https://doi.org/10.31891/2219-9365-2025-84-32
УДК 004.023

ПИРЧ Олена
Хмельницький національний університет
https://orcid.org/0009-0002-6485-4462
e-mail: oleksukolena@gmail.com
МОСТОВИЙ Сергій
Хмельницький національний університет
https://orcid.org/0000-0002-9505-3206
e-mail: serhii.mostovyi@khmnu.edu.ua

ПОРІВНЯЛЬНИЙ АНАЛІЗ КЛАСИЧНИХ І МАШИННИХ МЕТОДІВ ВИЯВЛЕННЯ АНОМАЛІЙ У СЛАБОНАВАНТАЖЕНИХ МЕРЕЖАХ

У даній статті проведено порівняльний аналіз сучасних методів виявлення аномалій у слабонавантажених комп'ютерних мережах. До аналізованих методів віднесено: алгоритм Isolation Forest, метод One-Class SVM, щільнісну кластеризацію DBSCAN та нейромережевий підхід на основі LSTM-Autoencoder. Порівняння методів проведено за групами характеристик: точність виявлення аномалій, обчислювальна складність, вимоги до навчальних даних та адаптивність до специфіки слабонавантажених мереж.

У результаті проведеного аналізу виявлено, що вибір оптимального методу залежить від конкретних умов застосування. Для систем реального часу з обмеженими ресурсами найбільш придатним є Isolation Forest, тоді як для складних багатовимірних аномалій перевагу має LSTM-Autoencoder.

Ключові слова: виявлення аномалій, слабонавантажені мережі, машинне навчання, кібербезпека, аналіз мережного трафіку.

PYRCH Olena, MOSTOVYI Serhii
Khmelnitskyi National University

COMPARATIVE ANALYSIS OF CLASSIC AND MACHINE METHODS FOR ANOMALIES DETECTION IN LIGHTLY LOADED NETWORKS

This article presents a comprehensive comparative analysis of classical and machine learning–based methods for anomaly detection in lightly loaded computer networks. Such networks, including small enterprise infrastructures, industrial Internet of Things systems, sensor networks, and remote monitoring environments, are characterized by low traffic intensity, limited statistical data, and constrained computational resources. These features significantly reduce the effectiveness of traditional anomaly detection approaches designed for high-load network environments.

The study examines four widely used methods: Isolation Forest, One-Class Support Vector Machine (One-Class SVM), DBSCAN density-based clustering, and an LSTM-based Autoencoder. The comparison is conducted according to key evaluation criteria, including anomaly detection accuracy, computational complexity, training data requirements, adaptability to low-data scenarios, and interpretability of results. Particular attention is paid to the ability of each method to operate under conditions of sparse observations and high variability of normal network behavior.

The analysis demonstrates that no single method is universally optimal for all lightly loaded network scenarios. Isolation Forest shows the best balance between detection efficiency and computational cost, making it suitable for real-time systems with limited resources. One-Class SVM provides high detection accuracy for complex decision boundaries but requires careful parameter tuning and greater computational effort. DBSCAN offers strong interpretability and effectively detects cluster-based anomalies, although its performance depends heavily on parameter selection. LSTM-Autoencoder achieves superior results in detecting complex temporal anomalies but demands substantial training data and computational resources, which limits its applicability in typical low-load environments.

The results highlight the importance of selecting anomaly detection methods based on specific operational constraints and data characteristics. The paper also emphasizes the potential of hybrid and ensemble approaches to improve robustness and detection reliability in lightly loaded networks. The findings contribute practical guidelines for designing efficient anomaly detection systems in resource-constrained network environments.

Keywords: anomaly detection, lightly loaded networks, machine learning, cybersecurity, network traffic analysis.

Стаття надійшла до редакції / Received 01.11.2025
Прийнята до друку / Accepted 02.12.2025

ПОСТАНОВКА ПРОБЛЕМИ У ЗАГАЛЬНОМУ ВИГЛЯДІ ТА ЇЇ ЗВ'ЯЗОК ІЗ ВАЖЛИВИМИ НАУКОВИМИ ЧИ ПРАКТИЧНИМИ ЗАВДАННЯМИ

Комп'ютерні мережі з низьким навантаженням утворюють особливу категорію інформаційних систем, що характеризуються низькою інтенсивністю передачі даних. Прикладами таких мереж є мережі малих підприємств, промислові системи Інтернету речей, сенсорні мережі критичної інфраструктури, телеметричні та дистанційні системи спостереження. Особливістю таких мереж є те, що обсяг переданих даних може становити від декількох десятків до сотень подій за період спостереження, що значно менше, ніж у типових корпоративних або телекомунікаційних мережах.

Проблема виявлення аномалій стає особливо актуальною в мережах з низьким навантаженням з

кількох причин: по-перше, недостатня кількість статистичних даних ускладнює використання традиційних статистичних підходів, які вимагають репрезентативної вибірки для точного моделювання нормальної поведінки системи; по-друге, окремі аномалії також можуть мати критичний вплив на роботу системи в разі недовантажених мереж, що підвищує вимоги до чутливості методів виявлення; нарешті, обмежені обчислювальні ресурси багатьох вузлів у таких мережах вимагають використання ефективних алгоритмів з низькою обчислювальною складністю.

Класично, системи виявлення вторгнень та аномалій, які були розроблені для мереж з великим обсягом даних, показали низьку ефективність у середовищах з невеликим навантаженням. Основними проблемами є високий рівень помилкових спрацьовувань через недостатню статистичну потужність, нездатність адаптуватися до швидких змін структури при невеликій кількості спостережень, труднощі з досягненням оптимально встановлених порогових значень в умовах високої варіативності нормальної поведінки.

