Phillip Wenig

Finding, Clustering, and Classifying Anomalies on Large and Multivariate Time Series

Zusammenfassung

Multivariate Zeitreihen sind eine Form von reellwertigen Sequenzdaten, die gleichzeitig verschiedene zeitabhängige Variablen aufzeichnen. Sie stammen meist aus Multi-Sensor-Konfigurationen und dienen einer Vielzahl wichtiger Analysezwecke, einschließlich der Erkennung von normalem und abnormalem Verhalten. Anomalien treten häufig in einzelnen Kanälen einer Zeitreihe auf, können aber auch in der Korrelation mehrerer Kanäle gefunden werden. Zwar gibt es wirksame Data-Mining-Algorithmen zur Erkennung anomaler und strukturell auffälliger Testaufzeichnungen, doch führen diese Algorithmen keine semantische Kennzeichnung durch. Daher verbringen Datenanalystinnen und -analysten viele Stunden damit, die großen Mengen automatisch extrahierter Beobachtungen mit den ihnen zugrunde liegenden Ursachen in Verbindung zu bringen. Die Komplexität, Menge und Vielfalt der extrahierten Zeitreihen macht diese Aufgabe nicht nur für Menschen, sondern auch für bestehende Algorithmen schwierig: Diese Algorithmen benötigen entweder Trainingsdaten für überwachtes Lernen, können nicht mit unterschiedlichen Zeitreihenlängen umgehen oder leiden unter außergewöhnlich langen Laufzeiten.

Um die Analyse von Anomalien in sehr großen Zeitreihen zu erleichtern, untersuchen wir in dieser Dissertation drei Arten von Algorithmen: Anomalie-Erkennung, Clustering und Klassifizierung. Genauer gesagt, geben wir einen Überblick über das Forschungsfeld der Erkennung von Zeitreihenanomalien und weisen auf Defizite bei den veröffentlichten Benchmarks hin. Anschließend schlagen wir einen neuartigen und skalierbaren Zeitreihen-Anomalie-Detektor vor, der Anomalien in den Korrelationen von Zeitreihenkanälen finden kann und aufzeigt, in welchen Kanälen Anomalien auftreten. Um die Berechnung der Anomalieerkennung zu verteilen, haben wir eine neuartige Bibliothek für den Aufbau reaktiver und verteilter Algorithmen entwickelt. Darüber hinaus schlagen wir ein schnelles und effektives Clustering-Verfahren für Zeitreihen mit unterschiedlichen Längen vor und präsentieren ein System, um extrem unbalancierten Datenpartitionen während des verteilten Trainings von Algorithmen für maschinelles Lernen entgegenzuwirken.

Abstract

Multivariate time series are a form of real-valued sequence data that simultaneously record different time-dependent variables. They originate mostly from multi-sensor setups and serve a variety of important analytical purposes, including the detection of normal and abnormal behavior. Anomalies often occur in individual channels of a time series, but can also be found in the correlation of multiple channels. While effective data mining algorithms exist for the detection of anomalous and structurally conspicuous test recordings, these algorithms do not perform any semantic labelling. So data analysts spend many hours connecting the large amounts of automatically extracted observations to their underlying root causes. The complexity, amount and variety of extracted time series make this task hard not only for humans, but also for existing algorithms: These algorithms either require training data for supervised learning, cannot deal with varying time series lengths, or suffer from exceptionally long runtimes.

To facilitate the analysis of anomalies in very large time series, we investigate three types of algorithms in this dissertation: Anomaly Detection, Clustering, and Classification. More precisely, we create an overview of the time series anomaly detection research field and point out short commings with published benchmarks. Then, we propose a novel and scalable time series anomaly detector that can find anomalies in the correlations of time series channels and reveal in which channels anomalies occur. To distribute the anomaly detection computation, we developed a novel library for building reactive and distributed algorithms. Moreover, we propose a fast and effective clustering technique for time series with varying lengths and introduce a framework for counteracting extremely skewed data partitions during the distributed training of machine learning algorithms.