DOI: 10.32347/2412-9933.2024.58.80-85
УДК 004.03+624.05

Цюцюра Микола Ігорович
Доктор технічних наук, професор, професор кафедри інженерії програмного забезпечення та кібербезпеки, *https://orcid.org/0000-0003-4713-7568*
Державний торговельно-економічний університет, Київ
Коваленко Андрій Юрійович
Магістрант кафедри програмної інженерії та кібербезпеки, *https://orcid.org/0009-0008-3869-6767*
Державний торговельно-економічний університет, Київ

ОЦІНКА АЛГОРИТМІВ ВИЯВЛЕННЯ АНОМАЛІЙ ЗА ДОПОМОГОЮ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Анотація. *Розглянуто значення виявлення аномалій як важливої проблеми у різних сферах діяльності програмних продуктів сучасного світу, виявлення аномалій може бути у нагоді в кібербезпеці, роботі інтернету речей, аналізі фінансових операцій. Насамперед аномалії можуть сигналізувати про необхідність вчинення будь-яких дій задля уникнення негативних наслідків. Крім того, досліджено значення виявлення аномалій для бізнес-аналітики та ризик-менеджменту. Приділено багато уваги дослідженню різних типів аномалій, зокрема точковим, контекстуальним та колективним, з наведенням прикладів у різних контекстах. Вказано на важливість використання інтелектуальних алгоритмів машинного навчання для виявлення аномалій у великих обсягах даних та швидкого опрацювання інформації з попередженням персоналу. Виявлення аномалій за допомогою машинного навчання є актуальною проблемою в сучасному світі при роботі з великими обсягами даних і постійно зростаючими загрозами у сфері кібербезпеки, фінансових шахрайств, медичної діагностики, виробничої безпеки та інших галузях. Завдяки поширенню інтернету речей (IoT) та великому обсягу даних, які вони генерують, виявлення незвичайних, аномальних або підозрілих подій стає все більш складною задачею для традиційних методів обробки даних. Машинне навчання уможливлює автоматизувати процес виявлення аномалій, використовуючи алгоритми для аналізу і класифікації даних. Це допомагає покращити ефективність і швидкість виявлення аномалій, зменшити витрати на ручний аналіз та сприяти більш точному і швидкому реагуванню на потенційні загрози або проблеми. З поглибленим розвитком технологій машинного навчання, таких як нейронні мережі, алгоритми глибокого навчання та постійне зростання моделей для навчання машини, можливості виявлення аномалій стають все більш точними та різноманітними. Це дає змогу виявляти аномалії у реальному часі та забезпечувати надійний рівень безпеки в різних сферах діяльності, що є надзвичайно важливим у сучасному цифровому світі. Виокремлюють три ситуації, в яких може застосовуватися алгоритм: контрольоване навчання, напівконтрольоване навчання та навчання без нагляду. Класифікація базується на алгоритмічному доступі, включаючи методи імовірнісні, методи вимірювання відстані та щільності, методи кластеризації, методи, що базуються на заняттях, методи реконструкції та спектральні методи. Для вибору оптимального підходу до виявлення аномалій важливо враховувати різні фактори. У статті наведено ілюстративні приклади роботи алгоритмів виявлення аномалій на основі реальних даних.*

Ключові слова: машинне навчання; види аномалій; алгоритми виявлення аномалій; дослідження та оцінка алгоритмів

Мета роботи

Метою статті є розгляд актуальності і значення виявлення аномалій за допомогою методів машинного навчання. Стаття спрямована на визначення важливості цієї проблеми в сучасному світі з великим обсягом даних та постійно зростаючими загрозами. Вона також має на меті висвітлення основних переваг використання машинного навчання для виявлення аномалій, таких як автоматизація процесу, покращення ефективності та швидкості реакції, а також зменшення витрат на ручний аналіз даних. Об’єктом дослідження є розробка програми для визначення аномалій за допомогою штучного інтелекту. Предмет дослідження – програми для визначення аномалій.

