Kapitel 4

Anomaliedetektion

Anomaliedetektion beschreibt die Aufgabe, Trends, Muster und Punkte in einem Datensatz zu finden, die nicht dem Normalzustand entsprechen [5]. Anders gesagt lautet das Ziel: die Punkte finden, die sich von den anderen Punkten im Datensatz stark unterscheiden [11, Kap. 10]. Diese andersartigen Datenpunkte oder -sequenzen werden in der Regel als Anomalie, Ausreißer oder Ausnahmen bezeichnet, wobei Anomalie der geläufigste Begriff ist. Anomaliedetektion findet große Verwendung in verschiedenen Anwendungsbereichen, wie z. B. in der Netzwerktechnik zur Erkennung von potenziellen Angriffen durch Eindringlinge in ein Netzwerk anhand von ungewöhnlichem Traffic [1]. Auch in der Medizin können nach einem Elektrokardiogramm (*EKG*) durch Anomaliedetektion Herzrhythmusstörungen erkannt werden [6], genau wie eine Bank ein Interesse an Anomalien im Kreditkartenverhalten ihrer Kunden hat, um Betrugsfälle zu erkennen [8, 4].

Die simpelste Herangehensweise zur Erkennung von Anomalien ist die, dass zuerst definiert wird, welche Punkte im Datensatz normalem Verhalten entsprechen und alle davon abweichenden Punkte als Anomalie zu kennzeichnen. Doch so einfach die Herangehensweise wirkt, so anfällig ist sie auch für Fehler. Dabei heben sich einige Herausforderungen hervor.

Zum Einen die Frage, wo genau die Grenze zwischen normalem und anomalem Verhalten liegen soll. Eine Region zu definieren, die jeden möglichen normalen Punkt beinhaltet und jedmöglichen anomalen Punkt ausschließt, ist nicht trivial und oft nicht präzise durchführbar. So ist es durchaus möglich, dass in manchen Fällen anomale Punkte als normal bezeichnet werden, und normale Punkte als anomal, je nachdem, wo die Grenze liegt.

Es stellt sich ebenfalls die Frage, ob eine Anomalie einer binären Natur unterliegt: Entweder es handelt sich um eine Anomalie oder einen Normalzustand. Doch die Wahrheit liegt oft in der Mitte. Weicht ein Punkt oder eine Sequenz bereits nur leicht vom Normal ab, so kann es bereits erste Hinweise auf mögliches zukünftiges anomales Verhalten in einer Zeitserie geben, bevor sich solche Datenpunkte als Anomalie zeigen. Deshalb ist es hilfreich, charakterisieren zu können, wie weit der Punkt oder die Sequenz vom Normal abweicht. Diese Charakterisierung kann dabei als *Anomaly Score* bezeichnet werden und beispielsweise eine Dezimalzahl zwischen 0 und 1 sein.

Normalzustände sind in Zeitserien oft zeitvariant und daher schwer festzuhalten bei einer kontinuierli-

chen Datenaufzeichnung. Zudem sind Normalzustände und Abweichungen davon in unterschiedlichen Bereichen auch unterschiedlich signifikant. Während beim menschlichen Körper eine geringe Abweichung der Körpertemperatur bereits gravierend sein kann, ist die gleiche relative Abweichung in einer anderen Domäne wie in einem Aktienkurs weniger drastisch und unterliegt dementsprechend auch einem Anpassungsbedarf, bevor es an die Erkennung möglicher Anomalien geht.

Daraus lässt sich direkt zum nächsten Problem übergehen. Die Unterscheidung zwischen globalen und lokalen Anomalien [3]. Hier ist der Kontext wichtig: Eine Person mit einer Körpergröße von mehr als 2 m ist in ihrer Nachbarschaft sicherlich eine Anomalie, während sie in einem Basketballteam kaum herausragt - im wahrsten Sinne des Wortes. Diese Art der Anomalie wird auch als kontextuelle Anomalie bezeichnet [13, S. 12].

