УДК 004.8:004.932 DOI https://doi.org/10.32782/1995-0519.2024.5.7

ІНФОРМАЦІЙНА СИСТЕМА ВИЯВЛЕННЯ АНОМАЛІЙ В ДАНИХ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Олексій Шушура
доктор технічних наук, доцент, професор кафедри цифрових технологій в енергетиці
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», просп. Берестейський, 37, Київ, Україна, 03056, shushura_oleksii@lll.kpi.ua
ORCID: 0000-0003-3200-720X

Єлизавета Мороз
магістр кафедри цифрових технологій в енергетиці
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», просп. Берестейський, 37, Київ, Україна, 03056, yelyzaveta.mor@gmail.com
ORCID: 0009-0006-8865-4173

Ірина Сегеда
кандидат економічних наук, доцент, доцент кафедри цифрових технологій в енергетиці
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», просп. Берестейський, 37, Київ, Україна, 03056, Segeda.Iryna@lll.kpi.ua
ORCID: 0000-0003-1958-4985

Людмила Асєєва
доктор філософії, доцент кафедри комп'ютерної інженерії
Державний університет інформаційно-комунікаційних технологій, вул. Солом'янська, 7, Київ, Україна, 03110, aseewal@i.ua
ORCID: 0000-0001-5954-4211

У статті розглянуто інформаційну систему для виявлення аномалій у великих наборах даних з використанням методів машинного навчання. Актуальність проблеми обумовлена зростанням обсягів даних і складністю їх аналізу, що вимагає розробки автоматизованих рішень для виявлення відхилень у даних, таких як шахрайство, несправності або інші нетипові ситуації. Виявлення аномалій є важливим інструментом у таких сферах, як кібербезпека, фінансовий моніторинг, промислова діагностика та медична аналітика.

Сформовано вимоги до системи виявлення аномалій за допомогою методів машинного навчання, які формалізовані у вигляді діаграми прецедентів UML, спроєктована її структура та обрані засоби розробки, створене програмне забезпечення на мові Python та проведене його тестування. Система базується на кількох алгоритмах машинного навчання, включаючи Isolation Forest, Local Outlier Factor та DBSCAN, які забезпечують ефективність і точність виявлення аномалій у різних прикладних задачах. Дослідження ефективності системи на контрольних вибірках даних показало високий рівень точності виявлення аномалій.

Запропонована система дозволяє користувачам аналізувати дані та виявляти аномалії без необхідності глибоких знань у програмуванні чи налаштуванні алгоритмів. Вона автоматично проводить аналіз, порівнюючи результати роботи різних моделей, та надає можливість візуалізації результатів для покращення розуміння виявлених аномалій. Результати дослідження показали, що система здатна швидко й точно ідентифікувати аномальні дані, що дозволяє значно скоротити час аналізу та підвищити ефективність прийняття рішень. Запропоноване рішення може стати важливим інструментом для автоматизації процесів виявлення аномалій у великих наборах даних, що є критично важливим у сучасних умовах зростання обсягів інформації. Результати досліджень можуть бути використані розробниками інформаційних технологій, що працюють в області аналізу даних.

Ключові слова: виявлення аномалій, машинне навчання, інформаційна система, Local Outlier Factor, Isolation Forest, DBSCAN, аналіз даних.

Актуальність роботи. Виявлення аномалій у даних є важливим завданням, яке знаходить застосування в багатьох сферах, таких як кібербезпека, фінансовий моніторинг, промислова діагностика та медична аналітика. Сьогодні значна кількість компаній використовує штучний інтелект для покращення та автоматизації процесу аналізу даних з метою виявлення незвичайних

патернів або поведінки. Через зростання обсягів інформації, що обробляється, та різноманіття структур даних, актуальною стає розробка інформаційних систем для автоматизованого виявлення аномалій на основі методів машинного навчання. Очікується, що ринок штучного інтелекту зросте з 241,8 мільярдів доларів США у 2023 році до майже 740 мільярдів доларів США у 2030 році, оскільки все більше компаній інвестують у ці технології [1].

