УДК 004.852 **DOI: 10.33042/2522-1809-2024-3-184-16-22**

О.В. Мойсеєнко

Івано-Франківський національний технічний університет нафти і газу, Україна

АДАПТАЦІЯ АЛГОРИТМУ ВИЯВЛЕННЯ АНОМАЛІЙ У ЧАСОВИХ РЯДАХ ДЛЯ НЕСТАЦІОНАРНИХ ПОТОКОВИХ ДАНИХ

У роботі запропоновано модифікацію алгоритму виявлення аномалій у часових рядах. Алгоритм забезпечує виявлення аномалій в нестаціонарних потокових даних. Прогнозна межа для типової поведінки системи визначається за допомогою теорії екстремальних значень. Запропонований алгоритм демонструє роботу в умовах зашумлених нестаціонарних даних у кількох класах часових рядів.

Ключові слова: *концептуальний дрейф, теорія екстремальних значень, багатовимірний часовий ряд, виявлення викидів.*

Постановка проблеми

Виявлення аномалій в потокових часових даних стало важливою темою дослідження завдяки широкому спектру можливих застосувань, серед яких виявлення екстремальних погодних умов, зловмисні атаки на об'єкти під охороною, контроль несанкціонованих витоків газу та нафти, незаконних під'єднань до трубопроводів, несправностей кабелю живлення, забруднення води та інших екологічних об'єктів. Швидке детектування нештатних режимів, виявлення цих критичних подій є важливим для захисту життів та/або активів. Тому розробка відповідних систем, методів детектування є актуальним завданням. Однак, оскільки ці програми витрачають більшу частину свого експлуатаційного часу в «типовому» стані і відповідні потокові дані отримують за допомогою мільйонів датчиків, ручний моніторинг є неефективним.

Виявлення аномалій є ключовою сферою для впровадження технологій штучного інтелекту (ШІ) та Інтернету речей і базується на принципі аналізу даних. В останні роки ШІ досягнув великих успіхів. У результаті дослідження в багатьох галузях, зокрема розпізнаванні зображень, автономне водіння, перетворення мови в текст і перетворення тексту в мову, дали вражальні результати.

Однак галузь виявлення аномалій ще не освоєна ШІ через притаманну їй складність і проблеми.

А. Лавін та інші [1] наводять три причини, чому виявлення аномалії є складним процесом:

1. Нескінченна множина аномальних моделей. Будь-який тип шаблону, крім звичайного шаблону даних (нормальних даних), можна назвати шаблоном аномальних даних. Існує лише кілька видів нормальних даних, але шаблони аномальних даних є дуже різноманітними (все, що не вписується в шаблон «нормальності»). Крім того, якщо дані є багатовимірним часовим рядом, кількість різних аномальних моделей зростатиме експоненціально. І з такою кількістю аномалій можна побудувати безліч аномальних моделей, які ніколи раніше не були отримані. Така різноманітність, зрозуміло, ускладнює проблему [2].

2. Дисбаланс даних. У типовому випадку кількість даних, що містять аномалії, є дуже малою порівняно з типовими даними. Такий дисбаланс ускладнює навчання моделей виявлення аномалій і, відповідно, фактичне їх виявлення.

Однією з найважливіших речей у навчанні моделям машинного та глибокого навчання є об'єм даних. Об'єм даних важливий для традиційних методів машинного навчання, але він також відіграє більший вплив на результат в методах глибокого навчання. З іншого боку, у задачі виявлення аномалії, незалежно від того, скільки нормальних даних є, важко отримати хорошу продуктивність виявлення аномалії, оскільки важливих даних аномалії занадто мало.

Дисбаланс даних також спричиняє труднощі у фактичному практичному застосуванні. Через характер проблеми виявлення аномалії «класифікація аномальних даних як нормальних (помилка 2-го типу, хибнонегативний результат)» є більш фатальною, ніж «класифікація нормальних даних як ненормальних (помилка 1-го типу, хибнопозитивний результат)». Це дуже складна проблема – знайти всі аномальні дані в часовому ряді, не пропустивши їх. Проте, якщо половина нормальних даних класифікується як аномалія, сенс виявлення аномалії втрачається.

3. Різні типи аномальних даних. Кожний тип аномальних даних має дуже різні характеристики, тому можна сказати, що важко виявити всі види аномальних даних.