4.1 Anomaliearten

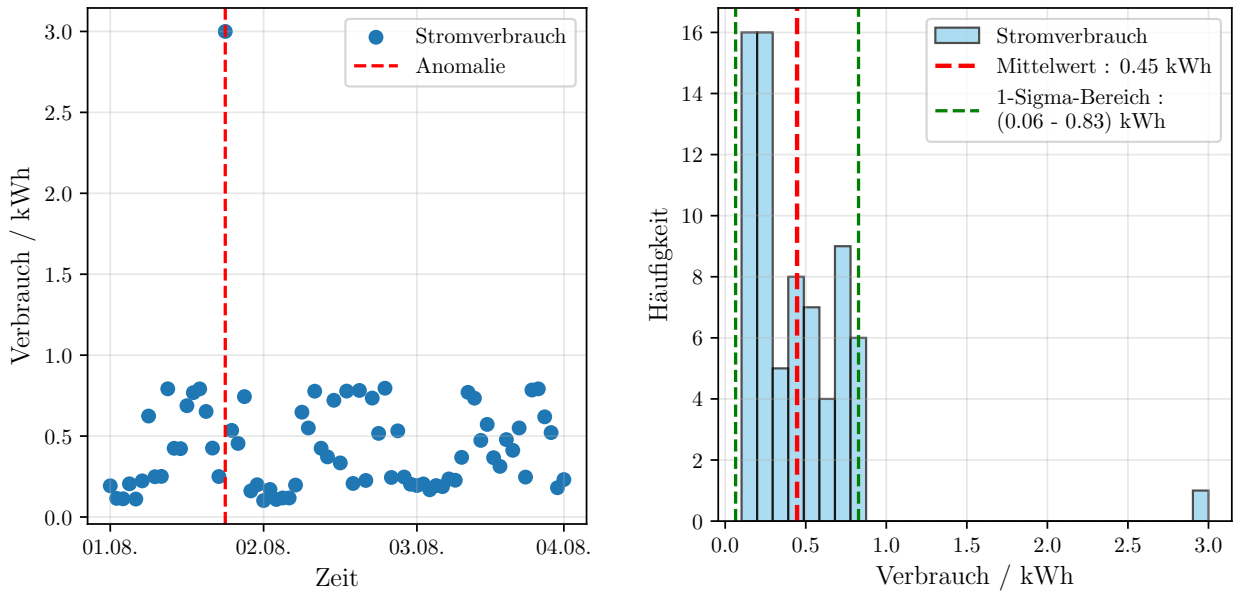
Doch bevor eine Auswahl an geeigneten Verfahren oder Algorithmen zur Anomaliedetektion getroffen wird, muss zuerst verstanden werden, welche verschiedenen Arten von Anomalien es gibt und wie sich diese voneinander unterscheiden. Auch wenn Studien zeigen, dass es durchaus Algorithmen gibt, die über mehrere verschiedene Kategorien gut abschneiden [13, S. 30 - 31] [10], so soll zunächst für jede Kategorie mindestens ein passender Kandidat gefunden werden. Diese werden dann in einem nächsten Schritt kreuzweise getestet, um auch solche Allrounder entdecken zu können. Dabei ist auch immer der Kontext der Anwendung wichtig. Wie eingangs erwähnt, sind für verschiedene Tätigkeitsfelder verschiedene Anforderungen an die Präzision oder Genauigkeit gestellt, weshalb immer die spezifischen Anforderung bedacht werden müssen, und nicht jeder Algorithmus gleich performant ist über mehrere Datensätze hinweg.

Für die Kategorien wird sich zunächst auf wenige, für diese Arbeit relevante, beschränkt: **Punktanomalien**, **Subsequenzanomalien** und **Korrelationsanomalien**, abgeleitet von Chandola et al. [5].

4.1.1 Punktanomalien

Ein einzelner Datenpunkt, der stark von den anderen Punkten im Datensatz abweicht, heißt Punktanomalie [5]. Genauer gesagt, wenn ein Datenpunkt weit außerhalb der Wahrscheinlichkeitsverteilung des Datensatzes liegt, ist er anomal [11, Kap. 10]. Punktanomalien können recht leicht erkannt werden, da Punktanomalien stark vom Mittelwert und vom Median des Datensatzes abweichen. Wenn von Ausreißern gesprochen wird, sind damit typischerweise Punktanomalien gemeint.