Таким чином, актуальним є завдання дослідження та порівняння існуючих методів виявлення аномалій з точки зору їх застосовності до специфічних умов слабонавантажених мереж, що і становить предмет даного дослідження.

ФОРМУЛЮВАННЯ ЦІЛЕЙ СТАТТІ

Метою статті є проведення комплексного порівняльного аналізу класичних і сучасних методів машинного навчання виявлення аномалій в умовах низьконавантажених комп'ютерних мереж.

– Для досягнення поставленої мети необхідно виконати наступні завдання:

– Надати теоретичні та математичні пояснення щодо принципів роботи методів Isolation Forest, One-Class SVM, DBSCAN та LSTM-Autoencoder.

– Дослідження властивостей кожного методу, що застосовується в умовах обмежених навчальних даних, типових для мереж з недостатніми ресурсами.

– Провести порівняльний аналіз методів за критеріями точності виявлення аномалій, обчислювальної складності, вимог до обсягу навчальних даних, здатності до адаптації та інтерпретовності результатів.

– Визначити переваги та недоліки кожного методу стосовно різних типів аномалій, поширених у мережах з низьким навантаженням.

Об'єктом дослідження є процес виявлення аномалій у мережевому трафіку слабонавантажених комп'ютерних мереж. Предметом дослідження є методи машинного навчання та їх характеристики в контексті застосування для виявлення аномалій при обмежених обсягах даних. Наукова новизна роботи полягає у систематизації та порівняльному аналізі сучасних методів виявлення аномалій саме для умов слабонавантажених мереж, що раніше не було предметом комплексного дослідження.

ОГЛЯД ІСНУЮЧИХ РІШЕНЬ

Методи виявлення аномалій можна класифікувати за різними критеріями. За типом навчання існують методи з наглядом, без нагляду та з частковим наглядом. Методи з наглядом вимагають маркованих даних як для нормальних, так і для аномальних зразків, що на практиці рідко досяжно через складність збору репрезентативної вибірки аномалій. Неконтрольовані методи не вимагають маркованих даних і базуються на припущенні, що аномалії значно відрізняються від основної маси даних. Напівконтрольовані методи навчаються тільки на нормальних даних, що є найбільш реалістичним сценарієм для практичного застосування.

Залежно від підходу до моделювання, методи можна розділити на статистичні, засновані на відстані, засновані на щільності, засновані на класифікації та засновані на ансамблі. Статистичні методи моделюють нормальну поведінку за допомогою розподілу ймовірностей і визначають аномалії як спостереження з низькою ймовірністю. Методи, засновані на відстані, використовують міри подібності між точками даних, припускаючи, що аномалії розташовані далеко від нормальних точок. Методи, засновані на щільності, визначають аномалії як точки в областях з низькою щільністю даних.

Залежно від виду виявлених аномалій, можна провести диференціацію на точкові аномалії, контекстні аномалії та колективні аномалії. Точкові аномалії - це поодинокі спостереження, які значно відрізняються від стандарту. Контекстна аномалія є аномальною лише в певному контексті. Колективні аномалії представляють групи пов'язаних спостережень, які в цілому демонструють аномальну поведінку.

Статистичні методи є основою виявлення аномалій. Параметричні статистичні підходи припускають, що нормальні дані генеруються з певного розподілу ймовірностей. Найпростішим з них є гаусівська модель, в якій дані моделюються як нормально розподілені з параметрами μ і σ, де аномалії — це ті точки, що виходять за межі $\mu \pm k\sigma$, для деякого малого значення k, зазвичай 2 або 3.

Непараметричні статистичні методи не роблять жодних припущень щодо розподілу даних. Такі методи включають підходи на основі гістограм, оцінку щільності ядра та тести на основі рангу. Підходи на основі гістограм поділяють простір ознак на комірки, обчислюють частоту точок, що потрапляють у комірку,

і позначають точки, що потрапляють у комірки з низькою частотою, як аномальні.

Основним недоліком статистичних методів для недовантажених мереж є необхідність наявності достатньої кількості даних для забезпечення надійної оцінки параметрів розподілу. При невеликих розмірах вибірки дисперсія статистичних оцінок є дуже великою, що призводить до високого рівня помилкових спрацьовувань.

Методи на основі відстані базуються на концепції близькості точок у наборі даних. k-найближчі сусіди обчислюють k-ту відстань до найближчого сусіда для кожної точки, позначаючи точки з найбільшими відстанями як аномалії. Цей метод дуже простий у реалізації і не передбачає жодних припущень щодо розподілу даних. Однак, будучи наївним методом реалізації, обчислювальна складність становить $O(n^2)$ з недоліками, пов'язаними з вибором параметра k.

Методи на основі локальної щільності порівнюють локальну щільність точки з щільністю її сусідів. Таким чином, аномалії можна виявити в регіонах з різною щільністю точок. LOF обчислює локальний фактор аномалії, який відображає, наскільки щільність об'єкта відрізняється від щільності його околиць.

Методи на основі відстаней можуть бути корисними для недовантажених мереж, враховуючи їх здатність працювати з невеликими наборами даних. З іншого боку, вони чутливі до вибору метрики, що використовується для відстані та параметрів алгоритму, що може бути проблематичним за відсутності достатніх даних для валідації.

Це, в свою чергу, призвело до розробки потужних методів виявлення винятків на основі машинного навчання. Метод однокласового опорного вектора створює межу, яка охоплює більшість нормальних даних у просторі ознак. Однокласний SVM знаходить гіперплощину з максимальним відступом, яка відокремлює дані від початку координат у просторі вищої розмірності, створеному картою ядра.

