно-орієнтованого методу кластеризації з впровадженням методів попередньої обробки даних. Об’єктом роботи є динамічні економічні системи, а предметом – моделі, методи та інформаційні технології кластеризації та обробки даних.

2 Аналіз літератури

Вирішення теоретичних і практичних проблем аналізу даних моніторингу динамічних економічних систем за допомогою методів машинного навчання становить великий інтерес для дослідників в Україні та поза її межами.

На сьогодні опубліковано низку робіт, що описують використання методів машинного навчання задля вирішення задач класифікації, кластеризації та прогнозування різноманітних даних. У роботі [1] розглядалося прогнозування економічного індексу діяльності фермерського господарства за допомогою методів Linear Regression, k-means, Random Forest, Gradient Boosting та Neural Networks. Також в роботі [1] зазначено, що нейромережі отримують менше значення середньоквадратичної помилки на відміну від методу кластеризації k-means, але вимагають значно більше обчислювальних ресурсів для навчання й, особливо, під час прогнозування. В роботі [2] показана розробка програмного забезпечення для мульти-об’єктного аналізу економічних, механічних та екологічних властивостей цементних композитів з використанням методів неінформованого навчання. В роботі [3] розглядається використання методу k-means для аналізу популяції риби й забезпечення сталого рибальства, показано, що k-means надає більшу точність у порівнянні з SVM методом. Автори статті [4] розробили прогнозну економічну модель з використанням методів машинного навчання. В роботі [5] за допомогою класичних методів і штучних нейромереж прогнозуються впливи соціальних та економічних факторів на ціну нафти.

Використання агентно-орієнтованого підходу кластеризації вже розглядалося в роботі [6] для удосконалення системи медичного моніторингу. Автори показали значне покращення якості кластеризації в порівнянні з традиційними методами та виділили декілька можливих метрик, що можуть поліпшити якість кластеризації для різних типів даних.

Як видно з аналізу літератури впровадження методів штучного інтелекту для аналізу динамічних економічних систем є актуальною проблемою, але у зазвичай використовуються прості методи без використання модифікацій [1–5]. Агентно-орієнтований підхід в кластеризації вже показав покращення якості кластеризації для даних систем медичного моніторингу [6], тому доцільно буде запровадити його для аналізу даних динамічних економічних систем.

3 Теоретичні відомості

Методи кластеризації є важливим інструментом у сфері обробки даних та машинного навчання. Методи можна розділити по типам формування кластерів. Основні з них це:

– центроїдні методи, які представляють кожен кластер єдиним вектором. До них відносяться популярні k-means, c-means та їх модифікації. Мають невисокі вимоги до продуктивності, але обмежені центроїдним уявленням кластерів;

– методи, засновані на щільності, наприклад DBSCAN та OPTICS, в яких області більшої щільності визначаються за кластери [7]. Найбільшою проблемою методів є складність визначення значущих кластерів в наборах даних різної щільності [7];

– графові методи, в яких кластери визначаються графами, що побудовані на елементах вибірок, до них можна віднести HCS. Серед недоліків слід зазначити високу складність й специфічність даних з якими методи дають якісні результати [8];

– нейронні методи, ще відомі як нейронні мережі неінформованого навчання. До них можна віднести нейромережі Кохонена або SOM. Серед недоліків – високі вимоги до обчислювальних потужностей, складності з визначенням початкових ваг та часте розділення одного кластеру [9].

Розглянемо центроїдні методи, а саме метод c-means. Основна ідея методу c-means полягає в тому, що кожен об'єкт у наборі даних належить до одного з певного числа кластерів. Кожен кластер характеризується центроїдом, який представляє середнє значення атрибутів об'єктів у кластері. Метод c-means намагається мінімізувати суму квадратів відстаней між об'єктами і центроїдами їхніх кластерів [6].

Серед основних переваг методу c-means слід зазначити:

– простота: є відносно простим у реалізації та розумінні, не вимагає складних математичних обчислень та алгоритмів;

– швидкість: метод виконується досить швидко, особливо на великих наборах даних. C-means може бути ефективно застосований для швидкої кластеризації;

– застосовується до різних типів даних: може бути використаний для кластеризації різних типів даних, включаючи числові, категоріальні та бінарні дані, але потребує попередньої обробки.

Але метод c-means також має певні недоліки:

– залежність від початкового вибору центроїдів. Результати методів c-means можуть значно змінюватись залежно від початкового вибору центрів. Це може впливати на якість кластеризації та може призвести до отримання різних результатів при кожному запуску алгоритму;

– чутливість до шуму. Випадкові аномалії або викиди можуть вплинути на формування кластерів та призвести до неправильної кластеризації;

– обмеження форми кластерів. Методи c-means передбачають, що кластери мають форму сфери і однаковий розмір. Це може бути недостатньо гнучким для деяких типів даних, де кластери можуть мати складніші форми та розподіл.

У попередніх роботах [6, 10] вже розглядалося усунення недоліків з використанням відстані Махаланобіса, модифікації методу навчання та використання різних метрик, що призвело до покращення показників якості розпізнавання й підвищення рівня нечутливості до шуму.

Агентно-орієнтовані методи є потужними інструментами в галузі штучного інтелекту та мультиагентних систем. Вони базуються на ідеї моделювання окремих "агентів" зі своїм власним станом, здатностями до сприйняття, взаємодії та прийняття рішень. Ці методи дозволяють досліджувати складні системи, в яких взаємодіють багато агентів з різними характеристиками та поведінкою [11].

