

The Elephant in the Room: Towards A Reliable Time-Series Anomaly Detection Benchmark

Zusammenfassung

January 18, 2025

1 Dataset

Wir haben 1070 hochwertige Zeitreihen, die aus 40 verschiedenen Datensätzen bestehen.

2 Evaluierungsmetrik

VUS-PR (Volume Under Surface of Precision-Recall) ist die zuverlässigste und genaueste Metrik für die Anomalieerkennung in Zeitreihen.

Es wurden 40 verschiedene Anomalieerkennungsalgorithmen evaluiert.

3 Ergebnisse der Studie

Die Autoren stellen fest, dass einfachere statistische Modelle und Architekturen oft bessere Ergebnisse liefern als komplexe neuronale Netzwerke. Allerdings könnten neuronale Netze gute Ergebnisse bei multivariaten Zeitreihen liefern.

4 Einfluss von IoT-Technologien

Die Einführung und Verbreitung von IoT-Technologien hat die Menge an sequentiellen Messungen drastisch erhöht. Daher haben wir auch viele Zeitreihen (also einen großen Datensatz). Sie spielen eine wesentliche Grundlage für viele Aufgaben in der Analyse.

Zentrale Aufgaben der Zeitreihenanalyse:

- **Abfragen (Querying):** Informationen aus Daten extrahieren.
- **Vorhersagen (Forecasting):** Prognosen über zukünftige Werte erstellen.
- **Klassifizierung:** Daten in Kategorien einteilen.
- **Clustering:** Ähnliche Muster gruppieren.

5 Definition von Zeitreihen

Eine Zeitreihe ist eine geordnete Sequenz von reellen Beobachtungswerten. Formal wird eine Zeitreihe durch

$$X = \{x_1, \dots, x_T\}$$

dargestellt, wobei $x_t \in \mathbb{R}^N$ die Werte eines Signals von N -Sensoren über T Zeitschritte repräsentiert.

5.1 Univariate Zeitreihen

Für univariate Zeitreihen gilt: $N = 1$ (eine Messung pro Zeitschritt).

5.2 Multivariate Zeitreihen

Für multivariate Zeitreihen gilt: $N > 1$ (mehrere Messungen pro Zeitschritt, z. B. mehrere Sensoren).

6 Arten von Anomalien

Anomalien können als einzelne Werte oder in Form von Sequenzen auftreten:

6.1 Punktbasierte Anomalien

- **Point Anomalies:** Einzelne Datenpunkte, die signifikant von der Mehrheit abweichen.
- **Contextual Anomalies:** Einzelne Datenpunkte, die innerhalb eines spezifischen Kontexts (z. B. Zeit oder Zustand) nicht dem erwarteten Muster entsprechen.

6.2 Sequenzbasierte Anomalien

- **Collective Anomalies:** Gruppen von Datenpunkten (Subsequenzen), die sich von typischen, zuvor beobachteten Mustern unterscheiden.

7 Algorithmen:

Die Algorithmen werden nach dem Grad des erforderlichen Vorwissens und der Art der Verarbeitung kategorisiert. Also wenn wir kein Vorwissen haben, also keine gelabelte Daten haben, dann werden unüberwachte Modelle genutzt. Wenn wir gelabelte Daten haben nutzen wir überwachtes Lernen (Dies wird aber nur selten verwendet, da Anomalien oft schwer verfügbar sind).

8 Art der Verarbeitung

Die Verarbeitung von Anomalien kann auf verschiedene Weisen erfolgen:

8.1 Distance-based

Misst die Ähnlichkeit oder Distanz zwischen Datenpunkten oder -sequenzen.

8.2 Density-based

Sucht nach Regionen mit ungewöhnlich niedriger Dichte im Merkmalsraum.

8.3 Prediction-based

Verwendet Modelle, um den nächsten Wert oder Zustand vorherzusagen, und erkennt Abweichungen zwischen den vorhergesagten und den tatsächlichen Werten.