Isolation Forest представляє новий підхід, заснований на ізоляції аномалій. Метод будує ансамбль дерев рішень, де кожне дерево створюється шляхом випадкового поділу простору ознак. Основна ідея полягає в тому, що аномалії легше ізолювати, тобто для їх відокремлення від основної маси даних потрібно менше розділів.

Підходи на основі кластеризації, включаючи DBSCAN, виявляють аномалії як точки даних, які не належать до жодного кластера. DBSCAN знаходить кластери як області високої щільності, відокремлені областями низької щільності, без апріорного визначення кількості кластерів.[1]

Глибоке навчання відкрило більше можливостей для виявлення аномалій у складних високорозмірних даних. Автокодери навчаються стискати вхідні дані в компактне латентне представлення і відтворювати їх з цього представлення. Передбачається, що автокодер, навчений на нормальних даних, буде погано відтворювати аномальні зразки, що дозволить виявляти їх з високою похибкою відтворення. Глибоке навчання відкрило більше можливостей для виявлення аномалій у складних високорозмірних даних. Автокодери навчаються стискати вхідні дані в компактне латентне представлення і відтворювати їх з цього представлення. Передбачається, що автокодер, навчений на нормальних даних, буде погано відтворювати аномальні зразки, що дозволить виявляти їх з високою похибкою відтворення.

LSTM-Autoencoder поєднує рекурентні нейронні мережі з архітектурою автокодера для обробки послідовностей даних. Шари LSTM здатні фіксувати довгострокові залежності в часових рядах, що робить їх особливо придатними для виявлення тимчасових аномалій у мережевому трафіку. Варіаційні автоенкодери розширюють концепцію класичних автоенкодерів, навчаючи імовірнісний розподіл латентного простору. Це дозволяє не тільки виявляти аномалії, але й оцінювати їх ймовірність, що корисно для інтерпретації результатів.

Генеративні суперечливі мережі також використовуються для виявлення аномалій. Ідея полягає в тому, щоб навчити генератор на нормальних даних, щоб аномалії виявлялися як зразки, які генератор не може відтворити або які дискримінатор може легко відрізнити від реальних даних.

Ключовою проблемою використання глибокого навчання в мережах з обмеженими ресурсами є необхідність великих обсягів навчальних даних. Глибокі нейронні мережі мають мільйони параметрів, і для уникнення перенавчання потрібно мати багато прикладів. Це обмежує їх застосування в сценаріях, де спостерігається нестача даних.

Гібридні підходи поєднують переваги різних методів для досягнення кращих результатів. Наприклад, статистичні методи можуть використовуватися для попередньої фільтрації даних, а потім застосовуються більш складні методи машинного навчання. Ансамблеві методи поєднують прогнози декількох базових моделей для підвищення надійності виявлення.

Активне навчання - це інтерактивне вдосконалення моделі шляхом запиту експертного маркування найбільш інформативних зразків. Це стає особливо корисним, коли маркування даних вимагає значних ресурсів. Перенесення навчання дозволяє використовувати знання, отримані під час виконання одного завдання або в одній області, для ефективного вирішення іншого завдання або в іншій області. Це корисно для мереж з обмеженими ресурсами, де легко використовувати моделі, попередньо навчені на величезних наборах даних із подібними областями. Методи виявлення аномалій, характерних для мереж з низьким

навантаженням, залишаються обмеженими. Більшість робіт, вже опублікованих у літературі, зосереджуються на корпоративних мережах з високим трафіком, магістральних комунікаційних каналах тощо. Однак деякі дослідники почали приділяти увагу особливим проблемам, що виникають у сценаріях з низькою частотою.

Особливості промислових мереж Інтернету речей, де трафік часто є спорадичним і передбачуваним, вимагають спеціалізованих підходів. Дослідження показують, що традиційні методи виявлення вторгнень на основі сигнатур є неефективними через специфіку протоколів і поведінки пристроїв.

Проблема холодного старту виникає, коли система має запускатися без достатньої кількості історичних даних. Загалом, це дуже актуальна проблема для мереж з недостатнім навантаженням. Методи повинні бути здатні робити надійні висновки на основі обмеженої інформації та швидко адаптуватися до змін. Енергоефективність та обчислювальна складність мають вирішальне значення для пристроїв з обмеженими ресурсами. Методи повинні забезпечувати баланс між точністю виявлення та вимогами до обчислювальних ресурсів і споживання енергії.[2]

Isolation Forest представляє парадигмальний зсув у підході до виявлення аномалій: замість моделювання нормальної поведінки, метод базується на припущенні, що аномалії є рідкісними і відрізняються від нормальних точок; таким чином, їх легше ізолювати. Алгоритм будує сукупність дерев ізоляції, де кожне дерево створюється шляхом рекурсивного і випадкового поділу простору ознак.

Математична основа методу полягає у вимірюванні глибини ізоляції точки. Для точки x глибина h(x) визначається як кількість розділень, необхідних для ізоляції цієї точки в дереві. Аномалії мають меншу середню глибину порівняно з нормальними точками, оскільки вони розташовані в менш щільних областях простору і вимагають менше розділень для ізоляції.

Оцінка аномалії s(x,n) нормалізується за допомогою середньої глибини невдалого пошуку в бінарному дереві пошуку, що залежить від розміру вибірки n. Це дає оцінку від 0 до 1. Значення, близькі до 1, вказують на те, що екземпляр x, ймовірно, є аномалією. Обчислювальна складність алгоритму становить $O(t \cdot \psi \cdot \log \psi)$ для навчання, де t це кількість дерев, а ψ це розмір підвибірки. Для прогнозування складність становить $O(t \cdot \log \psi)$, що значно нижче, ніж у методах, заснованих на відстані. Це робить Isolation Forest особливо привабливим для додатків, що працюють у режимі реального часу.

