УДК 004.9+574

Д.О. Шмундяк[1]
Н.С. Іжаковська[1]
Д.О. Литвиненко[1]
А.О. Судець[1]

АНАЛІЗ МОЖЛИВОСТЕЙ PYTHON-БІБЛІОТЕК ЩОДО ВИЯВЛЕННЯ АНОМАЛЬНИХ ДАНИХ У ЗАДАЧІ ПРОГНОЗУВАННЯ СТАНУ АТМОСФЕРНОГО ПОВІТРЯ

[1]Вінницький національний технічний університет, Україна

Анотація

Проведено порівняльний аналіз деяких бібліотек для мови програмування Python щодо їх можливостей виявлення аномальних даних. Описано основні принципи роботи кожної з бібліотек, вказано їх переваги та недоліки. Проведено практичні випробування та порівняння ефективності цих бібліотек для виявлення аномалій у даних громадського моніторингу якості атмосферного повітря мережі EcoCity.

Ключові слова: *якість атмосферного повітря, аномалії часових рядів, Python, машинне навчання, EcoCity.*

Abstract

The purpose of the paper is to conduct a comparative overview of some of the Python programming language libraries to understand their anomaly detection capabilities. The primary principles of every approach and algorithm were described, and their advantages and disadvantages were listed. The efficiency of these approaches was compared by applying algorithms to detect anomalies inside the dataset of the air quality monitoring system EcoCity.

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

Вступ

Прогнозування стану атмосферного повітря є важливою задачею, необхідною для прийняття рішень. Аналіз, проведений за даними громадського моніторингу цього стану, зібраними станціями мережі EcoCity [1], показав, що на точність такого прогнозування суттєво впливають аномальні значення. Отже, важливо їх чітко виявляти та мінімізувати такий вплив. Для цього існують ряд Python-бібліотек.

Метою даного дослідження є проведення порівняльного аналізу та систематизація можливостей Python-бібліотек scikit-learn, statsmodels та sesd щодо виявлення та мінімізації впливу аномальних даних на прогнозування стану атмосферного повітря.

Розв’язання задачі

Проаналізуємо можливості Python-бібліотеки sesd. Дана бібліотека реалізує Метод Seasonal Hybrid ESD. Це є одним із підходів для виявлення аномалій в часових рядах. Він комбінує сезонну адаптацію з експоненційно зваженою середньою (Exponential Smoothing) і використовує Extreme Studentized Deviate (ESD) для виявлення відхилень від очікуваного розподілу даних [3, 4]. Seasonal Hybrid ESD (SH-ESD), ґрунтується на тесті Generalized ESD - двоетапний процес дозволяє моделі виявляти як глобальні аномалії, що виходять за межі очікуваних сезонних мінімумів і максимумів, так і локальні аномалії, які інакше були б замасковані сезонністю. Це досягається за рахунок використання декомпозиції часових рядів та використання надійних статистичних показників, а саме медіани разом із ESD. Крім того, для довгих часових рядів алгоритм використовує кускову апроксимацію. Перевагами моделі є:

- врахування сезонності: враховує сезонні зміни в часових рядів, що дозволяє ефективно виявляти аномалії, які повторюються у певних періодах часу;
- гнучкість та налаштовування: метод може бути налаштований залежно від потреб користувача (період сезонності та порогові значення);

– підтримка широкого спектра даних: може бути застосований до різних типів даних, включаючи числові часові ряди, а також категоріальні та багатовимірні дані.

Серед недоліків можна виділити:

– обмежена робота з незвичайними аномаліями: метод краще працює з аномаліями, які досить схожі на звичайні сезонні зміни;

– вимоги до часу обчислень: алгоритм має не найбільшу швидкість, особливо для великих обсягів даних або довгих часових рядів.

