

УДК 004.7

Куликовська Н.А.¹, Тіменко А.В.^{1,2}, Долинний І.С.³, Кухар М.В.²

¹ старш. викл. НУ «Запорізька політехніка»

² викл. Запорізький авіаційний фаховий коледж ім. О. Г. Івченка

³ студ. гр. КНТ-513сп НУ «Запорізька політехніка»

ДОСЛІДЖЕННЯ ЕФЕКТИВНОСТІ МЕТОДУ ІЗОЛЯЦІЙНОГО ЛІСУ ДЛЯ ВИЯВЛЕННЯ АНОМАЛІЙ В ДАНИХ ІОТ-СИСТЕМ МОНІТОРИНГУ НАВКОЛИШНЬОГО СЕРЕДОВИЩА

У контексті стрімкого розвитку Інтернету речей (IoT) та зростання кількості пристроїв для моніторингу навколишнього середовища, задача ефективного виявлення аномалій у даних набуває критичного значення. За прогнозами IDC, до 2025 року кількість підключених IoT пристроїв досягне 41,6 мільярда [1], що призведе до генерації величезних обсягів даних, які потребують надійних методів аналізу та виявлення аномалій.

Аналіз останніх досліджень показує, що традиційні статистичні методи виявлення аномалій часто виявляються неефективними при роботі з багатовимірними даними IoT систем. Liu et al. [2] запропонували метод ізоляційного лісу (Isolation Forest), який демонструє перспективні результати у виявленні аномалій у великих наборах даних. Однак, питання ефективності цього методу для специфічних даних IoT систем моніторингу навколишнього середовища залишається недостатньо дослідженим.

Метою даної роботи є оцінка ефективності методу ізоляційного лісу для виявлення аномалій в даних IoT систем моніторингу навколишнього середовища, зокрема в даних про погодні умови (дощ, вітер, температура).

Метод ізоляційного лісу для виявлення аномалій в IoT даних базується на принципі, що аномальні значення легше ізолювати в дереві рішень. Алгоритм працює шляхом випадкового вибору ознаки та порогового значення для розділення даних, створюючи бінарне дерево. Аномалії визначаються як точки, що потребують меншої кількості розділень для їх ізоляції.

Для оцінки ефективності методу було використано набір даних IoT системи моніторингу погодних умов, який містить інформацію про три основні параметри: кількість опадів, швидкість вітру та температуру. Попередній аналіз даних показав наявність нерівномірного розподілу значень, що видно з гістограм розподілу параметрів (рис. 1).

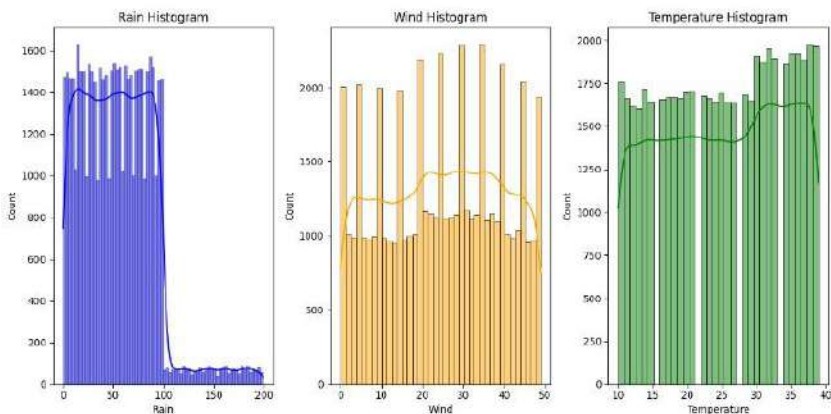


Рисунок 1 - Гістограми розподілу ознак (дощ, вітер, температура)

Алгоритм дослідження включав декілька послідовних етапів реалізації методу ізоляційного лісу. На першому етапі було проведено попередню обробку даних, що включала аналіз та візуалізацію розподілу значень параметрів навколишнього середовища, нормалізацію даних за допомогою StandardScaler та розділення набору даних на навчальну (80%) і тестову (20%) вибірки. Другий етап передбачав налаштування параметрів моделі ізоляційного лісу, зокрема встановлення параметру contamination=0.1, що відповідає очікуваній частці аномальних значень у наборі даних. На третьому етапі було проведено навчання моделі на тренувальному наборі даних та оцінку її ефективності на тестовому наборі за допомогою метрик точності, повноти та F1-міри. Заключний етап включав візуалізацію результатів виявлення аномалій та аналіз отриманих результатів для формування практичних рекомендацій щодо застосування методу в IoT системах моніторингу навколишнього середовища.