Основні переваги агентно-орієнтованих методів [11]:

– моделювання складних систем. Агентно-орієнтовані методи дозволяють моделювати складні системи, в яких агенти взаємодіють один з одним та з оточуючим середовищем. Це дозволяє вивчати властивості та поведінку систем, що складаються з багатьох взаємодіючих компонентів;

– гнучкість та адаптивність. Агенти в агентно-орієнтованих системах можуть бути гнучкими та адаптивними, здатними до зміни своєї стратегії та поведінки відповідно до змін у середовищі. Це дозволяє їм ефективно пристосовуватись до нових умов та вирішувати складні завдання;

– розподіленість. Агентно-орієнтовані методи підтримують розподілений підхід до обробки інформації та виконання розрахунків. Кожен агент може мати свою власну локальну інформацію та приймати рішення на основі цієї інформації. Це дозволяє ефективно розподіляти завдання та підвищувати швидкодію та масштабованість систем.

Серед недоліків агентно-орієнтованого методів зазначимо [11]:

– складність реалізації. Розробка агентно-орієнтованих систем може бути складною задачею, оскільки вимагає моделювання взаємодії та поведінки багатьох агентів. Вибір правильних стратегій та алгоритмів для агентів також може бути викликом;

– обмежена доступність даних. Агенти можуть мати обмежений доступ до інформації про стан системи та інших агентів. Це може призвести до обмежень у їх здатності приймати оптимальні рішення;

– проблеми координації. Взаємодія багатьох агентів може вимагати складної координації та комунікації між ними. Управління великою кількістю агентів та забезпечення синхронізації може бути викликом.

4 Агентно-орієнтована модифікація методу кластеризації

Для порівняння якості кластеризації при використанні різних метрик було вирішено обрати метрики, що використовувались в попередніх роботах [6, 10]:

$$d(x_{ij}, c_j) = \begin{cases} d_1(x_{ij}, c_j) & (I) \\ w_{ij}^{-1} d_1(x_{ij}, c_j) & (II) \\ -D_{KL}(x_{ij}, c_j) & (III) \\ p(x_{ij}, c_j^{t-1}) * \log_2 p(x_{ij}, c_j^t) & (IV) \end{cases}, \quad (1)$$

де I – це відстань Мангеттенська відстань, II – відстань Махаланобіса з оберненими значення функції приналежності, III – дивергенція Кульбака-Лейблера, IV – крос-ентропія.

Маючи відстань для визначення між-елементної відстані, отримаємо вираз для визначення функції витрат для кожного кластеру, тобто середню міру внутрішньокласової відстані:

$$cl\_loss(P_j) = \frac{1}{|P_j|} \sum_{i=1}^{|P_j|} d(x_{ij}, c_j). \quad (2)$$

Тоді за виразом (2) визначимо загальну функцію витрат для оцінки поточної якості кластеризації:

$$loss(X^t) = \frac{1}{K^t} \sum_{j=1}^{K^t} cl\_loss(P_j). \quad (3)$$

Відповідно до роботи [6] визначимо алгоритм агентно-орієнтованої модифікації методу c-means:

1. Визначити початкову кількість кластерів та встановити обмеження на кількість елементів в кожному кластері. Обрати центри кластерів серед елементів вхідної вибірки.
2. За обраною між-елементною відстанню визначити набір найближчих елементів до кожного з кластерів. Отримані центр та найближчі елементи є агент-кластером.
3. Обчислити значення приналежності до кожного з кластерів, та обчислити значення розподілу параметрів та обчислити нові центри кластерів.
4. За обраною між-елементною відстанню визначити набір найближчих елементів до кожного з кластерів.
5. Для кожного кластеру за виразом (2) обчислити значення функції витрат.
6. Оцінити поточну якість кластеризації за виразом (3). У випадку режиму роботи алгоритму в автопошуку оптимальної кількості кластерів, та збільшенні значення функції витрат зупинити алгоритм.
7. Провести відбір агентів-кластерів та відкинути агент-кластер з найбільшим значенням функції витрат.
8. Повернутися до 2 етапу, за умови, що цільова кількість кластерів не досягнута.

5 Набір даних

У дослідженні використаний для аналізу набір даних «Wholesale customers» [12]. Дані цього набору відносяться до клієнтів оптового дистриб'ютора, а сам набір даних включає річні витрати в грошових одиницях на різні категорії продуктів. Він складається з 440 записів із 8 змінними стану (параметрами), такими як:

- «Fresh» – річні витрати на свіжу продукцію;
- «Milk» – річні витрати на молочні продукти;
- «Grocery» – річні витрати на продовольчі товари;

- «Frozen» – річні витрати на заморожені продукти;
- «Detergents_Paper» – річні витрати на миючі засоби та паперові вироби;
- «Delicatessen» – річні витрати на делікатесні продукти;
- «Channel» – канал клієнтів, готель/ресторан/кафе або роздрібний канал;
- «Region» – регіон клієнта.

Розглядаючи задачу кластеризації даних, було сформовано три кластера, використовуючи змінну стану «Region». Ця змінна стану включає в себе три регіони (кластери) з загальною кількістю альтернатив позначених в дужках: Лісабон (77), Опорто (47), інший регіон (316). Враховуючи що змінна «Channel» є категоріальною, то в даній роботі вона не буде використовуватися.

6 Попередня обробка даних

Для виконання завдань попередньої обробки даних була використана мова програмування Python та допоміжні бібліотеки, такі як Pandas, SciPy, NumPy, matplotlib, seaborn, scikit-learn та інші. Мова програмування Python є потужним інструментом для попередньої обробки даних завдяки наявності бібліотек, які дозволяють з легкістю проводити експерименти, аналіз даних, будувати графіки, обробляти та трансформувати дані.

6.1 Пропущені значення

Дуже часто в реальних наборах даних присутні пропущені значення, які не дозволяють використовувати математичні моделі та методи машинного навчання. Через це з'являється необхідність у попередній перевірці та видаленні / заповненні пропущених значень. Незважаючи на те, що автор набору даних зазначив, що в поточному немає пропущених значень, була перевірена їх наявність за допомогою бібліотеки Pandas шляхом завантаження набору даних та виконання відповідної функції для пошуку пропущених значень. У результаті перевірки набору даних пропущених значень не виявлено, що дозволяє використовувати його в подальшому дослідженні.