Цей метод має кілька важливих переваг для мереж з невеликим навантаженням. По-перше, він не вимагає великих обсягів даних для навчання, оскільки випадковий розподіл простору не вимагає точної оцінки розподілу даних. По-друге, метод є стійким до високої розмірності завдяки випадковому вибору ознак для розподілу. По-третє, параметр забруднення дозволяє явно контролювати очікувану частку аномалій.

Одним із недоліків цього методу є те, що інтерпретація результатів є громіздкою, оскільки випадковий характер поділу ускладнює розуміння того, чому дана точка класифікується як аномальна. Крім того, цей метод виявиться менш ефективним у пошуку локальних аномалій у регіонах з неоднорідною щільністю.[3]

Однокласна SVM розширює класичний метод опорних векторів на завдання виявлення аномалій. Метод будує гіперплощину в просторі ознак (можливо, нескінченно вимірному через ядрове відображення), що відокремлює дані від початку координат з максимальним зазором. Точки, що потрапляють на «неправильний» бік гіперплощини, класифікуються як аномалії.

Математична формалізація задачі полягає в мінімізації норми вектора w, що визначає орієнтацію гіперплощини, з одночасною максимізацією зміщення ρ та мінімізацією порушень обмежень через slack змінні ξ_i. Параметр ν контролює компроміс між максимізацією зазору та мінімізацією помилок, визначаючи верхню межу для частки помилок навчання та нижню межу для частки опорних векторів. Ядерні функції відіграють критичну роль в ефективності методу. Радіальна базисна функція є найпопулярнішим вибором через її здатність моделювати складні нелінійні границі. Параметр γ ядра контролює «радіус впливу» окремих навчальних прикладів. Великі значення γ призводять до складніших границь, що можуть перенавчатися, тоді як малі значення створюють більш згладжені границі.

Обчислювальна складність One-Class SVM становить $O(n^2 \cdot d)$ для навчання в гіршому випадку, де d - розмірність простору ознак. Це може бути проблематичним для великих датасетів, хоча в практиці складність часто близька до $O(n \cdot s \cdot d)$, де s - кількість опорних векторів. Для передбачення складність становить $O(s \cdot d)$, що залежить від кількості опорних векторів.

Переваги методу для слабонавантажених мереж включають теоретичну обґрунтованість через зв'язок з теорією статистичного навчання, здатність моделювати складні нелінійні границі через ядрові функції, явний контроль над компромісом між помилками та складністю моделі через параметр ν. Метод також має добре розвинений математичний апарат для аналізу та оптимізації. Недоліки включають чутливість до вибору ядра та його параметрів, що вимагає ретельного налаштування. Метод також потребує нормалізації ознак, оскільки чутливий до їх масштабу. Інтерпретація результатів може бути складною, особливо при використанні нелінійних ядер. Нарешті, метод може бути обчислювально затратним для великих датасетів, хоча для слабонавантажених мереж це зазвичай не є проблемою.[4]

DBSCAN є представником підходу до пошуку винятків на основі кластеризації за щільністю. Цей метод не вимагає попередньої інформації про кількість кластерів і може виявляти кластери довільної форми. Основна ідея алгоритму полягає в тому, що кластери відповідають областям точок з високою щільністю, розділеним областями з низькою щільністю.

Алгоритм оперує двома ключовими параметрами: ε, що визначає радіус околиці точки, та minPts, що визначає мінімальну кількість точок в ε-околиці для того, щоб точка вважалася core point. Точки класифікуються на три категорії: core points, що мають принаймні minPts сусідів у радіусі ε; border points, що мають менше minPts сусідів, але належать ε-околиці якогось core point; noise points, що не є ні core, ні border points - саме ці точки класифікуються як аномалії.

Алгоритм працює ітеративно, починаючи з довільної необробленої точки. Якщо точка є core point, алгоритм формує новий кластер, додаючи всі точки з її ε-околиці. Потім для кожної нової доданої точки, що також є core point, процес рекурсивно повторюється. Після обробки всіх точок, що залишилися необробленими точками, стають noise points. Обчислювальна складність DBSCAN, в наївній реалізації, становить $O(n^2)$, оскільки для кожної точки необхідно обчислити відстань до всіх інших точок. Використовуючи просторові індекси, такі як k-d дерева або R-дерева, складність можна зменшити в середньому до $O(n \log n)$.

Визначення оптимальних параметрів є критичною задачею для ефективного застосування DBSCAN. Метод k-distance graph допомагає визначити параметр ε: будується графік відсортованих відстаней до k-го найближчого сусіда, і точка «коліна» на цьому графіку дає оцінку оптимального ε. Параметр minPts зазвичай вибирається як функція розмірності простору, типова евристика - $\text{minPts} \geq D + 1$. Переваги використання DBSCAN у мережах з низьким навантаженням полягають у тому, що немає необхідності вказувати кількість кластерів, він може виявляти кластери будь-якої форми і природно ідентифікує аномалії як точки шуму, що є відносно надійним для параметризації, коли вони правильно обрані. Метод також інтуїтивно зрозумілий і легкий для інтерпретації.

Недоліки методу включають чутливість до параметрів ε та minPts, особливо коли кластери мають різну щільність. У просторах високої розмірності концепція щільності стає менш значущою через «прокляття розмірності». Метод також може бути неефективним, якщо нормальна поведінка не формує чітких кластерів або якщо аномалії випадково потрапляють у щільні області.

LSTM-автокодер поєднує архітектуру автокодера з рекурентними нейронними мережами для обробки послідовностей даних. Цей метод є особливо ефективним для виявлення тимчасових аномалій у часових рядах, що робить його актуальним для аналізу мережевого трафіку.

