УДК 004.85:657.6

А. С. Лановик
А. А. Яровий

БАГАТОКОМПОНЕНТНА СИСТЕМА ВИЯВЛЕННЯ ТА АНАЛІЗУ АНОМАЛІЙ У ФІНАНСОВИХ ЗВІТАХ НА ОСНОВІ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

У роботі запропоновано гібридну архітектуру системи виявлення та аналізу аномалій у фінансових даних на основі інтеграцій методів машинного навчання. Розглянуто чотири класи підходів, критерії їх порівняння та логіка формування загального конвеєра, який поєднує в собі Isolation Forest, LSTM Autoencoder, DBSCAN та нечітку експертну систему ANFIS з механізмом перевірки аналітика.

Ключові слова: аномалії, фінансова звітність, штучний інтелект, Isolation Forest, LSTM Autoencoder, DBSCAN, нечітка експертна система ANFIS, human-in-the-loop, виявлення шахрайства, фінансовий аналіз.

Abstract

The work substantiates the choice and describes the principles of integration of machine learning methods for building a hybrid system for detecting and analyzing anomalies in financial data. Four classes of approaches, their comparison criteria and the logic of forming a common pipeline, which combines Isolation Forest, LSTM Autoencoder, DBSCAN and the fuzzy ANFIS expert system with the analyst verification mechanism, are considered.

Keywords: anomalies, financial reporting, artificial intelligence, Isolation Forest, LSTM Autoencoder, DBSCAN, fuzzy ANFIS expert system, human-in-the-loop, fraud detection, financial analysis.

Вступ

За даними звіту ACFE, у 2024 році організації втрачають у середньому 5% річного доходу через шахрайство, а фінансова звітність залишається одним із ключових векторів маніпуляцій [1]. Попередня система на основі одного детектора Isolation Forest мала три фундаментальні обмеження: відсутність урахування часових залежностей, неможливість розрізнити типи аномалій та відсутність механізму накопичення знань [2].

Метою даного дослідження є подолання цих обмежень шляхом розробки гібридної архітектури, що поєднує кілька взаємодоповнюючих методів. Вибір цих методів і принципи їх інтеграції обґрунтовані нижче.

Аналіз та обґрунтування вибору методів

Під час дослідження було проаналізовано чотири класи підходів до виявлення аномалій у фінансових часових рядах: статистичні методи, методи машинного навчання, методи глибинного навчання та підходи гібридних множин. Порівняння проводилося за такими критеріями: здатність виявляти точкові та контекстуальні аномалії, обчислювальна складність, вимоги до розмітки даних та можливість інтерпретації результатів [3].

Статистичні методи (Z-score, ARIMA, GARCH) базуються на моделюванні розподілу нормальних спостережень. Але усі три методи передбачають нормальність розподілу та лінійність залежностей. Фінансові ринки мають товсті хвости розподілу та групування волатильності (volatility clustering), що суперечить цим припущенням. Вказані методи не здатні виявляти нелінійні аномалії та погано масштабуються на багатовимірних даних.

Методи на основі щільності (LOF, One-Class SVM) порівнюють локальну щільність точки з її сусідами. One-Class SVM будує гіперплощину, що відокремлює нормальні спостереження від

аномалій. Обидва методи мають квадратичну обчислювальну складність $O(n^2)$, що є слабким місцем цих методів для великих фінансових датасетів. Крім того, вони чутливі до вибору гіперпараметрів і не враховують часову структуру даних.

Isolation Forest (IF) вибрано як перший базовий детектор через параметр лінійної складності O (n), нечутливість у масштабі ознак і відсутність потреби в позначених даних. Алгоритм виділяє аномалії шляхом випадкового розподілу простору ознак: аномальні точки виділяються швидше та отримують нижчу нормалізовану оцінку [4].