Постійно проводиться багато досліджень з метою порівняння й оцінювання різних алгоритмів

виявлення аномалій [1−12]. Деякі з ключових результатів цих досліджень [5−8; 10; 12]:

1. Не існує універсального алгоритму, який би найкраще підходив для всіх задач виявлення аномалій.

2. Навчальні алгоритми зазвичай дають кращі результати, ніж ненавчальні алгоритми, але їм потрібна більша кількість даних для навчання.

3. Вибір алгоритму залежить від конкретної задачі, доступних даних та обчислювальних ресурсів.

Деякі з найпоширеніших алгоритмів виявлення аномалій:

1. *k-Nearest Neighbors (kNN).* Цей алгоритм визначає аномальні дані як ті, які перебувають на значній відстані від більшості інших даних.

2. *Local Outlier Factor (LOF).* Цей алгоритм визначає аномальні дані як ті, які мають значно меншу щільність сусідів, ніж інші дані.

3. *Isolation Forest.* Цей алгоритм будує множину дерев рішень для ізоляції аномальних даних.

4. *One-Class Support Vector Machines (OC-SVM).* Цей алгоритм знаходить межу, яка відокремлює «нормальні» дані від аномальних даних.

Виклад основного матеріалу

Аномалія або виключення належить до певних даних, властивості яких настільки відрізняються від норми, що є підозра, що вони були згенеровані спеціальним механізмом. Однак це дещо кругове визначення вже підкреслює труднощі, властиві цій темі. Але аномальні події надзвичайно цінні для підприємницької діяльності: від шахрайської платіжної транзакції та частого виходу з ладу обладнання до особливого бажання платити клієнта з позитивної сторони, аномалії сигналізують про необхідність дії для компанії.

Саме в даних, які відрізняються від маси, можна знайти цікаві ділові операції і відкрити дещо нове або незвичайне [7−10] (ключові елементи виявлення аномалій) рис. 1.

З меншими обсягами даних і низькими вимогами до критичності в часі такі перевірки можна проводити вручну. Однак великі обсяги даних і швидке опрацювання вимагають підтримки інтелектуальних алгоритмів. Ці алгоритми також дають чіткі правила і рішення про те, що вважається аномальним, а отже, роблять їх доступними для кількісного аналізу в першу чергу.

1. Інтелектуальні автоматичні рішення. Окрім термінів «аномалія» та «викид», термін «*Novelty Detection»*для виявлення виняткових точок даних також поширений в англійській мові та переважно використовується як синонім.У більш вузькому сенсі виявлення новизни стосується порівняння точки даних із сукупністю, яка, як відомо, є нормальною, тоді як *виявлення* викидів стосується ідентифікації викидів у змішаній нормальній/аномальній сукупності.

Що стосується виявлення новизни, процес отримує трохи більше оцінки, на яку він насправді заслуговує: він не лише захищає від небажаних небезпек, але також дає можливість підвищити цінність даних для бізнес-аналітики і зробити їх доступними для тих, хто приймає рішення.

Дані, доступні для аналізу, природно, можуть бути різними. Виявлення аномалій було використано дуже рано у сфері IT-безпеки – саметут доступні великі обсяги даних у вигляді даних журналу. Наприклад, доступ до IT-системи з-за кордону в незвичайний час або з іншою схемою використання представлятиме аномальний сигнал, який може виявити аналіз.