Він має два ключові компоненти: кодер і декодер. Кодер обробляє вхідну послідовність і генерує компактне латентне представлення, яке узагальнює всю істотну інформацію послідовності. Потім декодер відновлює вихідну послідовність на основі латентного представлення. Обидва компоненти використовують шари LSTM, які можуть моделювати довгострокові залежності в даних. Клітина LSTM вирішує проблему зникаючого градієнта, характерну для класичних рекурентних мереж. Вона містить три ворота: ворота забуття, які визначають, яку інформацію з попереднього стану комірки слід забути; ворота введення, які контролюють, яку нову інформацію додати до стану комірки; та ворота виведення, які визначають, яку інформацію зі стану комірки передати далі. Ці механізми дозволяють мережі вибірково зберігати та використовувати інформацію в різних часових масштабах.

Процес навчання мінімізує помилку реконструкції між вхідною та відновленою послідовністю. Стандартна функція середньоквадратичної помилки є прикладом функції втрати, яка вимірює різницю між вхідною послідовністю та її реконструкцією. Оскільки навчання відбувається тільки на нормальних даних, мережа буде добре знати, як реконструювати нормальні послідовності. Виявлення аномалій базується на припущенні, що автокодер, навчений на нормальних даних, погано реконструює аномальні послідовності. Для кожної нової послідовності обчислюється помилка реконструкції, і якщо вона перевищує встановлений поріг, послідовність класифікується як аномальна. Поріг може бути визначений статистично на основі розподілу помилок реконструкції для навчальних даних.

Переваги LSTM-автокодера для мереж з невеликим навантаженням включають здатність фіксувати складні часові залежності, що особливо важливо для виявлення аномалій, які проявляються через зміни в часовій структурі трафіку. Метод здатний працювати в режимі напівнаглядового навчання, використовуючи тільки нормальні дані. Латентне представлення може бути використане для візуалізації та інтерпретації даних.

До недоліків методу належать високі вимоги до обчислювальних ресурсів як для навчання, так і для інференції. Метод вимагає значного обсягу даних для ефективного навчання через велику кількість параметрів мережі, що може бути проблематичним для мереж з невеликим навантаженням. Вибір архітектури мережі, довжини послідовності та порогу виявлення вимагає експертних знань та експериментів. Інтерпретація того, чому певна послідовність вважається аномальною, може бути складною через непрозорість нейронних мереж.

Для об'єктивного порівняння методів необхідно враховувати кілька критеріїв на різних рівнях. До

них можуть належати, серед іншого, такі категорії, як ефективність виявлення, обчислювальні вимоги, вимоги до даних, адаптивність та інтерпретованість.

Ефективність виявлення оцінюється за допомогою традиційних метрик класифікації. Точність показує частку правильно ідентифікованих аномалій серед усіх позначених як аномалії, що є критично важливим для мінімізації помилкових спрацьовувань. Відтворюваність відображає здатність методу виявляти всі справжні аномалії, що є важливим для безпеки. Показник F1 представляє гармонійне середнє значення точності та відтворюваності. AUC-ROC характеризує здатність методу розділяти класи за різними порогами класифікації.

Вимоги до обчислювальних ресурсів включають складність навчання та прогнозування, використання пам'яті, можливості паралелізації та здатність виконувати завдання в режимі реального часу. У мережах з недостатнім навантаженням, особливо в сценаріях IoT, енергоефективність також є важливим питанням.

Вимоги до даних включають мінімальну кількість навчальних прикладів, необхідність маркованих даних, стійкість до шуму та винятків, а також здатність обробляти неповні дані. Також можуть бути вимоги щодо збалансованості класів.

Адаптивність тут включає можливість онлайн-навчання, чутливість до параметрів, інкрементне оновлення моделі та здатність адаптуватися до зміщення концепції або змін у розподілі даних.

Інтерпретованість визначає, наскільки легко зрозуміти, чому певна точка класифікується як аномалія, що є важливим для практичного застосування та налагодження системи.[5]

Isolation Forest демонструє високу ефективність для точкових аномалій, особливо коли аномалії значно відрізняються від нормальних точок у глобальному масштабі. Метод показує хороші результати навіть при високій розмірності даних завдяки випадковому вибору ознак. Однак він може бути менш ефективним для локальних аномалій та аномалій, які проявляються через незначні зміни в кореляціях між ознаками. Типова точність становить 0,75-0,85, повнота - 0,70-0,80 на збалансованих наборах даних.

One-Class SVM демонструє відмінні результати для даних з чіткою межею між нормальними та аномальними зразками. При правильному виборі ядра та параметрів метод здатний моделювати складні нелінійні межі. Ядро RBF особливо ефективне для даних з кластерною структурою. Метод чутливий до вибору параметра ν, який контролює компроміс між точністю та повнотою. Типова точність становить 0,70-0,90, повнота - 0,65-0,85, залежно від налаштувань.

DBSCAN особливо ефективний, коли нормальна поведінка утворює чіткі щільні кластери, а аномалії є ізольованими точками. Метод природно виявляє групові аномалії як невеликі кластери або ізольовані точки. Однак його ефективність сильно залежить від вибору параметрів ε і minPts. При неоднорідній щільності кластерів метод може класифікувати точки з розріджених кластерів як аномалії. Типова точність становить 0,60-0,85, повнота - 0,55-0,75.