У різних статтях є різні способи класифікації аномалій часових рядів, але можна виокремити три основні глобальні типи:

1) точкова аномалія. Дані, які не підпадають

під розподіл нормальних даних. Вони є найбільш вивченими;

2) контекстуальна аномалія (умовна аномалія). Це дані, в яких немає нічого незвичного в розподілі самих даних, але потік або контекст даних є ненормальним;

3) групова аномалія (колективні аномалії) – це дані, які на перший погляд виглядають нормально, якщо розглядати їх ізольовано. Однак це тип аномалій, який виглядає ненормальним, якщо розглядати його як групу даних.

Виявлення групових аномалій викликає найбільше труднощів, ніж точкових аномалій або контекстних аномалій, але останнім часом було проведено багато досліджень для їх вирішення за допомогою методів глибокого навчання, як-от варіаційні автокодери (VAE), змагальні автокодери (AAE) і генеративні змагальні мережі (GAN).

Аналіз останніх досліджень і публікацій

Існує багато методів, які були використані та досліджені в області виявлення аномалій у даних часових рядів для проблем, викладених вище.

Для виявлення аномалій у даних часових рядів у минулому використовувалось багато підходів, але найкращими вважаються наступні три методи: 3-сигма, Boxplot і ARIMA. Однак усі ці методи призначені для одновимірних часових рядів і мають той недолік, що Boxplot і 3-сигма можуть виявляти лише точкові аномалії [3, 4].

Розглянемо методи навчання на основі даних без маркування – «навчання без нагляду».

Найпоширенішим методом з них є модель на основі Autoencoder [5]. Під час використання Autoencoder припускається, що існує простір нижчої розмірності (множина), у якому дані великої розмірності можуть бути представлені простіше. Аномальні дані відображаються автокодером і перетворюються на різноманіття нормальних даних. Це різноманіття означає «ключові характеристики всіх навчальних даних», і важко включити особливості аномальних даних, яких є дуже невелика кількість.

Якщо використовувати модель на основі цих автокодерів, вона може аналізувати будь-які дані, які не мають звичайного шаблону, як аномальні дані, незалежно від типу чи шаблону аномалії. Однак колектор залежить від багатьох факторів, зокрема від структури моделі, що призводить до підйому та падіння продуктивності. Це означає, що продуктивність може значно відрізнятися залежно від налаштувань моделі (гіперпараметрів) і може бути важко знайти найкращу модель. Крім того, якщо аномальні дані близькі до нормальних даних або можуть бути представлені тим самим колектором, Autoencoder може успішно відновити навіть аномальні дані. Якщо це станеться, модель виявлення аномалії не зможе її добре виявити.

Однією з напівконтрольованих моделей виявлення аномалій на основі навчання є Deep SVDD [6]. Це модель, яка застосовує глибоке навчання до SVDD (Support Vector Data Description). Ця модель перетворює дані в простір ознак і вивчає дані, щоб знайти оптимальну гіперсферу, що оточує нормальні об'єкти. Дані всередині цієї гіперсфери нормальні, а дані зовні є аномальними.

Проводиться багато досліджень щодо виявлення аномалій в різних типах задач [7–10]. Деякі з них:

– Out-of-Distribution (OOD): задачі, які класифікують різні види стійких станів, але все ще дозволяють ідентифікувати дані як невідомі, коли ви стикаєтеся з ними вперше;

– контрастне навчання: метод неконтрольованого навчання, який вивчає подібність даних без міток, наприклад, нормальне / ненормальне, для досягнення точності, порівнянної з моделями контрольованого навчання;

– генеративна модель: наприклад, VAE (варіаційний автокодер) або GAN (генеративна змагальна мережа), підвищує точність шляхом вивчення розподілу навчальних даних;

– Transformer: моделі на основі Transformer, які добре показали себе в обробці природної мови (NLP), використовують для виявлення аномалій.

Проте значна кількість цих досліджень спрямована на виявлення аномалій у даних зображень, а не в даних часових рядів. Розглядаючи виявлення аномальних підпослідовностей у межах часового ряду, основну увагу зосереджують не на окремих спостереженнях, а на підпослідовностях, які значно відрізняються від решти послідовностей.

Алгоритм, запропонований К. Шварцем [11] з використанням теорії екстремальних значень, здатний виявляти як викиди такого характеру, так і адитивні викиди, і походить від роботи П. Берріджа та А. Тейлора [12].