Наступною розглянемо бібліотеку scikit-learn, а саме її модуль Isolation Forest. Isolation Forest - це алгоритм виявлення аномалій, який працює шляхом ізоляції аномалій у даних [5]. Алгоритм полягає у побудові випадкового бінарного дерева. Коренем дерева є весь простір ознак; у черговому вузлі вибирається випадкова ознака і випадковий поріг розбиття. Критерієм зупинки є тотожний збіг всіх об'єктів у вузлі, тобто вирішальне дерево будується повністю. Значення anomaly_score для алгоритма є глибиною листка в побудованому дереві. Переваги моделі:

– простота використання: Scikit-learn дозволяє легко використовувати алгоритм без складних конфігурацій;

– ефективність: є швидким алгоритмом. Він використовує випадкові дерева для розбиття даних, що дозволяє ефективно працювати з великими наборами даних;

– стійкість: має добру стійкість до викидів та шуму в даних;

Недоліки моделі:

– вразливість до перекошених даних: як і більшість алгоритмів машинного навчання, Isolation Forest може бути вразливим до перекошених даних. Якщо кількість аномалій значно перевищує кількість нормальних зразків або навпаки, алгоритм може видавати неточні результати;

– неефективність для високо-вимірних даних: Isolation Forest може стикатися з проблемою, коли кількість ознак у даних є дуже великою. У таких випадках алгоритм може втратити ефективність і потребувати більше обчислювальних ресурсів;

– відсутність різноманітних додаткових функцій: Бібліотека Scikit-learn, включаючи реалізацію Isolation Forest, може бути обмеженою у функціональності порівняно з іншими спеціалізованими бібліотеками.

Наостанок розглянемо бібліотеку statsmodels. У statsmodels немає вбудованих функцій для прямого виявлення та фільтрації аномалій у часових рядах. Однак, можна використовувати інші функції та методи з бібліотеки для реалізації такого аналізу [6]. Один з підходів - це аналіз залишків. Можна використовувати статистичні моделі, такі як ARIMA або ETS, для прогнозування часового ряду. Порівнюючи фактичні значення з прогнозованими, можна виявити аномалії, коли значення значно відрізняються від прогнозу. Інший підхід - це використання статистичних методів. Це включає встановлення порогових значень, використання стандартного відхилення, Z-перетворення та інших методів для виявлення значень, які перевищують певний поріг.

Переваги використання statsmodels включають наявність статистичних моделей та гнучкість у виборі підходу для виявлення аномалій. Однак, для ефективного використання бібліотеки потрібні знання статистики та часових рядів.

Прогнозування даних стану атмосферного повітря за даними мережі EcoCity

Здійснено прогнозування даних стану атмосферного повітря за даними мережі EcoCity з використанням описаних раніше бібліотек. Дані для дослідження отримані за допомогою сервісу «Кабінет дослідника» [2], до якого автори мають доступ, завдяки угоді між EcoCity і ВНТУ. Це - веб-система, яка дозволяє отримати доступ та використовувати у своїх дослідженнях інформацію, отриману від станцій моніторингу атмосферного повітря. Для прогнозування було обрано показник «PM10» (пил, розміром 10 мкм і менше), отриманих з однієї зі станцій у Вінницькій області. На рисунку 1 зображено графік часового ряду зазначеного показника.

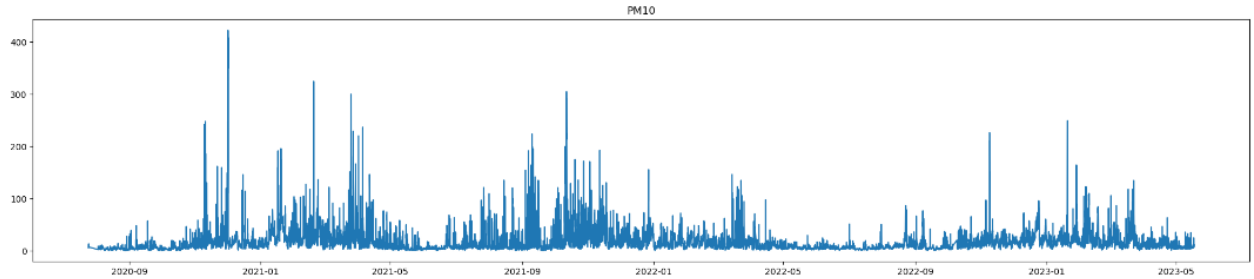

Рис. 1. Вхідні дані для передбачення аномалій

Приклад коду для виявлення аномалій, використовуючи бібліотеку sesd (Seasonal Hybrid ESD) наведено на рисунку 2, а результат його роботи – на рисунку 3.

```
outliers_indices = sesd.seasonal_esd(anomaly_df,periodicity = 10, hybrid=True, max_anomalies=1000, alpha = 3)
anomalies = df.loc[outliers_indices]
```

