УДК 004.6

АЛГОРИТМЫ ОБНАРУЖЕНИЯ АНОМАЛИЙ В РАЗНОРОДНЫХ НАБОРАХ ДАННЫХ

Шукуров Н.А.
магистр, Таджикский технический университет имени М.С. Осими (Душанбе, Таджикистан)

ANOMALY DETECTION ALGORITHMS IN HETEROGENEOUS DATASETS

Shukurov N.
master's degree, Tajik Technical University named after M.S. Osimi (Dushanbe, Tajikistan)

Аннотация

Статья посвящена исследованию алгоритмов обнаружения аномалий в разнородных наборах данных. Описаны основные методы, включая статистические подходы, алгоритмы машинного обучения и глубокого обучения, такие как изоляционный лес, автокодировщики и рекуррентные нейронные сети (РНС), а также методы временных рядов. Приведены примеры применения изоляционного леса для выявления аномальных точек данных. Статья подчеркивает актуальность использования этих алгоритмов в задачах предсказания сбоев и повышения безопасности. Заключение акцентирует внимание на важности комбинации различных подходов для достижения высокой точности и эффективности.

Ключевые слова: обнаружение аномалий, разнородные данные, машинное обучение, изоляционный лес, глубокое обучение.

Abstract

The article explores anomaly detection algorithms for heterogeneous data sets. It describes key methods, including statistical approaches, machine learning algorithms, and deep learning techniques such as Isolation Forest, autoencoders, and recurrent neural networks (RNNs), as well as time series methods. An example using Isolation Forest to detect anomalous data points is provided. The article emphasizes the relevance of these algorithms in failure prediction and security enhancement tasks. The conclusion highlights the importance of combining different approaches to achieve high accuracy and efficiency.

Keywords: anomaly detection, heterogeneous data, machine learning, Isolation Forest, deep learning.

Введение

Современные наборы данных, используемые в различных отраслях, часто обладают высокой степенью разнородности, что затрудняет их анализ и обработку стандартными методами. Разнородные данные включают как структурированные, так и неструктурированные данные, которые могут поступать из множества источников, таких как социальные сети, датчики, финансовые системы и веб-ресурсы. Из-за сложности этих данных и их различной природы возникает потребность в эффективных алгоритмах, способных обнаруживать аномалии и отклонения от нормы, которые могут сигнализировать о проблемах в работе системы, безопасности или данных.

Обнаружение аномалий в разнородных данных важно для различных применений, таких как предсказание неисправностей в производственных системах, обнаружение мошенничества

в банковской сфере, анализ поведения пользователей и мониторинг состояния инфраструктуры. Традиционные методы анализа данных, ориентированные на однородные наборы, часто не справляются с задачей анализа разнородных источников, что повышает актуальность разработки специализированных алгоритмов. Эти алгоритмы способны учитывать сложные взаимосвязи между параметрами, которые отличаются по структуре и временным характеристикам.

Цель данной статьи заключается в рассмотрении существующих алгоритмов обнаружения аномалий, применяемых к разнородным наборам данных, а также в анализе их эффективности и возможностей. В статье будут описаны различные подходы, включая методы машинного обучения и глубокого обучения, а также подходы, основанные на анализе временных рядов и статистическом моделировании.

Основная часть. Основные алгоритмы обнаружения аномалий

Обнаружение аномалий в разнородных наборах данных требует применения различных подходов, которые могут учитывать как структурированные, так и неструктурированные данные. Наиболее эффективные методы включают статистические подходы, алгоритмы машинного обучения (МО) и глубокого обучения, а также методы, основанные на временных рядах [1].

Статистические методы

Статистические методы являются одним из классических подходов к обнаружению аномалий и основываются на анализе распределений данных. Например, метод межквартильного размаха (IQR) позволяет выявлять аномалии в числовых данных, сравнивая значения с межквартильным диапазоном. Другим методом является использование гипотезы нормального распределения, где аномалии определяются как значения, находящиеся за пределами нескольких стандартных отклонений от среднего. Статистические методы эффективны для анализа небольших и однородных данных, но часто имеют ограничения при работе с разнородными наборами, где распределение данных может быть неоднородным и непредсказуемым.