6.2 Перевірка асиметрії даних

Однією з важливих частин для отримання високих результатів моделей машинного навчання є приведення даних до нормального розподілу та робота з асиметрією даних. Для роботи з асиметрією даних та візуалізації графіків були використані бібліотеки SciPy та seaborn відповідно.

Були побудовані діаграми ящиків з вусами (рис. 1) для оригінальних даних. Не важко бачити, що дані мають досить широкий розподіл, викиди та асиметрію. Додатково були визначені коефіцієнти асиметрії для кожної змінної (табл. 1).

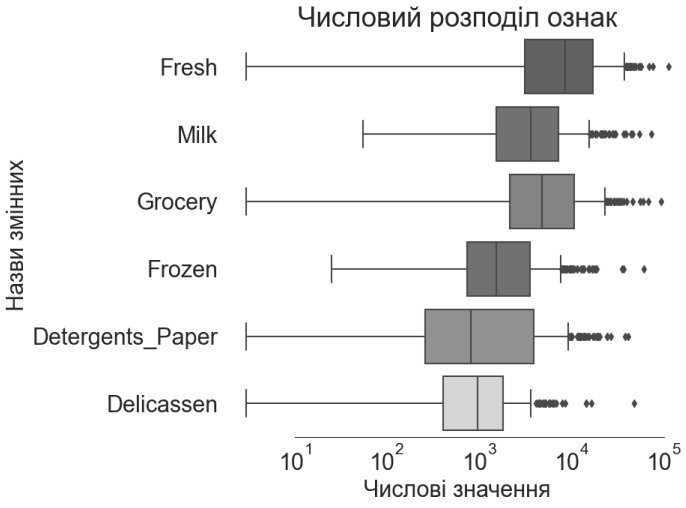

Рис. 1 – Діаграми ящиків з вусами для оригінальних даних

Таблиця 1. Коефіцієнти асиметрії

Змінна	**Коефіцієнт асиметрії**
Delicassen	11.11
Frozen	5.89
Milk	4.04
Detergents_Paper	3.62
Grocery	3.58
Fresh	2.55

Результати досліджень вказують на наявність асиметрії, тому для зменшення асиметрії та приведення розподілів змінних до нормального розподілу було використане перетворення Бокса-Кокса [13]. Після цього перетворення коефіцієнти асиметрії (табл. 2) значно знизилися та дані були приведені до нормального розподілу.

Таблиця 2. Порівняння коефіцієнтів асиметрії

Змінна	**Коефіцієнт асиметрії (до)**	**Коефіцієнт асиметрії (після)**
Delicassen	11.11	0.02
Frozen	5.89	0.003
Milk	4.04	0.01
Detergents_Paper	3.62	-0.03
Grocery	3.58	0.10
Fresh	2.55	-0.01

Про це також свідчать діаграми ящиків з вусами та гістограми розподілів, які відображені на рис 2.

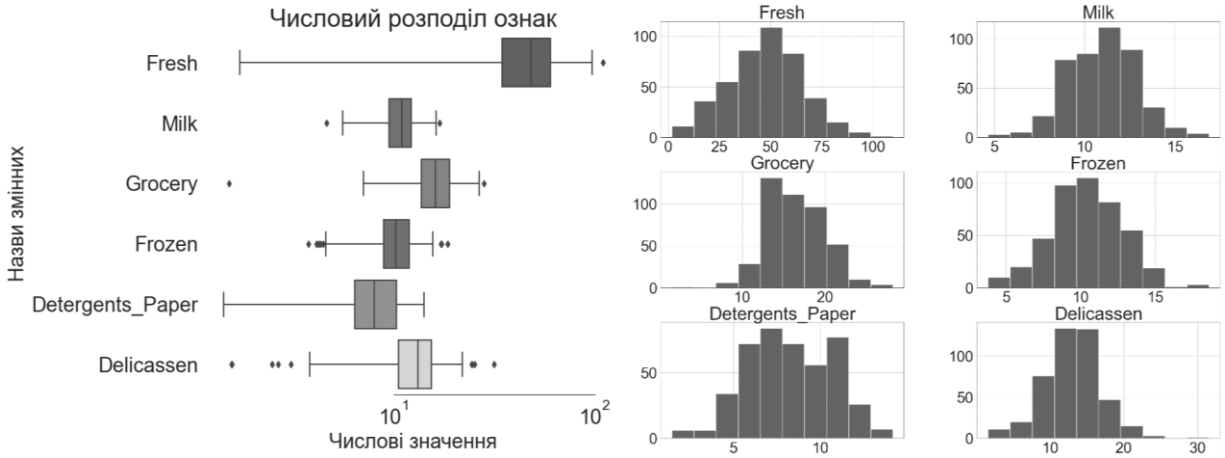

Рис. 2 – Діаграми ящиків з вусами та гістограми розподілів

6.3 Нормалізація даних

Існує багато підходів до нормалізації даних, які знаходять своє застосування у різних задачах, включаючи кластеризацію даних. Для розглядуваної задачі є досить важливим відстань між альтернативами, для кожної альтернативи повинен бути однаковий мінімум та максимум значень змінних стану. Тому в даному випадку найкращим методом нормалізації є мін-макс нормалізація (масштабування). Цей метод дозволяє нормалізувати кожну змінну альтернативи до певного діапазону значень. У роботі дані були нормалізовані до діапазону від 0 до 1 включно. Загальна формула для мін-макс нормалізації до діапазону [0, 1]

$$x_{res} = \frac{x - \min(x)}{\max(x) - \min(x)}, \tag{4}$$

де x є оригінальним значенням, а x_{res} – нормалізоване значення.