LSTM-Autoencoder найбільш ефективний для тимчасових аномалій, які проявляються через зміни в послідовності подій. Метод здатний виявляти складні аномалії, які не є очевидними при аналізі окремих точок. Він особливо корисний для контекстних аномалій, де нормальність залежить від історії подій. Однак метод вимагає ретельного налаштування архітектури та порогу виявлення. Типова точність становить 0,80-0,95, повнота - 0,75-0,90 при наявності достатньої кількості даних для навчання. Isolation Forest має найнижчу обчислювальну складність серед розглянутих методів. Складність навчання $O(t \cdot \psi \cdot \log \psi)$, де t - кількість дерев (зазвичай 100), а ψ - розмір підвибірки (зазвичай 256), робить метод дуже швидким навіть для великих наборів даних. Час інференції $O(t \cdot \log \psi)$ дозволяє обробляти нові точки майже миттєво. Використання пам'яті залежить від кількості та глибини дерев, але залишається помірним. Метод легко паралелізується, оскільки дерева будуються незалежно.

One-Class SVM має вищу обчислювальну складність. Навчання вимагає від $O(n^2 \cdot d)$ до $O(n^3 \cdot d)$ операцій залежно від алгоритму оптимізації, що може бути проблематичним для великих наборів даних. Для мереж з невеликим навантаженням і невеликими обсягами даних це зазвичай є прийнятним. Час інференції $O(s \cdot d)$, де s це кількість опорних векторів, зазвичай невеликий. Використання пам'яті залежить від кількості опорних векторів і розмірності матриці ядра.

DBSCAN має складність $O(n^2)$ для наївного впровадження, але може бути оптимізований до $O(n \log n)$ при використанні просторових індексів. Для мереж з невеликим навантаженням і малим n навіть наївне впровадження працює досить швидко. Метод не вимагає окремої фази прогнозування для навчальних даних, але для класифікації нових точок необхідно обчислити відстані до всіх основних точок. Використання пам'яті залежить від типу просторового індексу.

LSTM-Autoencoder має найвищу обчислювальну складність. Навчання вимагає багатьох епох проходження через весь набір даних, що вимагає значних обчислювальних ресурсів. Складність залежить від розміру мережі, довжини послідовностей і кількості епох. Використання GPU значно прискорює навчання. Час інференції також може бути значним для довгих послідовностей. Використання пам'яті включає зберігання параметрів мережі і проміжних активацій.

Isolation Forest має найнижчі вимоги до обсягу даних для навчання. Метод може ефективно працювати навіть з декількома сотнями прикладів завдяки випадковому характеру побудови дерева. Метод не вимагає маркованих даних і працює в повністю неконтрольованому режимі навчання. Стійкість до шуму та винятків у навчальних даних є високою, оскільки вони автоматично ізолюються. Метод не чутливий до масштабу ознак.

One-Class SVM вимагає помірної кількості навчальних даних, принаймні декількох сотень прикладів для надійного визначення меж. Метод працює в режимі напівконтрольованого навчання, використовуючи тільки нормальні дані. Він чутливий до масштабу ознак, що вимагає попередньої нормалізації. Наявність винятків у навчальних даних може значно вплинути на результат, хоча параметр ν дозволяє частково контролювати цей ефект. DBSCAN має середні вимоги до обсягу даних. Для надійного визначення щільності кластерів потрібна достатня кількість точок у кожному кластері. Метод працює в режимі неконтрольованого навчання. Чутливий до масштабу ознак через використання відстаней. Може бути стійким до невеликої кількості викидів, але великі викиди можуть вплинути на визначення параметрів.

LSTM-Autoencoder має найвищі вимоги до обсягу навчальних даних. Глибокі нейронні мережі потребують тисяч прикладів для уникнення перенавчання, що може бути проблематичним для слабонавантажених мереж. Методи регуляризації та data augmentation можуть частково компенсувати недостатню кількість даних. Метод працює в режимі напівконтрольованого навчання. Чутливий до шуму в даних, що може призвести до навчання неправильних патернів.

Isolation Forest має обмежену адаптивність. Метод не підтримує онлайн-навчання в класичному вигляді, модель потрібно повністю перенавчати для адаптації до нових даних. Однак низька обчислювальна складність робить перенавчання практично можливим. Метод має лише декілька параметрів (кількість дерев, contamination), що спрощує налаштування. Стійкість до зміни розподілу даних помірна.

One-Class SVM також має обмежену адаптивність. Існують інкрементальні версії SVM, але вони рідко використовуються на практиці. Метод чутливий до вибору ядра та його параметрів, що вимагає ретельного налаштування. При зміні характеру даних може знадобитися повне перенавчання та підбір нових параметрів. Адаптація до concept drift вимагає періодичного перенавчання.

DBSCAN є статичним методом для фіксованого датасету. Для адаптації до нових даних потрібно перезапускати алгоритм на об'єднаному датасеті. Існують incremental версії DBSCAN, але вони складніші в реалізації. Вибір параметрів ε та minPts є критично важливим і може потребувати коригування при зміні характеру даних. Метод може бути чутливим до порядку обробки точок у деяких реалізаціях.

LSTM-Autoencoder має найкращий потенціал для адаптації. Нейронні мережі підтримують онлайн-навчання та fine-tuning на нових даних. Можна періодично дообучувати модель на найсвіжіших даних для адаптації до змін. Однак це потребує ретельного контролю learning rate та regularization для уникнення catastrophic forgetting. Метод має багато гіперпараметрів, що ускладнює налаштування, але надає більше можливостей для оптимізації.

Isolation Forest має середню інтерпретованість. Можна пояснити, що точка є аномалією, тому що її легко ізолювати, але конкретні ознаки, що призвели до цього, визначити складно через випадковість побудови дерев. Середня глибина ізоляції дає кількісну міру аномальності. Важливість ознак можна оцінити статистично, аналізуючи, які ознаки частіше використовуються для ізоляції аномалій.