Методы МО

МО-алгоритмы, такие как метод опорных векторов (SVM) и случайный лес (Random Forest), применяются для обнаружения аномалий путем обучения модели на исторических данных. SVM, например, выделяет гиперплоскость, которая разделяет нормальные и аномальные точки в пространстве признаков. Случайный лес строит несколько деревьев решений для анализа данных, и аномальные точки определяются на основе частоты их классификации в качестве «выбросов». МО позволяет использовать как супервизорные, так и несупервизорные алгоритмы, что делает их гибкими в использовании для различных наборов данных. Эти методы хорошо работают с разнородными наборами данных, но для достижения высокой точности требуют достаточно больших объемов данных для обучения [2].

Изоляционные методы

Изоляционный лес (Isolation Forest) — метод, специально разработанный для обнаружения аномалий. Он строит несколько деревьев, где данные, которые легче изолируются, считаются аномалиями. Этот метод эффективен для анализа высокоразмерных и разнородных данных, поскольку он не требует гипотезы о распределении данных и работает быстро на больших объемах. Изоляционный лес особенно подходит для применения в наборах данных, где аномалии представляют собой редкие и значительно отличающиеся от нормы объекты [3].

Глубокое обучение

Методы глубокого обучения, такие как автокодировщики и генеративно-состязательные сети (GAN), обеспечивают высокую гибкость в анализе сложных и разнородных данных. Автокодировщики представляют собой нейронные сети, обучающиеся воспроизводить входные данные на выходе. При этом они эффективно кодируют нормальные данные, но «проваливаются» на аномальных, что делает их обнаружение возможным на основе высокой ошибки восстановления. GAN, с другой стороны, создают синтетические данные и

сравнивают их с реальными, что позволяет выявлять аномалии как отклонения от синтетической нормы. Глубокое обучение эффективно для работы с разнородными данными, особенно в случаях, когда набор данных обладает сложной внутренней структурой и большим объемом [4, 5].

Методы анализа временных рядов

Для данных, которые включают временные зависимости, такие как сенсорные данные или финансовые транзакции, используются методы анализа временных рядов. Рекуррентные нейронные сети (РНС), такие как LSTM, позволяют моделировать временные зависимости и выявлять аномалии на основе отклонений от предсказанных значений. Также используются методы, такие как ARIMA и STL, которые выявляют аномалии путем анализа изменений и выбросов в данных на временной шкале. Методы временных рядов особенно полезны для прогнозирования изменений во времени и обнаружения аномалий, связанных с сезонностью и трендами [6].

Пример использования изоляционного леса для обнаружения аномалий

Изоляционный лес (Isolation Forest) – один из популярных методов для обнаружения аномалий, особенно в разнородных и высокоразмерных наборах данных. Этот алгоритм работает на основе принципа изоляции: данные, которые легче изолировать, считаются аномальными [7]. Изоляционный лес строит несколько деревьев, и если точка данных изолируется на ранних уровнях дерева, то она считается выбросом. Ниже приведен пример кода на Python, демонстрирующий использование изоляционного леса для обнаружения аномалий. Библиотека sklearn предоставляет готовую реализацию этого алгоритма.

```
from sklearn.ensemble import IsolationForest
import numpy as np

# Пример данных, содержащих нормальные точки и выбросы
data = np.array([[10, 5], [12, 6], [11, 5], [9, 5], [15, 10], [100, 2], [20, 1], [10, 4], [11, 6], [110, 2]])

# Инициализация модели изоляционного леса с указанием доли выбросов
iso_forest = IsolationForest(contamination=0.2, random_state=42)
iso_forest.fit(data)

# Предсказание аномалий (-1 - аномалия, 1 - нормальная точка)
predictions = iso_forest.predict(data)
print("Результаты предсказания:", predictions)

# Вывод результатов
for i, prediction in enumerate(predictions):
    if prediction == -1:
        print(f"Точка {data[i]} считается аномалией")
    else:
        print(f"Точка {data[i]} считается нормальной")
```

Описание работы примера

Подготовка данных: Массив data содержит нормальные точки и несколько выбросов, которые мы хотим обнаружить.