Також побудуємо діаграми ящиків з вусами для даних після нормалізації. Ці діаграми показані на рис. 3.

Рис. 3 – Діаграми ящиків з вусами для нормалізованих даних

На діаграмах ящиків з вусами видно, що розподіл став більш широкий з меншою кількістю викидів. Для більш наглядної перевірки та візуалізації кластерів, використаємо методи зменшення розмірності, такі як PCA [14] та t-SNE [15] (рис. 4).

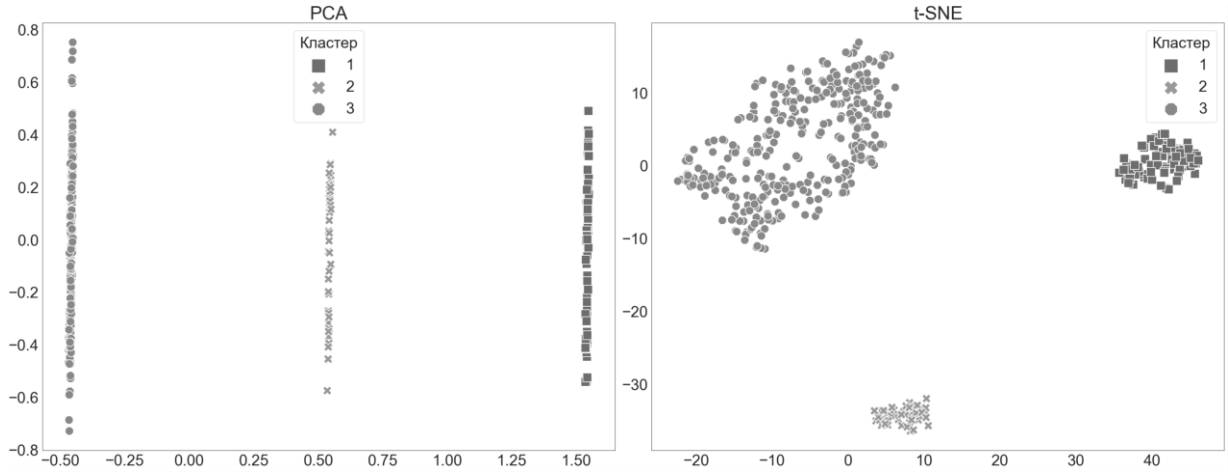

Рис. 4 – Візуалізація нормалізованих даних

У результаті було отримано досить чітке розділення даних на кластери в двовимірних просторах, які були отримані за допомогою методів зменшення розмірності.

6.4 Аномальні значення

Аномальні значення зазвичай дуже сильно впливають на результати кластеризації через те, що центр кластера зміщується у сторону аномалій, що призводить до зменшення якості кластеризації. У роботі [16] були розглянуті різні методи виявлення аномальних значень та визначено, що найбільш ефективними серед них є ізоляційний ліс і генеративні змагальні мережі. В даній роботі використаємо ізоляційний ліс для виявлення аномалій в поточних даних. Цю модель було створено за допомогою бібліотеки scikit-learn. У результаті були побудовані двовимірні графіки за допомогою даних зменшеної розмірності. Ці графіки відображені на рис. 5.

З графіків можна побачити, що ізоляційний ліс визначив значення, які розташовуються на краях, аномальними (значення –1). Хоча є випадки, коли дані, які знаходяться в середині кластеру були визначені аномальними. Це може бути пов’язано з методами зменшення розмірності, які додають певну похибку у даних зменшеної розмірності.

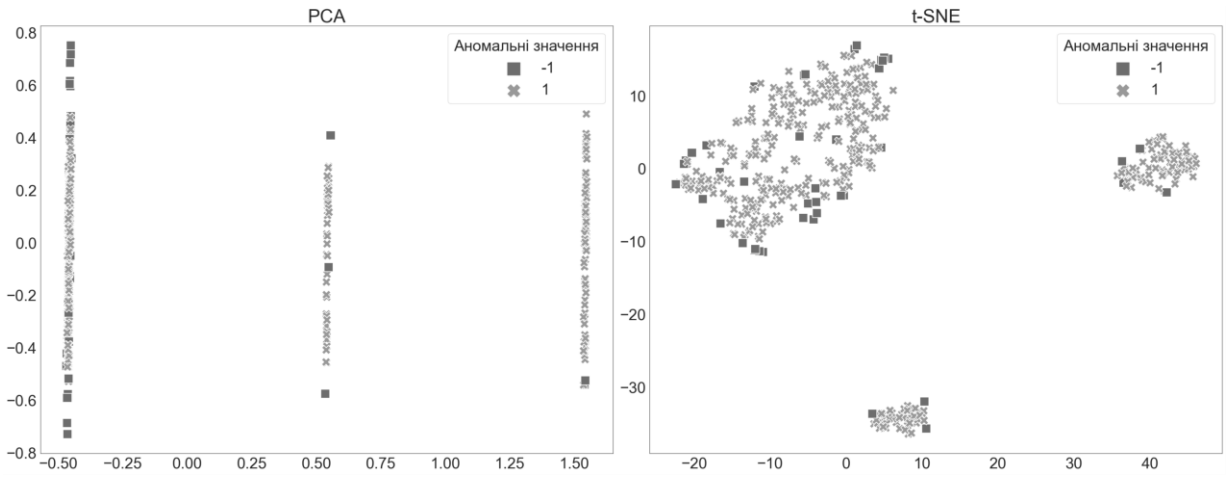

Рис. 5 – Візуалізація аномальних значень в даних

6.5 Агрегування даних

У попередніх підрозділах показане використання методів зменшення розмірності даних для візуалізації результатів, такі як PCA і t-SNE. Зазвичай метод t-SNE використовується тільки для візуалізації даних, він не зберігає структуру даних для подальшого прогнозування або кластеризації даних. Тому для агрегування даних буде використаний лише метод головних компонент (PCA). Також, згідно з [17], серед математичних моделей глибокого навчання визначено, що найкращим є стандартний автокодувальник.

