ISSN 2522-9818 (print)
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2024. No. 1 (27)*

УДК 004.8:006.91 DOI: https://doi.org/10.30837/ITSSI.2024.27.236

В. Ащепков

ВИКОРИСТАННЯ МОДЕЛІ *ISOLATION FOREST* ДЛЯ ВИЯВЛЕННЯ АНОМАЛІЙ У ДАНИХ ВИМІРЮВАНЬ

Предметом дослідження є модель ізольованого лісу, яка є потужним та ефективним інструментом для виявлення аномалій у вимірюваних показниках та викидів, що може застосовуватися в різних сферах, де важливо забезпечити високу точність і надійність вимірювань. **Мета роботи** – застосування моделі ізольованого лісу для виявлення незвичайних або аномальних зразків, що відрізняються від типових патернів у вихідних показниках. Це досягається з допомогою ізоляції аномальних зразків від нормальних з допомогою побудови багатьох різних дерев рішень. **Завданням статті** є виявлення викидів у результатах, які були отримані в процесі дослідження з підготовки до міжнародних порівнянь на державному первинному еталоні масової та об'ємної витрати рідини, маси та об'єму рідини, що протікає по трубопроводу, з допомогою вимірювання коріолісового витратоміра. Показники, зібрані під час метрологічних досліджень, обробляються моделлю для виявлення аномалій. Ця модель аналізує результати та визначає аномальні або викидні значення, що можуть свідчити про систематичні або випадкові помилки вимірювань. Вона дає змогу швидко та ефективно виявити навіть найменші відхилення в показниках, що допомагає підтримувати високу точність і достовірність результатів вимірювань. Основними **методами** вияву викидів у статистичному аналізі, які не залежать від розподілу показників, є критерій Граббса, міжквартильний розподіл, середньоквадратичне відхилення. Вони чутливі до розміру вибірки, але є простими та зрозумілими інструментами. Проте модель ізольованого лісу також має обмеження, зокрема вона може бути вимогливою до обчислювальних ресурсів за умови великих обсягів інформації. Крім того, необхідно брати до уваги, що використання моделі вимагає належного налаштування параметрів для досягнення оптимальних результатів. **Результатом дослідження** є оцінка ефективності моделі ізольованого лісу способом порівняння її з традиційними методами виявлення викидів. Порівняльний аналіз результатів різних підходів до одного завдання є ефективним методом оцінювання ефективності роботи моделі. **Висновки**. Наприкінці статті сформульовано перспективу подальшого дослідження з окресленого напряму. Робота буде спрямована на впровадження методів виявлення аномалій у вимірюваних показниках і покращення точності та достовірності результатів вимірювань у різних галузях, що може широко застосовуватися в науці та промисловості.

Ключові слова: невизначеність; виявлення аномалій; вимірювання; метрологія; оброблення даних; алгоритми машинного навчання; статистичні методи.

Вступ

Актуальність дослідження випливає із реальної потреби в автоматизації процесів та підвищення точності вимірювань у сучасних умовах. Завдання виявлення аномалій в метрології є актуальним, що пояснюється важливістю раннього виявлення та усунення помилок у вимірюваних даних. Оскільки точність вимірювань є ключовою в багатьох сферах, зокрема промисловості, науці, технології та медицині, необхідність розроблення ефективних методів виявлення аномалій стає надзвичайно важливою.

Автоматизовані системи вимірювання збільшують обсяг та швидкість збору даних, але водночас підвищують ризик виникнення в них помилок та аномалій. Тому важливо розробляти методи та інструменти, що дають змогу вчасно виявляти ці аномалії та усувати їх на ранніх етапах вимірювань.

У цьому контексті модель ізольованого лісу є потужним інструментом для виявлення аномалій у вимірюваних даних. Її застосування може покращити якість та достовірність результатів вимірювань, а також сприяти підвищенню ефективності процесів метрологічного контролю та управління якістю.

У статті досліджено застосування моделі ізольованого лісу для виявлення аномалій у результатах вимірювань, які проводились на державному первинному еталоні масової та об'ємної витрати рідини, маси та об'єму рідини, що протікає по трубопроводу (далі – ДЕТУ 03-04-04). У роботі порушено питання ефективності та можливості застосування запропонованої моделі в метрологічних задачах і визначено переваги її використання порівняно з традиційними методами виявлення аномалій.

ISSN 2522-9818 (print)
Сучасний стан наукових досліджень та технологій в промисловості. 2024. №1 (27) ISSN 2524-2296 (online)

Постановка завдання

Міжнародні звірення в метрології відіграють вирішальну роль у забезпеченні надійності, точності та порівняності вимірювань по всьому світу. Вони спрямовані на розроблення та прийняття міжнародних стандартів і національних еталонів, що необхідно для стандартизації методів вимірювань у різних сферах.

У період з 2002 до 2024 рр. у світі було проведено або проводиться 27 міжнародних звірень за напрямом вимірювання витрати рідини та об'єму рідини в межах *EUROMET, APMP (Asian Pasific Metrology Program), SIM (System of Inter-American Metrology)* та *CCM (Consultative Committee for Mass and Related Quantities)* [1–3].

У процесі підготовки до міжнародних звірень за напрямом масової витрати рідини на ДЕТУ 03-04-04 досліджувалися складові частини еталона, зокрема стабільність та повторюваність масової витрати рідини, та стандартна невизначеність вимірювань похибки коріолісових витратомірів. Після проведення вимірювань коріолісових витратомірів виникла необхідність у дослідженні випадкової та систематичної похибки вимірювань. Для цього було впроваджено модель машинного навчання "Ізольований ліс", яка має виявляти аномалії в результатах вимірювання та відтворювати випадкову та систематичну похибки вимірювань.