Автоматизація процесу виявлення аномалій в даних за допомогою методів машинного навчання полягає у застосуванні алгоритмів для автоматичного аналізу великих масивів даних і виявлення у них відхилень від норми [2]. Це допомагає оптимізувати роботу аналітичних систем та запобігати потенційним загрозам чи помилкам. Використання цих методів забезпечує точніше та швидше виявлення аномалій, що особливо важливо для компаній, які працюють із великими обсягами даних або в умовах, де швидкість прийняття рішень є критичною [3]. Наприклад, застосування методів машинного навчання допомагає виявляти аномальні показники у рівні популяції фітопланктону, який є фундаментальною складовою екосистеми Землі. Відстеження таких аномалій допомагає дослідникам реагувати на зміни, щоб зберігати показники на прийнятному рівні [4].

Сучасні дослідження у галузі виявлення аномалій зосереджені на розробці спеціалізованих алгоритмів, таких як метод опорних векторів (SVM), методи кластеризації (наприклад, DBSCAN), методи локального виявлення викидів (LOF) та ізольовані лісові алгоритми (IF). Ці методи довели свою ефективність у різних прикладних задачах, таких як виявлення шахрайства в банківській сфері, аналіз промислових даних для передбачення відмов обладнання та моніторинг кіберзагроз у реальному часі [5, 6].

Для автоматизації процесів виявлення аномалій застосовуються програмні системи, що поєднують різні алгоритми та методи аналізу даних. Наприклад, бібліотеки машинного навчання, такі як Scikit-learn та PyOD [5], надають інструменти для роботи з різними алгоритмами виявлення аномалій, включаючи методи на основі дерев рішень, методи ядрових оцінок та нейронні мережі. Основним викликом цих систем є необхідність вибору оптимального методу для конкретної задачі, а також підбір відповідних гіперпараметрів для досягнення максимальних результатів.

Глибокі нейронні мережі також знаходять застосування у завданнях виявлення аномалій. Особливо популярними є архітектури ConvNet, ResNet та автоенкодери, що використовуються для виявлення відхилень у складних багатовимірних даних [7]. Застосування методів попереднього навчання та transfer learning дає можливість ефективно працювати з обмеженими наборами даних, що є особливо важливим у задачах кібербезпеки та аналізу рідкісних подій [8]. Вибір між різними методами та архітектурами залежить від специфіки даних та мети аналізу [9]. Наприклад, методи на основі ізольованого лісу є ефективними для великих наборів даних з високою розмірністю, тоді як DBSCAN краще підходить для кластеризації даних з відомими локальними аномаліями. Для більш складних випадків можуть використовуватися комбінації методів або гібридні підходи, що обумовлює актуальність подальших досліджень в цій області та розробки спеціалізованих інформаційних систем.

Метою даної роботи є розробка інформаційної системи для виявлення аномалій у даних на основі методів машинного навчання. Для підвищення точності та швидкості процесу виявлення аномалій необхідно застосувати сучасні алгоритми машинного навчання, такі як методи кластеризації, методи локального виявлення викидів та ізольовані лісові алгоритми. Дослідження передбачає визначення вимог до системи виявлення аномалій, розробку її архітектури, вибір відповідних засобів розробки, а також створення програмного забезпечення з подальшим тестуванням на реальних наборах даних.

Матеріали і результати досліджень. Функціональні вимоги до інформаційної системи виявлення аномалій в наборах даних у формальному вигляді представлено діаграмою прецедентів, що наведена на рисунку 1.

На рисунку 1 видно, що використання системи передбачає одну роль користувача без додаткових обмежень у функціональних можливостях. Користувач має можливість переглядати, візуалізувати, конкатенувати та редагувати дані, а також проводити аналіз на аномальні значення.

Для автоматизації процесу виявлення аномалій та підвищення продуктивності інформаційної системи було обрано кілька методів машинного навчання: Local Outlier Factor (LOF), DBSCAN та Isolation Forest (IF). Кожен із цих методів має свої переваги у специфічних типах задач. Розглянемо їх більш детально.

Рис. 1. Діаграма прецедентів системи виявлення аномалій

Метод LOF використовується для виявлення локальних аномалій з використанням щільності зразків, представлених у багатовимірному просторі. Він дозволяє ідентифікувати ті об'єкти, які суттєво відрізняються від їх найближчих сусідів, що дає змогу ефективно виявляти аномалії, що мають локальну природу. Алгоритм реалізації методу LOF в інформаційній системі наведено на рисунку 2.

Метод DBSCAN є методом кластеризації, що дозволяє виділяти аномалії у вигляді точок, що не належать жодному кластеру. Його перевагою є можливість виявляти аномалії у випадках, коли надані дані мають чітку кластерну структуру.