Autoencoder LSTM (LSTM AE) вибрано як другий базовий детектор для подолання основного обмеження IF – нездатності врахувати часові залежності. Архітектура поєднує повторювані рівні LSTM (кодування часової послідовності у векторі фіксованої довжини) і декодер (реконструкція вихідної послідовності). Модель тренується виключно на нормальних даних: контекстні аномалії – шаблони, в яких кожне спостереження не є відхиленням, але послідовність є ненормальною, вони отримують набагато більшу помилку реконструкції [5]. Таким чином, IF і Autoencoder LSTM виявляють принципово різні класи аномалій, що є основним аргументом на користь їх поєднання.

Узагальнений порівняльний аналіз розглянутих методів наведено в табл. 1.

Таблиця 1 – Порівняльний аналіз методів виявлення аномалій

Критерій	**Z-score / ARIMA**	**LOF / OCSVM**	**Isolation Forest**	**LSTM Autoencoder**	**Ансамбль IF + LSTM AE**
Точкові аномалії	+	+	+	–	+
Часові (контекстуальні) аномалії	±	–	–	+	+
Обчислювальна складність	O(n)	$O(n^2)$	O(n)	O(n·w)	O(n·w)
Потреба у мічених даних	–	–	–	–	–
Нечутливість до масштабу	–	±	+	+	+
Інтерпретованість результатів	+	±	±	–	+ (ANFIS)

Проте, виявлення аномалій є лише першим етапом: щоб зрозуміти їх природу та систематизувати за типами, виявлені відхилення необхідно згрупувати. Для цього до системи включено алгоритм кластеризації DBSCAN. На відміну від K-Means, DBSCAN не потребує попередньо визначеної кількості кластерів і природним чином ізолює аномалії шуму, які не належать до жодного кластера [6]. Косинусна подібність, яка є більш стійкою в багатовимірних просторах, ніж евклідова відстань, використовується як метрика для відстані між аномальними точками в просторі 17-вимірних ознак.

Після того як аномалії виявлено та згруповано, постає задача їх інтерпретації: необхідно автоматично встановити, які комбінації фінансових показників характерні для кожного типу відхилення. Для цього використовується алгоритм Apriori – алгоритм пошуку частих наборів елементів у даних. Кластери аномалій є основою для автоматичної генерації правил бази знань методом Апріорі: дискретизовані характеристики кожного кластера вважаються транзакціями і алгоритм виявляє їх найбільш характерні комбінації.

Автоматична генерація інтерпретованих правил виду "умова характеристик – тип аномалії" усуває третє обмеження попередньої системи, а саме: відсутність механізму накопичення знань. Правила слугують основою для навчання системи ANFIS без потреби у розмічених даних.

Ключовою перевагою ANFIS перед класифікаторами нейронних мереж є поєднання здатності нейронних мереж апроксимувати нелінійні залежності з лінгвістичною інтерпретацією нечіткої логіки: кожне правило в системі відповідає зрозумілій умові форми "якщо волатильність висока і RSI перевищено, тоді тип аномалії – є волатильний сплеск" [7].

У роботі параметри функцій належності ANFIS визначаються на основі порогів фінансової літератури та технічного аналізу, а адаптація відбувається не шляхом градієнтного навчання (що вимагало б мічених даних), а через зворотний зв'язок аналітика в механізмі "human-in-the-loop". Це принципова відмінність від традиційного впровадження ANFIS і робить систему придатною для використання на немаркованих фінансових даних.

Адаптивна база знань формується з правил форми "умови характеристики – тип аномалії – рекомендація". Нові правила автоматично генеруються алгоритмом Apriori на основі кластерів DBSCAN, після чого аналітик підтверджує або відхиляє їх шляхом діалогу. Ваги правил оновлюються зваженим методом: правила, підтверджені аналітиком, отримують більшу вагу, відхилені правила – вага зменшується. Після досягнення порогової кількості підтверджень система переходить на автоматичну класифікацію без втручання аналітика.

