УДК 004.942

ПОШУК АНОМАЛІЙ У ЧАСОВИХ РЯДАХ

О. О. Колодяжна[1], А. М. Родіонов[1]

[1] *Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», Фізико-технічний інститут*

Анотація

Кожного дня ми маємо справу з все більшими обсягами інформації. Проблемами, які можуть виникати у даних, є, наприклад, недопустимі, екстремальні значення. Тому задача пошуку аномалій у даних наразі стоїть гостро. У даній роботі проведено порівняльний аналіз статистичних методів та методів машинного навчання для виявлення аномалій у часових рядах. За результатами встановлено, що методи машинного навчання можуть дати кращі результати, ніж статистичні.

Ключові слова: аномалія, машинне навчання, нейронні мережі, часовий ряд

Вступ

Дані стають дедалі важливішими у всіх сферах нашого життя. Вони є ключовими для діяльності багатьох установ. Завдяки доступу до все більшого обсягу інформації недостатньо просто збирати та зберігати її, дані потрібно обробляти та аналізувати. Однією з важливих задач є виявлення аномалій у наборах даних. Мета виявлення аномалій — знаходження випадків, які є незвичними, екстремальними для певного набору даних. Як результат, ми можемо відслідкувати випадки шахрайства, вторгнення в мережу, поломку якого-небудь приладу на виробництві тощо. У даній роботі виконано порівняльний аналіз методів виявлення аномалій у часових рядах.

1. Поняття часового ряду

Часовий ряд – це впорядкована послідовність певних спостережень та значень яких-небудь їх параметрів. Зазвичай дані спостережень узяті при однакових інтервалах часу. Кожному значенню в часовому ряді повинен відповідати час вимірювання або номер за порядком. Цей тип даних використовується для прогнозування наступних значень та виявлення аномалій.

У даній роботі будемо розглядати одновимірні часові ряди.

2. Поняття аномалії

Аномалії — це закономірності в даних, які не відповідають чітко визначеному поняттю нормальної поведінки [1].

Формально виявлення аномалій можна описати як функцію φ:

$$\varphi : \mathbb{R}^n \to \mathbb{R}, \quad \varphi(x) \mapsto \gamma$$

де γ — міра аномальності екземплярів даних, $x \in X \subset \mathbb{R}^n$, X — набір даних. Для того, щоб визначити, відноситься даний об'єкт до аномальних чи ні, вводиться певне порогове значення $\delta \in \mathbb{R}$. Ті точки даних, що відповідають значенням $\gamma > \delta$, класифікуються як аномалії.

3. Виявлення аномалій в одновимірних часових рядах

Виявлення аномалій у часових рядах тісно пов'язане з аналізом та прогнозуванням часових рядів. Зазвичай прогнозуючу модель навчають на тренувальних даних і далі застосовують для отримання результату на тестових даних. Часто також при застосуванні алгоритмів використовуються ковзні вікна.

Міру аномальності певного об'єкту можна оцінювати за допомогою функції відстані. Тобто аномальність екземпляра даних будемо визначати як відстань між передбаченим значенням цього екземпляра та його істинним значенням:

$$\varepsilon_i = d(x_i, \hat{x}_i)$$

де x_i — істинне значення, $\hat{x}_i$ — передбачене значення, d — функція відстані, наприклад, евклідова відстань. Якщо значення ε_i перевищує певний поріг, тобто $\varepsilon_i > \delta$, $\delta \in \mathbb{R}$, то дана точка вважається аномалією.

3.1. Методи виявлення аномалій

Загалом усі методи виявлення аномалій можна розділити на 3 групи: статистичні методи, алгоритми машинного навчання та алгоритми глибинного навчання. Розгляньмо коротко кожну з підгруп.

1) Статистичний підхід
   Статистичні методи виявлення аномалій спираються на припущення, що «нормальні екземпляри даних трапляються в областях високої ймовірності стохастичної моделі, тоді як аномалії мають місце в областях низької ймовірності стохастичної моделі» [1].

Основним та найпростішим методом є так зване правило 3σ, яке ґрунтується на припущенні, що дані розподілені за нормальним законом. У такому випадку аномаліями вважаються ті об'єкти, які знаходяться на відстані більше 3σ від середнього значення, де σ — стандартне відхилення. Для прогнозування часового ряду розглядають також авторегресивні моделі (AR), моделі ковзного середнього (MA) та модель ARIMA, що є узагальненою комбінацією двох попередніх моделей.

