technische universität
dortmund

Bachelorarbeit

Detektion von Zeitreihenanomalien in der Niederspannung

Joël Haubold
Monat der Abgabe

Gutachter:
Prof. Dr. Rudolph
Dr.-Ing. Sebastian Ruthe

Technische Universität Dortmund
Fakultät für Informatik
Lehrstuhl für Computational Intelligence (LS-11)
https://ls11-www.cs.tu-dortmund.de/

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Motivation und Hintergrund

1.1.1 Anomaliererkennung auf Zeitreihen

Anomalieerkennung ist ein weitreichendes Forschungsgebiet, welches seit langem XDDDD

1.1.2 Daten aus der Niederspannung

Das deutsche Verteilnetz wurde ursprünglich mit dem Ziel gebaut, den in Großkraftwerken produzierten Strom und über das Transportnetz in die einzelnen Regionen Deutschlands transportiert wird, regional an die Endkunden (sowohl Industrie- und Gewerbekunden als auch Haushalte) zu verteilen. Das Verteilnetz ist dabei baumartig strukturiert und besteht aus der Hochspannungsebene die den Übergabepunkt des Transportnetz enthält und sich hin zur Mittelspannungsebene, Niederspannungsebene und schließlich den Endkunden verzweigt.
Mit zunehmender Integration von Erneuerbaren Energien wie Wind- und PV-Anlagen in die Mittel- und Niederspannungsebene steigt auch die Dynamik in den unteren Spannungsebenen. Lastflüsse die vorher stets von oben (Hochspannung) nach unten (Mittel-, Niederspannung) gerichtet waren, kehren sich in Teilen um und können zu einer lokal höheren Auslastung des Netzes führen. Hinzu kommen neue Verbraucher wie z.B. Elektrofahrzeuge die insbesondere in den frühen Abendstunden und über die Nacht verteilt das Netz stärker belasten.
Um diese Effekte erkennen und analysieren zu können, müssen die Niederspannungsebene zunächst messtechnisch erfasst werden. Die Firma PPC hat ein Messgerät entwickelt, welches sich in Ortsnetzstationen (Übergabepunkt von Mittel- zu Niederspannung) einbauen lässt und dort eine dreiphasige Spannungsmessung durchführen kann. Zusätzlich verfügt das Messgerät über eine Kommunikationsanbindung mit der sich die Daten abrufen und an einem zentralen Punkt aggregieren und auswerten lassen.

1.2 Aufbau der Arbeit

In dieser Arbeit werden zuerst in Kapitel 2 die Grundsätze von Zeitreihenanomalieerkennung, sowie die zwei in ihr eingesetzten Verfahren Robust Random Cut Forest", und Öne Dimensional Support Vector Machineërläutert. In Kapitel 3 wird auf die im Rahmen dieser Arbeit angewendete Implementierung und deren Ergebnisse eingegangen. In Kapitel 5 werden, auf der Basis dieser Ergebnisse, die beiden implementierten Verfahren miteinander verglichen.

Kapitel 2

Grundlagen

2.1 Notationen

Die in dieser Arbeit verwendeten Notationen lehnen sich an die in dem Papier [4] verwendeten an:

- :DDDDD

2.2 Anomalien

In einem gegebenen Datensatz Z an Punkten, wird einer dieser Punkte $x \in Z$ als Outlier bezeichnet, falls er sich signifikant in einen oder mehreren seiner Merkmale von den Punkten $Z - \{x\}$ unterscheidet. Seien Y alle anomalen Punkte aus Z. Ein Modell welches Z darstellt ist entsprechend wesentlich komplexer als ein Modell welches $Z - Y$, also nur die nicht-normalen *Inliners* von Z darstellt. Wie stark sich x in seinen Merkmalen von anderen Punkten in Z unterscheiden muss, beziehungsweise wie stark x die Komplexität des Modells von Z erhöht, damit x als Anomalie gesehen wird ist hängt oft von der jeweiligen Zielsetzung ab.
Die meisten Applikationen erzeugen ihre Daten über einen oder mehreren generierenden Prozessen, beispielsweise durch die Beobachtung von Nutzeraktivität, oder durch das Ablesen von externen Daten. Dementsprechend lassen sich über das Erkennen dieser Anomalien Informationen über die jeweilige Applikationen sammeln. [1]

2.2.1 Komplikationen

Die Diversität von möglichen Datensätzen und deren Merkmalen macht es generell nicht möglich, ein allgemeines Vorgehen für die Erkennung von Anomalien zu bestimmen. Dazu kommen mögliche Eigenschaften die dies weiterhin erschweren, oder es bestimmten Vorge-