1. Невизначеність вимірювань

Невизначеність в метрології є одним з найважливіших понять у цій галузі, оскільки відтворює ступінь неконкретності результатів вимірювань. У метрологічній практиці, навіть за використання найсучасніших методик та інструментів, неможливо уникнути впливу різних факторів, що можуть спричиняти похибки та невизначеність у результати. Ці фактори можуть бути як випадковими помилками, так і систематичними неточностями, а також іншими чинниками, такими як зміни в середовищі, умови експлуатації обладнання тощо.

У контексті метрології невизначеність зазвичай виражається у вигляді діапазону значень, у межах якого міститься справжнє значення вимірюваної величини з певною ймовірністю. Цей діапазон відтворює ступінь упевненості в результаті та дозволяє брати до уваги різні джерела невизначеності, які можуть впливати на вимірювання. Однак у метрології також існує поняття розширеної невизначеності, що передбачає не лише стандартну невизначеність вимірювання, а й інші фактори, які додають невизначеність у результати. Це можуть бути, наприклад, нестандартні умови експлуатації обладнання, відхилення від технічних характеристик вимірювальних приладів або процесів вимірювання, а також інші чинники, які можуть впливати на достовірність результатів [4].

2. Типи невизначеності вимірювань

У метрології виокремлюють два основних типи невизначеності: тип А і тип Б.

Експериментальну дисперсію, що характеризує складник невизначеності, отриману внаслідок оцінювання за типом А, знаходять із рядів повторних спостережень, і вона є статистичною оцінкою дисперсії. Експериментальне стандартне відхилення отримують як додатний квадратний корінь із дисперсії, позначають як u_A і для зручності називають стандартною невизначеністю типу А. Оцінення компонентів стандартної невизначеності за типом А основане на розподілах частості. Тому для оцінювання стандартної невизначеності за типом А необхідно провести n незалежних спостережень вимірюваної величини q в умовах повторюваності. Здебільшого найкращою доступною оцінкою математичного сподівання чи очікуваного значення μ_q величини q, що змінюється випадково, є середнє арифметичне або середнє значення q з n спостережень [5]:

$$\bar{q} = \frac{1}{n}\sum_{k=1}^{n} q_k \ . \qquad (1)$$

Експериментальне стандартне відхилення середнього значення u_A $(\bar{q})$ розраховується за формулою [2, 3]

$$u_A(\bar{q}) = \sqrt{\frac{\sum_{i=1}^{n}(q_k - \bar{q})^2}{n(n-1)}} \ . \qquad (2)$$

Тип Б, або систематична невизначеність, пов'язаний із систематичними джерелами помилок або неточностей, що впливають на результати вимірювань. Оцінювання цього типу невизначеності часто вимагає складніших методів, оскільки систематичні помилки можуть бути менш передбачуваними. Приклади типу Б містять

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 1 (27)

систематичні неточності, що виникають через неоднорідність обладнання, дрейф калібрування або інші фактори, які можуть спричинити постійну похибку у вимірюваннях.

Обидва типи невизначеності відіграють важливу роль в оцінюванні загальної розширеної невизначеності результатів вимірювань, і для точного оцінювання розширеної невизначеності необхідно брати до уваги обидва типи та проводити відповідний аналіз і корекцію.

Для оцінювання невизначеностей типів А і Б застосовуються різні методи й моделі, такі як аналіз часових рядів і регресійний аналіз, що дають змогу виявити та оцінити різні джерела невизначеності. У контексті метрології також упроваджуються методи машинного навчання, що можуть допомогти в цьому процесі [6]. Однак варто зауважити, що методи машинного навчання не завжди здатні безпосередньо оцінювати типи невизначеності, але вони можуть бути корисними для аналізу складних результатів і виявлення патернів, що зі свого боку допомагає ідентифікувати потенційні джерела невизначеності для подальшого управління ними та для їх зменшення.

3. Модель ізольованого лісу

Модель ізольованого лісу (*Isolation Forest*) належить до алгоритмів машинного навчання, що використовуються для виявлення аномалій у даних. Машинне навчання – це галузь штучного інтелекту, яка створює та розробляє алгоритми й моделі, здатні "навчатися" на основі даних та здійснювати прогнози або приймати рішення на підставі цього навчання. Модель ізольованого лісу є частиною сімейства методів машинного навчання, основаних на ідеї "дерев прийняття рішень". Її основним завданням є виявлення аномалій у даних, тобто об’єктах або подях, що суттєво відрізняються від інших даних або не відповідають загальному шаблону. Це може бути корисно в багатьох галузях, зокрема метрології, фінансах, медицині тощо, де важливе виявлення аномальних або потенційно небезпечних ситуацій [7–13].

Ізольований ліс працює на основі простого принципу: він будує аномальні дерева прийняття рішень, розподіляючи дані на різні підгрупи доти, доки не вдасться ізолювати аномалії в невелику кількість розподілень. Оскільки аномалії зазвичай потребують меншої кількості розподілень для їх виділення, вони матимуть коротший шлях до кореня дерева порівняно з нормальними об’єктами даних. Отже, модель ізольованого лісу виконує завдання виявлення аномалій в даних на підставі їх ізоляції від нормальних об’єктів у невелику кількість розподілень [14].

4. Використання моделі

Модель ізольованого лісу відрізняється від традиційних методів розрахунку невизначеності тим, що аналізує кожне значення у вибірці незалежно одне від одного. У звичайних методах оцінювання невизначеності, зокрема методах найменших квадратів або максимальної правдоподібності, береться до уваги середнє значення і розкид даних для отримання оцінки невизначеності. Такий підхід дозволяє моделі ізольованого лісу виявляти аномалії на підставі їх ізоляції від нормальних об’єктів у вибірці, без необхідності знання про розподіл даних або середнє значення. Отже, вона може бути ефективним інструментом для виявлення аномалій у різних галузях, де важливе виявлення незвичайних або потенційно небезпечних ситуацій.