Архітектура гібридної системи

Загальна архітектура системи побудована як послідовний конвеєр (pipeline), де кожен компонент виконує строго визначену функцію та передає результати наступному етапу. Стратегія агрегування детекторів базується на зваженій сумі нормалізованих оцінок аномальності (вага 0,5 + 0,5), що зберігає кількісну міру довіри кожного детектора на відміну від бінарного голосування більшістю. Теоретична оцінка ефективності запропонованої архітектури базується на порівнянні з попередньою системою (Silhouette Score = 0,68, 17 виявлених аномалій на 327 торгових днях). Ансамблеве поєднання IF та LSTM AE теоретично забезпечує повне покриття обох класів аномалій: IF виявляє точкові відхилення з лінійною складністю O(n), тоді як LSTM AE опрацьовує часові залежності з складністю O(n·w), недоступні для статичних детекторів. Розширення простору ознак з 7 до 17 очікувано підвищить розрізнювальну здатність системи. Верифікація запропонованих покращень на реальних даних фондового ринку є предметом подальшого тестування в межах розвитку даного дослідження.

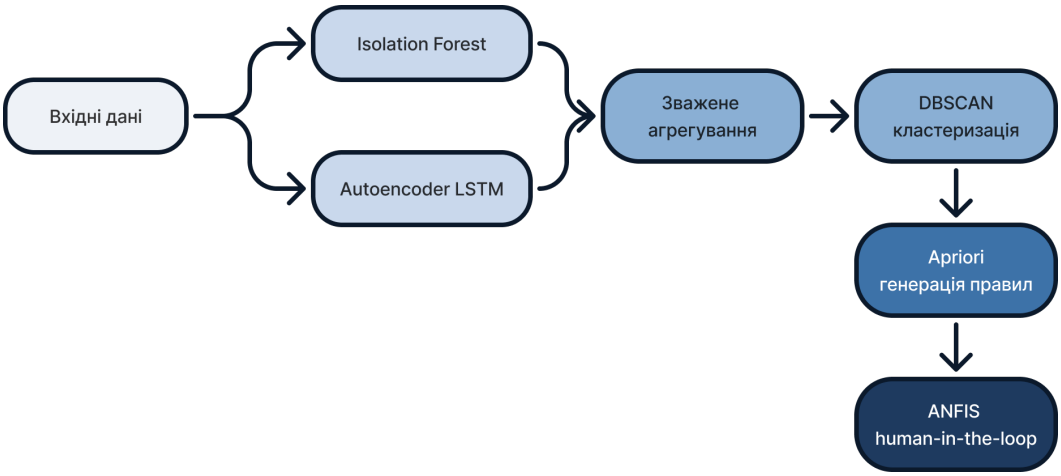

Рисунок 1 – Структурно-логічна схема системи виявлення та аналізу аномалій у фінансових звітах

На рисунку 1 показано, що кожен блок конвеєра виконує певну роль:

- Вхідні фінансові дані – нормалізовані часові ряди у 17-вимірному просторі ознак (без попередньої розмітки).
- Isolation Forest – виявляє точкові статистичні відхилення, формує оцінку аномальності для кожного спостереження.
- LSTM Autoencoder – виявляє контекстуальні часові аномалії через помилку реконструкції послідовності.
- Зважене агрегування – об'єднує оцінки двох детекторів (0,5 + 0,5); спостереження з комбінованою оцінкою вище порогового значення позначається як аномалія.
- DBSCAN – кластеризує виявлені аномалії за косинусною подібністю; шумові точки ізолюються окремо.
- Apriori, ANFIS (Human-in-the-Loop) – алгоритм Apriori генерує інтерпретовані правила з кластерів; аналітик підтверджує правила через діалог; ANFIS класифікує тип аномалії та формує рекомендацію.