Побудуємо візуалізацію нормалізованих даних за допомогою стандартного автокодувальника для порівняння з побудованими візуалізаціями. Результат зображено на рис. 6.

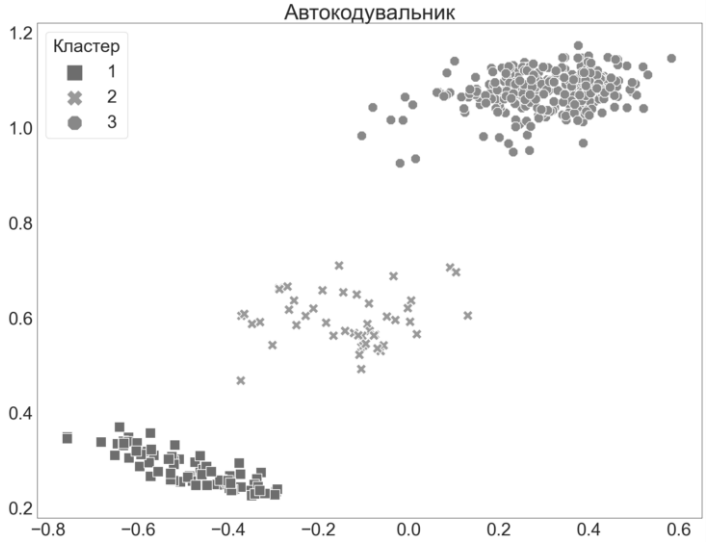

Рис. 6 – Візуалізація нормалізованих даних за допомогою стандартного автокодувальника

Автокодувальник чітко розділив дані на три кластери та більш згрупував їх, що дозволяє математичній моделі кластеризації даних виконувати кластеризації з більшою якістю. Слід зазначити, що при візуалізації кластерів була використана цільова змінна кластеру. Тому з'являється потреба перевірити результати агрегування даних без неї.

Проведемо агрегування даних за допомогою методу головних компонент і автокодувальника без використання цільової змінної. Результати побудови графіків відображено на рис. 7.

Аналізуючи результати, слід зазначити, що кількість змінних стану не є великою. Як наслідок, застосування математичних моделей і методів агрегування даних може призвести до втрати цінної інформації для кластеризації. Так, на прикладі агрегування даних до розмірності 2, неможливо розділити дані на кластери. Тому агрегування даних буде використано тільки для візуалізації результатів кластеризації.

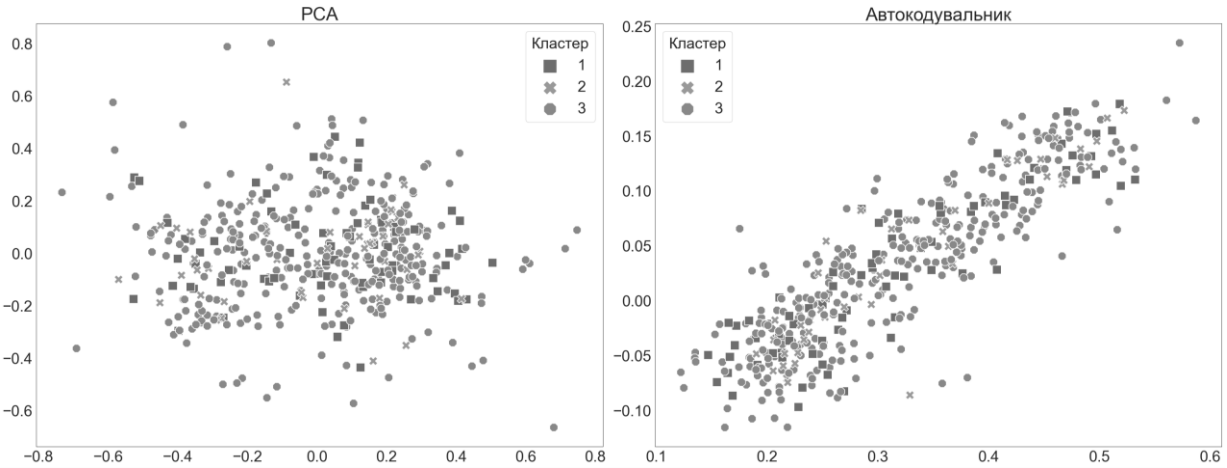

Рис. 7 – Агрегування нормалізованих даних без використання цільової змінної

7 Застосування методу на даних оптового дистриб'ютора

Метод був застосований для кластеризації даних оптового дистриб'ютора використовуючи такі міри між-елементної відстані: Мангеттенська відстань (I), відстань Махаланобіса з оберненим значення функції приналежності (II), дивергенція Кульбака-Лейблера (III), крос-ентропія (IV). У таблиці 3 вказані значення точності за якими оцінювалась якість кластеризації.

Таблиця 3. Результати кластеризації даних оптової дистрибуції з використанням різних мір між-елементної відстані

	Використана міра			
	I	**II**	**III**	**IV**
Точність	0.52	0.62	0.8	0.57

Беручи до уваги результати обчислення кожного з варіантів між-класової відстані, було обрано найкращу міру (міра III – дивергенція Кульбака-Лейблера) та для неї було побудовано матрицю конфузів, ROC (рис. 8) та LF-криві для кращого аналізу отриманих результатів.