У процесі досліджень з підготовки ДЕТУ 03-04-04 до міжнародних звірень виконано вимірювання коріолісового витратоміра та розраховано відносну похибку вимірювань витратоміра. Відносна похибка обчислювалася за такою формулою [15]:

$$\varepsilon = \frac{\triangle m}{m_{ref}} \cdot 100\%; \quad (3)$$

$$\triangle m = m_v - m_{ref}, \quad (4)$$

де $\triangle m$ – похибка вимірювань;

m_{ref} – значення маси рідини еталона;

m_v – значення маси рідини коріолісового витратоміра.

Значення похибок витратоміра використовувалися як вхідні показники для моделі ізольованого лісу. Вимірювання проводились на трьох точках витрати: 5 т/г, 25 т/г, 45 т/г. Кількість вимірів в одній точці не нормоване, тому відрізняється залежно від умов проведення вимірювань. У підсумку отримуємо три вибірки похибки вимірювань на трьох значеннях витрати рідини. Після оброблення вибірки алгоритмом отримуємо значення ступеня аномальності (рис. 1).

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №1 (27)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

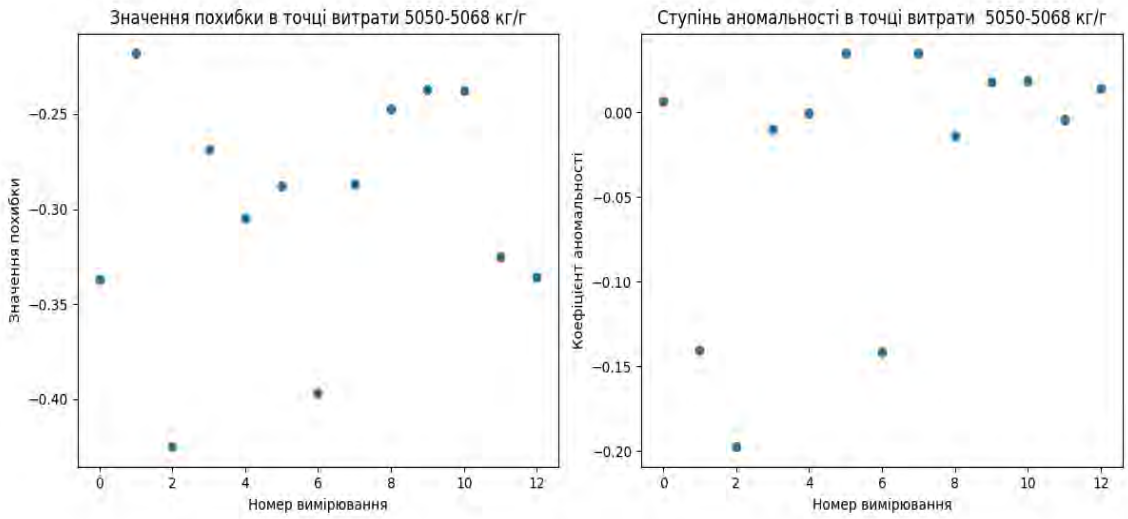

Рис. 1. Результат роботи алгоритму в першій точці витрати

Результати подані у вигляді графіків. Після застосування моделі ізольованого лісу до вхідних показників, зображених на лівому графіку (рис. 1), отримуємо графік значень ступеня аномальності для кожного вимірювання, зображеного праворуч (рис. 1). Цей ступінь аномальності дає змогу визначити, наскільки модель вважає відповідні вихідні значення вимірювання аномальними.

"Аномальність" у контексті метрології належить до незвичайних значень, що відрізняються від очікуваних значень результатів і виникають з різних причин, тобто відхилення, що можуть бути спричинені систематичною або випадковою похибкою, похибкою витратоміра та іншими невідомими факторами впливу.

5. Налаштування моделі

У налаштуванні моделі ізольованого лісу необхідно зважати на кілька параметрів, що впливають на її роботу й результати. Один із таких параметрів – кількість розподілів, що визначає, на скільки підгруп ми розподілимо дані в процесі побудови дерев рішень. Наприклад, збільшення кількості розподілів може сприяти вищій точності виявлення аномалій, але водночас може збільшити час обчислень. Максимальна кількість об'єктів у листі дерев впливає на глибину дерев і може впливати на здатність моделі розрізняти аномальні та нормальні значення. Розмір вибірки також має значення: великі вибірки здатні допомогти уникнути перенавчання моделі, але можуть збільшити час навчання.

Параметр "кількість припущених аномалій" у моделі ізольованого лісу визначає, скільки аномальних об'єктів очікується в навчальній вибірці. Цей параметр впливає на те, як модель розпізнає аномалії, та може бути корисним, якщо заздалегідь відомо або припущено, що велика кількість аномалій у навчальних даних відсутня. Що вищий цей параметр, то більш чутливою до аномалій буде модель. За умови максимального значення цього параметра (50%) всі значення ступеня аномальності визначаються в діапазоні від –1 до 1 (рис. 2). На рис. 2 зображено криву передбачення класів, яка розподіляє значення на нормальні чи аномальні.

"Максимальна глибина дерева" визначає максимальну кількість рівнів у кожному дереві, які модель буде створювати. Цей параметр впливає на здатність моделі виявляти складні взаємозв'язки між даними. Занадто мала глибина може призвести до недооцювання аномалій, а занадто велика – до перенавчання моделі та неефективного виявлення аномалій. У процесі налаштування моделі в цьому дослідженні цей параметр дорівнює 1 000, тобто модель будує 1 тис. дерев. Одне дерево зображено на рис. 3.

Дерева рішень утворюються способом випадкового вибору підмножини ознак та поділу даних на дві частини на кожному рівні побудови. Кожне дерево рішень починається з кореневого вузла, що містить усі доступні ознаки. За допомогою випадкового вибору підмножини ознак береться певна кількість ознак для розгляду на цьому рівні. Потім дані розподіляються на дві групи відповідно до значень обраних ознак і певного порога. Об'єкти,

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 1 (27)

що мають значення ознаки менше за поріг, потрапляють до однієї групи, тоді як об'єкти з більшим значенням або таким, що дорівнює порогу, – до іншої.