Als Beispielszenario dient ein Smart Meter, das den stündlichen Stromverbrauch misst. In Abb. 4.1a ist der gemessene Stromverbrauch dargestellt mit einer klar erkennbaren Punktanomalie am 01.08. um 18 Uhr. Die Anomalie wird mit bloßem Auge deutlich und kann auch mit statistischen Größen nachgewiesen werden, wie in Abb. 4.1b anhand der Häufigkeitsverteilung und dem Mittelwert sowie dem Median zu sehen ist. Das Histogramm dient als gute Approximation für die Wahrscheinlichkeitsverteilung der Messwerte, und zeigt entsprechend die Eindeutigkeit des Ausreißers.



(a) Stündliche Smart Meter Messdaten

(b) Histogramm des gemessenen Stromverbrauchs

Abbildung 4.1: Beispielszenario einer Punktanomalie: Stromverbrauch eines Haushaltes über den Zeitraum von drei Tagen. Anhand des Histogramms wird die Anomalie verdeutlicht.

Um nun eine Aussage treffen zu können, ist es wichtig den Kontext der vorliegenden Daten zu kennen. Wenn Daten für ein weitaus größeres Zeitfenster vorliegen, z. B. für eine Woche oder einen Monat, könnte sich möglicherweise zeigen, dass der hohe Verbrauch öfter und regelmäßiger vorkommt als im gezeigten Zeitraum von drei Tagen. Ob eine globale oder lediglich eine lokale Anomalie vorliegt, wird mit einem größeren Datensatz besser erkennbar. Die Anomalie könnte beispielsweise auf das gelegentliche Betreiben einer Sauna im Haus zurückführbar sein, dann würde es sich lediglich um eine lokale Anomalie handeln und in einem größeren Zeitraum in bestimmten Abständen öfter vorkommen, und wäre somit keine globale Anomalie [11, Kap. 10].

Punktanomalien sind im Kontext dieser Arbeit tendenziell weniger relevant, sollen aber aufgrund ihrer grundsätzlichen Bedeutung bzgl. Anomaliedetektion als einfachste Kategorie trotzdem beleuchtet werden, um entsprechende Algorithmen, die der Erkennung solcher Punktanomalien zuzuordnen sind, auch gegenüber anderen Anomalien zu testen.

4.1.2 Subsequenzanomalien

Eine Zeitserie wird gem. Gl. 3.1 bereits als eine Menge definiert. Demnach wird eine Subsequenz $S_{i,j} = \{ S_i, \ldots, S_j \} \subseteq S$ von der Zeitserie S umfasst, mit der Länge oder Mächtigkeit $|S_{i,j}| = j - i + 1$ und $|S_{i,j}| \geq 1$ [10] und stellt somit einen Ausschnitt der urpsrünglichen Zeitserie dar. Subsequenzanomalien sind Muster in Zeitreihen, die von anderen Mustern innerhalb der gleichen Zeitreihe abweichen [5][13, S. 12]. Im Gegensatz zu Punktanomalien beziehen sich Subsequenzanomalien auf mehrere konsekutive Datenpunkte, die ein ungewöhnliches Muster bilden. Eine anomale Subsequenz kann also bedeuten, dass die Datenpunkte innerhalb der Subsequenz Werte in einem normalen, zu erwartenden Bereich annehmen, aber der zu Grunde liegende Trend ungewöhnlich ist [5][2, S. 17]. Solche

ungewöhnlichen oder einzigartigen Trends und Entwicklungen können auf zukünftig auftretende Probleme hindeuten, die sonst unentdeckt bleiben würden.

