Bachelorarbeit
Einsatz von Prozessanalysen für maschinelle Lernverfahren zur Anomalieerkennung in Multibeam Daten

-

Sommersemester 2022

Erstbetreuer: Tobias Ziolkowski

Manuel Krebs

20. August 2022

Inhaltsverzeichnis

1 Einführung **1**

2 Literaturübersicht **2**

3 Methodik **3**
3.1 Interquartilsabstand . . . 3
3.1.1 Funktionsweise . . . 3
3.1.2 Anwendung . . . 4
3.2 Isolation Forest . . . 5
3.2.1 Funktionsweise . . . 5
3.2.2 Anwendung . . . 5
3.3 Local Outlier Factor . . . 6
3.3.1 Funktionsweise . . . 6
3.3.2 Anwendung . . . 6
3.4 One-Class Support Vector Machine . . . 7
3.4.1 Funktionsweise . . . 7
3.4.2 Anwendung . . . 7

4 Ergebnisse **8**

5 Diskussion **9**

6 Fazit **10**

Kapitel 1

Einführung

- Multibeam Daten -¿ Definition
- Prozessanalysen für maschinelle Lernverfahren -¿ Definition und warum wichtig
- Anomalien -¿ Definition, Beispiele und Erklärung, dass Anomalie = Outlier
- Sagen, was ich mit dem Paper erreichen möchte -¿ Forschungsfrage

Outliers are the patterns which are not in the range of normal behavior [2]
Moving window pattern! Chandola, V., Banerjee, A., Kumar, V.: Outlier detection: a survey

Kapitel 2

Literaturübersicht

Andere Anwendungsgebiete!

Kapitel 3

Methodik

3.1 Interquartilsabstand

Der Interquartilsabstand (im folgenden mit IQR angekürzt) ist ein Streuungsmaß, welcher in statistischen Analysen dabei hilft, durch Verteilungen Rückschlüsse über einen Datensatz zu ziehen. Visuell werden diese meist durch einen Boxplot dargestellt, welche zu den am weitesten verbreiteten Werkzeugen in der statistischen Praxis, insbesondere in der Phase der explorativen Datenanalyse [1].

3.1.1 Funktionsweise

Zunächst wird das erste und das dritte Quartil berechnet. Der IQR bildet sich nun aus der Differenz des der beiden Quartile, also IQR = Quartil 3 - Quartil 1. Um dies als graphische Repräsentation zu verdeutlichen zeigt Grafik 3.1 einen Boxplot, welcher die 'z'-Koordinate der ersten 1000 Messdaten veranschaulicht. Der IQR lässt sich nun durch die blau eingefärbte Fläche zeigen, welche 50 Prozent der Messdaten enthält. Alle Messwerte, welche außerhalb oberen und unteren Grenze liegen, lassen sich nun als potenzielle Outlier identifizieren. Diese Grenzen werden typischerweise auf das 1,5-fache des IQR gesetzt. Die untere Grenze lässt sich also durch Quartil 1 - (1.5 * IQR) und die obere Grenze durch Quartil 3 + (1.5 * IQR) berechnen [2]. Wie in der Grafik 3.1 nun zu erkennen ist, befinden sich einige der Messwerte nur knapp außerhalb der Grenzen. bei den Werten weit außerhalb der Grenzen ist die Wahrscheinlichkeit am höchsten, dass diese Ausreißer sind.

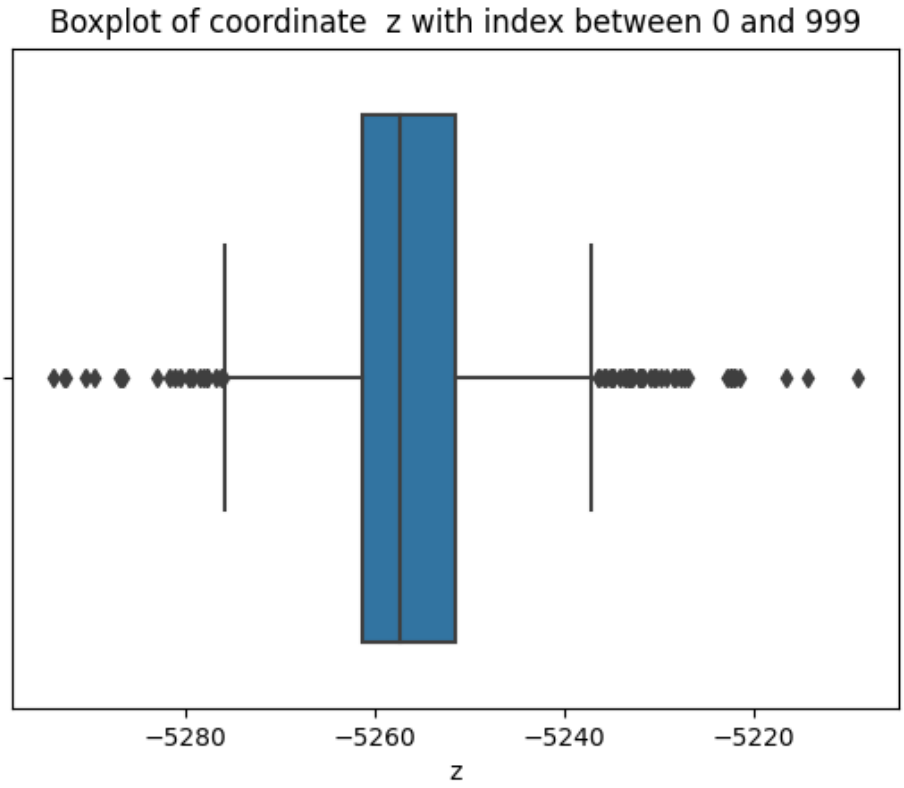

Abbildung 3.1: Example Boxplot

3.1.2 Anwendung

Bevor die Methodik angewendet werden kann muss der Datensatz angepasst werden. Dies geschieht mithilfe dem zuvor beschrieben Ansatz des Moving-Window-Patterns. Nachdem der Datensatz aufgeteilt wurde, wird die Methode mit den einzelnen Chunks aufgerufen. Hier hat man nun mehrere Möglichkeiten:

- Händische Identifizierung von Outliern mithilfe der Ausgabe mehreren Boxplots.
- Berechnung der oberen und unteren Grenze und automatische Filterung aller Werte, welche sich außerhalb befinden.

3.2 Isolation Forest

Zhang, J., Zulkernine, M.: Anomaly based network intrusion detection with unsupervised outlier detection

3.2.1 Funktionsweise

3.2.2 Anwendung

3.3 Local Outlier Factor

Jabez, J., Muthukumar, B.: Intrusion detection system: anomaly detection using outlier detection approach. ICCC, 338–346 (2015)

3.3.1 Funktionsweise

3.3.2 Anwendung

3.4 One-Class Support Vector Machine

3.4.1 Funktionsweise

3.4.2 Anwendung

Kapitel 4

Ergebnisse

Kapitel 5

Diskussion

Kapitel 6

Fazit

Literaturverzeichnis

[1] Y. H. Dovoedo and S. Chakraborti. Boxplot-Based Outlier Detection for the Location-Scale Family. 44(6):1492–1513.

[2] H. P. Vinutha, B. Poornima, and B. M. Sagar. Detection of Outliers Using Interquartile Range Technique from Intrusion Dataset. In Suresh Chandra Satapathy, Joao Manuel R.S. Tavares, Vikrant Bhateja, and J. R. Mohanty, editors, *Information and Decision Sciences*, volume 701 of *Advances in Intelligent Systems and Computing*, pages 511–518. Springer Singapore.