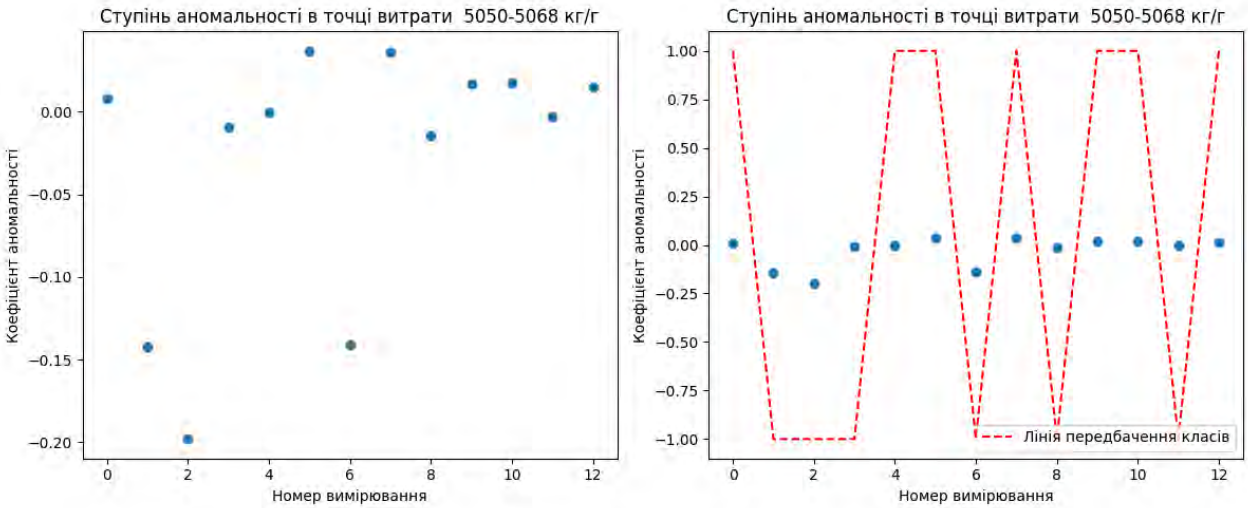

Рис. 2. Передбачення класів

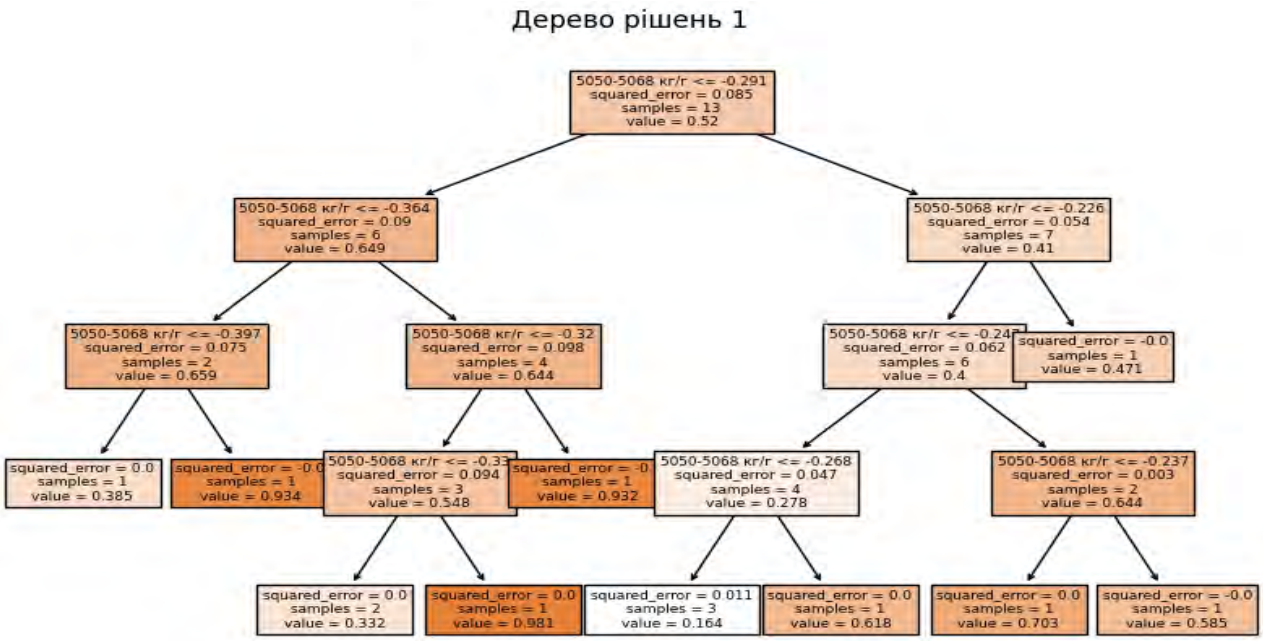

Рис. 3. Перше дерево рішень

Після розподілення даних кожна з отриманих груп стає новим піддеревом, і процес побудови повторюється для кожної з цих груп рекурсивно. Кожен новий вузол є випадковою ознакою та порогом, за яким дані розподіляються на дві частини. Цей процес триває до досягнення критерію зупинки, наприклад, досягнення максимальної глибини дерева або досягнення мінімальної кількості об'єктів у вузлі.

У дослідженні критерієм зупинки був параметр "максимальна кількість прикладів". Цей параметр визначає максимальну кількість прикладів, які можуть бути обрані для навчання кожного дерева в лісі. Зазвичай це стосується вибору підмножини даних із великого їх набору.

В алгоритмі цей параметр визначається як *"auto"*, тобто кількість прикладів для навчання кожного дерева дорівнює кількості загальних прикладів у наборі даних. Це означає, що для кожного дерева буде випадково обрано стільки прикладів, скільки доступно в наборі даних.

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №1 (27)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

Установлення іншого значення дозволяє контролювати розмір кожної підвибірки для навчання дерева. Наприклад, якщо встановлено значення 0.5, то для кожного дерева буде випадково обрано половину прикладів із загального набору даних для навчання. Це може бути корисним для зменшення обсягу інформації та прискорення процесу навчання, особливо для великих наборів даних.

