Realtime Anomaly Detection in Time Series Data

Studiengang: BSc in Informatik | Vertiefung: Mobile Computing
Betreuer: Prof. Dr. Andreas Danuser
Experte: Dr. Federico Flückiger (Eidg. Finanzdepartement EFD)

Digitale Daten sind der Rohstoff des 21. Jahrhunderts und die Basis für viele neue Geschäftsmodelle. Die Menge an diesen Daten ist in den letzten Jahren exponentiell gewachsen. Dieser Trend wird sich nicht zuletzt wegen der Milliarden an vernetzten IoT-Geräten fortsetzen. Um die Qualität dieser grossen Datenmengen hoch zu halten, ist es wichtig, Datenanomalien automatisiert zu identifizieren. Anomalien sind Muster in Daten, die nicht dem erwarteten Verlauf entsprechen.

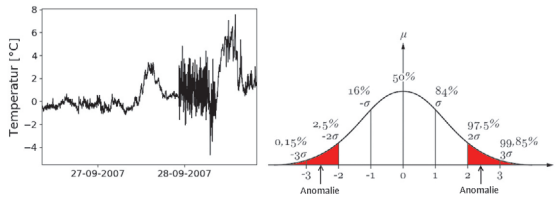
Ausgangslage und Zielsetzung

Für das Erkennen von Anomalien in Zeitreihen gibt es bereits viele bestehende Verfahren und Algorithmen. Die Mehrheit davon wurde für statische Datensets entwickelt, in denen alle Daten gleichzeitig verfügbar sind. Schätzungen zufolge müssen in Zukunft bis zu 30% der Daten in Echtzeit verarbeitet werden. Dies führt dazu, dass auch die Anomalieerkennung vermehrt in Echtzeit durchgeführt werden muss. Zudem liegt der Fokus bestehender Ansätze meist auf der Erkennung von punktuellen Anomalien wie Ausreissern.
Das Ziel dieser Thesis ist die Implementierung dreier verschiedener Verfahren, mit denen sowohl punktuelle als auch kollektive Anomalien in Sensordaten in Echtzeit erkannt werden können.

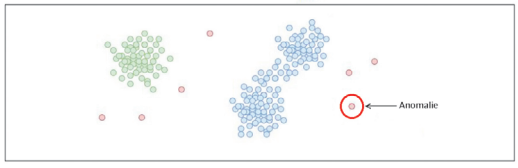
Umsetzung

Beim ersten Verfahren handelt es sich um ein rein **statistisches Verfahren** kombiniert mit dem **Sliding Window** Konzept. Die Grundannahme ist dabei, dass sich normale Dateninstanzen in einem Gebiet mit hoher Wahrscheinlichkeit eines stochastischen Modells befinden, während Anomalien eine geringe Wahrscheinlichkeit haben.
Das zweite Verfahren basiert auf einem **Clustering-Algorithmus** und dem **Sliding Window** Konzept. Die Grundannahme ist dabei, dass normale Daten zu einem Cluster gehören, während Anomalien keinem Cluster angehören.
Das dritte Verfahren basiert auf Vorhersagen und wird mit einem **Künstlichen Neuronalen Netz (KNN)** implementiert. Überschreitet die Differenz zwischen dem vorhergesagten Wert und dem tatsächlichen Wert einen Grenzwert, so wird der Wert als Anomalie klassifiziert. Der vorhergesagte Wert kann zudem als Korrektur des anomalen Wertes verwendet werden.

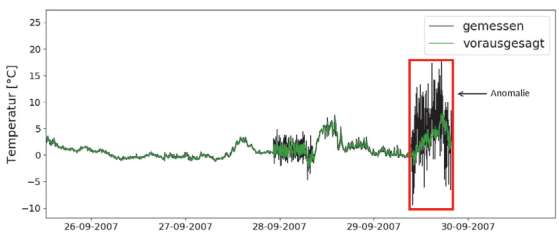
Statistisches Verfahren



Clustering-Verfahren



KNN-Verfahren



Resultate

Die Ergebnisse haben gezeigt, dass nicht zwingend auf einen komplexen Algorithmus oder künstliche Intelligenz zurückgegriffen werden muss, um gute Ergebnisse zu erzielen. Im Gegenteil, gewisse Anomalietypen werden mithilfe des statistischen Verfahrens sogar besser erkannt. Jeder Anomalietyp konnte von mindestens einem der implementierten Verfahren mit Genauigkeiten von über 90% erkannt werden.
Es bleibt zu erwähnen, dass die in dieser Thesis implementierten Verfahren nur einen Ausschnitt der zahlreichen Möglichkeiten darstellen, um Anomalien in Zeitreihen zu erkennen. Auch bieten die von uns implementierten Verfahren keine universelle Lösung für die Erkennung von Anomalien in verschiedensten Kontexten. Die Konfiguration der Toleranzfaktoren und Fenstergrössen muss immer abhängig von den erwarteten Daten vorgenommen werden.

Mathias Rudolf
mathias.rudolf@outlook.com

Elisa Schnabel
elisa.schnabel@gmx.ch