Виявленню аномалій в потокових даних приділено дуже мало уваги порівняно із задачами виявлення двох інших типів аномалій. Виявленню аномалій третього типу наразі присвячена робота [13], в якій запропоновано метод, що використовує аналіз головних компонент, застосований до характеристик часових рядів разом із областями найвищої щільності та α-оболонками, щоб ідентифікувати незвичайні часові ряди у великій колекції часових рядів. Остання робота Л. Вілкінсона [14] також має здатність вирішувати проблеми такого характеру.

Мета статті

Метою роботи є модифікація алгоритму прогнозування часових рядів для раннього виявлення аномалій в нестаціонарних зашумлених потокових даних.

Виклад основного матеріалу

Підґрунтя для розробки адаптованого алгоритму виявлення аномалій.

Підходи до розв'язування задачі виявлення аномалій для часових даних можна розділити на два основні сценарії:

1) пакетна обробка;
2) аналіз потокових даних.

При пакетній обробці передбачається, що весь набір даних доступний до аналізу, і його метою є виявлення всіх наявних аномалій.

Сценарій потокових даних створює багато додаткових проблем через його складний характер і те, як дані змінюються з часом. Проблеми виникають за наявності великого об'єму й високої швидкості потокових даних, шумів і нестаціонарності сигналу (або «дрейфу концепції»). Останнє суттєво ускладнює ідентифікацію відмінностей між новою «типовою» поведінкою та аномальними подіями. Вирішення цієї проблеми вимагає, щоб алгоритм виявлення міг навчатися та адаптуватися до мінливих умов. Ці проблеми ускладнили наявні розв'язки для раннього виявлення аномалій у разі пакетного сценарію у контексті потокових даних.

Алгоритм, запропонований нами, базується на теорії екстремальних значень, розділі теорії ймовірностей, який стосується статистичної поведінки випадкових даних екстремального порядку у вибірці.

На сьогодні аномалії в основному визначаються за різницею / відхиленням або щільністю. Коли аномалії визначаються з погляду різниці, можна побачити відносно великі відхилення між типовими даними та аномаліями. У [11, 12, 14] наводять кілька прикладів цього підходу, коли спостереження (або кластери спостережень) з великими відхиленнями від найближчих сусідів визначаються як аномалії. При такому підході в процесі побудови моделі використовувалася «теорема інтервалу» [11] у теорії екстремальних значень. На противагу, визначення аномалії в термінах щільності спостережень означає, що аномалія – це спостереження (або кластер спостережень), яке має дуже низьку ймовірність виникнення. Робота [15], на якій базується метод [12], згадує можливість використання теорії екстремальних значень і непараметричних оцінок поведінки хвоста. В [16, 17] є декілька прикладів, коли теорія екстремальних значень була використана для пошуку спостережень, які мають екстремальну щільність. Основна увага цих методів була зосереджена на визначенні порога для щільності точок даних, щоб розрізняти аномалії та типові спостереження.

Модифікований нами алгоритм містить вдосконалення, які усувають два обмеження методів [13] і [14].

Метод, запропонований в [13], визначає найбільш незвичайний ряд у великій колекції часових рядів, незалежно від того, чи є якісь із них справді аномальними. Тоді як в нашому методі нештатна ситуація буде визначена лише за наявності аномальної події. Визначення межі типової поведінки та моніторинг нових точок даних, які потрапляють за межі, дозволить нам усунути цей недолік.

Метод «HDoutliers», запропонований Л. Вілкінсоном [14], ґрунтується на припущенні, що відстані найближчих сусідніх аномальних точок (або кластерів точок) будуть значно більшими за відстані між типовими точками даних.

Однак деякі дослідження не показують великих значень відхилень між типовими спостереженнями та аномаліями. Натомість аномалії відхиляються від більшості значень, або від області типових даних, поступово, без різкого великого відхилення між типовими та аномальними спостереженнями. Це особливо притаманно часовим рядам, а ще більше – часовим рядам з автокореляцією.

Запропонований нами алгоритм вимагає репрезентативного набору даних про типову поведінку системи. Сучасні технології збору та зберігання даних дають змогу багатьом організаціям накопичувати великі об'єми даних.

Оскільки, за визначенням, аномалії є винятком, як порівняти з типовою поведінкою системи, більшість доступних збережених даних мають відображати типову поведінку певної системи. Не обов'язково мати репрезентативні зразки всіх можливих типів типової поведінки цієї системи для того, щоб алгоритм працював добре. Основна ідея полягає в тому, щоб мати набір даних для розігріву, з якого можна отримати початкові значення параметрів моделі прийняття рішень.