Також важливо зауважити, що алгоритм знаходить ділянку в даних, де концентрація "звичайних" або нормальних значень переважає над аномальними. Це можна уявити як пошук ділянки на графіку, де найбільше точок з "нормальними" значеннями, водночас точки, які значно відрізняються від цієї ділянки, можуть бути розглянуті як аномалії або викиди. Наприклад, якщо на графіку точки згруповані у верхній частині, то значення в цій ділянці можуть вважатися "нормальними", тоді як точка, розташована посередині графіка, де незначна концентрація значень, може бути визначена як аномалія. Цей підхід відрізняється від традиційного методу розрахунку невизначеності, де немає такого акценту на аналізі концентрації значень. Замість цього, традиційні методи часто базуються на статистичних мірах центральної тенденції та розкиду даних.

Для зменшення випадковості результатів моделі додатково встановлюється початкове значення генератора випадкових чисел, що використовується для ініціалізації внутрішніх випадкових процесів в алгоритмі. Якщо значення встановлене на "1000", це генерує випадкове число від 0 до 999 включно щоразу, коли виконується алгоритм. Проте важливо зазначити, що якщо не зберегти це випадкове число та не застосувати його в подальших запусках моделі, кожен новий запуск генеруватиме нове випадкове число, що призведе до різних результатів. Однак, якщо обсяг даних не великий та не дуже глибоке дерево, вплив цих змін може бути незначним.

Додатково, щоб зменшити випадковість результатів, алгоритм запускає модель ізольованого лісу 10 разів, після чого результати усереднюються та подається середній результат запусків. Це дає змогу підвищити точність роботи моделі.

Як було сказано раніше, алгоритм будує 1 тис. дерев, з яких складається ліс (рис. 4).

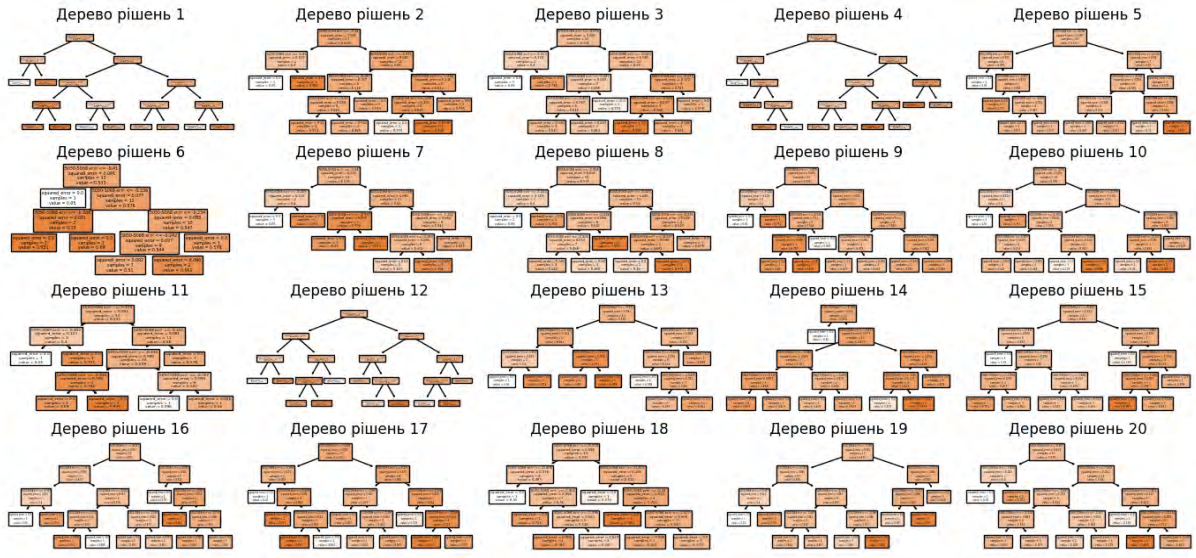

Рис. 4. Перші 20 дерев у лісі

Щоразу, коли модель будує дерево рішень, вона обирає випадкову підвибірку з початкових значень (у першій точці витрати цих значень – 11). Потім вона ізолює ці 11 значень у випадковому порядку на основі випадкової ознаки та будує дерево рішень. Цей процес повторюється 1 тис. разів, і щоразу він здійснюється саме на цих 11 значень у випадковому порядку.

Необхідно збалансовувати всі параметри, щоб досягти оптимальної ефективності алгоритму.

Крім того, важливо обирати відповідні значення для цих параметрів, що забезпечать оптимальну продуктивність моделі та якість її роботи. Різні значення параметрів можуть призвести до різних результатів і впливати на здатність моделі виявляти аномалії в даних.

ISSN 2522-9818 (print)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2024. No. 1 (27)

6. Результати дослідження

Оскільки значення отримані з одного й того самого коріолісового витратоміра, точність якого не змінюється залежно від витрати рідини, можемо об'єднати значення розкиду різних точок витрати. Такий підхід дасть змогу відтворити всі аномалії та викиди на одному зведеному графіку для зручності аналізу (рис. 5).