Рис. 2. Приклад використання бібліотеки sesd

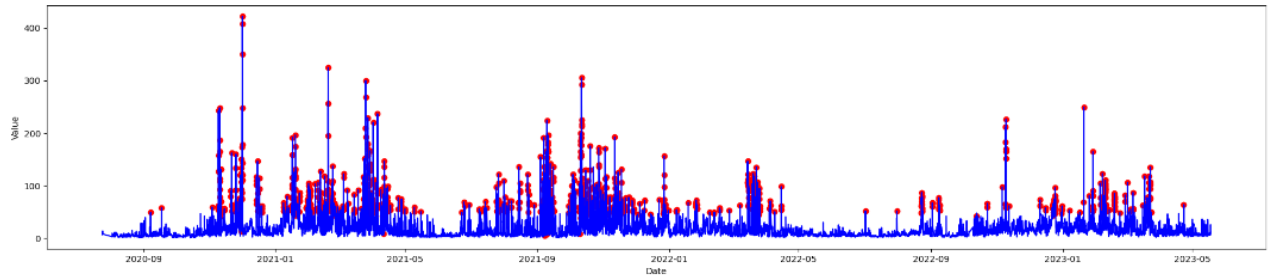

Рис. 3. Результат роботи бібліотеки sesd

Тепер спробуємо застосувати метод Isolation Forest бібліотеки Scikit-learn. Приклад коду наведено на рисунку 4, а результат роботи – на рисунку 5.

```
model = IsolationForest(contamination=0.05)
model.fit(df['y'].values.reshape(-1, 1))
y_pred = model.predict(df['y'].values.reshape(-1, 1))
```

Рис. 4. Приклад використання бібліотеки Isolation Forest

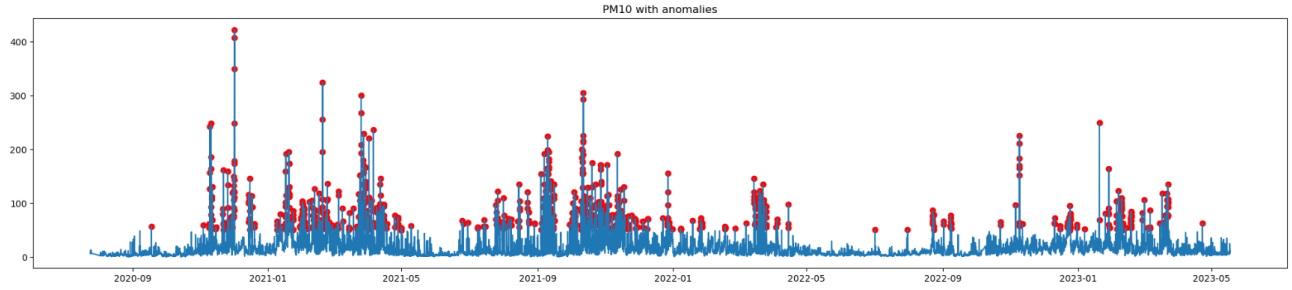

Рис. 5. Результат роботи Isolation Forest

Також перевіримо роботу з пошуком аномалій, використовуючи бібліотеку statsmodel. Використовуючи розклад часових рядів, як показано на рисунку 6, отримуємо результат, зображений на рисунку 7.

```
period = 24
result = seasonal_decompose(df_anomaly['y'], model='additive', period=period)
residual = result.resid
threshold = 200.0
anomalies = df_anomaly[residual.abs() > threshold]
df_filtered = df_anomaly.drop(anomalies.index)
```