Опис алгоритму.

Алгоритм 1 (Офлайн фаза: Побудова моделі типової поведінки).

Вхідні дані: D_{norm}, набір із n часових рядів (які можуть мати однакову або різну довжину), що генеруються за типової поведінки системи.

Вихідні дані: t^*, аномальний поріг.

І етап:

Обчислюємо m ознак (процедурно як в [18] і [13]) з кожного часового ряду в D_{norm}. Це створює матрицю ознак M розміром $n \times m$. Кожен рядок M відповідає часовому ряду, а кожен стовпець M відповідає типу ознаки.

Таке представлення часових рядів на основі ознак дозволяє нам обробляти часові ряди різної довжини та/або початкових точок, оскільки воно може перетворювати часові ряди будь-якої довжини чи початкових точок у вектор ознак фіксованого розміру. Це зменшує розмірність початкових багатоваріантних часових рядів за допомогою функцій, які інкапсулюють динамічні властивості окремих

часових рядів.

Усього визначаємо 14 ознак. Дев'ять ознак ми беремо аналогічно як в [13] (середнє значення, дисперсія, змінна дисперсія в залишку, зсув рівня з використанням рухомого вікна, зміна дисперсії, сила лінійності, сила кривизни і сила гостроти).

Ще п'ять ознак ми пропонуємо додатково: викиди часового ряду на основі фактора Фано (мінімальне та максимальне значення), відношення міжквартильного середнього до середнього арифметичного, відношення середніх значень даних, які більші та менші генерального середнього.

II етап:

Оскільки різні набори створюють ознаки в різних діапазонах, необхідно нормалізувати стовпці результівної матриці ознак M. Нехай M^* – нормалізована матриця ознак $n \times m$.

III етап:

Застосовуємо аналіз головних компонент до матриці ознак M^*.

IV етап:

Визначаємо двовимірний простір 2D за допомогою перших двох головних компонентів (PC) з кроку 3 (аналогічно як в [13]).

Кожна точка даних у цьому двовимірному просторі відповідає часовому ряду в D_{norm}. Це було зроблено з метою максимізації наших шансів на отримання інформації за допомогою візуалізації [19]. Результівний двовимірний простір – *2D PC space*.

V етап:

Визначаємо щільність ймовірності *2D PC space* за допомогою оцінки щільності ядра з двовимірним ядром Гауса.

VI етап:

Визначаємо кількість N екстремумів.

VII етап:

Підбираємо розподіл Гумбеля до отриманих значень. Значення параметрів розподілу Гумбеля обчислюємо за допомогою оцінки максимальної правдоподібності. Оскільки значення щільності ймовірності були розраховані за допомогою оцінки щільності ядра з двовимірним ядром Гауса, то можна довести, що $f\hat{}$ потрапляє в область максимального тяжіння розподілу Вейбулла.

VIII етап:

Використовуючи розподіл, отриманий на кроці 7, визначаємо розташування аномального порога *f2e*, прирівнявши відповідну функцію розподілу, яка є одновимірною в Ψ-просторі, до деякого значення достовірної ймовірності, наприклад, 0,999.

Згідно з рекомендаціями [20], для обробки потокових даних необхідно використати модель ковзного вікна. Після врахування w і t, які представляють довжину ковзного вікна та поточну часову точку, відповідно, наша мета полягає в тому, щоб ідентифікувати часові ряди, які є аномальними відносно типової поведінки системи. Ковзне вікно продовжує рухатися вперед з поточною точкою часу, зберігаючи фіксовану довжину вікна w. У результаті модель ігнорує всі дані, отримані до часу $(t - w)$. Крім того, термін дії кожного елемента даних закінчується рівно через w часових кроків.

Алгоритм 2 (Онлайн фаза: Тестування потокових даних).

Вхідні дані: $W[t - w, t]$, поточне ковзне вікно з n часовими рядами, t^*, аномальний поріг (результат Алгоритму 1).

Вихідні дані: вектор індексів ряду аномальних даних в часовому вікні $W[t - w, t]$.

1. Обчислюємо m ознак (ознаки, визначені на кроці 1 Алгоритму 1) для кожного з n часових рядів у $W[t - w, t]$. Отримаємо матрицю ознак M_{test} розміром $n \times m$.

