

ЗАСТОСУВАННЯ БІБЛІОТЕКИ DARTS ДЛЯ ПОШУКУ АНОМАЛІЙ У ЧАСОВИХ РЯДАХ СТАНУ АТМОСФЕРНОГО ПОВІТРЯ

Вінницький національний технічний університет

Анотація

У роботі розглянуто застосування бібліотеки Darts як інструменту для пошуку аномалій у часових рядах якості атмосферного повітря. Здійснено огляд відповідного модуля бібліотеки та його основних компонентів, які дозволяють будувати гнучкі конвеєри для ідентифікації аномальних значень. Проаналізовано зручність використання Darts у середовищі Python, її інтеграцію з іншими популярними бібліотеками. Отримані результати на даних моніторингу PM2.5 мережі EcoCity свідчать про доцільність використання бібліотеки Darts для автоматизованого виявлення аномалій у задачах аналізу даних.

Ключові слова: машинне навчання, аномалії часових рядів, якість повітря, часові ряди, Darts, Python.

Abstract

This paper examines the application of the Darts library as a tool for anomaly detection in air quality time series. An overview of the library's anomaly detection module and its core components is provided, enabling the construction of flexible pipelines for identifying anomalous values. The usability of Darts within the Python environment and its integration with other popular libraries are analyzed. The results obtained on PM2.5 monitoring data from the EcoCity network indicate that the Darts library is effective for automated anomaly detection in data analysis tasks.

Keywords: machine learning, time series anomalies, air quality, time series, Darts, Python.

Вступ

Аналіз часових рядів є важливим науковим та практичним напрямом, який охоплює велику кількість сфер. Часові ряди широко використовуються у різних галузях, зокрема в економіці, екології та інформаційних технологіях. У сучасних умовах зростання об'ємів даних особливо актуальним є застосування інструментів, які дозволяють ефективно аналізувати ці дані. Для різних задач аналізу часових рядів активно застосовуються програмні бібліотеки, що реалізують різноманітні статистичні методи та підходи машинного навчання. Однією з таких бібліотек є Darts, розроблена для роботи з часовими рядами у середовищі Python. Вона забезпечує уніфікований інтерфейс для використання як класичних статистичних моделей, так і сучасних моделей машинного навчання. Завдяки цьому Darts є зручним інструментом для проведення досліджень і практичного застосування методів аналізу та прогнозування часових рядів.

Можливості бібліотеки

Бібліотека Darts є сучасним інструментом для аналізу, моделювання та прогнозування часових рядів у середовищі Python. Вона надає уніфікований інтерфейс для роботи з різними моделями та методами, що значно спрощує процес дослідження та побудови прогнозних моделей. Однією з важливих функцій Darts є зручне представлення часових рядів за допомогою спеціального об'єкта TimeSeries, який дозволяє працювати як з одновимірними, так і з багатовимірними даними. Це забезпечує ефективну організацію даних, підтримку часових індексів та можливість виконання різних операцій над часовими рядами, зокрема об'єднання, масштабування, фільтрації та трансформації. Бібліотека підтримує широкий набір моделей прогнозування. До них належать класичні статистичні моделі, такі як ARIMA, SARIMA, експоненційне згладжування, а також сучасні методи машинного навчання. Це дозволяє будувати більш точні моделі, враховуючи зовнішні змінні. Крім того, Darts підтримує побудову ансамблевих моделей, що поєднують прогнози кількох алгоритмів для підвищення точності прогнозування.

Методи пошуку аномалій

Одним з важливих кроків аналізу часових рядів є пошук аномалій. Бібліотека Darts містить окремий модуль, що надає набір інструментів для вирішення цієї задачі. Практичну перевірку роботи бібліотеки здійснено на даних моніторингу стану атмосферного повітря у Вінницькій області за 2022–2026 роки, отриманих від мережі громадського моніторингу EcoCity.

Модуль складається з чотирьох елементів, які можуть працювати як окремо, так і в комбінації та будувати конвеєри, що дозволяє досліднику варіювати методи оцінювання аномалій та стратегії ансамблювання:

1. Scorer – це компонент бібліотеки, який обчислює числову оцінку аномалії для кожної точки часового ряду. Цей компонент повертає числову оцінку аномалії у вигляді часового ряду (рис. 1). Далі цей ряд може використовуватися іншими компонентами для визначення того чи є точка аномальною. Прикладами таких компонентів є:

- a) NormScorer – обчислює відхилення фактичного значення від прогнозу моделі.
- b) KMeansScorer – використовує метод KMeans на фрагментах ряду для оцінки аномальності.