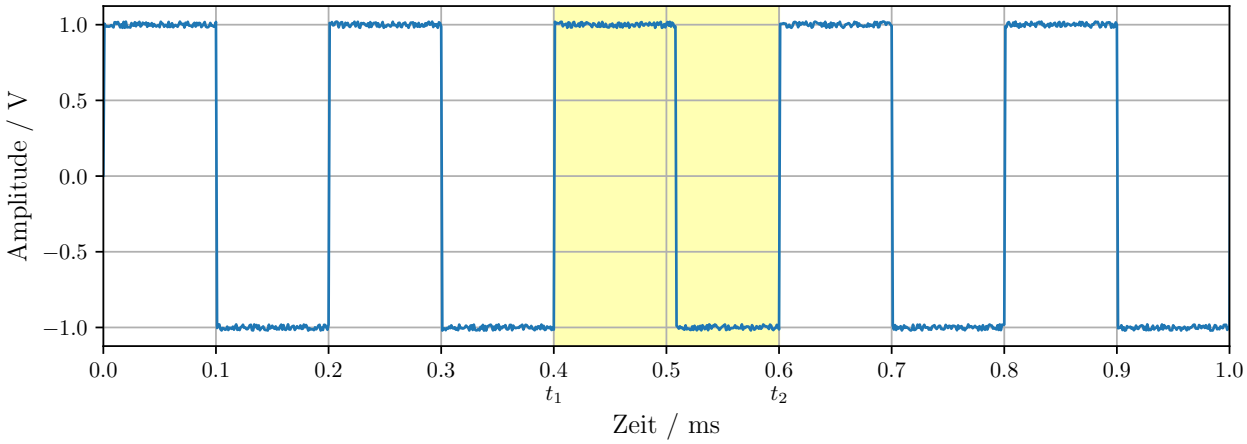


Abbildung 4.2: Einfaches Beispiel einer Subsequenzanomalie: Rechteckspannung, die zwischen -1 und +1 V oszilliert mit einer Frequenz von 5 kHz. Auffällig ist die Periode zwischen $t_1 = 0.4\,ms$ und $t_2 = 0.6\,ms$, bei der eine verspätete abfallende Flanke zu beobachten ist.

Das Beispiel in Abb. 4.2 zeigt eine sichtbare Subsequenzanomalie, die verspätete abfallende Flanke einer gemessenen Rechteckspannung. Das Muster zwischen t_1 und t_2 ist also merklich anders verglichen zu den restlichen 0,2 ms langen Perioden und daher eine Anomalie.

Bei der Analyse von EKG Daten spielen Subsequenzanomalien eine wichtige Rolle und können wertvolle Rückschlüsse auf die Herzgesundheit liefern [6]. Abb. 4.3 zeigt EKG-Daten eines Patienten mit monomorpher ventrikulärer Tachykardie. Diese kann zu Kammerflimmern übergehen, welches unbehandelt sogar zu einem Herzstillstand führen kann [12][7, S. 131 ff.].

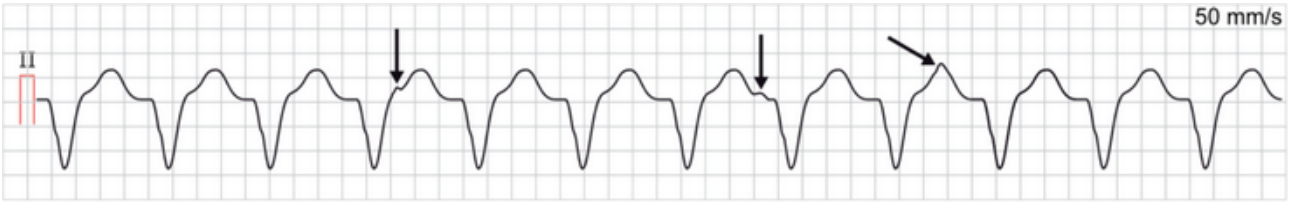


Abbildung 4.3: EKG Kanal mit Diagnose: Ventrikuläre Tachykardie [12]

Sichtbar sind die einzelnen Unregelmäßigkeiten im EKG Verlauf. Die Pfeile kennzeichnen die sog. P-Wellen, die Informationen darüber liefern, dass Vorhöfe und Herzkammern nicht synchron schlagen [12][7, S. 31 f.]. Durch die Irregularitäten lässt sich also erkennen, dass für den untersuchten Patienten eine Behandlung notwendig ist und betont die Wichtigkeit, diese Anomalien zu erkennen, um wesentlich Schlimmeres zu verhindern.