Висновки

Обрана комбінація методів та принципи їх інтеграції забезпечують комплексне вирішення проблеми виявлення та аналізу аномалій у фінансових даних. Кожен компонент системи заповнює певний проміжок, який не може заповнити жоден інший елемент у наборі:

- Isolation Forest – ефективне та неконтрольоване виявлення точкових статистичних відхилень лінійної складності;
- LSTM Autoencoder – виявлення контекстних часових аномалій, недоступних для статичних детекторів;
- DBSCAN – групування аномалій без апріорного регулювання кількості кластерів з вибором точок шуму;
- Apriori – автоматична генерація інтерпретованих правил бази знань на основі кластерів;
- ANFIS – лінгвістична інтерпретація та класифікація аномалій з адаптацією через відгуки аналітиків.

Жодне із передбачених комерційних рішень (IBM Cognos Analytics, SAP Financial Compliance, Oracle OFSAA) не реалізує всі згадані можливості одночасно [8 – 10]. Розроблена архітектура є відкритою та розширюваною, що дозволяє адаптувати її до різних фінансових інструментів і ринків без необхідності попереднього рейтингу навчальної вибірки.

Розроблена система орієнтована на аудиторські підрозділи та compliance-відділи фінансових установ, що працюють з немаркованими часовими рядами.

Список використаних джерел

1. Association of Certified Fraud Examiners. Occupational Fraud 2024: A Report to the Nations [Електронний ресурс]. – Austin : ACFE, 2024. – 120 с. – Режим доступу: https://www.acfe.com/-/media/files/acfe/pdfs/rttn/2024/2024-report-to-the-nations.pdf.
2. Лановик А. С., Яровий А. А. Виявлення та аналіз аномалій у фінансових звітах на основі гібридного ансамблевого підходу Isolation Forest та LSTM Autoencoders / А. С. Лановик, А. А. Яровий : Матеріали LV Всеукраїнської науково-технічної конференції факультету інтелектуальних інформаційних технологій та автоматизації. – В. : ВНТУ, 2026. – [Електронний ресурс] – Режим доступу: https://conferences.vntu.edu.ua/index.php/all-fksa/all-fksa-2026/paper/view/28322.
3. Chandola V., Banerjee A., Kumar V. Anomaly Detection: A Survey. ACM Computing Surveys. – 2009. – Vol. 41, No. 3. – Article 15.
4. Malhotra P. et al. Long Short Term Memory Networks for Anomaly Detection in Time Series. Proceedings of ESANN 2015. – P. 89–94.
5. Ester M., Kriegel H.-P., Sander J., Xu X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. – Portland, 1996. – P. 226 – 231.

6. Jang J.-S. R. ANFIS: Adaptive-Network-Based Fuzzy Inference System. IEEE Transactions on Systems, Man, and Cybernetics. – 1993. – Vol. 23, No. 3. – P. 665–685.
7. Agrawal R., Srikant R. Fast Algorithms for Mining Association Rules. Proceedings of the 20th International Conference on Very Large Data Bases (VLDB'94). – San Francisco, 1994. – P. 487 – 499.
8. IBM Cognos Analytics. Forecasting and anomaly detection [Електронний ресурс]. – 2024. – Режим доступу: https://www.ibm.com/docs/en/cognos-analytics.
9. SAP. Finance NXT Anomaly Detection by Bosch Global Software Technologies [Електронний ресурс]. – Режим доступу: https://www.sap.com/products/technology-platform/partners/bosch-global-software-technologies-boscol-collaboration-app-for-sap-cloud-products.html#features-section.
10. Oracle. Anomaly Detection [Електронний ресурс]. – Режим доступу: https://docs.oracle.com/en/database/oracle/machine-learning/oml4sql/23/dmcon/anomaly-detection.html

Лановик Анастасія Сергіївна – студентка групи КН-25М, факультет інформаційних інтелектуальних технологій та автоматизації, Вінницький національний технічний університет, Вінниця.

Яровий Андрій Анатолійович – д-р техн. наук, професор кафедри комп'ютерних наук, Вінницький національний технічний університет, м. Вінниця.