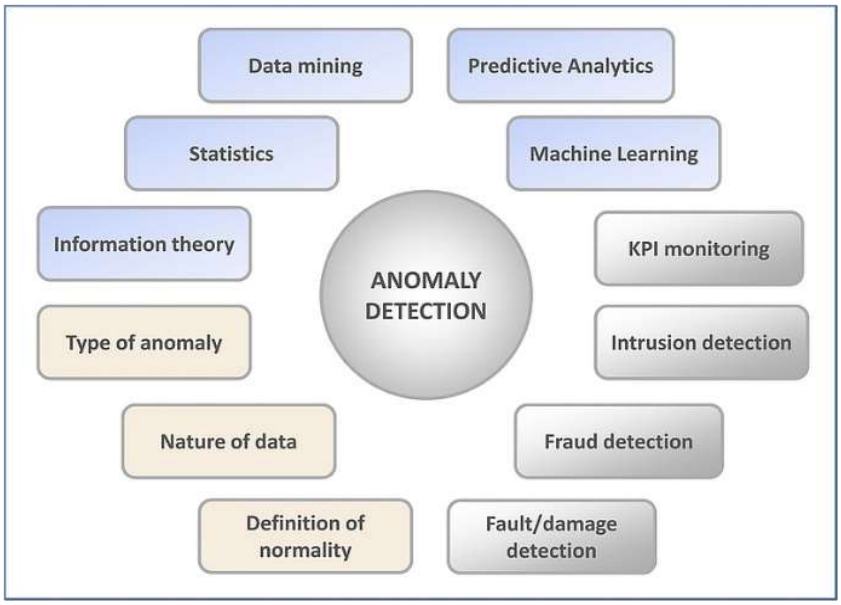

Рисунок 1 – Ключові елементи виявлення аномалій: методи (синій), визначення (помаранчевий) і області застосування (сірий) [1]

Ще один аспект моніторингу аномалій можна знайти в області Інтернету речей (IoT). Дані, що надаються датчиками, надають інформацію про стан машин, IT-пристроїв та інших активів і допомагають здійснювати профілактичне *обслуговування*. Це також дає можливість своєчасно реагувати на умови, що відхиляються від норми.

Подібна методологія виникає у сфері *виявлення шахрайства*. Ця методологія відіграє центральну роль, особливо в платіжних операціях – але не тільки там. Показовий приклад – вкрадена кредитна картка. Купівельна поведінка злочинця настільки сильно відхиляється від норми щодо обсягу продажів, місця розташування та частоти, що спрацьовує сигнал тривоги, тож шахрайській продажі можна запобігти. Виявлення спроб шахрайства також є важливим заходом у решті сфери послуг.

На додаток до цих «класичних» випадків використання, із цифровізацією всіх бізнес-процесів також зростає потреба виявляти з часом аномалії в ключових показниках ефективності (KPE), що мають відношення до управління бізнесом. До них належать, наприклад, падіння продажів, зміни в платіжній поведінці клієнтів або (в онлайн-бізнесі), зниження рейтингу кліків. Проактивне звітування про такі аномалії доповнює класичну бізнес-аналітику і може скоротити час для відповідної реакції.

Інтелектуальна автоматизація виявлення аномалій дає змогу збільшити деталізацію спостережуваних показників. Різноманітні канали, які в класичній BI розглядаються лише як єдине ціле, можна розглянути частинами. Наприклад, продажі компанії можна відстежувати на рівні категорій продуктів, продуктів і каналів збуту – завдання, яке можна досягти лише з великими зусиллями шляхом ручної перевірки.

Для того щоб конкретизувати визначення аномалії як «підозрілої» точки даних, представлене на початку, першим кроком є огляд щодо різних типів аномалій [1]. На підставі цієї першої класифікації вже зрозуміло, що ретельне завчасне визначення цілі є важливим для всіх застосувань автоматизованого виявлення аномалій.

2. Точкова аномалія. Вибіркова аномалія виникає, коли окреме значення є винятковим. Слід розрізняти однофакторні та багатофакторні точкові аномалії [4].

Однофакторна аномалія проявляється в одному вимірі даних. Наприклад, якщо вимірюють учнів у початковій школі, вирізняється висота 1,80 м – підозра в тому, що вимірювали і вчителя, або помилка запису. Отже, ця точка даних була «згенерована» іншим механізмом.

Однофакторні аномалії зазвичай помічають швидко, навіть при поверхневому аналізі. Більш складні аномалії, з іншого боку, проявляються лише тоді, коли кілька метрик розглядаються разом і тому є багатовимірними.

Якщо розглядати один вимір ізольовано, то таких відхилень не виявлено. Якщо виміряти кількість учнів загальноосвітньої школи, ні зріст 1,70 м, ні вага 25 кг не будуть особливо дивними, оскільки 10-річна дитина може важити 25 кг, а 16-річний підліток також може мати зріст 1,70 м.