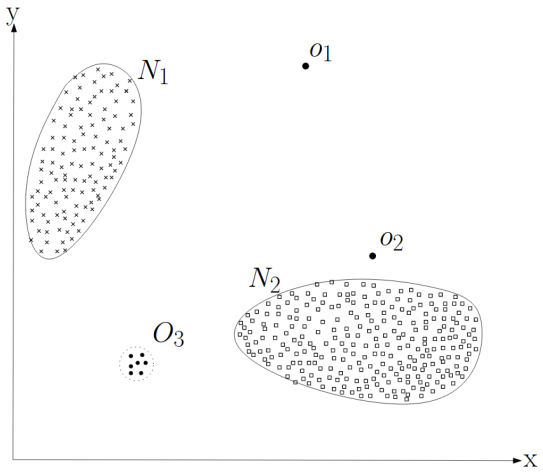


Abbildung 2.1: Ein Beispieldatensatz mit zwei Anomalien o_1 und o_2, sowie eine Punktegruppe O_3 von 7 Anomalien. Die Gruppen N_1 und N_2 stellen die Inliner des Datensatzes da. Quelle: [2]

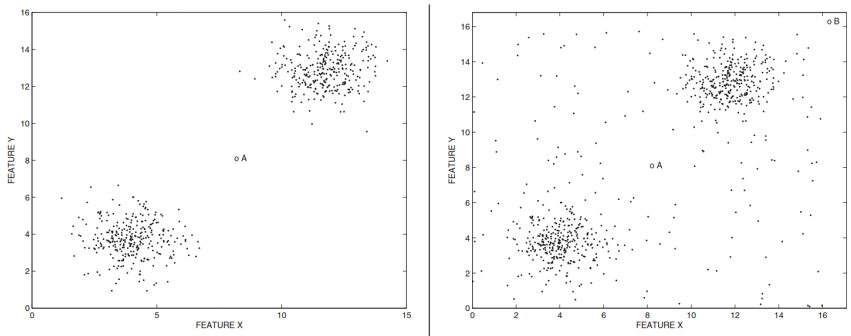
hen sogar unmöglich machen, Anomalie von Inliner zu unterscheiden. Ein Überblick über einige dieser ist hier aufgeführt:

Kontextabhängigkeit

Es ist zu beachten das bei zwei anomalen Punkten nicht die gleichen Grenzwerte für die einzelnen Merkwerte gelten müssen, es kommt vielmehr auf die Kombination der Merkmale an. Ein einfaches Beispiel ist ein über die Zeit stetig zunehmender Messwert. Ein Punkt dessen Wert zu Beginn aus der Zeitreihe nach oben ausreißt, ist wahrscheinlich anomal. Die Punkte die später durch den Trend der Zeitreihe diesen Wert überschreiten, sind deswegen aber nicht zwingend selber anomal, noch invalidieren sie den Status des Ausreißers als Anomalie. [5]

Duplikate

Erschwerend für die Anomalieerkennung kann es sein falls sich mehrere Anomalien eines Datensatzes ähneln, wie in Abbildung 2.1. Während sich die Punkte in O_3 eindeutig von den beiden Inliner-Punktegruppen N_1 und N_2 abgrenzen, so haben sie alleinstehend betrachtet dennoch untereinander eine starke Ähnlichkeit., ein Modell des dargestellten Datensatzes vereinfacht sich durch die einzelne Entfernung eines Punktes aus O_3 nicht. [4] Sollen die Punkte in O_3 von einem Anomalieerkennungsverfahren als Anomalie eingestuft werden, so muss entweder dem Verfahren mitgeteilt werden das Inliner Ähnlichkeiten zu den Punkten in N_1 und N_2 haben müssen, oder es muss so kalibriert werden, dass eine Ansammlung von 7 ähnlichen Punkten noch nicht als Inlinergruppe gesehen wird. Mehr dazu in Sektion 2.3.1



(a) No noise (b) With noise

Abbildung 2.2: Der Einfluss von Rauschen auf einen Datensatz bestehend aus zwei Inlinergruppen und einem anomalen Punkt A. Quelle: [1]

Rauschen

Je nach generierenden Prozess des Datensatzes kann es sein das in diesem neben der zu beobachtenden Größe, weitere Punkte aufgenommen werden, welche sich in ihren Merkmalen stark von den Inlinern unterscheiden, aber nicht von Relevanz für den Beobachter des Prozesses sind. [1] In den beiden Abbildungen 2.2 ist die Schwierigkeit die Rauschen bei der Anomalieerkennung mit sich bringt zu sehen. In Abbildung 2.2 (a) ist der Punkt A offensichtlich anomal. In 2.2 (b) könnte dieser allerdings Teil des Rauschens sein. Um den Punkt A als anomal markieren zu können, aber nicht den Rest des uninteressanten Rauschens, muss dem Anomalieerkennungsverfahren mitgeteilt werden das Punkte mit seinen Merkmalen als anomal gelten.