В авторегресії значення залежної змінної прогнозується, використовуючи лінійну комбінацію її попередніх значень.

$$X_t = c + \sum_{1}^{p} \alpha_i X_{t-1} + \varepsilon_t$$

де p — порядок авторегресивної моделі, ε_t — білий шум. Значення ε_t визначають ступінь аномальності об'єкта та можуть бути знайдені як різниця прогнозованого та істинного значень [2]. Моделі ковзного середнього використовують залежність між залишковими значеннями для прогнозування значень у наступні періоди часу. Такі моделі допомагають підготуватися до непередбачуваних подій, катастрофічних подій.

$$X_t = \mu + \sum_{1}^{q} \theta_i X_{t-1} + \varepsilon_t$$

де q – порядок моделі, μ – середнє значення часового ряду. Для виявлення аномалій у наборі даних застосовується той же принцип, що і для авторегресивної моделі.

Зазвичай часові ряди не є стаціонарними, а моделі, які були згадані, застосовуються саме для стаціонарних часових рядів. На допомогу у випадках нестаціонарності приходить модель ARIMA. ARIMA(p, d, q) включає в себе модель AR p-го порядку, MA q-го порядку та додатковий параметр d, який допомагає зробити ряд стаціонарним перед прогнозуванням. Параметр d визначає, скільки разів потрібно знайти різницю часового ряду. Тобто якщо $d = 1$, то $X_t^{'} = X_t - X_{t-1}$. Такий підхід позбавляє часові ряди їх тренду та повертає ряд із постійним середнім.

2) Алгоритми машинного навчання

Наразі існує досить багато різних методів машинного навчання, які застосовуються для виявлення аномалій у даних. Одним із прикладів є Isolation Forest [3], проте зазвичай даний алгоритм застосовується до даних великої розмірності. Ще одним алгоритмом, який не був розрахованим для застосування у часових рядах, є LOF. Проте в роботі [4] було розширено його область застосування. Для виявлення аномалій застосовується також кластерний метод на основі щільності DBSCAN [5]. Даний алгоритм досить гарно справляється з даними великої розмірності, проте його застосовують і для одновимірних часових рядів, наприклад, у роботі [6] DBSCAN був використаний для даних щодо середньодобових показників температури протягом 33 років.

У даній роботі буде розглянуто результати роботи алгоритму DBSCAN. Даний метод розділяє точки на кластери, кількість яких встановлюється самим алгоритмом. Також після закінчення роботи алгоритму не всі точки можуть мати свій клас. Саме ті точки, які залишаються, розуміють як аномальні. DBSCAN містить у собі два гіперпараметри: ε та k, де ε — відстань, на якій розглядаються сусіди даної точки, а гіперпараметр k відповідає за мінімальну кількість точок, яка потрібна для утворення кластера.

3) Алгоритми глибинного навчання

Останнім часом нейронні мережі стали дуже популярними. Вони показали неймовірні результати в різних задачах комп'ютерого зору, обробки мови тощо. Також виникає інтерес до застосування нейронних мереж для задач прогнозування часових рядів. Багато дослідників у своїх роботах порівнюють результати, отримані за допомогою класичних моделей прогнозування, таких як ARIMA, з результатами, отриманими за допомогою нейронних мереж. Наприклад, у роботі [7] автори порівнюють нейронні мережі з ARIMA, зосередившись на 16 часових рядах різної складності.

У даній статті буде розглянуто результати роботи згорткових нейронних мереж (CNN) та вентильних рекурентних вузлів (GRU). Зазвичай згорткові нейронні мережі застосовують у випадках, коли вхідними даними є картинки. Але якщо заглибитися в них, то побачимо, що вони шукають певні шаблони, закономірності в зображеннях. Часовим рядам теж притаманні певні закономірності, тому CNN є доречною моделлю для прогнозування часових рядів. До картинок застосовують двомірні (2D) згортки. Тоді як до послідовностей застосовуються одновимірні (1D) згортки. Архітектура GRU є дуже подібною до LSTM, проте має тільки 2 вентилі замість 3. Вона не містить вентиль виходу. Саме це спрощення дозволяє GRU мати менше параметрів та швидше працювати. Оскільки у даній роботі маємо справу з не дуже великим набором даних, то перевага була надана вентильним рекурентним вузлам.

4. Набір даних

Для оцінки наведених алгоритмів було обрано 3 різні датасети з одновимірними часовими рядами. Обрані набори даних зібрані службою AmazonCloudwatch та містять показники щодо використання ЦП. Кожен набір складається з 4032 екземплярів даних. Значення зафіксовані з інтервалом у 5 хвилин. Дані узяті з Numenta Anomaly Benchmark (NAB), що є еталоном оцінки алгоритмів виявлення аномалій, особливо на потокових даних. Усі аномалії в наборах даних є промаркованими ав-

торами. Два датасети містять по 2 аномалії, і третій набір даних містить у собі 1 аномальну точку.

Усі 3 набори даних були розбиті на тренувальні й тестові набори в пропорції 80%/20% відповідно. На тренувальних наборах налаштовувалися всі параметри моделей. Також перед застосуванням алгоритмів датасети були стандартизовані.