2. Проєктуємо нову матрицю функцій M_{test}, на *2D PC space* із типовими даними, отриманими за допомогою часових рядів у D_{norm}.

Нехай $y_1, y_2, \ldots, y_n$ – це точки даних, отримані шляхом проєктування M_{test} на *2D PC space*.

3. Розраховуємо значення щільності ймовірності для вибірки $y_1, y_2, \ldots, y_n$ як $f\hat{}$ з кроку 5 Алгоритму 1.

4. Знаходимо будь-які y_j, які задовольняють умову $f\hat{}(y_j) < t^*$, де $j = 1, 2, \ldots, n$, і позначаємо відповідний часовий ряд (якщо такий є) як аномальний у часовому вікні $W[t - w, t]$.

5. Повторюємо кроки 1–4 фази онлайн для кожного нового часового вікна, яке генерується в поточний момент часу t.

Адаптація алгоритму для нестаціонарних часових рядів.

Нестаціонарність може проявлятися в багатьох різних формах. В економетричній літературі нестаціонарність іноді класифікують або як «структурний злам», або як «зміну в часі», еволюційну зміну [21]. У літературі з машинного навчання нестаціонарність відома як «дрейф концепції», а в [22] і [23] визначаються чотири класи нестаціонарності: раптові, наростальні, поступові та повторювані.

Згідно з [22], існує два підходи, які можуть бути використані для адаптації моделей, щоб впоратися з дрейфом концепції: сліпий та інформативний. При сліпому підході модель рішення оновлюється через регулярні проміжки часу без урахування реальності (незалежно від того, чи дійсно відбулася зміна чи ні). На противагу, інформативний підхід оновлює модель прийняття рішення лише в тому разі, якщо виявлено випадки «дрейфу концепції» [23].

Дотримуючись визначення в [24], ми пропонуємо підхід для раннього виявлення нестаціонарності, який використовує статистичні визначення різниці для обчислення відмінностей між розподілами, отриманими із колекції типових часових рядів, у

яких остання модель визначена і згенерована із типової серії в поточному тестовому вікні. Це дозволяє нам визначити, чи є якась значна різниця між останньою типовою поведінкою та новою типовою поведінкою.

Алгоритм 3 (Виявлення нестаціонарності).

Вхідні дані: w, довжина рухомого вікна, D_{t0}, сукупність n часових рядів довжиною w, які генеруються з n датчиків за останньої типової поведінки цієї системи, в якій визначено поточну модель прийняття рішень, W – тестовий потік.

Вихідні дані: вектор індексів аномального ряду в кожному вікні.

1. Обчислюємо f_{t0} – щільність ймовірності *2D PC space* для D_{t0}, використовуючи оцінку щільності ядра з двовимірним ядром Гауса.

2. Нехай $W[t - w, t]$ – поточне тестове вікно для n часових рядів довжиною w. Обчислюємо m ознак (крок 1 Алгоритму 1) для кожного з цих n часових рядів у $W[t - w, t]$. Отримаємо матрицю ознак M_{test} розміром $n \times m$.

3. Зіставляємо M_{test} на *2D PC space* D_{t0}. Нехай спроєктовані точки даних утворюють Y_{tt}.

4. Визначаємо точки у *2D PC space*, які відповідають типовому ряду в $W[t - w, t]$, з використанням аномального порога (результат Алгоритму 1), визначеного за допомогою D_{t0}. Нехай $Y_{ttnorm}(\subseteq Y_{tt})\}$ – набір точок даних у *2D PC space*, які відповідають типовому ряду $W[t - w, t]$, і $W[t - w, t]_{norm}$ $(\subseteq W[t - w, t])$ буде відповідати наборам типових часових рядів у $W[t - w, t]$.

5. Нехай p – частка аномалій, виявлених у $W[t - w, t]$. Якщо $p < p^*$, де $p^* > 0{,}5$, переходимо до кроку а); інакше переходимо до кроку б).

Ми приймаємо $p^* = 0{,}5$ за простим «правилом більшості». Однак також можна вибрати значення відмінне від 0,5 за замовчуванням, щоб максимізувати точність або запобігти ризикам невірної класифікації.

а) Обчислюємо f_{tt}, функцію щільності ймовірності набору Y_{ttnorm}, використовуючи оцінку щільності ядра з двовимірним ядром Гауса. Позначимо через $\hat{f}_{tt}$ отриману функцію щільності ймовірності.