Mehrdimensionalität

Hat der zu untersuchende Datensatz eine hohe Dimensionalität in seinen Merkmalen, führt dies zu weiteren Problemen bei der Anomalieerkennung. Mit zunehmender Anzahl an Merkmalsdimensionen erhöhen sich die möglichen Kombinationen an Dimensionen auf denen nach anomalen Merkmalen gesucht werden kann exponentiell, womit der Aufwand der Anomalieerkennung ansteigen kann. Weiterhin führt diese Zunahme der möglichen Dimensionskombinationen auf denen gesucht werden kann, dass es immer wahrscheinlicher wird, für jeden Punkt mindestens eine solche Kombination zu finden, dass er auf dieser anomal ist. Umgekehrt wird es mit zunehmenden Dimensionen, auf denen man nach anomalen Ausprägungen suchen kann, schwieriger die relevanten Dimensionen zu finden. Es entsteht effektiv ein Rauschen, da die relevanten Dimensionen gegenüber den nicht relevanten untergehen. [3]

2.3 Anomalieerkennung durch maschinelles Lernen

Ein Anomalieerkennungsverfahren bietet generalisiert die Funktion auf einem Datensatz Anomalien zu erkennen. Dabei eignen sich nicht alle Verfahren für alle Datensätze, sei es weil sie für eine bestimmte Eigenschaft des Datensatzes nicht geeignet sind, oder umgekehrt weil sie zur Leistungsverbesserung bestimmte Eigenschaften im Datensatz voraussetzen.

2.3.1 Überwachtes und unüberwachtes Lernen

Generell lassen sich alle maschinellen Lernverfahren durch maschinelles Lernen in zwei Bereiche teilen:

Unüberwachtes Lernen

Überwachtes Lernen

2.3.2 Robustheit

Die Robustheit eines Algorithmus beschreibt seine Stabilität gegenüber Anomalien im Trainingsdatensatzes und gegenüber ungewollten Unterschieden zwischen dem Trainingsdatensatz und dem Testdatensatz. Weiterhin kann ein Anomalieerkennungsverfahren besonders Robust gegenüber einer Eigenschaft von Datensätzen, wie zum Beispiel Rauschen oder Mehrdimensionalität, sein, die sich allgemein negativ auf die Performance von auf ihrem Datensatz ausgeführten Algorithmen auswirkt.

2.3.3 Streaming Data

Space Time Anpassung des Modells, Live Ergebnisse

2.3.4 Kriterien zur Performancebeurteilung

2.3.5 F-Measure

2.4 Anomalierkennung durch Random Forests

Anhang A

Weitere Informationen

Abbildungsverzeichnis

Algorithmenverzeichnis

Literaturverzeichnis

[1] Aggarwal, Charu C: *Outlier analysis*. In: *Data mining*. Springer, 2015.

[2] Chandola, Varun, Arindam Banerjee und Vipin Kumar: *Anomaly detection: A survey*. ACM computing surveys (CSUR), 41(3):1–58, 2009.

[3] Erfani, Sarah M, Sutharshan Rajasegarar, Shanika Karunasekera und Christopher Leckie: *High-dimensional and large-scale anomaly detection using a linear one-class SVM with deep learning*. Pattern Recognition, 58:121–134, 2016.

[4] Guha, Sudipto, Nina Mishra, Gourav Roy und Okke Schrijvers: *Robust random cut forest based anomaly detection on streams*. In: *International conference on machine learning*, Seiten 2712–2721, 2016.

[5] Tan, Swee Chuan, Kai Ming Ting und Tony Fei Liu: *Fast anomaly detection for streaming data*. In: *Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

Hiermit versichere ich, dass ich die vorliegende Arbeit selbstständig verfasst habe und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet sowie Zitate kenntlich gemacht habe.

Dortmund, den 20. Mai 2020

Muster Mustermann