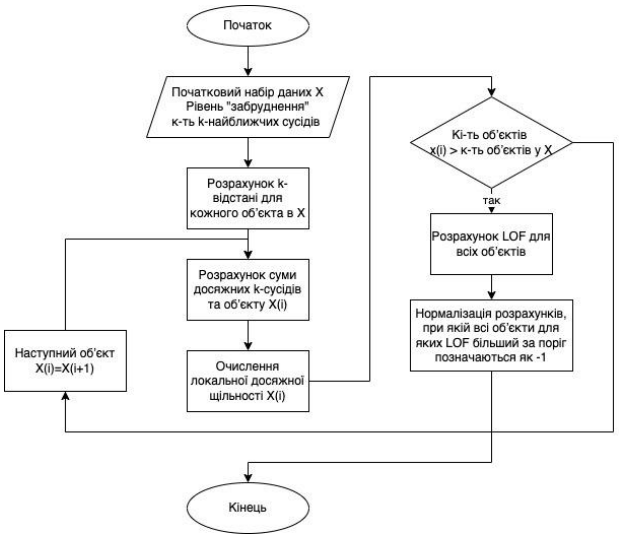

Рис. 2. Схематичне представлення алгоритму LOF

Метод Isolation Forest (IF) базується на деревоподібних моделях, що «ізолюють» аномальні точки шляхом утворення дерева рішень. Це дозволяє ефективно виявляти аномалії за рахунок швидкого відкидання аномальних об'єктів з масиву даних.

В інформаційній системі виявлення аномалій реалізовано всі три зазначені методи. Слід відмітити, що ці алгоритми машинного навчання досить схожі за принципом дії: вони виявляють об'єкти, які за певним критерієм мають аномальні значення, та використовуючи нормалізацію позначають їх числом, що виходить за межі діапазону [0; 1].

Архітектура системи не передбачає використання систем управління базами даних, вся необхідна інформація зберігається у файлах. Програмне забезпечення системи реалізовано на мові програмування Python з використанням її бібліотек.

Система дозволяє користувачеві налаштовувати параметри моделі та порівнювати методи виявлення аномалій для вибору найкращого з них для певного набору даних. Коли система отримує вхідні дані у вигляді файлу користувачеві надається вибір методу виявлення аномалій, можливість визначення параметрів моделі, а після її спрацювання виводиться візуалізація виявлених аномалій та перелік аномальних значень. Приклад завдання параметрів для визначення аномальних значень наведено на рисунку 3.

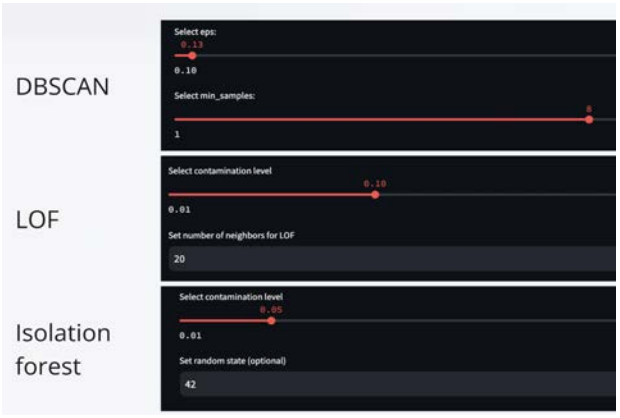

Рис. 3. Приклад завдання параметрів для виявлення аномальних значень

Тестування роботи системи, проведене на відкритому наборі даних Iris [10], показало її готовність до практичного застосування.

Приклад результатів застосування методу LOF показано на рисунку 4.

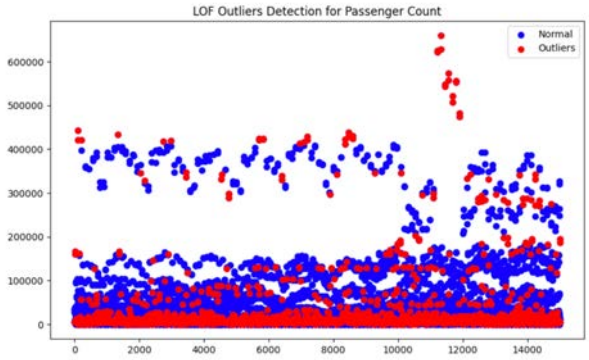

Рис. 4. Візуалізація аномалій, виявлених методом LOF

Зручний інтерфейс системи забезпечує гнучкість у налаштуванні параметрів моделей, дозволяє знижувати навантаження на аналітиків та підвищувати швидкість прийняття рішень.