Однак поєднання обох вимірювань (25 кг, 1,70 м) в одній дитині практично неможливо. Отже, таке вимірювання було б багатовимірною аномалією, механізмом генерації якої, ймовірно, є помилка запису.

3. Контекстуальна аномалія. Ще один тип аномалії – це контекстуальні аномалії. Ці точки даних стають помітними, лише якщо розглядати їх у більш широкому контексті. Візьмемо для прикладу випадок із IT-безпеки: мережевий трафік компанії значно коливається вдень і вночі. Великий обсяг даних, який зазвичай спостерігається в робочий час, може свідчити про несанкціонований доступ до даних компанії вночі. Існує контекстна аномалія, яка стосується безпеки.

4. Колективна аномалія. Останнім і, мабуть, найскладнішим типом аномалії є сфера так званих колективних аномалій. При цьому окремі точки даних не помітні. Натомість аномалія виявляється лише тоді, коли розглядається група даних. Сигнал на рис. 2 показує електрокардіограму, на якій кожне серцебиття виглядає нормальним.

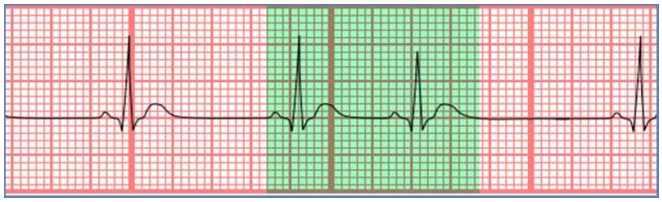

Рисунок 2 – Сукупні аномалії.
Екстрасистоли на електрокардіограмі

Тільки нерегулярність додаткового удару (екстрасистолія) визначає аномальний, як це називається в медичному контексті, стан.

5. Часовий ряд. Особливе місце займають часові ряди, в яких крім однофакторних часто присутні контекстуальні або колективні аномалії (рис. 3). З одного боку, оновлення часових рядів у майбутнє як частина прогнозної аналітики дає змогу виявити майбутні аномалії до їх виникнення. Для цього використовуються статистичні методи, такі як ARIMA, або алгоритми машинного навчання, такі як нейронні мережі (LSTM).

З іншого боку, прогнозовані значення та їх довірчі інтервали можуть бути використані як визначення нормальності для виявлення аномалій, коли вони виникають [3].

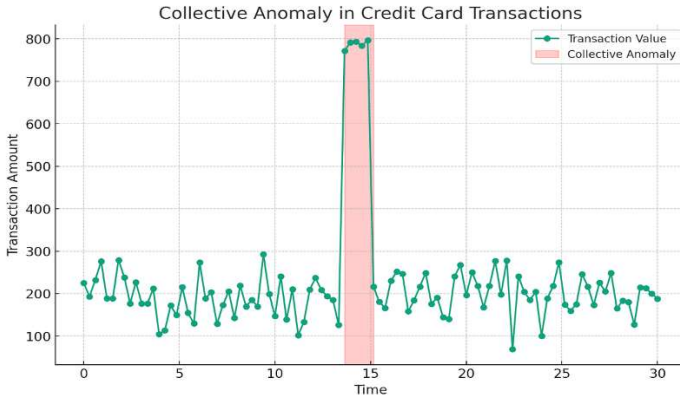

Рисунок 3 – Приклади часових рядів

6. Алгоритми виявлення аномалій. Відомі різні підходи до алгоритмічної ідентифікації викидів, вибір яких має залежати не лише від конкретних властивостей питання, але й від загальних властивостей даних.

Можна виокремити три ситуації:

1. На додаток до звичайно позначених даних існують також відомі аномальні дані (контрольоване навчання).

2. Є лише дані, позначені як нормальні, але жодних аномалій не позначено (напівконтрольоване навчання).

3. Є лише немарковані дані (навчання без нагляду).

Друга класифікація базується на алгоритмічному доступі.