5. Метрики оцінки роботи алгоритмів

Для оцінки алгоритмів використаємо F_2-score. F_β-score є комбінацією precision та recall. Дана оцінка обчислюється за формулою:

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{\beta^2 \cdot \text{Precision} + \text{Recall}}$$

У нашому випадку $\beta = 2$.

Recall показує, яку долю аномалій розпізнав алгоритм. Дана метрика обраховується за формулою:

$$\text{Recall} = \frac{TP}{TP + FN}$$

де TP — кількість правильно класифікованих об'єктів, FN — кількість пропущених алгоритмом аномалій.

Precision показує, наскільки точною є наша модель, та обчислюється за формулою:

$$\text{Precision} = \frac{TP}{TP + FP}$$

де FP — кількість об'єктів, які не є аномаліями, проте алгоритм позначив їх аномаліями.

6. Результати

У табл. 1 наведені оцінки результатів роботи кожного з алгоритмів для кожного з набору даних. Візуально результати представлені на рис. 1.

Name	**AR**	**MA**	**ARIMA**
First dataset	0.5	0.5556	0.5556
Second dataset	1.0	1.0	0.9091
Third dataset	0.7143	0.8334	0.625
	DBSCAN	**CNN**	**GRU**
First dataset	1.0	0.9091	0.4348
Second dataset	0.5556	1.0	1.0
Third dataset	1.0	0.625	0.625

Табл. 1. F_2-score для кожного датасету

З табл. 1 та рис. 1 бачимо, що всі алгоритми непогано справилися з завданням. Моделі MA та ARIMA у першому датасеті пропускають одну з аномалій, а AR помічає досить багато зайвих аномалій. DBSCAN показав себе найкраще на даних першого та третього датасетів.

Висновки

У роботі проведено аналіз статистичних методів, методів машинного і глибинного навчання для виявлення аномалій у часових рядах. За результатами

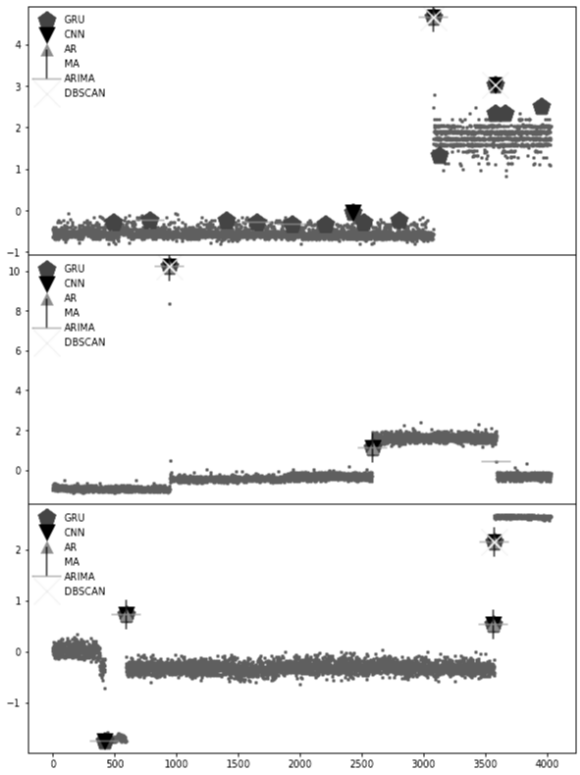

Рис. 1. Результати роботи усіх алгоритмів

експериментів з'ясовано, що в середньому методи машинного та глибинного навчання дають кращу точність, ніж статистичні методи на основі прогнозування часових рядів.

Перелік використаних джерел

1. Chandola V., Banerjee A., Kumar V. Anomaly Detection: A Survey // ACM Comput. Surv. — 2009. — Vol. 41, no. 3. — P. 2–33.
2. Chandola V. Anomaly Detection for Symbolic Sequences and Time Series Data : Ph. D. thesis. — University of Minnesota, 2009.
3. Liu F. T., Ting K. M., Zhou Z.-H. Isolation Forest // 2008 Eighth IEEE International Conference on Data Mining. — P. 413–422.
4. Oehmcke S., Zielinski O., Kramer O. Event Detection in Marine Time Series Data // KI 2015: Advances in Artificial Intelligence. — Springer. — P. 279–286.
5. A density-based algorithm for discovering clusters in large spatial databases with noise / Ester M., Kriegel H.-P., Sander J., and Xu X. // Second International Conference on Knowledge Discovery and Data Mining. — 1996. — P. 226–231. — Access mode: `https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf`.
6. Çelik M., Dadaşer-Çelik F., Dokuz A. Ş. Anomaly detection in temperature data using DBSCAN algorithm // 2011 International Symposium on Innovations in Intelligent Systems and Applications. — P. 91–95.
7. Tang Z., Fishwick P. A. Feedforward Neural Nets as Models for Time Series Forecasting // ORSA Journal on Computing. — 1993. — Vol. 5, no. 4. — P. 374–385.