Darin liegt auch eine der Herausforderungen der Subsequenzanomaliedetektion: Ab wann ist ein Trend, der so noch nicht aufgetreten ist, Grund genug, um Maßnahmen zu ergreifen? Es bedarf also menschlicher Expertise zur Einordnung und Interpretation von Anomalien, eben wie bei EKG Daten.

4.1.3 Korrelationsanomalien

Während Punkt- und Subsequenzanomalien sowohl für univariate als auch multivariate Datensätze und Zeitserien auftreten können, sind Korrelationsanomalien nur möglich bei zwei oder mehr Dimensionen einer Zeitreihe und betrachten die Interaktionen zwischen verschiedenen Kanälen. Von einer Korrelationsanomalie spricht man bei Abweichungen dieser Beziehung zwischen zwei oder mehreren Kanälen [13, S.12-13] [14].

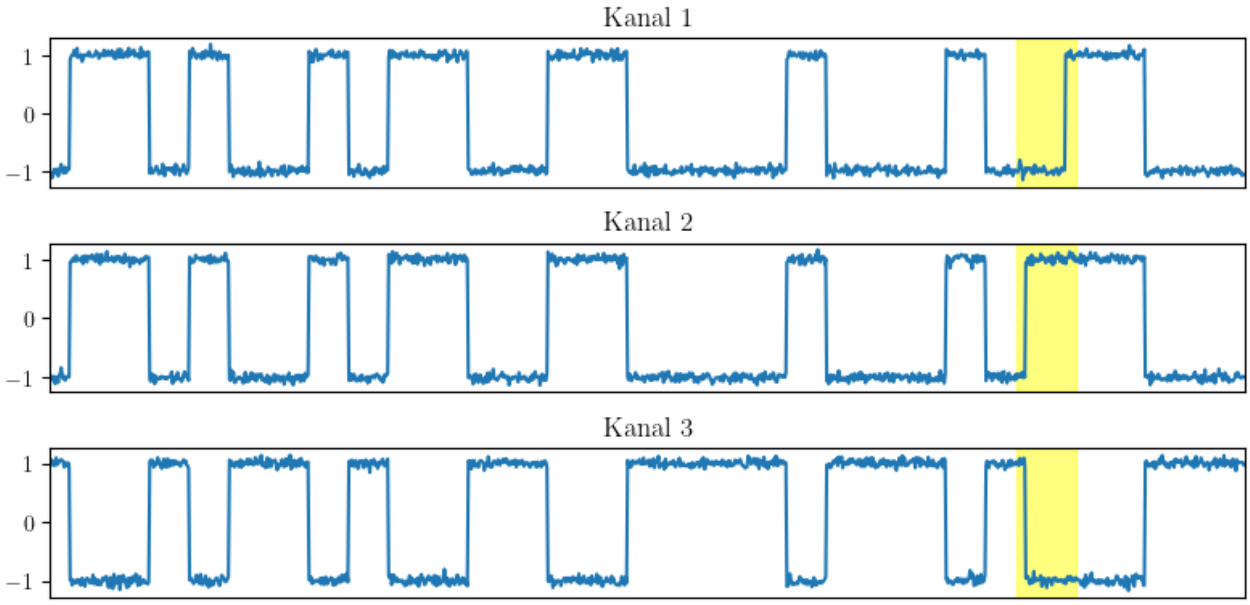


Abbildung 4.4: Korrelationsanomalie zwischen Kanal 1 und den Kanälen 2 und 3 im gelb markierten Bereich. Quelle: Datensatz *CoMuT* [9]

Im vorliegenden Beispiel in Abb. 4.4 ist ein Auszug aus dem Datensatz *CoMuT* - **Co**rrelated **Mu**ltivariate **T**ime Series [9] dargestellt. Die Zeitreihe besteht aus drei Kanälen, die zu zufälligen Zeitpunkten sprungartig ihren Wert zwischen -1 und 1 wechseln und jeweils leicht verrauscht sind. Kanal 1 und 2 sind stark korreliert, während Kanal 3 stark antikorreliert zu den beiden ersten Kanälen ist. Diese Korrelation wird im markierten Bereich verletzt, da Kanal 1 später springt als die anderen beiden Kanäle – somit liegt eine Korrelationsanomalie vor.