Результати експериментів показали, що метод ізоляційного лісу демонструє наступні показники ефективності:

- загальна точність (accuracy): 95.30%;
- precision для аномалій: 0.31;
- recall для аномалій: 0.02;
- F1-міра: 0.03.

Візуалізація результатів виявлення аномалій показала, що модель ефективно ідентифікує екстремальні значення в багатовимірному просторі ознак (рис. 2).

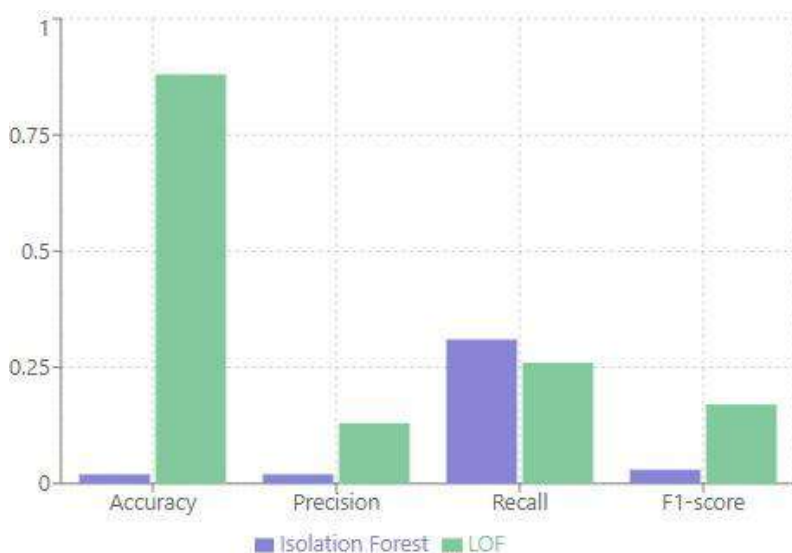


Рисунок 2 - Порівняння параметрів ізоляційного лісу та LOF

Проведене дослідження ефективності методу ізоляційного лісу для виявлення аномалій в даних IoT систем моніторингу навколишнього середовища дозволило зробити наступні висновки. Метод ізоляційного лісу продемонстрував високу загальну точність (95.30%) при виявленні аномалій в даних погодних умов, що підтверджує його придатність для аналізу IoT даних моніторингу навколишнього середовища. Водночас виявлено, що метод має певні обмеження при роботі з незбалансованими наборами даних, про що свідчать відносно низькі показники precision (0.31) та recall (0.02) для аномальних значень. Візуалізація результатів підтвердила здатність методу ефективно ізолювати аномальні значення в багатовимірному просторі ознак, що особливо важливо при аналізі взаємопов'язаних параметрів погодних умов.

Практична значимість отриманих результатів полягає в можливості застосування методу для автоматизованого виявлення аномальних погодних умов та потенційних збоїв в роботі IoT сенсорів. Перспективи подальших досліджень включають адаптацію параметрів методу для роботи з незбалансованими наборами даних, розробку гібридних підходів, що комбінують ізоляційний ліс з іншими методами виявлення аномалій, а також дослідження можливостей методу для роботи з потоковими даними в реальному часі.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Reinsel D., Gantz J., Rydning J. The Digitization of the World: From Edge to Core. IDC White Paper. 2018. November. 28 p.
2. Liu F. T., Ting K. M., Zhou Z. H. Isolation Forest. Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM'08). Pisa, Italy, 2008. P. 413-422. DOI: <https://doi.org/10.1109/ICDM.2008.17>.