One-Class SVM має низьку інтерпретованість при використанні нелінійних ядер. У просторі вищої розмірності, створеному ядровим відображенням, складно інтерпретувати межу класифікації. При використанні лінійного ядра можна аналізувати коефіцієнти вектора w, але це рідко дає оптимальні результати. Відстань до границі дає міру аномальності. Опорні вектори можуть бути інтерпретовані як прикордонні точки нормальної області.

DBSCAN має найвищу інтерпретованість. Аномалії чітко ідентифікуються як точки шуму, що не належать до жодного кластера. Можна візуалізувати кластери та аномалії у двовимірному або тривимірному просторі. Відстань до найближчого кластера дає міру аномальності. Легко пояснити результати експертам домену через інтуїтивність концепції щільності.

LSTM-Autoencoder має низьку інтерпретованість через «чорну скриньку» природу глибоких нейронних мереж. Помилка реконструкції показує, наскільки аномальна послідовність, але не пояснює чому. Методи візуалізації активацій та attention mechanisms можуть частково допомогти в інтерпретації. Аналіз латентного простору може виявити структуру в даних. Порівняння оригінальної та реконструйованої послідовностей може вказати на аномальні елементи.

Вибір оптимального методу виявлення аномалій для конкретного застосування в слабонавантажених мережах залежить від багатьох факторів. Необхідно враховувати специфічні вимоги та обмеження системи, характеристики даних та типи очікуваних аномалій.

Обсяг доступних даних є першочерговим критерієм. При дуже малих обсягах (сотні прикладів) перевагу мають Isolation Forest та DBSCAN. При помірних обсягах (тисячі прикладів) можна розглядати One-Class SVM. LSTM-Autoencoder доцільний лише при наявності десятків тисяч прикладів або можливості використання трансферного навчання.

Обчислювальні ресурси визначають можливість використання складних методів. Для пристроїв з обмеженими ресурсами (IoT, вбудовані системи) найкращим вибором є Isolation Forest завдяки низькій обчислювальній складності. One-Class SVM та DBSCAN мають помірні вимоги. LSTM-Autoencoder вимагає значних ресурсів і більш придатний для централізованої обробки.

Вимоги до латентності впливають на вибір методу для систем реального часу. Isolation Forest забезпечує найшвидший інференс. One-Class SVM та DBSCAN мають помірну латентність. LSTM-Autoencoder може мати помітну затримку, особливо для довгих послідовностей.

Тип аномалій є критично важливим для вибору. Для точкових аномалій ефективні всі методи, з перевагою Isolation Forest. Для локальних аномалій краще підходять DBSCAN та One-Class SVM з локальними ядрами. Для темпоральних аномалій найкращим є LSTM-Autoencoder. Для колективних аномалій ефективні DBSCAN та LSTM-Autoencoder.

Вимоги до інтерпретованості можуть бути критичними в деяких застосуваннях. DBSCAN надає найбільш зрозумілі результати. Isolation Forest має середню інтерпретованість. One-Class SVM та LSTM-Autoencoder складні для інтерпретації.

Промислові IoT-мережі характеризуються періодичною передачею телеметрії з датчиків, обмеженими обчислювальними ресурсами пристроїв, критичністю виявлення відхилень у роботі обладнання та необхідністю роботи в режимі реального часу. Для таких систем рекомендується Isolation Forest як основний метод завдяки низькій обчислювальній складності та здатності швидко виявляти точкові аномалії. Можна комбінувати з простими статистичними методами для попередньої фільтрації.

Корпоративні мережі малого бізнесу мають обмежений персонал для моніторингу безпеки, необхідність виявлення різноманітних загроз, помірні обчислювальні ресурси та потребу в зрозумілих поясненнях виявлених аномалій. Рекомендується DBSCAN для базового моніторингу з можливістю візуалізації результатів, або комбінація Isolation Forest для швидкої детекції та One-Class SVM для більш точного аналізу підозрілих подій.

Системи моніторингу критичної інфраструктури вимагають високої надійності виявлення, низького рівня помилкових спрацьовувань, здатності виявляти складні атаки та можливості аудиту та пояснення рішень. Рекомендується ансамблевий підхід, що комбінує кілька методів для підвищення надійності. Основа - Isolation Forest для швидкої детекції, One-Class SVM для уточнення, LSTM-Autoencoder для виявлення складних темпоральних аномалій при наявності достатніх ресурсів.

Сенсорні мережі в агросекторі характеризуються дуже низькою частотою передачі даних, високою залежністю від зовнішніх факторів, суворими енергетичними обмеженнями та простотою очікуваних аномалій. Рекомендується Isolation Forest з агресивною фільтрацією для мінімізації енергоспоживання, або прості статистичні методи для найбільш критичних обмежень ресурсів.

Системи віддаленого медичного моніторингу потребують високої точності та мінімізації хибних тривог, здатності виявляти тонкі зміни в стані пацієнта, збереження конфіденційності даних та можливості пояснення медичним персоналом. Рекомендується LSTM-Autoencoder при наявності достатніх даних для захоплення складних темпоральних патернів, або One-Class SVM для більш інтерпретованих результатів при обмежених даних.[4]

Часто оптимальним рішенням є не вибір одного методу, а поєднання сильних сторін кількох підходів. Гібридні системи можуть запропонувати найкращу загальну продуктивність. Каскадна архітектура використовує швидкий метод для первинної фільтрації, після чого більш точний, але повільний метод аналізує підозрілі події. Наприклад: Isolation Forest для швидкої фільтрації більшості нормальних подій, One-Class SVM або LSTM-Autoencoder для детального аналізу кандидатів на аномалії, експертна система або ручний аналіз для остаточного рішення щодо критичних випадків.