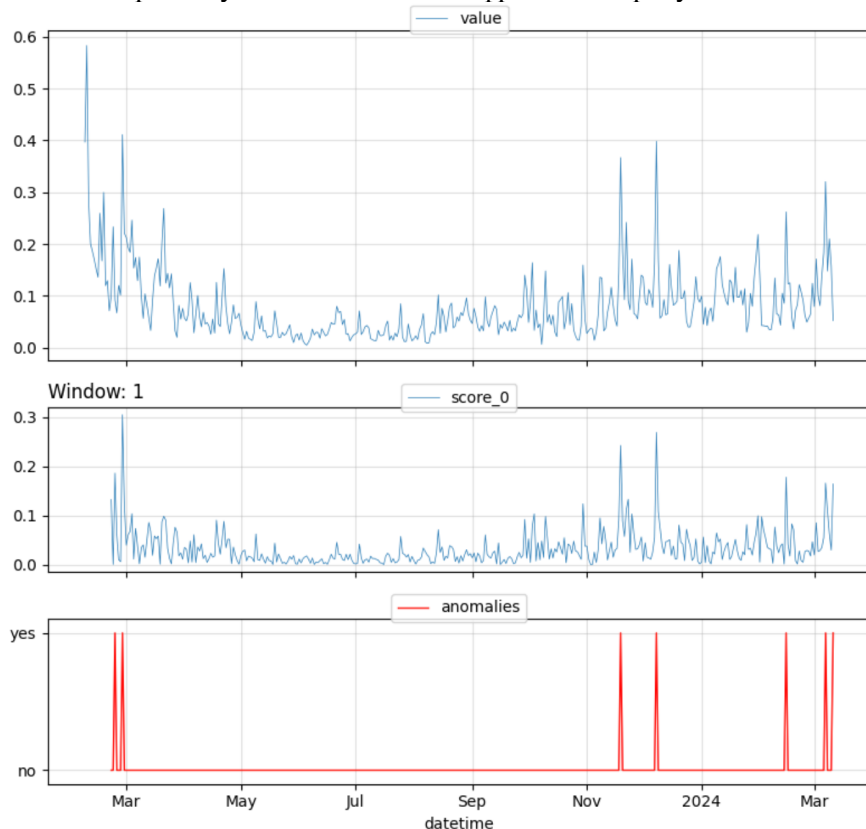


Рис. 1. Приклад використання KMeansScorer для значень показника якості повітря PM2.5

2. Detector – це компонент бібліотеки, який перетворює числові оцінки аномалій у бінарні мітки. Він визначає, які точки ряду слід вважати аномальними на основі порогів або різних статистичних критеріїв. Прикладом є метод QuantileDetector, який позначає точки як аномальні, якщо їх оцінка перевищує певний квантиль (рис. 2).

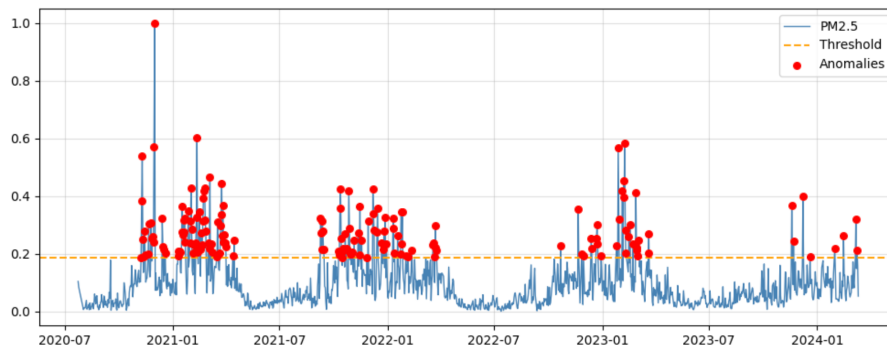


Рис. 2. Приклад використання QuantileDetector для виявлення аномальних значень показника якості повітря PM2.5

3. Anomaly Models – це об’єднані моделі, які поєднують прогнозу модель часових рядів та оцінки для автоматичного виявлення аномалій. Такі моделі дозволяють оцінювати відхилення фактичних значень від очікуваних, враховуючи тренди, сезонність та локальні коливання даних. Інтерфейс бібліотеки дозволяє інтегрувати відомі моделі, серед яких, ARIMA, BlockRNN, ExponentialSmoothing та інші.

4. Aggregators – це компоненти бібліотеки, які дозволяють об’єднувати результати використання інших компонентів в єдину послідовність оцінок або бінарних міток аномалій. Це забезпечує більш надійне та точне виявлення аномалій, особливо коли ряд складний або має неоднорідну структуру.

Висновки

В роботі було проаналізовано використання бібліотеки Darts для роботи з часовими рядами та виявлення аномалій у цих рядах. Було розглянуто основні можливості бібліотеки: використання компонентів для оцінки відхилень точок ряду; застосування компонентів для позначення аномалій; а також комбінування результатів роботи цих компонентів.

Було продемонстровано, що Darts дозволяє виділяти аномальні значення показника PM2.5 на даних моніторингу якості атмосферного повітря від мережі громадського моніторингу EcoCity. Отримані результати підтверджують, що бібліотека Darts є потужним інструментом для аналізу часових рядів та автоматизованого пошуку аномалій, і може застосовуватися в наукових і практичних задачах.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Julien Herzen, Francesco Lässig, Samuele Giuliano Piazzetta, Thomas Neuer, Léo Tafti, Guillaume Raille, Tomas Van Pottelbergh, Marek Pasieka, Andrzej Skrodzki, Nicolas Huguenin, Maxime Dumonal, Jan Kościsz, Dennis Bader, Frédérick Gusset, Mounir Benheddi, Camila Williamson, Michal Kosinski, Matej Petrik, and Gaël Grosch. 2022. Darts: user-friendly modern machine learning for time series. J. Mach. Learn. Res. 23, 1, Article 124 (January 2022). – [Електронний ресурс]. Режим доступу: <https://jmlr.org/papers/v23/21-1177.html>
2. Unit8 SA. Darts Anomaly Detection API Reference. – [Електронний ресурс]. Режим доступу: https://unit8co.github.io/darts/generated_api/darts.ad.html
3. Air Quality Monitoring from EcoCity (Vinnytsia). Kaggle. – [Електронний ресурс]. Режим доступу: <https://www.kaggle.com/datasets/vbmokin/air-quality-monitoring-from-ecocity>

Шмундяк Дмитро Олександрович – PhD, асистент кафедри системного аналізу та інформаційних технологій, e-mail: dmytro.shmundiak@vntu.edu.ua

Shmundiak Dmytro O. – PhD, teaching assistant of the Chair of System Analysis and Information Technologies, e-mail: dmytro.shmundiak@vntu.edu.ua