9 TSB-AD Benchmark im Vergleich zu allen ältern Benchmarks

9.1 Datensätze

- TSB-AD umfasst die bisher größte Sammlung von Datensätzen zur Anomalieerkennung in Zeitreihen. Die Sammlung umfasst fast doppelt so viele Datensätze wie die bisher größte Sammlung.
- SB-AD kombiniert menschliche Wahrnehmung und algorithmische Unterstützung, um Datensätze zu kuratieren.
- Während viele frühere Benchmarks sich auf univariate Zeitreihen konzentrieren, integriert TSB-AD multivariate Zeitreihen, um realistischere Szenarien abzubilden.

9.2 Algorithmen

- TSB-AD deckt eine Vielzahl von Algorithmen ab, darunter statistische Methoden, neuronale Netzwerke und Foundation Models.

9.3 Evaluationsmethoden

- Ein zuverlässiges und regelmäßig aktualisiertes Testbed für den Vergleich der Modellleistung sowie Optimierung der Hyperparameter verschiedener Algorithmen.

10 Probleme durch Fehlkennzeichnungen (Mislabeling Issues)

10.1 Inkonsistente Labeling-Standards

- Ähnliche Muster in Zeitreihen werden inkonsistent gekennzeichnet, z. B. als Anomalie oder als normal. z.B. Ein zweiter Spike in einer Zeitreihe, der einem zuvor markierten Anomaliefall ähnelt, wird nicht als Anomalie markiert (falsches Negativ) und umgekehrt kann eine markierte Anomalie ohne signifikante Merkmale auftreten (falsches Positiv).
- Nicht offengelegte Zusatzdaten: Datensatz-Ersteller hatten möglicherweise Zugang zu zusätzlichen Informationen, die nicht mit den Daten geteilt wurden, was zu Fehlkennzeichnungen führen kann.

10.2 Verzerrungen in den Datensätzen (Bias in Datasets)

- Anomalien treten überwiegend gegen Ende einer Zeitreihe auf, z. B. im Yahoo-Datensatz. Solche Verzerrungen können Algorithmen bevorzugen, die die letzten Datenpunkte einer Serie als anomal vorhersagen.
- Eingeschränkte Anzahl von Anomalien: Datensätze wie der UCR-Datensatz ([106]) enthalten oft nur eine Anomalie pro Zeitreihe, basierend auf der Annahme, dass dies ideal ist. Diese Annahme ignoriert reale Szenarien, in denen mehrere oder weniger offensichtliche Anomalien auftreten können.

Dies hat die Folge: Algorithmen könnten aufgrund der Verzerrungen und Fehlkennzeichnungen in den Datensätzen fälschlicherweise als effektiver bewertet werden.

11 Flaws in Evaluation Measures

Schwächen traditioneller Bewertungsmethoden bei der Anomalieerkennung bei Zeitreihen

- Ungleichgewicht in den Datensätzen: Anomalien sind oft selten, was die Zuverlässigkeit bestimmter Metriken (z. B. F1-Score) beeinträchtigen kann.
- zeitlicher Abhängigkeit: Metriken wie der F1-Score behandeln jeden Zeitschritt unabhängig und ignorieren die sequenzielle Natur von Zeitreihen.

- AUC-ROC: Dieser misst die Fläche unter der Kurve, die die TPR true positive rate gegen die FPR darstellt. Bei Anomalieerkennung sind die FPR-Werte oft sehr niedrig, da es viele True negative gibt. Dies macht große Teile der ROC-Kurve irrelevant
- AUC-PR ist eine gute Alternative, da sie das Verhältnis von Precision und Recall betont. Außerdem wird empfohlen noch kombinierte Bewertungen anzuwenden

11.1 Probleme der Punktbasierten Metriken

- Punktbasierte Metriken wie AUC-ROC und AUC-PR betrachten jede Zeitreihe-Datenpunkt unabhängig voneinander. Jeder Beitrag wird gleich stark gewertet, unabhängig von Kontext oder zeitlicher Nähe.
- sie sind anfällig für leichte Verzögerungen in den Scores

12 Dataset Overview

Der Bau des TSB-AD-Datensatzes erfolgt in drei Schritten. Dies stellt sicher, dass der Datensatz groß ist als auch eine hohe Qualität hat.