б) Обчислюємо f_{tt}, функцію щільності ймовірності набору Y_{tt}, використовуючи оцінку щільності ядра з двовимірним ядром Гауса. Нехай $\hat{f}_{tt}$ – отримана функція щільності ймовірності. У разі «раптової» зміни всі (або більшість) точок у Y_{tt} можуть лежати за межами аномальної межі, визначеної D_{t0}. У результаті всі (або більшість) із цих точок у Y_{tt} будуть позначені як аномалії. Це може означати початок нової типової поведінки. Тож у цій ситуації рекомендується оновити модель рішення з використанням усіх рядів у поточному вікні (замість лише типового виявлені серії, які тепер представляють меншість), у такий спосіб дозволяючи моделі автоматично адаптуватися до мінливого середовища.

6. Використовуючи міру відмінності (наприклад, відстань Кульбака-Лейблера, відстань Хеллінгера, загальну варіаційну відстань або відстань Дженсена-Шеннона), перевіряємо нульову гіпотезу H_0: $f_{t0} = f_{tt}$.

Оскільки розподіл цих вимірювань відстані невідомий, для визначення критичних точок для тесту можна використовувати методи початкового завантаження, як пропонується в [25]. Однак ці обчислювальні методи не придатні для виявлення швидкої зміни в розподілах, що є фундаментальною вимогою для нашого аналізу потокових даних. Тому використовуємо підхід з [26]: перевіряємо нульову гіпотезу H_0: $f_{t0} = f_{tt}$, використовуючи квадрат міри розбіжності:

$$T = R\,[\,f_{t0}(x) - f_{tt}(x)]^2 dx\,, \qquad (1)$$

запропонованої у [25]. Оскільки тестова статистика, заснована на інтегральному квадраті відстані між двома оцінками щільності на основі ядра, є асимптотично нормальною за нульовою гіпотезою, вона дозволяє нам обійти громіздкі обчислення, які використовуються при звичайних методах повторної вибірки для обчислення критичних квантилів нульового розподілу.

7. Якщо H_0 відхиляється і $p < p^*$, D_{t0} встановлюється на $W[t - w, t]_{norm}$. Якщо H_0 відхиляється і $p > p^*$, D_{t0} встановлюється на $W[t - w, t]$.

8. Повторюємо кроки 1–7 для кожного нового часового вікна, яке генерується поточною точкою часу t.

Результати.

Для порівняння ефективності модифікації алгоритму обрано метод «HDoutliers», запропонований Л. Вілкінсоном [14].

Як було зазначено, запропонований алгоритм для тестування вимагає репрезентативної вибірки типової поведінки кожного з наборів даних для отримання початкового значення аномального порога. Однак для цих прикладів, взятих з відкритих джерел в інтернеті [7], ми не мали репрезентативних зразків типової поведінки відповідних систем. Тому ми обрали ковзне вікно $W[1, 100]$ і $W[1, 50]$ як репрезентативну вибірку типової поведінки, щоб отримати початкове значення для аномального порога.

Незважаючи на те, що для жодного з цих випадків не було належного репрезентативного зразка типової поведінки, запропонований нами Алгоритм 3 для виявлення нестаціонарності дозволив моделі адаптуватися до типової поведінки системи з часом. На рис. 1 подано відповідні значення p для перевірки гіпотези H_0: $f_{t0} = f_{tt}$, описаної на кроці 6 Алгоритму 3. Для кожного з цих прикладів рівень значущості ми встановили на 0,05 (95 % ймовірність).

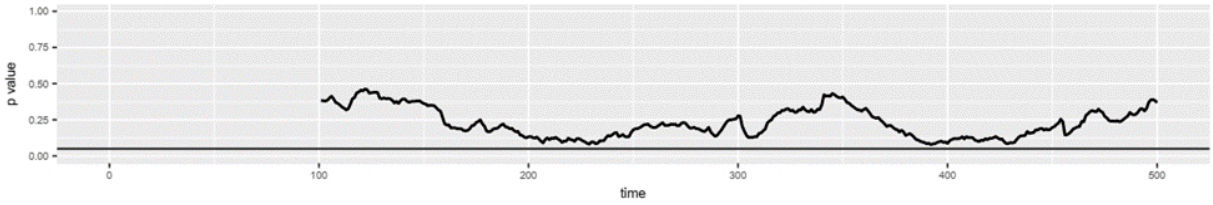

Рис. 1. Ймовірність виявлення дрейфу концепції

Наша запропонована структура (поєднання Алгоритмів 1, 2, 3) показує середній рівень точності 0,95 за весь період, зберігаючи при цьому низькі хибнопозитивний (в середньому 0,047) і хибнонегативний (0,000 у середньому) результати протягом періоду часу, що розглядається.