Таблиця 4. Матриця конфузів для класифікації даних оптового дистриб'ютора за використання дивергенції Кульбака-Лейблера

		Передбачений клас		
		Лісабон	**Опорто**	**Інші**
Актуальний кластер	Лісабон	16	0	61
	Опорто	0	19	28
	Інші	0	1	316

Можна помітити, що один з кластерів майже повністю вірно класифікований, але, зважаючи на великий перекіс даних, інші кластери були визначені набагато гірше, що може бути суттєвою проблемою для даних з недостатньою кількістю прецедентів для деяких класів. Також слід зазначити, що значення функції витрат для даної кластеризації досягло значення -0.0574 (рис. 8).

За результатами отриманими на вибірці оптової дистрибуції виявлено, що за використання дивергенції Кульбака-Лейблера в якості між-об'єктної відсутні алгоритм дає найкращі результати в точності, тому має сенс використати саме цю міру для тестування алгоритму в режимі автоматичного визначення кількості кластерів. Як видно з графіку (рис. 9) мінімальне значення точності досягнуто для кількості 3 й 4, що відповідає дійсності.

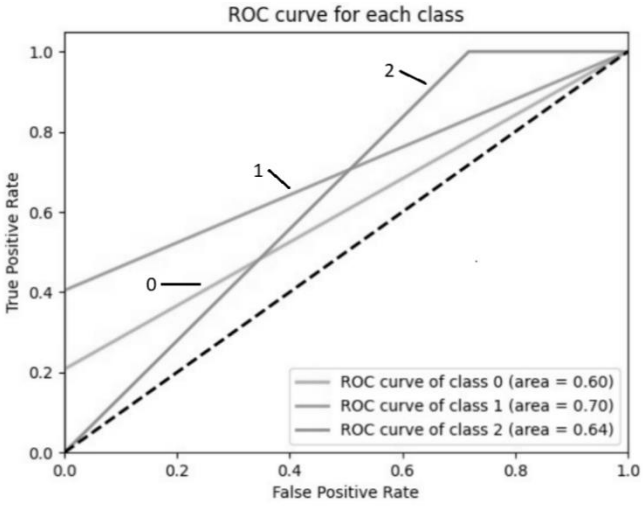

Рис. 8. ROC-криві для кожного з класів вибірки оптового дистриб'ютора для дивергенції Кульбака-Лейблера

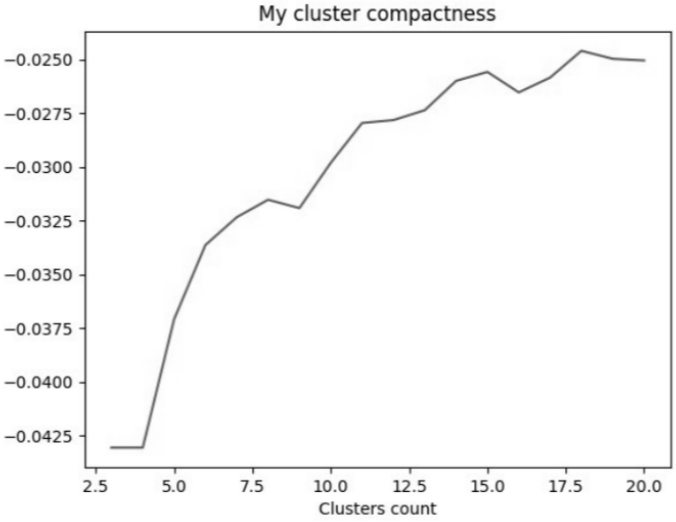

Рис. 9. Відношення кількості кластерів до значенню функції витрат, що отримані під час тренування моделі на вибірці оптового дистриб'ютора за використання дивергенції Кульбака-Лейблера.

8 Висновки

Результатом роботи стало підвищенні точності кластеризації даних, та показано точне визначення цільової кількості кластерів даних, генерованих динамічними економічними системами, за допомогою використання агентно-орієнтованого методу кластеризації з впровадженням методів попередньої обробки даних. Використання методів попередньої обробки даних показало явну присутність 3-х цільових кластерів, що було підтверджено методом.

Розроблений метод показав високі результати точності кластеризації за рахунок проведеної обробки даних, правильно обраної міри елементної відстані та використання агентно-орієнтованого підходу. Цей метод можна використовувати для покращення якості кластеризації даних динамічних економічних систем, але метод вимагає доопрацювання в збільшенні гнучкості щодо визначення розміру агентів-кластерів.

ЛІТЕРАТУРА

1. J. Wełeszczuk, B. Kosińska-Selbi, P. Cholewińska. Prediction of Polish Holstein's economical index and calving interval using machine learning. *Livestock Science. October 2022. Volume 2.* DOI: https://doi.org/10.1016/j.livsci.2022.105039 (дата звернення 25.06.2023).