12.1 Schritt 1: Dataset-Sammlung

- Es werden 13 univariate und 20 multivariate Datensätze zur Anomalie-Erkennung verwendet
- Um die Vielfalt und Größe des Datensatzes zu erhöhen, werden multivariate Zeitreihen in univariate Formate umgewandelt, indem jede Kanal als eigenständige Zeitreihe betrachtet wird.
- Diese Transformation wurde gemacht, da in einigen multivariaten Datensätzen nur bestimmte Kanäle (meistens nur einer) wertvolle Informationen für die Anomalie-Erkennung liefern, während andere Kanäle unbrauchbare Infos haben.
- Eine Korrelationsanalyse zeigt, dass bestimmte Kanäle eine stärkere Korrelation mit den Anomalien als andere haben. Diese Kanäle werden beibehalten und in univariate Zeitreihen umgewandelt, um die Anomalie-Erkennung zu verbessern.
- Für jede Zeitreihe wird der höchste Wert der Bewertung über alle Detektoren hinweg aufgezeichnet. Die besten 40% der Zeitreihen nach diesen Bewertungsergebnissen werden ausgewählt.

Nach dieser Auswahl wurden insgesamt 46 Zeitreihe Datensätze (13 univariate und 33 multivariate) für die TSB-AD-Datenbank zusammengetragen.

12.2 Schritt 2: Überprüfung der Labels

Da es keine klare Definition für Anomalien gibt und rein algorithmische Beurteilungen der Label-Qualität oft unzureichend sind, kombiniert dieser Schritt automatisierte Bewertungen mit manuellen Inspektionen. Dies stellt sicher, dass nur qualitativ hochwertige Zeitreihen mit geeigneten Labels im TSB-AD-Datensatz verbleiben.

12.3 Schritt 3: Algorithmengestützte Verifizierung der Label-Qualität

Das Beurteilen der Eignung eines Datensatzes für die Anomaliedetektion und die Validierung der Anomalie-Labels ist schwer, wenn man das manuell machen möchte, daher werden Algorithmen verwendet, um die Qualität der Labels zu überprüfen.

Der resultierende Datensatz umfasst hochwertige, mit Anomalien gekennzeichnete Zeitreihen:

- **TSB-AD-U:** Für univariate Zeitreihen.
- **TSB-AD-M:** Für multivariate Zeitreihen.

- **Eval-Set:** Für die Bewertung der Modelle.
- **Tuning-Set:** Zur Optimierung von Hyperparametern.

Univariate Datensätze wie UCR und YAHOO enthalten deutlich mehr Zeitreihen als andere Datensätze. Um eine Dominanz solcher Datensätze zu vermeiden, wurden strategische Sampling-Techniken angewendet, um eine ausgewogenere Verteilung für **TSB-AD-U-Eval** sicherzustellen.

13 Kategorien von Algorithmen zur Anomalieerkennung (40 Algorithmen)

- Statistische Methoden: Fokus auf klassische Ansätze, die auf mathematischen Modellen basieren, um Anomalien durch Abweichungen von erwarteten Mustern zu identifizieren.
- Neuronale Netzwerk-basierte Methoden: Diese Methoden lernen typische Muster in historischen Trainingsdaten, die keine Anomalien enthalten. Abweichungen in neuen Testdaten werden als potenzielle Anomalien erkannt.
- Foundation-Model-basierte Methoden: Diese Modelle zeichnen sich durch beeindruckende Few-Shot- und Zero-Shot-Fähigkeiten aus. Hier kann man LLMs adaptieren wie z.B. OFA Modell oder MOMENT (Modell für Zeitreihen)

Um eine faire Vergleichsbasis zwischen den verschiedenen Modellen sicherzustellen, werden standardisierte Metriken wie der Mean Squared Error (MSE) verwendet.

14 Evaluation Measures

Es werden sowohl punktbasierte als auch bereichsbasierte Metriken verwendet.