Висновки

Запропоновано модифікацію алгоритму для виявлення аномальних рядів у великій колекції потокових часових рядів за допомогою теорії екстремальних значень. Ми визначаємо аномалію як спостереження, яке є дуже малоймовірним, з огляду на розподіл типової поведінки такої системи.

Адаптація алгоритму для роботи з нестаціонарними даними здійснюється за рахунок застосування ковзного вікна порівняння щільності ознак, що дозволяє моделі прийняття рішень автоматично пристосовуватись до мінливого середовища відповідно до виявлених змін.

Попередні дослідження, проведені як на синтетичних даних, так і даних, отриманих від оптичних кабелів, показують, що запропонована структура (Алгоритми 1, 2 і 3) може добре працювати при наявності нестаціонарних і зашумлених часових рядів мультимодальних типових класів.

References

1. Lavin A. Evaluating Real-Time Anomaly Detection Algorithms–The Numenta Anomaly Benchmark / Lavin A, Ahmad S. // Machine Learning and Applications (ICMLA), IEEE 14th International Conference on. IEEE, 2015, pp. 38–44.
2. Russakovsky O. Imagenet large scale visual recognition challenge / Russakovsky O // International journal of computer vision 115.3, 2015: 211–252.
3. Makridakis S. The M5 accuracy competition: Results, findings and conclusions / S.Makridakis, E. Spiliotis, V. Assimakopoulos // Int J Forecast, 2020.
4. Makridakis S. The M5 accuracy competition: Results, findings and conclusions / S.Makridakis, E. Spiliotis, V. Assimakopoulos // International Journal of Forecasting (2020): 1–24.
5. Tziolas, T., Papageorgiou, K., Theodosiou, T., Papageorgiou, E., Mastos, T., & Papadopoulos, A. (2022). Autoencoders for Anomaly Detection in an Industrial Multivariate Time Series Dataset. *Engineering Proceedings*, *18*(1), 23. https://doi.org/10.3390/engproc2022018023
6. Ruff L. Deep semi-supervised anomaly detection / Ruff L. // arXiv preprint arXiv:1906.02694, - 2019.
7. Krohn D. Fiber optic sensors: fundamentals and applications / D.A Krohn, T. MacDougall, A. Mendez // Isa, - 2000.
8. Catalano A. An intrusion detection system for the protection of railway assets using Fiber Bragg Grating sensors / A Catalano, FA Bruno, M Pisco, A Cutolo, A Cusano // Sensors 14(10), 2014, 18268–18285.
9. Gupta M. Outlier detection for temporal data: A survey / M Gupta, J Gao, C Aggarwal, J Han // IEEE Transactions on Knowledge and Data Engineering 26(9), 2014, 2250–2267.
10. Hayes M. Contextual anomaly detection framework for big sensor data / M. Hayes, M. Capretz // Journal of Big Data 2(1), 2015.
11. Schwarz K. Wind dispersion of carbon dioxide leaking from underground sequestration, and outlier detection in eddy covariance data using extreme value theory / K. Schwarz // ProQuest. 2008.
12. Burridge P. Additive Outlier Detection Via Extreme-Value Theory / P. Burridge, A. Taylor // Journal of Time Series Analysis 27(5), 2066, p. 685–701.
13. Hyndman RJ. Large-scale unusual time series detection. In: Data Mining Workshop (ICDMW)/ RJ Hyndman, E Wang, N Laptev // IEEE International Conference on. IEEE, - 2015, pp. 1616–1619.
14. Wilkinson L. Visualizing Big Data Outliers through Distributed Aggregation / L. Wilkinson // IEEE transactions on visualization and computer graphics 24(1), 2018, p. 256–266.
15. Perron P. Searching for additive outliers in nonstationary time series / P. Perron, G Rodríguez // Journal of Time Series Analysis 24(2), 2009, p. 193–220.
16. Sundaram S. Aircraft engine health monitoring using density modelling and extreme value statistics. / Sundaram S, IGD Strachan, DA Clifton, L Tarassenko, S King // Proceedings of the 6th International Conference on Condition Monitoring and Machine Failure Prevention Technologies, 2009.
17. S. Hugueny. Novelty detection with extreme value theory in vital-sign monitoring. PhD thesis. University of Oxford, 2013.
18. Fulcher BD. Highly comparative time-series analysis. PhD thesis. University of Oxford, 2012.
19. Kang Y. Visualising forecasting algorithm performance using time series instance spaces / Y. Kang, RJ. Hyndman, K. Smith-Miles // International Journal of Forecasting 33(2), 2017, p. 345–358.
20. Jin R. Frequent pattern mining in data streams / R. Jin, G. Agrawal // Data Streams. Springer, 2007, pp. 61–84.
21. Rapach DE. Structural breaks and GARCH models of exchange rate volatility / DE. Rapach, JK. Strauss // Journal of Applied Econometrics 23(1), 2008, p. 65–90.
22. Gama J. A survey on concept drift ˙ adaptation / J. Gama, I. Žliobaite, A. Bifet, M. Pechenizkiy, A Bouchachia // ACM Computing Surveys (CSUR) 46(4), 2014, p. 44.
23. Faria ER. Novelty detection in data streams / ER. Faria, IJ. Gonçalves, AC. de Carvalho, J. Gama // Artificial Intelligence Review 45(2), 2016, p.235–269.
24. Dries A. Adaptive concept drift detection / A. Dries, U. Rückert // Statistical Analysis and Data Mining 2(5-6), 2009, p. 311–327.
25. Anderson NH. Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates / NH Ander-