Висновки. В результаті дослідження розроблено інформаційну систему для виявлення аномалій у даних на основі методів машинного навчання. Формалізовано вимоги до системи, сформована її архітектура та розроблене програмне забезпечення, що інтегрує кілька популярних алгоритмів виявлення аномалій, таких як Isolation Forest, Local Outlier Factor (LOF) та DBSCAN, які забезпечують її адаптивність до різних типів даних для досягення високої точності виявлення аномалій.

Розроблена інформаційна система виявлення аномалій може стати важливим інструментом для автоматизації аналізу даних у різних галузях, таких як кібербезпека, фінанси, промисловість та екологічний моніторинг. Її використання дозволяє зменшити ризики, пов'язані з аномальними подіями, та підвищити ефективність аналітичних процесів у сучасних умовах роботи з великими обсягами даних. В якості напрямків подальших досліджень можна виділити застосування методів ансамблевого навчання та інтеграцію додаткових алгоритмів виявлення аномалій, реалізацію автоматичної оптимізації параметрів алгоритмів.

Для впровадження системи необхідно провести її адаптацію для розгортання на хмарних платформах, що дозволить масштабувати її для великих обсягів даних та інтегрувати з іншими інформаційними системами.

ЛІТЕРАТУРА

1. Shutenko V. AI Anomaly Detection: Best Tools And Use Cases. TechMagic. URL: https://www.techmagic.co/blog/ai-anomaly-detection/ (date of access: 12.10.2024).

2. Шушура О., Левченко Л., Савчук А. Аналіз даних на основі нейронних мереж із використанням мікросервісної архітектури. *Вісник КрНУ імені Михайла Остроградського*. 2023. № 3. С. 80–85. https://doi.org/10.32782/1995-0519.2023.3.10 (дата звернення: 03.04.24)

3. Segeda I., Kotsiuba V., Shushura O., Bokovets V., Koval N., Kalizhanova A. Decentralized platform for financing charity projects. *Informatyka, Automatyka, Pomiary W Gospodarce I Ochronie Środowiska*, 2024. Vol. 14 No 3, P. 129–134. https://doi.org/10.35784/iapgos.6140 (date of access: 02.10.2024).

4. Ciranni M., Odone F., Pastore V. P. Anomaly detection in feature space for detecting changes in phytoplankton populations. *Frontiers in Marine Science*. 2023. Vol. 10. https://doi.org/10.3389/fmars.2023.1283265 (date of access: 05.06.2024).

5. Raza K. Nature-Inspired Intelligent Computing Techniques in Bioinformatics. Singapore: Springer, 2022. 328 c.

6. Vanini P. Online payment fraud: from anomaly detection to risk management. *Financial Innovation*. 2023. Vol. 9. https://doi.org/10.1186/s40854-023-00470-w (date of access: 23.08.2024).

7. Hosseinzadeh M. Improving security using SVM-based anomaly detection: issues and challenges. *Soft Computing*. 2021. Vol. 25. P. 3195–3223. https://doi.org/10.1007/s00500-020-05373-x (date of access: 18.06.2024).

8. Zanatta Bruno G., B. Chaves Rodrigues K., Vieira Cardoso K., Luz Correa S., Bonato Both C. Anomaly Detection in Cloud-native B5G Systems using Observability and Machine Learning COTS Solutions. *Journal of Internet Services and Applications*. 2023. Vol. 14(1). P. 189–199. https://doi.org/10.5753/jisa (date of access: 28.06.2024).

9. Andrews J. Comparing Outlier Detection Methods. Medium. URL: https://towardsdatascience.com/comparing-outlier-detection-methods-956f4b097061 (date of access: 10.08.2024).

10. Iris Species. Kaggle: Your Machine Learning and Data Science Community. URL: https://www.kaggle.com/datasets/uciml/iris (date of access: 13.10.2024).