У *імовірнісних методах* статистична модель адаптується до даних. Імовірність генерації точки даних з цієї моделі оцінюється для виявлення викидів. Статистичні моделі мають бути вибрані належним чином і в разі необхідності параметризовані.

З іншого боку, *методи вимірювання відстані та щільності*, такі як алгоритм k-NN, розглядають кожну точку даних у контексті її оточення або її подібності до інших точок даних. Якщо для екземпляра існує достатньо велика кількість подібних даних, метод оцінює точку даних як нормальну.

Методи кластеризації працюють за подібним принципом, у якому алгоритми машинного навчання, такі як k-середні, використовуються для поділу даних на групи. Екземпляри, які перебувають далеко від усіх груп, визначаються як викиди.

Методи, що базуються на заняттях, вимагають принаймні частково класифікованого набору навчальних даних (контрольоване або напівконтрольоване навчання). Класифікатор машинного навчання навчається за допомогою навчальних даних, щоб передбачити, чи належить точка даних до класу. Широко використовуються однокласові опорні векторні машини (SVM), які визначають межу між нормальністю та аномаліями, а тому також називаються методами домену.

За допомогою *методів реконструкції та спектральних методів* дані переносяться до нижчої розмірності й таким чином стискаються. Екземпляри, які не можуть бути добре зіставлені в цьому процесі стиснення, вважаються аномаліями. Ці методи включають аналіз головних компонентів (PCA) і реплікаторні нейронні мережі. Подібний метод виникає *вінформаційно-теоретичних процедурах,* у яких оцінюються такі параметри, як ентропія та складність за методом Колмогорова.

7. Використання алгоритмів

Вибір відповідного підходу до виявлення аномалії залежить від багатьох факторів. На цьому етапі деякі ілюстративні приклади покажуть можливості алгоритмів. На рис. 4 показано вміст фенолу у винах із трьох різних сортів винограду.

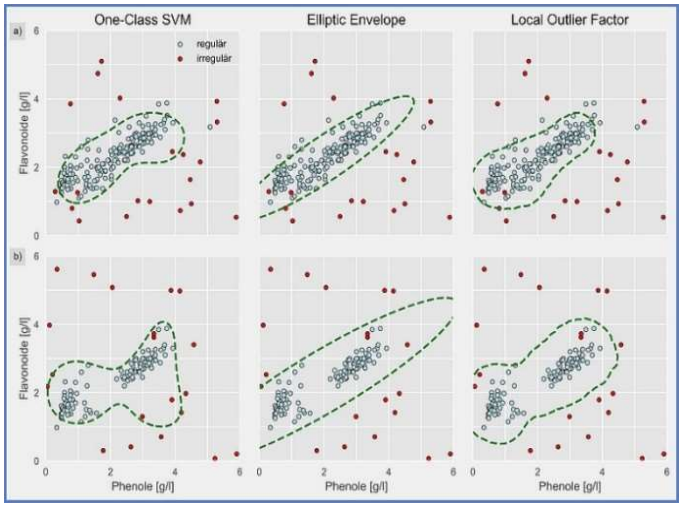

Рисунок 4 – Порівняння алгоритмів виявлення аномалій: наоснові класів (One-Class SVM), імовірнісних (Eliptic Envelope) і на основі щільності (фактор локального викиду)

На осі ординат можна побачити вміст флавоноїдів, які передусім відповідають за колір вина. На осі абсцис відкладено загальний вміст усіх фенолів. Вони мають значний вплив на смак вина. На додаток до виміряних точок даних показано випадкові забруднення, які алгоритми мають розпізнавати як аномалії.

Щоб знайти область нормальності, позначену пунктирною лінією на рис. 4, використовувався алгоритм зі сфери класових, імовірнісних методів і методів щільності [2]. Загалом якість алгоритмів порівняно висока.

Однак для виявлення аномалії на рис. 4 були доступні лише дані з двох сортів винограду як визначення нормальності, які розбиваються на два кластери. Враховуючи цю складнішу ситуацію з даними, метод щільності особливо підходить. Слід зазначити, що на практиці зазвичай існують дані

більшої розмірності, які дають змогу алгоритмам чіткіше розрізняти нормальність і аномалію.