14.1 Punktbasierte

- AUC-ROC, AUC-PR, F1, PA-F1, Event-based-F1

14.2 Bereichsbasierte Metriken

- R-based-F1, Affiliation-F1, Volume Under the Surface (VUS), PATE (Proximity-Aware Temporal Evaluation),

Bereichsbasierte Metriken sind robuster, indem sie zeitliche Natur von Zeitreihen und ihre konsistenten Muster berücksichtigen

15 Benchmark Evaluation and Analysis

15.1 Tuning/Evaluation Dataset Splitting

- Splitting Strategy: 15% der Daten jedes Datensatzes werden für die Hyperparameter-Tuning-Phase reserviert. (Dies ist aus jedem Datensatz)
- Die verbleibenden 85% dienen zur Evaluierung und zum Leistungsvergleich der Algorithmen.

15.2 Hyperparameter Tuning

- Sicherstellen, dass alle Algorithmen unter ihren optimalen Konfigurationen verglichen werden.
- Insgesamt ergeben sich 15 mögliche Modelle, von denen das beste auf dem Tuning-Set ausgewählt wird.
- Bei neuronalen Netzen werden Hyperparameter wie Lernrate, Anzahl der versteckten Schichten untersucht.

Es wurden Über 450 Varianten aus den 40 verschiedenen Erkennungsalgorithmen generiert.

16 Experimentelle Ergebnisse und Diskussion

- VUS-PR zeigte im Vergleich zu anderen Maßnahmen erhebliche Robustheit gegenüber Verzögerungen (Abbildung 6a).
- Im Gegensatz dazu führt PATE, obwohl es die Prinzipien von VUS erweitert, zu neuen Herausforderungen, die die Bewertung komplizieren und die Rechenanforderungen erheblich erhöhen.

17 Benchmark Accuracy Evaluation

- Die Top 12 Methoden bestehen überwiegend aus statistischen Ansätzen, wobei Sub-PCA auf Platz 1 ist.
- Es sind nur wenige neuronale Netzwerkmethoden (USAD, CNN)
- Die feinabgestimmte Version von MOMENT besser abschneidet als die Zero-Shot-Version
- In TSB-AD-M zeigen neuronale Netzwerkmethoden wie CNN und OmniAnomaly bessere Ergebnisse, aber auch statistische Methoden bleiben in multivariaten Fällen sehr effektiv. PCA, CNN und USAD gehören zu den konstant besten Methoden in beiden Datensätzen.

18 Ergebnisse der Algorithmen je nach Anomalietyp

Die Ergebnisse zeigen, dass Foundation-Model-basierte Ansätze starke Leistung bei der Erkennung von Punktanomalien zeigen (TimesFM und Chronos).

Bei Sequenzanomalien sind jedoch statistische Methoden die besten.

Die neuronale Netzwerke zeigen insgesamt eine höhere Effektivität bei der Erkennung von Punktanomalien. In Szenarien mit einzelnen Anomalien schnitten jedoch keine neuronalen Netzwerkmethoden als Top-Kandidaten ab.

In komplexeren Szenarien mit mehreren Anomalien konnte jedoch MOMENT seine Effektivität unter Beweis stellen.

19 Diskussion und Zusammenfassung

Statistische Methoden zeigen insgesamt eine robuste Leistung. Im Gegensatz dazu schneiden neuronale Netzwerke nicht so gut ab, wie oft erwartet.

Neuronale Netze und die Foundation-Modelle können bei der Erkennung von Punktanomalien und in multivariaten Szenarien weiterhin gut abschneiden.

Foundation-Modelle sind ganz gut bei der Erkennung von Punktanomalien, haben jedoch Schwierigkeiten mit Sequenzanomalien. Dies liegt an ihrem prädiktiven Mechanismus, der nur einen Wert pro Schritt unter Verwendung eines begrenzten Zeitrahmens schätzt, was zu verringerten Leistungen und veräuschten Ergebnissen führt.

Der Versuch mit LLMs führte zu unbefriedigenden Ergebnissen, was auf eine signifikante Forschungslücke in diesem Bereich hinweist.

CNN und OmniAnomaly zeigen in multivariaten Szenarien herausragende Ergebnisse. Dies deutet darauf hin, dass komplexe Szenarien in multivariaten Zeitreihen eine stärkere Modellkapazität erfordern, die häufig über die der statistischen Methoden hinausgeht.

Diese Arbeit wurde teilweise von Meta und Cisco Systems unterstützt. (Das Paper)