2. Soroush Mahjoubi, Rojyar Barhemat, Pengwei Guo, Weina Meng, Yi Bao. Prediction and multi-objective optimization of mechanical, economical, and environmental properties for strain-hardening cementitious composites (SHCC) based on automated machine learning and metaheuristic algorithms. *Journal of Cleaner Production. 20 December 2021. Volume 329.* DOI: https://doi.org/10.1016/j.jclepro.2021.129665 (дата звернення 25.06.2023).
3. Yasemin Gültepe. Analysis of Alburnus tarichi population by machine learning classification methods for sustainable fisheries. *SLAS Technology. 2022. Volume 27. Issue 4. Pages 261-266.* DOI: https://doi.org/10.1016/j.slast.2022.03.005 (дата звернення 25.06.2023).
4. Benjamin Decardi-Nelson, Jinfeng Liu. Robust Economic Model Predictive Control with Zone Control. *IFAC-PapersOnLine. 2021. Volume 54. Issue 3. Pages 237-242.* DOI: https://doi.org/10.1016/j.ifacol.2021.08.248 (дата звернення 25.06.2023).
5. Muhammad Mohsin, Fouad Jamaani. Green finance and the socio-politico-economic factors' impact on the future oil prices: Evidence from machine learning. *Resources Policy. 2023. Volume 85. Part A.* DOI: https://doi.org/10.1016/j.resourpol.2023.103780 (дата звернення 25.06.2023).
6. Strilets V., Donets V., Ugryumov M., Zelenskyi R., Goncharova T. Agent-Oriented data clustering for medical monitoring. *Radioelectronic and Computer Systems, 2022, № 1, P. 103–114.* DOI: https://doi.org/10.32620/reks.2022.1.08 (дата звернення 25.06.2023).
7. Johannes Schneider, Michail Vlachos. Fast parameterless density-based clustering via random projections. *CIKM '13: Proceedings of the 22nd ACM international conference on Information & Knowledge Management. October 2013. Pages 861–866.* DOI: https://doi.org/10.1145/2505515.2505590 (дата звернення 25.06.2023).
8. Erez Hartuv, Ron Shamir. A clustering algorithm based on graph connectivity, *Information Processing Letters. 2000 Volume 76. Issues 4–6. Pages 175-181.* DOI: https://doi.org/10.1016/S0020-0190(00)00142-3 (дата звернення 25.06.2023).
9. Wui Lee Chang, Lie Meng Pang, Kai Meng Tay. Application of self-organizing map to failure modes and effects analysis methodology. Neurocomputing. 2017. Volume 249. Pages 314-320. DOI: https://doi.org/10.1016/j.neucom.2016.04.073 (дата звернення 25.06.2023).
10. Donets V., Ugryumov M., Strilets V. A Measure Of Compactness For Fuzzy Clustering Based On Entropy. *Науковий збірник праці міжнародної науково-технічної конференції «Комп'ютерне моделювання у наукоємних технологіях (КМНТ -2022)».*
11. Jun Liu, Guobin Yang, Nan Zhou, Kaiyu Qin, Badong Chen, Yonghong Wu, Kup-Sze Choi. Event-triggered consensus control based on maximum correntropy criterion for discrete-time multi-agent systems. *Neurocomputing. 2023. Volume 545.* DOI: https://doi.org/10.1016/j.neucom.2023.126323 (дата звернення 25.06.2023).
12. Margarida Cardoso. Wholesale customers. *UCI Machine Learning Repository. 2014.* DOI: https://doi.org/10.24432/C5030X (дата звернення: 25.06.2023).
13. Sakia R.M. The box-cox transformation technique: A Review. *The Statistician. 1992. T. 41. № 2. C. 169.* DOI: https://doi.org/10.2307/2348250 (дата звернення: 25.06.2023).
14. Maćkiewicz A., Ratajczak W. Principal Components Analysis (PCA). *Computers & Geosciences. 1993. T. 19. № 3. C. 303–342.* DOI: https://doi.org/10.1016/0098-3004(93)90090-R (дата звернення: 25.06.2023).
15. L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research. 9 Nov 2008.* URL: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf (дата звернення: 25.06.2023).
16. Лихач О., Угрюмов М., Шевченко Д., Шматков С. Методи виявлення викидів в пробних вибірках при управлінні процесами в системах за станом. *Вісник Харківського національного університету імені В.Н. Каразіна, серія «Математичне моделювання. Інформаційні технології. Автоматизовані системи управління». 2022. (53). С. 21-40.*

17. Shevchenko D., Ugryumov M., Artiukh S. Monitoring data aggregation of dynamic systems using information technologies. *Innovative Technologies and Scientific Solutions for Industries. 2023. No. 1 (23), P. 123–131.* DOI: https://doi.org/10.30837/ITSSI.2023.23.123 (дата звернення: 25.06.2023).

