УДК 004.62

ПОРІВНЯЛЬНИЙ АНАЛІЗ АЛГОРИТМІВ ДЛЯ ВИЯВЛЕННЯ АНОМАЛІЙ У ЧАСОВИХ РЯДАХ МЕТРИК ІНФОРМАЦІЙНИХ СИСТЕМ

Вітрук Д.О. [1]; Платонов В.В. [2] ; Платонова Є.В. [3]
[1,2,3]Одеський національний університет імені І.І. Мечникова
[1,2,3] Україна, Одеса
[1] dmytro.vitruk@stud.onu.edu.ua; [2]v.platonov@onu.edu.ua;[3]jane.platonova@onu.edu.ua

Анотація. *У роботі проведено комплексне порівняння шести методів виявлення аномалій у часових рядах системних метрик (CPU, RAM, disk, network). Для оцінки ефективності були використані статистичні підходи (Moving Average), методи машинного навчання (Isolation Forest, Local Outlier Factor), кластеризація (K-Means) та рекурентні нейронні мережі (RNN). Алгоритми тестувалися на трьох типах датасетів: контрольованому, синтетичному та реальному. Результати показали, що RNN демонструє найвищу чутливість та надійність у виявленні складних часових залежностей, що робить цей метод пріоритетним для моніторингу критичних систем.*

Ключові слова: *аналіз часових рядів; виявлення аномалій; системні метрики; машинне навчання; Isolation Forest; RNN; моніторинг продуктивності.*

Вступна частина. Моніторинг метрик інформаційних систем, таких як завантаження процесора, пам'яті та мережевого трафіку, є критичним завданням для забезпечення їх стабільності та безпеки. Аномалії в цих метриках можуть слугувати ранніми індикаторами збоїв, проблем з продуктивністю або навіть кібератак. Ефективність систем моніторингу безпосередньо залежить від обраного алгоритму виявлення аномалій, проте вибір оптимального методу ускладнюється різноманітністю підходів та специфікою даних.

Метою дослідження є порівняння результативності різних класів алгоритмів для виявлення аномалій у часових рядах метрик інформаційної системи та розробка рекомендацій щодо їх застосування. Для досягнення мети було вирішено такі завдання:

1. Проаналізовано шість алгоритмів: коефіцієнти кореляції, Isolation Forest, Local Outlier Factor (LOF), Moving Average, K-Means та рекурентні нейронні мережі (RNN).
2. Проведено тестування алгоритмів на контрольованому, синтетичному та реальному датасетах.
3. Надано критерії оцінки та рекомендації щодо вибору оптимального алгоритму.

Основна частина. Для тестування було використано три датасети системних метрик: **контрольований** (згенерований за математичними формулами), **синтетичний** (на основі випадкових значень) та **реальний** (зібраний під час моніторингу системи з інтервалом 30 секунд).

Кожен із шести алгоритмів було застосовано для аналізу часових рядів метрик cpu, cpu idle, ram, disk, network rx та network tx.

Аналіз показав суттєві відмінності в ефективності методів:

- **Коефіцієнти кореляції** виявились ефективними для виявлення змін у взаємозв'язках між метриками на нестабільних даних, але малоінформативними у стабільних системах.

• **Isolation Forest** продемонстрував високу ефективність у виявленні як пікових навантажень, так і нетипових спадів активності, підтвердивши свою надійність для даних різної природи.

• **LOF** добре ідентифікував локальні аномалії, пов'язані з різкими змінами, але показав надмірну чутливість до масштабування даних у реальному датасеті.

• **Moving Average** виявився простим та швидким методом для фіксації значних відхилень від тренду, проте згладжував короткочасні аномалії.

• **K-Means** ефективно кластеризував мережеві дані, виявляючи аномальні режими роботи, але його точність залежить від чіткості структури даних.

• **Recurrent Neural Networks (RNN)** показали найвищу чутливість та точність на всіх трьох датасетах. Завдяки здатності враховувати часові залежності, RNN коректно ідентифікували як явні, так і приховані аномалії, які інші методи пропускали.

Висновки. Проведене дослідження підтвердило, що вибір методу для виявлення аномалій повинен ґрунтуватися на характері даних та вимогах до точності. Для критично важливих систем, де необхідно враховувати складні часові залежності, **рекурентні нейронні мережі (RNN) є найбільш ефективним та надійним інструментом**. Для менш складних завдань або для первинного аналізу можуть бути рекомендовані **Isolation Forest** як універсальний та швидкий метод, і **Moving Average** для базового моніторингу.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

[1] Яровий, А. Т., & Страхов, Є. М. (2019). *Аналіз часових рядів: навчальні матеріали ОНУ ім. І. І. Мечникова*. Retrieved from https://onu.edu.ua/pub/bank/userfiles/files/fmfit/nachalni_materiali/navch_mat_111_mag/analiz_chasovykh_ryadiv.pdf

[2] DataCamp. (2024, September 25). *Isolation Forest Guide: Explanation and Python Implementation*. Retrieved from https://www.datacamp.com/tutorial/isolation-forest

[3] DataCamp. (2022, March 16). *Recurrent Neural Network Tutorial (RNN)*. Retrieved from https://www.datacamp.com/tutorial/tutorial-for-recurrent-neural-network

Vitruk D.O., Platonov V.V.,Platonova Y.V.

A comparative analysis of algorithms for anomaly detection in time series of information system metrics

Abstract. *This paper provides a comprehensive comparison of six methods for anomaly detection in a time series of system metrics (CPU, RAM, disk, network). To evaluate their effectiveness, statistical approaches (Moving Average), machine learning methods (Isolation Forest, Local Outlier Factor), clustering (K-Means), and Recurrent Neural Networks (RNN) were used. The algorithms were tested on three types of datasets: controlled, synthetic, and real. The results showed that RNN demonstrates the highest sensitivity and reliability in detecting complex temporal dependencies, making it a priority method for monitoring critical systems.*

Keywords: *time series analysis; anomaly detection; system metrics; machine learning; Isolation Forest; RNN; performance monitoring.*