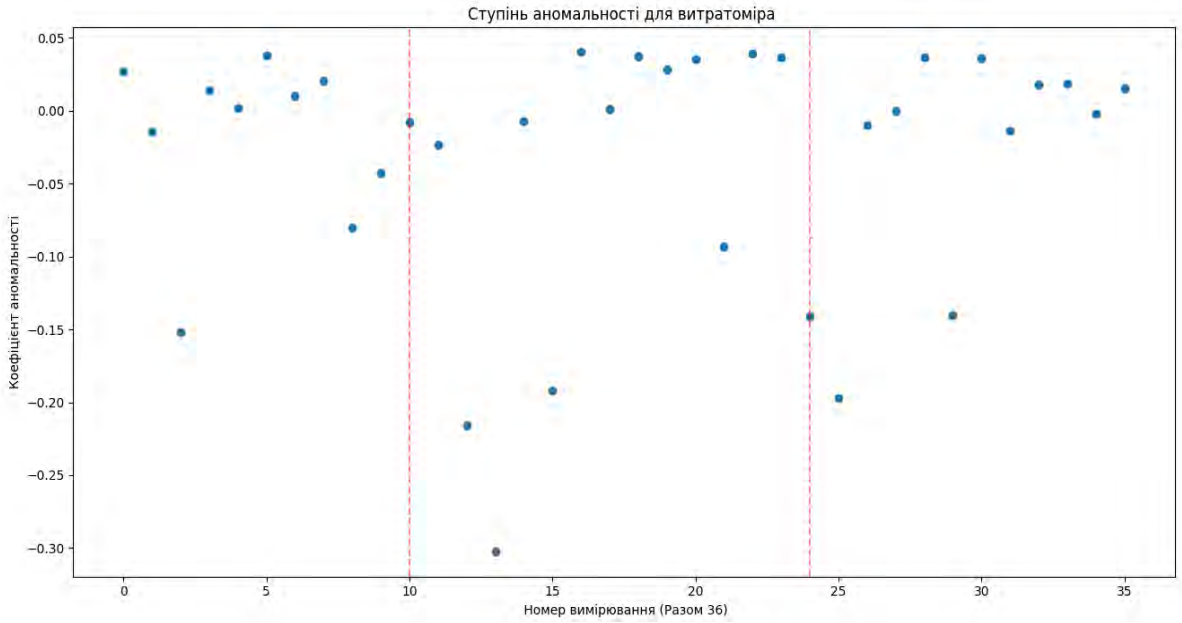

Рис. 5. Ступінь аномальності вимірювань витратоміра

На цьому зведеному графіку можна спостерігати розподіл аномальних значень для вимірювань одним витратоміром. Цей графік відтворює як систематичні, так і випадкові похибки, а також похибку самого витратоміра. Значення у верхній частині графіка показують нестабільність роботи обладнання еталона (систематична похибка) та неточність самого коріолісового витратоміра. Значення, віддалені від скупчення точок верхньої частини графіка, відтворюють випадкові похибки вимірювання. Значення, нижчі за 0.00 на шкалі ступеня аномальності, але не віддалені від скупчення, можуть містити як систематичні, так і випадкові помилки. Цей аналіз є лише візуальною оцінкою роботи моделі. Унаслідок роботи модель вважає значення, нижчі за 0.00, аномальними, а вищі – нормальними. Модель налаштована саме на виявлення випадкових помилок (аномалій), але дає змогу й надалі досліджувати результати.

7. Оцінка ефективності моделі

Традиційні методи виявлення викидів, такі як критерій Граббса, міжквартильний розподіл, середньоквадратичне відхилення тощо, можуть визначати точку як викид, якщо вона занадто віддалена від значень вибірки. У використанні цих критеріїв як оцінки ефективності моделі ізольованого лісу було виявлено лише одну точку як викид (лише критерієм Граббса, інші критерії не виявили у вибірці викидів), це є 13-та точка (найнижча на зведеному графіку). Розрахунок проводився для трьох вибірок аналогічно роботі алгоритму. Викид був знайдений в одній вибірці, в інших двох не було знайдено викидів. Порівняно з цим модель виявила дев'ять точок як аномальні (якщо не брати до уваги значення дуже наближенні до 0.00) з трьох вибірок окремо. Це вказує на те, що модель є більш ефективним інструментом для виявлення аномалій.

Крім того, якщо порівняти підхід моделі з традиційними методами встановлення невизначеності вимірювань, можна сказати, що розрахунки невизначеності надають кількісне значення систематичних і випадкових помилок, а модель дає розподіл цих помилок щодо їх місця у вибірці.

Однією з переваг моделі є можливості об'єднати результати значень однієї вибірки з іншими

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №1 (27)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

результатами. Це дає змогу бачити більше інформації, порівнювати результати та змінювати початкові умови проведення вимірювань. Це дозволить виявляти закономірності та причини виникнення систематичних та випадкових помилок. Подальший розвиток цього дослідження буде спрямовано саме на пошук цих закономірностей для підвищення точності та стабільності роботи на еталоні.

Висновки

Використання моделі ізольованого лісу є ефективним інструментом для виявлення аномалій у даних вимірювань, особливо в контексті метрології. Цей підхід дає змогу не лише ідентифікувати аномалії та вилучати їх для поліпшення точності вимірювань, але й порівнювати отримані значення між собою, незалежно від абсолютних значень. Завдяки цьому виникає більш повна картина розподілу аномалій та викидів у вимірюваних даних. Порівняно з традиційними методами виявлення викидів, модель ізольованого лісу дозволяє ефективніше й точніше виявляти аномалії, що робить її корисним інструментом для вирішення завдань метрології та інших галузей, де важлива висока точність вимірювань. Крім того, можливість порівняння різних вибірок даних та виявлення закономірностей у вимірювальних процесах дає змогу зменшувати невизначеність та покращувати якість результатів вимірювань.

Модель ізольованого лісу забезпечує інтуїтивно зрозумілий метод виявлення аномалій, що корисний для аналізу даних у метрології. Це розв'язок, який не потребує припущень про розподіл даних або структуру вибірки. Викиди можна не брати до уваги та вилучати з протоколу вимірювань як випадкові помилки, що допомагає зменшити невизначеність і покращити точність результатів. Цей підхід мінімізує вплив аномальних значень на загальний результат і підвищує достовірність отриманих даних. Способом вилучення викидів усувається викривлення, які можуть виникнути в статистичних характеристиках даних, таких як середнє значення або стандартне відхилення. Це сприяє більш точному оцінюванню невизначеності та покращує якість вимірювань.