Рис. 6. Приклад використання бібліотеки statsmodels

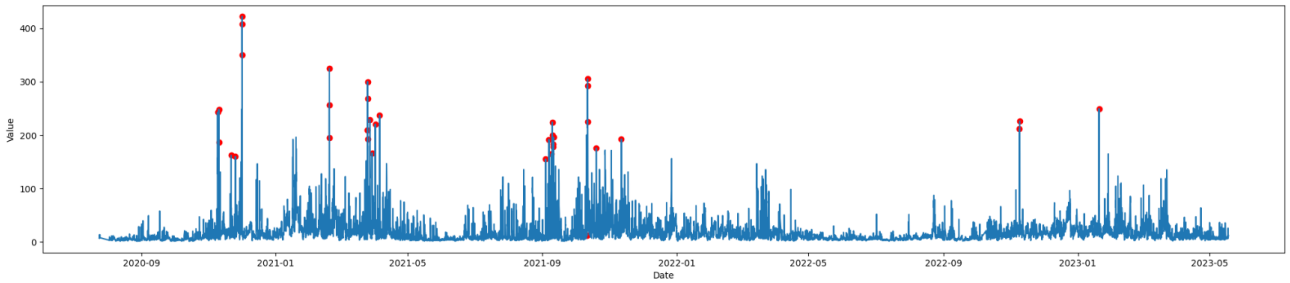

Рис. 7. Результати, отримані за допомогою бібліотеки statsmodels

Висновки

Розглянуто деякі популярні бібліотеки та їх модулі для пошуку аномалій в наборах даних. Кожна з бібліотек має свої певні переваги та недоліки, проте кожна надає зручний та потужний інструментарій, який як окремо, так і в комбінації, може успішно застосовуватися у дослідженнях, пов'язаних з пошуком аномалій. Кожну з описаних бібліотек було практично застосовано до набору даних моніторингу якості атмосферного повітря «EcoCity» та знайдено аномалії для показника «PM10» (пил, розміром 10 мкм і менше). Аналізуючи результати, можна помітити що sesd та scikit-learn показали дещо схожі результати, в той час як підхід з statsmodels знайшов набагато менше аномалій. З одного боку, знайдені ним аномалії це дійсно – явні аномалії. З іншого боку, можна сказати що безпосередньо застосований підхід не є досить ефективним.

Отримані результати є корисними та будуть використовуватися у подальших дослідженнях авторів.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Eco-City Громадський моніторинг стану якості повітря [Електронний ресурс] – Режим доступу до ресурсу: https://eco-city.org.ua/.
2. Eco-City Кабінет дослідника [Електронний ресурс] – Режим доступу до ресурсу: https://archive.eco-city.org.ua/.
3. AnomalyDetection: Anomaly Detection Using Seasonal Hybrid Extreme Studentized Deviate Test [Електронний ресурс] – Режим доступу до ресурсу: https://rdrr.io/github/twitter/AnomalyDetection/f/README.md.
4. Anomaly Detection: Seasonal ESD [Електронний ресурс] – Режим доступу до ресурсу: https://github.com/nachonavarro/seasonal-esd-anomaly-detection/blob/master/README.md.
5. Isolation Forest [Електронний ресурс] – Режим доступу до ресурсу: https://scikit-learn.org/stable/modules/outlier_detection.html#isolation-forest.
6. Statsmodels [Електронний ресурс] – Режим доступу до ресурсу: https://www.statsmodels.org/stable/index.html.

Шмундяк Дмитро Олександрович — аспірант кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, dimashmund@gmail.com.

Іжаковська Наталя Сергіївна — студентка кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, nataffkaizh@gmail.com.

Литвиненко Данило Олександрович — студент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, lytdanya@gmail.com.

Судець Анна Олександрівна — студентка кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, Вінниця, annasudets2.0@gmail.com.

Shmundiak Dmytro O. – Postgraduate student of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, dimashmund@gmail.com.

Izhakovska Natalia S. –student of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, nataffkaizh@gmail.com

Lytvynenko Danylo O. –student of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, lytdanya@gmail.com

Sudets Anna O. – student of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, annasudets2.0@gmail.com.