4.2 Algorithmen zur Anomaliedetektion

Um nun eine geeignete Auswahl an Algorithmen zu treffen

Literaturverzeichnis

[1] Jarosław Bernacki and Grzegorz Kołaczek. "Anomaly Detection in Network Traffic Using Selected Methods of Time Series Analysis". In: *International Journal of Computer Network and Information Security* 7.9 (Aug. 2015), pp. 10–18. ISSN: 2074-9104. DOI: 10.5815/ijcnis.2015.09.02.

[2] Paul Boniol and Themis Palpanas. "Detection of anomalies and identification of their precursors in large data series collections". PhD thesis. 2021. URL: http://www.theses.fr/2021UNIP5206/document.

[3] Markus M. Breunig et al. "LOF: Identifying Density-Based Local Outliers". In: *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*. SIGMOD/PODS00. ACM, May 2000, pp. 93–104. DOI: 10.1145/342009.335388.

[4] V. Ceronmani Sharmila et al. "Credit Card Fraud Detection Using Anomaly Techniques". In: *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*. IEEE, Apr. 2019, pp. 1–6. DOI: 10.1109/iciict1.2019.8741421.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. "Anomaly detection: A survey". In: *ACM Computing Surveys* 41.3 (July 2009), pp. 1–58. ISSN: 1557-7341. DOI: 10.1145/1541880.1541882.

[6] Mooi Choo Chuah and Fen Fu. "ECG Anomaly Detection via Time Series Analysis". In: *Frontiers of High Performance Computing and Networking ISPA 2007 Workshops*. Springer Berlin Heidelberg, 2007, pp. 123–135. ISBN: 9783540747673. DOI: 10.1007/978-3-540-74767-3_14.

[7] Alan Davies and Alwyn Scott. *Starting to Read ECGs: A Comprehensive Guide to Theory and Practice*. Springer London, 2015. ISBN: 9781447149651. DOI: 10.1007/978-1-4471-4965-1.

[8] Shanshan Jiang et al. "Credit Card Fraud Detection Based on Unsupervised Attentional Anomaly Detection Network". In: *Systems* 11.6 (June 2023), p. 305. ISSN: 2079-8954. DOI: 10.3390/systems11060305.

[9] Felix Naumann. *CoMuT - Correlated Multivariate Time Series*. 2024. URL: https://hpi.de/naumann/s/comut.

[10] Sebastian Schmidl, Phillip Wenig, and Thorsten Papenbrock. "Anomaly detection in time series: a comprehensive evaluation". In: *Proceedings of the VLDB Endowment* 15.9 (May 2022), pp. 1779–1797. ISSN: 2150-8097. DOI: 10.14778/3538598.3538602.

[11] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. New internat. edition. Always learning. Harlow: Pearson, 2014. 732 pp. ISBN: 9781292026152.

[12] *Ventrikuläre Tachykardie (VT)*. Aufgerufen: 08.01.2025. URL: https://ekgecho.de/thema/ventrikulaere-tachykardie-vt-ekg-kriterien-ursachen-klassifikation-behandlung-management/.

[13] Phillip Wenig. "Finding, Clustering, and Classifying Anomalies on Large and Multivariate Time Series". en. PhD thesis. 2024. DOI: 10.25932/PUBLISHUP-66043.

[14] Phillip Wenig, Sebastian Schmidl, and Thorsten Papenbrock. "Anomaly Detectors for Multivariate Time Series: The Proof of the Pudding is in the Eating". In: *2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW)*. IEEE, May 2024, pp. 96–101. DOI: 10.1109/icdew61823.2024.00018.