Ансамблеве голосування комбінує передбачення декількох методів для підвищення надійності. Кожен метод голосує за нормальність або аномальність точки, остаточне рішення приймається на основі більшості голосів або зваженої комбінації. Це підвищує точність та знижує ймовірність помилок окремих методів.

Спеціалізація за типами аномалій використовує різні методи для виявлення різних типів аномалій. DBSCAN для виявлення масових атак, Isolation Forest для точкових аномалій, LSTM-Autoencoder для темпоральних аномалій та комбінування результатів у загальну оцінку безпеки.

Адаптивний вибір методу динамічно вибирає найбільш придатний метод залежно від поточних умов. При низькому навантаженні використовуються більш складні методи, при високому - швидкі методи. При виявленні змін у розподілі даних автоматично перемикається на більш адаптивні методи.

ВИСНОВКИ З ДАНОГО ДОСЛІДЖЕННЯ І ПЕРСПЕКТИВИ ПОДАЛЬШИХ РОЗВІДОК У ДАНОМУ НАПРЯМІ

В дані статті проведено комплексний порівняльний аналіз чотирьох методів виявлення аномалій у контексті слабонавантажених комп'ютерних мереж: Isolation Forest, One-Class SVM, DBSCAN та LSTM-Autoencoder.

Дослідження показало, що кожен метод має свої переваги та недоліки, і вибір оптимального підходу залежить від специфічних вимог конкретного застосування. Isolation Forest демонструє найкращий баланс між ефективністю виявлення та обчислювальною складністю, що робить його оптимальним вибором для систем з обмеженими ресурсами та вимогами реального часу. One-Class SVM забезпечує високу точність при правильному налаштуванні, але вимагає більших обчислювальних ресурсів та ретельного підбору параметрів.

DBSCAN виявився найбільш інтерпретованим методом, що критично важливо для систем, де необхідно пояснювати рішення експертам домену. Однак його ефективність сильно залежить від вибору параметрів та характеру розподілу даних. LSTM-Autoencoder показав найкращі результати для виявлення складних темпоральних аномалій, але вимагає значних обсягів навчальних даних та обчислювальних ресурсів, що обмежує його застосовність у типових сценаріях слабонавантажених мереж.

Особлива увага приділена специфіці слабонавантажених мереж, де обмежена кількість даних, висока варіативність нормальної поведінки та обмежені обчислювальні ресурси створюють унікальні виклики. Показано, що традиційні методи, розроблені для високонавантажених мереж, потребують адаптації для ефективної роботи в таких умовах.

Практичне значення роботи полягає у наданні систематизованих рекомендацій щодо вибору методу залежно від характеристик конкретної системи. Визначено сценарії, в яких кожен метод є найбільш придатним, та запропоновано гібридні підходи для підвищення загальної ефективності систем виявлення аномалій.

Перспективи подальших досліджень включають розробку спеціалізованих модифікацій існуючих методів для слабонавантажених мереж, дослідження ефективності трансферного навчання для зменшення вимог до обсягу навчальних даних, розробку адаптивних систем, які автоматично вибирають оптимальний метод залежно від поточних умов, та створення еталонних наборів даних, що відображають специфіку слабонавантажених мереж для стандартизованого порівняння методів.

Література

1. Такс Д. М., Дуїн Р. П. Опис даних за допомогою опорних векторів // Машинне навчання. 2004. Т. 54, № 1. С. 45-66. [Електронний ресурс] – Режим доступу: https://link.springer.com/article/10.1023/B:MACH.0000008084.60811.49

2. Сакурада М., Яірі Т. Виявлення аномалій за допомогою автоенкодерів з нелінійним зменшенням розмірності // Матеріали 2-го семінару MLSDA 2014 з машинного навчання для аналізу сенсорних даних. 2014. С. 4-11.: https://doi.org/10.1145/2689746.268974.

3. Естер М., Крігель Х. П., Сандер Й., Сюй Х. Алгоритм на основі щільності для виявлення кластерів у великих просторових базах даних з шумом // Матеріали 2-ї Міжнародної конференції з виявлення знань та аналізу даних. 1996. С. 226-231. [Електронний ресурс] – Режим доступу https://file.biolab.si/papers/1996-DBSCAN-KDD.pdf.

4. Лю Ф. Т., Тінг К. М., Чжоу З. Х. Isolation Forest // Матеріали 8-ї Міжнародної конференції IEEE з інтелектуального аналізу даних. 2008. С. 413-422. . https://doi.org/10.1109/ICDM.2008.17

5. Чалапаті Р., Чавла С. Глибоке навчання для виявлення аномалій: опитування 2019 р./ https://doi.org/10.48550/arXiv.1901.03407

References

1. Tax D. M., Duin R. P. Support vector data description // Machine Learning. 2004. Vol. 54, No. 1. P. 45-66. [Elektronnyi resurs] – Rezhym dostupu: https://link.springer.com/article/10.1023/B:MACH.0000008084.60811.49

2. Sakurada M., Yairi T. Anomaly detection using autoencoders with nonlinear dimensionality reduction // Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis. 2014. P. 4-11.: https://doi.org/10.1145/2689746.268974

3. Ester M., Kriegel H. P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise // Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining. 1996. P. 226-231. [Elektronnyi resurs] – Rezhym dostupe https://file.biolab.si/papers/1996-DBSCAN-KDD.pdf.

4. Liu F. T., Ting K. M., Zhou Z. H. Isolation forest // Proceedings of the 8th IEEE International Conference on Data Mining. 2008. P. 413-422. . https://doi.org/10.1109/ICDM.2008.17

5. Chalapathy R., Chawla S. Deep learning for anomaly detection: A survey 2019./ https://doi.org/10.48550/arXiv.1901.03407