REFERENCES

1. J. Wełeszczuk, B. Kosińska-Selbi, P. Cholewińska. Prediction of Polish Holstein's economical index and calving interval using machine learning. *Livestock Science. October 2022. Volume 2.* DOI: https://doi.org/10.1016/j.livsci.2022.105039 (дата звернення 25.06.2023).
2. Soroush Mahjoubi, Rojyar Barhemat, Pengwei Guo, Weina Meng, Yi Bao. Prediction and multi-objective optimization of mechanical, economical, and environmental properties for strain-hardening cementitious composites (SHCC) based on automated machine learning and metaheuristic algorithms. *Journal of Cleaner Production. 20 December 2021. Volume 329.* DOI: https://doi.org/10.1016/j.jclepro.2021.129665 (дата звернення 25.06.2023).
3. Yasemin Gültepe. Analysis of Alburnus tarichi population by machine learning classification methods for sustainable fisheries. *SLAS Technology. 2022. Volume 27. Issue 4. Pages 261-266.* DOI: https://doi.org/10.1016/j.slast.2022.03.005 (дата звернення 25.06.2023).
4. Benjamin Decardi-Nelson, Jinfeng Liu. Robust Economic Model Predictive Control with Zone Control. *IFAC-PapersOnLine. 2021. Volume 54. Issue 3. Pages 237-242.* DOI: https://doi.org/10.1016/j.ifacol.2021.08.248 (дата звернення 25.06.2023).
5. Muhammad Mohsin, Fouad Jamaani. Green finance and the socio-politico-economic factors' impact on the future oil prices: Evidence from machine learning. *Resources Policy. 2023. Volume 85. Part A.* DOI: https://doi.org/10.1016/j.resourpol.2023.103780 (дата звернення 25.06.2023).
6. Strilets V., Donets V., Ugryumov M., Zelenskyi R., Goncharova T. Agent-Oriented data clustering for medical monitoring. *Radioelectronic and Computer Systems, 2022, № 1, P. 103–114.* DOI: https://doi.org/10.32620/reks.2022.1.08 (дата звернення 25.06.2023).
7. Johannes Schneider, Michail Vlachos. Fast parameterless density-based clustering via random projections. *CIKM '13: Proceedings of the 22nd ACM international conference on Information & Knowledge Management. October 2013. Pages 861–866.* DOI: https://doi.org/10.1145/2505515.2505590 (дата звернення 25.06.2023).
8. Erez Hartuv, Ron Shamir. A clustering algorithm based on graph connectivity, *Information Processing Letters. 2000 Volume 76. Issues 4–6. Pages 175-181.* DOI: https://doi.org/10.1016/S0020-0190(00)00142-3 (дата звернення 25.06.2023).
9. Wui Lee Chang, Lie Meng Pang, Kai Meng Tay. Application of self-organizing map to failure modes and effects analysis methodology. Neurocomputing. 2017. Volume 249. Pages 314-320. DOI: https://doi.org/10.1016/j.neucom.2016.04.073 (дата звернення 25.06.2023).
10. Donets V., Ugryumov M., Strilets V. A Measure Of Compactness For Fuzzy Clustering Based On Entropy. *Scientific collection of works of the international scientific and technical conference "Computer modeling in science-intensive technologies (KMNT -2022)".*
11. Jun Liu, Guobin Yang, Nan Zhou, Kaiyu Qin, Badong Chen, Yonghong Wu, Kup-Sze Choi. Event-triggered consensus control based on maximum correntropy criterion for discrete-time multi-agent systems. *Neurocomputing. 2023. Volume 545.* DOI: https://doi.org/10.1016/j.neucom.2023.126323 (дата звернення 25.06.2023).
12. Margarida Cardoso. Wholesale customers. *UCI Machine Learning Repository. 2014.* DOI: https://doi.org/10.24432/C5030X (дата звернення: 25.06.2023).
13. Sakia R.M. The box-cox transformation technique: A Review. *The Statistician. 1992. T. 41. № 2. C. 169.* DOI: https://doi.org/10.2307/2348250 (дата звернення: 25.06.2023).

14. Maćkiewicz A., Ratajczak W. Principal Components Analysis (PCA). *Computers & Geosciences. 1993. T. 19. № 3. C. 303–342.* DOI: https://doi.org/10.1016/0098-3004(93)90090-R (дата звернення: 25.06.2023).
15. L.J.P. van der Maaten and G.E. Hinton. Visualizing High-Dimensional Data Using t-SNE. *Journal of Machine Learning Research. 9 Nov 2008.* URL: https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf (дата звернення: 25.06.2023).
16. Lykhach O., Ugryumov M., Shevchenko D., Shmatkov S. Methods of detecting emissions in test samples during process control in state-based systems. Bulletin of Kharkiv National University named after V.N. Karazin, series "Mathematical modeling. Information Technology. Automated control systems". 2022. (53). C. 21-40. [In Ukrainian]
17. Shevchenko D., Ugryumov M., Artiukh S. Monitoring data aggregation of dynamic systems using information technologies. *Innovative Technologies and Scientific Solutions for Industries. 2023. No. 1 (23), P. 123–131.* DOI: https://doi.org/10.30837/ITSSI.2023.23.123 (дата звернення: 25.06.2023).

Donets Volodymyr *PhD student;*
V.N. Karazin Kharkiv National University, Svobody Sq 6, Kharkiv, Ukraine, 61022
Strilets Viktoriia *Ph.D, associate professor of the theoretical and applied system engineering;*
V.N. Karazin Kharkiv National University, Svobody Sq 6, Kharkiv, Ukraine, 61022
Shevchenko Dmytro *PhD student;*
V.N. Karazin Kharkiv National University, Svobody Sq 6, Kharkiv, Ukraine, 61022
Shmatkov Serhiy *Doctor of Engineering Sciences, professor, Head of Theoretical and Applied Systems Engineering Department;*
V.N. Karazin Kharkiv National University, Svobody Sq 6, Kharkiv, Ukraine, 61022

Agent-oriented method of clustering the wholesale distributor data

The **purpose** of the research is to improve the accuracy of data clustering and to determine the target number of data clusters generated by dynamic economic systems, using an agent-oriented clustering method with the introduction of data preprocessing methods.

Research methods: data processing and preparation methods, elemental distance measures, and clustering methods have been used. The software is developed by using the Python language. The following libraries have also been used: scikit-learn, NumPy, SciPy, Pandas, PyTorch and others.

As a **result** of the research, the data of the wholesale distributor have been processed by the data pre-processing methods such as the determination of missing values, the determination of asymmetry and the Box-Cox transformation. The normalization of the data with the min-max normalization method and the dimensionality reduction with the PCA and t-SNE methods have been carried out. Afterwards, the agent-oriented clustering method has been applied with the Manhattan distance, Mahalanobis distance with the inverse value of the membership function, Kullback-Leibler divergence and cross-entropy metrics. Kullback-Leibler divergence has shown the best accuracy results and has been chosen for the further testing. The ability of the agent-oriented method to determine the number of clusters has been tested. The use of data preprocessing methods shows the clear presence of 3 target clusters, which was confirmed by the method. **Conclusions**: The developed method allows for high clustering accuracy due to the performed data processing, the correctly selected measure of elemental distance and the use of an agent-oriented approach. This method can be used to improve the quality of data clustering of dynamic economic systems, but the method requires improvement in order to increase flexibility in determining the size of cluster agents

Keywords: *fuzzy clustering, multi-agent approach, data processing, Box-Cox transformation, PCA method, t-SNE method, autoencoder, Kullback-Leibler divergence, Mahalanobis distance, Manhattan distance.*

Надійшла у першій редакції 01.09.2022, в останній - 07.10.2022.