Висновки

Обсяг даних, які генеруються в сучасних компаніях, і пов'язана з ними деталізація даних роблять використання алгоритмів машинного навчання все більш привабливим. У поєднанні з автоматизованим аналізом у реальному часі ці підходи пропонують помітну додаткову цінність порівняно з класичними інструментами аналізу. Однак до вибору та використання правильних інструментів слід підходити ретельно, навіть при автоматизованих процесах. Важливо знайти потрібний рівень деталізації, за якого аномалії надійно виявляються і водночас мінімізуються помилкові тривоги. Лише за цих умов автоматизоване рішення може набути визнання в компанії.

Список літератури

1. Чандола, В., Банерджі, А., Кумар, В. Виявлення аномалій: опитування. У: ACM Computing Surveys 41, 2009.
2. Дуа, Д., Карра Таніскіду, Е. Репозиторій машинного навчання UCI. Каліфорнійський університет, Школа інформації та комп'ютерних наук, 2017.
3. Гупта, М. Виявлення викидів для часових даних: опитування. IEEE Transactions on Knowledge and Data Engineering 26, 2014.
4. Джолліфф, І. Т. : Аналіз основних компонентів. Нью-Йорк, Берлін, Гейдельберг: Springer 2002.
5. ML.NET Documentation URL: https://learn.microsoft.com/en-us/dotnet/machine-learning/
6. Guide to Intrusion Detection and Prevention Systems (IDPS). Technical report, National Institute of Standards and Technology, U.S. Department of Commerce, 2014. С. 215–249.
7. Christina Warrender, Stephanie Forrest, and Barak Pearlmutter. Detecting intrusion using system calls: alternative data models. In Proceedings of the IEEE Symposium on Security and Privacy, 1999.
8. Zengyou He, Xiaofei Xu, and Shengchun Deng. Discovering cluster-based local outliers. Pattern Recogn. Lett., 24 (9-10):1641-1650, June 2003.
9. Hawkins, Douglas M. Identification of Outliers. Chapman and Hall London; New York, 1980.
10. Bram Steenwinckel. Adaptive Anomaly Detection and Root Cause Analysis by Fusing Semantics and Machine Learning. European Semantic Web Conference. 2018.
11. Tsiutsiura M. I., Tsiutsiura S. V., and Kryvoruchko O. V. Information technologies for the development of the content of education: monographi. *CP "Comprint",* 2019. Kyiv: 118 p. ISBN -978-966-929-967-9.
12. Цюцюра М. І., Єрукаєв А. В., Гоц В. В., Костишина Н. В. Реалізація генетичного алгоритму шляхом застосування продукційних правил. *Управління розвитком складних систем.* Київ, 2019. № 39. С. 64 – 68. DOI: 10.6084/m9.figshare.11340653.

Стаття надійшла до редколегії 05.05.2024

Tsiutsiura Mykola
DSc (Eng.), Professor, Professor of the Department of Software Engineering and Cybersecurity,
https://orcid.org/0000-0003-4713-7568
Kyiv National University of Trade and Economics, Kyiv
Kovalenko Andriy
MSc student of the Department of Software Engineering and Cybersecurity,
https://orcid.org/0009-0008-3869-6767
State University of Trade and Economics, Kyiv

EVALUATION OF ANOMALY DETECTION ALGORITHMS USING MACHINE LEARNING METHODS

Abstract. *This article discusses how anomaly detection is an important problem in various areas of software products in the modern world, anomaly detection can be useful in cybersecurity, the Internet of Things, and financial transaction analysis. Above all, anomalies can signal the need to take some action to avoid negative consequences. In addition, the importance of anomaly detection for business intelligence and risk management is explored. Much attention is paid to the study of different types of anomalies, including point, contextual and collective, with examples in different contexts. The importance of using intelligent machine learning algorithms to detect anomalies in large amounts of data and quickly process information with warnings to staff is emphasized. Anomaly detection using machine learning is an urgent problem in the modern world when dealing with large amounts of data and ever-growing threats in the field of cybersecurity, financial fraud, medical diagnostics, industrial safety and other industries. With the proliferation of the Internet of Things (IoT) and the large amount of data it generates, detecting unusual, anomalous, or suspicious events is becoming increasingly challenging for traditional data processing methods. Machine learning automates the anomaly detection process by using algorithms to analyze and classify data. This improves the efficiency and speed of anomaly detection, reduces the cost of manual analysis, and facilitates a more accurate and rapid response to potential threats*