son, P Hall, DM Titterington // Journal of Multivariate Analysis 50(1), 1994, p.41–54.
26. Duong T. Closed-form density-based framework for automatic detection of cellular morphology changes/ T. Duong, B. Goud, K. Schauer // Proceedings of the National Academy of Sciences 109(22), 2012, p. 8382–8387.

Рецензент: д-р техн. наук, проф. С.І. Мельничук, Івано-Франківський національний технічний університет нафти і газу, Україна.

Автор: МОЙСЕЄНКО Олена Володимирівна
кандидат технічних наук, доцент, доцент кафедри комп'ютерних систем і мереж
Івано-Франківський національний технічний університет нафти і газу
E-mail – olena.moiseienko@nung.edu.ua
ID ORCID: https://orcid.org/0000-0002-7995-2949

ADAPTATION OF THE ALGORITHM FOR DETECTING ANOMALIES IN TIME SERIES FOR NON-STATIONARY STREAMING DATA

O. Moiseienko
Ivano-Frankivsk National Technical University of Oil and Gas, Ukraine

Detection of anomalies in streaming time series data has become an important research topic due to the wide range of possible applications, including the detection of extreme weather conditions, malicious attacks on protected facilities, monitoring unauthorised gas and oil leaks, illegal pipeline connections, power cable faults, and water and other environmental pollution. Rapid detection of abnormal conditions and identifying these critical events is essential to protect lives and assets. Therefore, developing appropriate systems and detection methods is an urgent task.

Anomaly detection in streaming data is challenging due to its large volume and high speed, presence of noise, and non-stationarity of the signal (or 'concept drift'). The latter significantly complicates the identification of differences between new 'typical' behaviour and abnormal events.

Solving this problem requires the algorithm for processing such data to learn and adapt to changing conditions.

The paper proposes a modification of the algorithm for detecting anomalies in time series. This algorithm provides early detection of abnormal series in a massive collection of non-stationary streaming time series data. Anomalies are observations that are highly improbable given the previous time series values. The proposed approach is based on the primary detection of the predictive limit for the typical system behaviour using the theory of extreme values, followed by checking for the abnormality of the following series using the sliding window technique. The time series parameters are used as input data and compared by density distribution to detect any significant changes in the distribution of the characteristics. It allows the decision-making model to automatically adapt to the changing environment according to detected changes.

Since anomalies are, by definition, exceptions to the typical behaviour of a system, most of the available stored data should reflect that typical behaviour of the system in question. It is unnecessary to have representative samples of all possible types of the standard behaviour of a given system for the algorithm to work well. The basic idea is to have a warm-up data set to obtain initial values for the decision model parameters. It makes it possible to determine if there is any significant difference between the last typical behaviour and the new typical behaviour.

The proposed algorithm demonstrates its performance under conditions of noisy non-stationary data in several time series classes.

Keywords: *concept drift, extreme value theory, multivariate time series, outlier detection.*