Розвиток моделі ізольованого лісу відбувається в напрямі розширення сфери застосування та покращення її ефективності. Завдання додаткових досліджень – удосконалення алгоритмів виявлення аномалій та оптимізація технічних можливостей моделі для різних вимог індустрії. Такий підхід може забезпечити необхідну точність та достовірність результатів у різних сферах застосування, що робить модель ізольованого лісу важливим інструментом для подальшого розвитку метрології та наукових досліджень.

Список літератури

1. Chun S., Furuichi N. Final report of the APMP water flow supplementary comparison (APMP.M.FF-S1), *Metrologia,* Vol. 59, 2022. DOI: 10.1088/0026-1394/59/1A/07004
2. Frahm E., Arias R., Maldonado M., Vargas J., Mendoza J., Arredondo A., Silvosa M. Supplementary comparison SIM.M.FF-S9.2016 for water flow measurement, *Metrologia*, Vol. 61, 2024. DOI: 10.1088/0026-1394/61/1A/07001
3. Huovinen M., Frahm E. EURAMET.M.FF-S13 final report, *Metrologia*, Vol. 59, 2022. DOI: 10.1088/0026-1394/59/1A/07010.
4. ДСТУ-Н РМГ 43:2006 Метрологія. Застосування. Посібники з вираження невизначеності вимірювань, 2006.
5. Zakharov I., Serhiienko M., Chunikhina T. Measurement uncertainty evaluation by kurtosis method at calibration of a household water meter*, Metrology and Metrology Assurance* (*MMA*). P. 83–86. 2020. DOI: 10.1109/MMA49863.2020.9254260
6. Vallejo M., Espriella C., Gómez-Santamaría J., Ramírez-Barrera A., Delgado-Trejos E. Soft metrology based on machine learning: a review, *Measurement Science and Technology*, Vol. 31, No. 3. P. 1–16. 2019. DOI:10.1088/1361-6501/ab4b39
7. Kebir S., Tabia K. Anomaly Detection in Real Scarce Data: A Case Study on Monitoring Elderly's Physical Activity and Sleep, *IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)*, 2023, P. 385–392, DOI: 10.1109/BIBE60311.2023.00069
8. Yu B., Yu Y., Xu J., Xiang G., Yang Z. MAG: A Novel Approach for Effective Anomaly Detection in Spacecraft Telemetry Data, *IEEE Transactions on Industrial Informatics*, Vol. 20, No. 3, P. 3891–3899. 2014. DOI: 10.1109/TII.2023.3314852
9. Li Z., Wang P., Wang Z., Zhan D. FlowGANAnomaly: Flow-Based Anomaly Network Intrusion Detection with Adversarial Learning, *Chinese Journal of Electronics,* Vol. 33, No. 1, 2022. P. 58–71. DOI: 10.23919/cje.2022.00.173

10. Barbieri L., Brambilla M., Stefanutti M., Romano C., Carlo N., Roveri M. A Tiny Transformer-Based Anomaly Detection Framework for IoT Solutions, *IEEE Open Journal of Signal Processing*, Vol. 4, 2023. P. 462-478. DOI: 10.1109/OJSP.2023.3333756.
11. Guo N., Lin C., Yan H., Zang J., Xiong M. Real-Time Pantograph Anomaly Detection Using Unsupervised Deep Learning and K-Nearest Neighbor Classification, *IEEE Transactions on Instrumentation and Measurement*, Vol. 73, 2024. P. 1–13. DOI: 10.1109/TIM.2024.3370747
12. Occorso M., An M., Olsen R., Perry V.Anomaly Detection as a Data Reduction Approach for Test Event Analysis at the Edge, *IEEE International Conference on Big Data (BigData),* 2023. P. 3863–3867, DOI: 10.1109/BigData59044.2023.10386215
13. Xiang H., Zhang X., Dras M., Beheshti A., Dou W., Xu X. Deep Optimal Isolation Forest with Genetic Algorithm for Anomaly Detection, *IEEE International Conference on Data Mining (ICDM)*, 2023 P. 678–687, DOI: 10.1109/ICDM58522.2023.00077
14. Liu F., Ting K., Zhou Z. Isolation Forest, *IEEE International Conference on Data Mining*, 2008. P. 413–422, DOI: 10.1109/ICDM.2008.17
15. Jurado K., Ludvigson S., Ng S. Measuring Uncertainty, *American Economic Review*, Vol. 105 (3). 2015. P. 1177–1216**.** DOI: 10.1257/aer.20131193