or issues. With the in-depth development of machine learning technologies such as neural networks, deep learning algorithms, and the constant growth of machine learning models, anomaly detection capabilities are becoming more accurate and diverse. This makes it possible to detect anomalies in real time and ensure a reliable level of security in various fields of activity, which is extremely important in today's digital world. There are three situations in which the algorithm can be applied: supervised learning, semi-supervised learning, and unsupervised learning. The classification is based on algorithmic access, including probabilistic methods, distance and density methods, clustering methods, activity-based methods, and reconstruction and spectral methods. To choose the best approach to anomaly detection, it is important to consider various factors. The article provides illustrative examples of anomaly detection algorithms based on real data.

Keywords: machine learning; types of anomalies; anomaly detection algorithms; research and evaluation of algorithms

References

1. Chandola, V., Banerjee, A., Kumar, V. (2009). Anomaly detection: a survey. In: ACM Computing Surveys, 41.
2. Dua, D., Carra Taniskidou, E. (2017). UCI Machine Learning Repository. University of California, School of Information and Computer Sciences.
3. Gupta, M. (2014). Outlier detection for temporal data: a survey. IEEE Transactions on Knowledge and Data Engineering, 26.
4. Jolliffe, I. T. (2002). Principal component analysis. New York, Berlin, Heidelberg: Springer.
5. ML.NET Documentation URL: https://learn.microsoft.com/en-us/dotnet/machine-learning/
6. Guide to Intrusion Detection and Prevention Systems (IDPS). (2014). Technical report, National Institute of Standards and Technology, U.S. *Department of Commerce*, 215–249.
7. Warrender, Christina, Forrest, Stephanie and Pearlmutter, Barak. (1999). Detecting intrusion using system calls: alternative data models. In Proceedings of the IEEE Symposium on Security and Privacy.
8. Zengyou, He, Xiaofei, Xu and Shengchun, Deng. (2003). Discovering cluster-based local outliers. *Pattern Recogn. Lett.,* 24 (9-10):1641–1650.
9. Hawkins, Douglas M. (1980). Identification of Outliers. Chapman and Hall London; New York.
10. Steenwinckel, Bram. (2018). Adaptive Anomaly Detection and Root Cause Analysis by Fusing Semantics and Machine Learning. European Semantic Web Conference.
11. Tsiutsiura, M. I., Tsiutsiura, S. V. and Kryvoruchko, O. V. (2019). Information technologies for the development of the content of education. Monograph. CP "Comprint". Kyiv: 118. ISBN -978-966-929-967-9.
12. Tsiutsiura, Mykola, Yerukaiev, Andrii, Hots, Vladyslav & Kostyshyna, Nataliia. (2019). Implementation of a genetic algorithm using product rules. *Management of Development of Complex Systems*, 39, 64–68. [in Ukrainian]; dx.doi.org\10.6084/m9.figshare.11340653.

Посиланя на публікацію

APA Tsiutsiura, M. & Kovalenko, A. (2024). Evaluation of anomaly detection algorithms using machine learning methods. *Management of Development of Complex Systems*, 58, 80–85, dx.doi.org\10.32347/2412-9933.2024.58.80-85.

ДСТУ Цюцюра М. І., Коваленко А. Ю. Оцінка алгоритмів виявлення аномалій за допомогою методів машинного навчання. *Управління розвитком складних систем.* Київ, 2024. № 58. С. 80 – 85, dx.doi.org\10.32347/2412-9933.2024.58.80-85.