INFORMATION SYSTEM FOR ANOMALY DETECTION IN DATA BASED ON MACHINE LEARNING METHODS

Oleksii Shushura
Doctor of Technical Sciences, Associate Professor, Professor at the Department of Digital Technologies in Energy
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37 Beresteyskyi ave., Kyiv, Ukraine, 03056, shushura_oleksii@lll.kpi.ua
ORCID: 0000-0003-3200-720X

Yelyzaveta Moroz
Master at the Department of Digital Technologies in Energy
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37 Beresteyskyi ave., Kyiv, Ukraine, 03056, yelyzaveta.mor@gmail.com
ORCID: 0009-0006-8865-4173

Iryna Segeda
Candidate of Science (Eng.), Associate Professor at the Department of Digital Technologies in Energy
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute", 37 Beresteyskyi ave., Kyiv, Ukraine, 03056, Segeda.Iryna@lll.kpi.ua
ORCID: 0000-0003-1958-4985

Liudmyla Asieieva
PhD, Associate Professor at the Department of Computer Engineering
State University of Information and Communication Technologies, 7 Solomianska str., Kyiv, Ukraine, 03110, aseewal@i.ua
ORCID: 0000-0001-5954-4211

The article discusses an information system for anomaly detection in large datasets using machine learning methods. The relevance of the problem is determined by the growing volumes of data and the complexity of their analysis, which necessitates the development of automated solutions for identifying deviations in data, such as fraud, malfunctions, or other atypical situations. Anomaly detection is a crucial tool in areas such as cybersecurity, financial monitoring, industrial diagnostics, and medical analytics.

Requirements for the anomaly detection system using machine learning methods have been formulated, which are formalized in the form of a UML use case diagram. Its structure has been designed, development tools selected, software developed in Python, and testing conducted. The system is based on several machine learning algorithms, including Isolation Forest, Local Outlier Factor, and DBSCAN, which ensure the effectiveness and accuracy of anomaly detection in various application tasks. The effectiveness of the system tested on benchmark datasets showed a high level of accuracy in detecting anomalies.

The proposed system allows users to analyze data and detect anomalies without the need for deep knowledge in programming or algorithm configuration. It automatically performs analysis by comparing the results of different models and provides visualization options for better understanding of the detected anomalies. The results of the study demonstrated that the system can quickly and accurately identify anomalous data, significantly reducing analysis time and enhancing decision-making efficiency. The proposed solution can become an important tool for automating anomaly detection processes in large datasets, which is critically important in the modern context of increasing information volumes. The findings can be utilized by information technology developers working in the field of data analysis.

Key words: anomaly detection, machine learning, information system, Local Outlier Factor, Isolation Forest, DBSCAN, data analysis.

REFERENCES

1. Shutenko, V. (2024). AI anomaly detection: Best tools and use cases. TechMagic. Retrieved from https://www.techmagic.co/blog/ai-anomaly-detection/.

2. Shushura, O., Levchenko, L., & Savchuk, A. (2023). Data analysis based on neural networks using microservice architecture. *Bulletin of KrNU named after Mykhailo Ostrohradskyi,* 3(140), 80–85. https://doi.org/10.32782/1995-0519.2023.3.10

3. Segeda, I., Kotsiuba, V., Shushura, O., Bokovets, V., Koval, N., & Kalizhanova, A. (2024). Decentralized platform for financing charity projects. *Informatics, Automation, Measurements in Economy and Environmental Protection*, 14(3), 129–134. https://doi.org/10.35784/iapgos.6140.

4. Ciranni, M., Odone, F., & Pastore, V. P. (2023). Anomaly detection in feature space for detecting changes in phytoplankton populations. *Frontiers in Marine Science*, 10. https://doi.org/10.3389/fmars.2023.1283265.

5. Raza, K. (2022). Nature-inspired intelligent computing techniques in bioinformatics. Singapore: Springer.

6. Vanini, P. (2023). Online payment fraud: From anomaly detection to risk management. *Financial Innovation*, 9. https://doi.org/10.1186/s40854-023-00470-w.

7. Hosseinzadeh, M. (2021). Improving security using SVM-based anomaly detection: Issues and challenges. *Soft Computing*, 25, 3195–3223. https://doi.org/10.1007/s00500-020-05373-x.

8. Zanatta, B. G., Chaves Rodrigues, K. B., Vieira Cardoso, K., Luz Correa, S., & Bonato Both, C. (2023). Anomaly detection in cloud-native B5G systems using observability and machine learning COTS solutions. *Journal of Internet Services and Applications*, 14(1), 189–199. https://doi.org/10.5753/jisa.

9. Andrews, J. (2024). Comparing outlier detection methods. Medium. Retrieved from https://towardsdatascience.com/comparing-outlier-detection-methods-956f4b097061.

10. Iris Species. (2024). Kaggle: Your machine learning and data science community. Retrieved from https://www.kaggle.com/datasets/uciml/iris.

Стаття надійшла 24.09.2024