References

1. Chun, S., Furuichi, N. (2022), "Final report of the APMP water flow supplementary comparison (APMP.M.FF-S1)" *Metrologia*, Vol. 59. DOI: 10.1088/0026-1394/59/1A/07004
2. Frahm, E., Arias, R., Maldonado, M., Vargas, J., Mendoza, J., Arredondo, A., Silvosa, M. (2024), "Supplementary comparison SIM.M.FF-S9.2016 for water flow measurement" *Metrologia*, Vol. 61, DOI: 10.1088/0026-1394/61/1A/07001
3. Huovinen, M., Frahm, E. (2022), "EURAMET.M.FF-S13 final report", *Metrologia*, Vol. 59, DOI: 10.1088/0026-1394/59/1A/07010.
4. DSTU-N RMG 43:2006 Metrology. Guidance on expressing measurement uncertainty [Metrolohiia. Kerivni vkazivky z vyrazhennia nevyznachennosti vymiriuvannia], 2006.
5. Zakharov, I., Serhiienko, M., Chunikhina, T. (2020), "Measurement uncertainty evaluation by kurtosis method at calibration of a household water meter", *Metrology and Metrology Assurance* (*MMA*) P. 83–86. DOI: 10.1109/MMA49863.2020.9254260
6. Vallejo, M., Espriella, C., Gómez-Santamaría, J., Ramírez-Barrera, A., Delgado-Trejos, E. (2019), "Soft metrology based on machine learning: a review", *Measurement Science and Technology*, Vol. 31, No. 3. P. 1–16. DOI: 10.1088/1361-6501/ab4b39
7. Kebir, S., Tabia, K. (2023), "Anomaly Detection in Real Scarce Data: A Case Study on Monitoring Elderly's Physical Activity and Sleep", *IEEE 23rd International Conference on Bioinformatics and Bioengineering (BIBE)*, P. 385–392, DOI: 10.1109/BIBE60311.2023.00069
8. Yu, B., Yu, Y., Xu, J., Xiang, G., Yang, Z. (2014), "MAG: A Novel Approach for Effective Anomaly Detection in Spacecraft Telemetry Data", *IEEE Transactions on Industrial Informatics*, Vol. 20, No. 3, P. 3891–3899, DOI: 10.1109/TII.2023.3314852
9. Li, Z., Wang, P., Wang, Z., Zhan, D., (2022), "FlowGANAnomaly: Flow-Based Anomaly Network Intrusion Detection with Adversarial Learning", *Chinese Journal of Electronics,* Vol. 33, No. 1, P. 58–71, DOI: 10.23919/cje.2022.00.173
10. Barbieri, L., Brambilla, M., Stefanutti, M., Romano, C., Carlo, N., Roveri, M. (2023), "A Tiny Transformer-Based Anomaly Detection Framework for IoT Solutions", *IEEE Open Journal of Signal Processing*, Vol. 4, P. 462–478, DOI: 10.1109/OJSP.2023.3333756
11. Guo, N., Lin, C., Yan, H., Zang, J., Xiong, M. (2024), "Real-Time Pantograph Anomaly Detection Using Unsupervised Deep Learning and K-Nearest Neighbor Classification", *IEEE Transactions on Instrumentation and Measurement*, Vol. 73, P. 1–13, DOI: 10.1109/TIM.2024.3370747
12. Occorso, M., An, M., Olsen, R., Perry, V. (2023), "Anomaly Detection as a Data Reduction Approach for Test Event Analysis at the Edge", *IEEE International Conference on Big Data (BigData),* P. 3863–3867, DOI: 10.1109/BigData59044.2023.10386215
13. Xiang, H., Zhang, X., Dras, M., Beheshti, A., Dou, W., Xu, X. (2023), "Deep Optimal Isolation Forest with Genetic Algorithm for Anomaly Detection", *IEEE International Conference on Data Mining (ICDM)*, P. 678–687, DOI: 10.1109/ICDM58522.2023.00077

Сучасний стан наукових досліджень та технологій в промисловості. 2024. №1 (27)
ISSN 2522-9818 (print)
ISSN 2524-2296 (online)

14. Liu, F., Ting, K., Zhou, Z. (2008), "Isolation Forest", *IEEE International Conference on Data Mining*, P. 413–422, DOI: 10.1109/ICDM.2008.17
15. Jurado, K., Ludvigson, S., Ng, S. (2015), "Measuring Uncertainty", *American Economic Review*, Vol. 105 (3). P. 1177–1216. DOI: 10.1257/aer.20131193

Надійшла 01.03.2024

Відомості про авторів / About the Authors

Ащепков Валерій Олегович – Харківського національного університету радіоелектроніки, молодший науковий співробітник Національного наукового центру "Інститут метрології", аспірант кафедри інформаційно-вимірювальної техніки, Харків, Україна; e-mail: ashhepkovvalera@gmail.com; ORCID ID: 0000-0003-3827-3445

Aschepkov Valeriy – Kharkiv National University of Radio Electronics, Junior Research Fellow at the National Scientific Center "Institute of Metrology", Postgraduate Student at the Department of Information Measurement Technology, Kharkiv, Ukraine.

THE USE OF THE ISOLATION FOREST MODEL FOR ANOMALY DETECTION IN MEASUREMENT DATA

The **subject** of the research is the Isolation Forest model, which is a powerful and efficient tool for detecting anomalies in measurement data and outliers, applicable in various fields where ensuring high accuracy and reliability of measurements is important. The **goal** of the study is to apply the Isolation Forest model to identify unusual or anomalous patterns that differ from typical patterns in the output data. This is achieved by isolating anomalous patterns from normal ones through the construction of multiple different decision trees. The **task** of the research is to detect outliers in data obtained during the preparation for international comparisons on the state primary standard for mass and volume flow rate of fluid, mass and volume of fluid flowing through a pipeline, by measuring with a coriolis flowmeter. Data collected during metrological studies undergo processing by the model to detect anomalies. This model analyzes the data and identifies anomalous or outlier values that may indicate systematic or random measurement errors. It enables quick and efficient detection of even the smallest deviations in the data, helping to maintain high accuracy and reliability of measurement results. The main **methods** for detecting outliers in statistical analysis, which are distribution-independent, are the Grubbs' criterion, interquartile range distribution, and standard deviation. They are sensitive to sample size but are simple and understandable tools. However, the Isolation Forest model also has its limitations, particularly it can be resource-demanding for large datasets. Additionally, it is necessary to consider that using the model requires proper parameter tuning to achieve optimal results. The **results** of the research include assessment of the Isolation Forest model's effectiveness by comparing it with traditional outlier detection methods. Comparative analysis of the results of different approaches to the same task is an effective method for evaluating the model's performance. **Conclusion.** The article concludes with the perspective of further research development in this direction. The work will focus on further developing methods for detecting anomalies in measurement data and improving the accuracy and reliability of measurement results in various application fields, which can find broad applications in science and industry.

Keywords: uncertainty; anomaly detection; measurement; metrology; data processing; machine learning algorithms; statistical methods.

Бібліографічні описи / Bibliographic descriptions

Ащепков В. О. Використання моделі *Isolation Forest* для виявлення аномалій у даних вимірювань. *Сучасний стан наукових досліджень та технологій в промисловості*. 2024. № 1 (27). С. 236–245. DOI: https://doi.org/10.30837/ITSSI.2024.27.236

Aschepkov, V. (2024), "The use of the Isolation Forest model for anomaly detection in measurement data", *Innovative Technologies and Scientific Solutions for Industries*, No. 1 (27), P. 236–245. DOI: https://doi.org/10.30837/ITSSI.2024.27.236