Создание модели: Создается модель IsolationForest, где параметр contamination=0.2 указывает на ожидаемую долю выбросов в данных (в данном случае 20%).

Предсказание аномалий: Метод predict() возвращает массив, где -1 указывает на аномалию, а 1 — на нормальную точку. Таким образом, алгоритм определяет выбросы, изолируя их на ранних уровнях дерева.

Вывод результата: Каждая точка данных проверяется на принадлежность к норме или аномалии, и результаты выводятся на экран.

Интерпретация результатов

В данном примере изоляционный лес обнаруживает точки, которые отличаются от нормальных значений, таких как (100, 2) и (110, 2). Эти точки имеют значения, которые значительно отклоняются от остальных точек, что и делает их аномалиями. Использование изоляционного леса позволяет эффективно выделять выбросы даже в разнородных наборах данных, делая его подходящим инструментом для различных прикладных задач анализа данных [8].

Заключение

Внедрение алгоритмов обнаружения аномалий в разнородные наборы данных предоставляет организациям возможность автоматизировать процессы мониторинга, повышая надежность работы систем и предотвращая потенциальные сбои и нарушения безопасности. Обзор основных алгоритмов показал, что статистические методы и методы МО, такие как изоляционный лес и методы глубокого обучения, могут быть адаптированы для анализа данных различной структуры и временных характеристик. Однако выбор оптимального метода зависит от конкретного типа данных и задачи. Методы, такие как изоляционный лес и автокодировщики, доказали свою эффективность для работы с высокоразмерными и разнородными данными. Эти алгоритмы обеспечивают высокую гибкость и масштабируемость, что позволяет использовать их в реальных условиях, где часто требуется оперативное обнаружение аномалий. Кроме того, применение временных моделей, таких как рекуррентные нейронные сети, позволяет учитывать временные зависимости, которые особенно важны для мониторинга динамических систем. Таким образом, комбинирование различных алгоритмов, таких как статистические подходы, МО и методы временных рядов, дает наиболее точные результаты и позволяет создавать устойчивые решения для обнаружения аномалий. Эти технологии обеспечивают новый уровень безопасности и контроля, позволяя компаниям своевременно реагировать на потенциальные угрозы и оптимизировать работу своих систем.

Список литературы

1. Багутдинов Р.А. Подход к обработке, классификации и обнаружению новых классов и аномалий в разнородных и разномасштабных потоках данных // Вестник Дагестанского государственного технического университета. Технические науки. 2018. Т. 45. №3. С. 85-93.
2. Чесноков М.Ю. Поиск аномалий во временных рядах на основе ансамблей алгоритмов DBSCAN // Искусственный интеллект и принятие решений. 2018. №1. С. 99-107.
3. Багутдинов Р.А., Саргсян Н.А., Красноплахтыч М.А. Аналитика, инструменты и интеллектуальный анализ больших разнородных и разномасштабных данных // Экономика. Информатика. 2020. Т. 47. №4. С. 792-802.
4. Вишератин А.А., Мухина К.Д., Струков А.М. Разработка интеллектуальной платформы мониторинга состояния городской среды на основе разнородных данных // Альманах научных работ молодых ученых Университета ИТМО. 2019. С. 146-150.
5. Щеглеватых Р.В., Сысоев А.С. Математическая модель обнаружения аномальных наблюдений с использованием анализа чувствительности нейронной сети // Моделирование, оптимизация и информационные технологии. 2020. Т. 8. №1. С. 14-15.
6. Симанков В.С., Колодий А.С. Подход к построению архитектуры интеллектуальной системы обнаружения и устранения сетевых аномалий // Вестник Адыгейского государственного университета. Серия 4: Естественно-математические и технические науки. 2012. №4 (110). С. 191-196.
7. Dang Y., Wang B., Brant R., Zhang Z., Alqallaf M., Wu Z. Anomaly detection for data streams in large-scale distributed heterogeneous computing environments // ICMLG2017 5th International Conference on Management Leadership and Governance. 2017. P. 121.
8. Zhu J., Ding C., Tian Y., Pang G. Anomaly heterogeneity learning for open-set supervised anomaly detection // Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2024. P. 17616-17626.