APPROACHES FOR TIME SERIES ANOMALY DETECTION

LENA LINDNER (MASTERSTUDIUM INFORMATIK)

Betreuer: Prof. Dr. Marcel Tilly, Prof. Dr. Markus Breunig

1 Time Series and the Significance of Anomalies

Time series are encountered in a variety of different applications. One of the best known ones would likely be weather forecasting: predicting tomorrow's weather based on data from previous days, like temperature or wind speed. A time series is defined as an ordered sequence of data points in which the order is based on time. The data points are usually recorded at discrete, equally-spaced intervals and most often they depend on each other in some way. Like when tracking the position of a car, the car's current position will depend on the one it had previously and the one before that and so on.

The dynamic nature of time series brings with it a number of particularities to look out for. For one, data points have an implicit order and should be observed in the context of their neighbouring data points. Algorithms also need to encapsulate special patterns in the data like seasonality and trends.

Weather forecasting falls into the category of time series prediction – where an algorithm continues the time series into the future and raises expectations for possible trends and up-coming events. Time series anomaly detection, on the other hand, shifts its focus fully on detecting abnormal behaviour in the time series data. Perhaps there is a sudden spike in temperature or an unexpected drop in air humidity? Some anomalies can be explained as simple measuring errors, but others may have significance behind them and analysing them mayprovide useful insights about the underlying system.

2 Detecting Correlation Anomalies

Correlation anomalies are a special case of anomalies which can appear in time series of two or more dimensions. In correlation anomalies, multiple time series may seem perfectly regular when observed independently, but in combination they present an abnormal event. Sticking with the weather analogy: Say the weather forecast for tomorrow were to predict heavy snowfall but also a temperature over thirty degrees Celsius; one would certainly wonder how such a combination is possible.

Figure 1a depicts two time series, which replicate each other almost perfectly. Figure 1b, in comparison, depicts the same two signals but the blue one shows anomalous behaviour: it starts off as regular until it suddenly falls and stays close to zero for the rest of the time span while its orange counterpart falls off much later.

Certain machine learning algorithms are able to detect the kind of time series anomalies portrait in Figure 1b. One of the more popular choices is a deep learning model called an Autoencoder. It consists of an Encoder and a Decoder part: The Encoder takes the input time series and compresses it to a very small representation of itself, retaining only the most essential features. The Decoder then takes this compressed representation and tries to reconstruct the input data from it with the goal to make it as similar to the original time series as possible. The better the Encoder becomes at capturing the important characteristics of the data, the better the Decoder will be at truthfully reconstructing the data.

When used for anomaly detection, Autoencoders are trained on regular samples only and evaluated on regular and abnormal samples. It is assumed that the Autoencoder will have trouble reconstructing abnormal samples, since they are very different from the samples it has seen during training. If the reconstruction loss of the Autoencoder is above a threshold, samples are predicted to be anomalies.

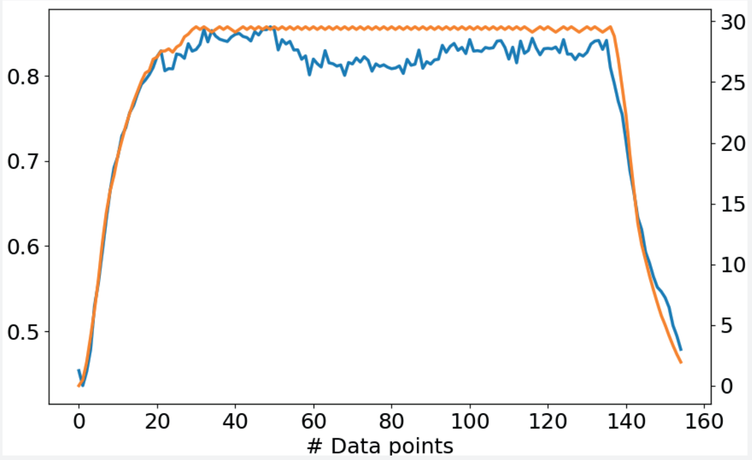
Autoencoders, naturally, are not the only machine learning models which can successfully be applied to detect time series anomalies. Other examples include Support Vector Machines (with a specialized case being One-Class Support Vector Machines), Local Outlier Factor, Isolation Forest (a variation of Random Forests), Convolutional Neural Networks, Long Short Term Memory Networks, or even large Transformer models.

3 Conclusion

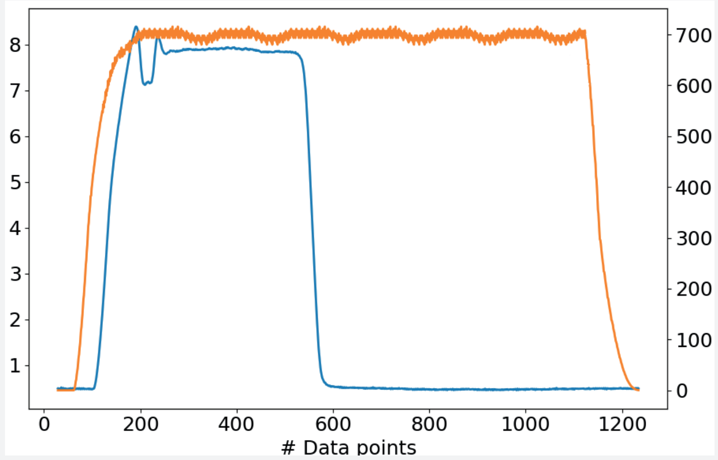
Since the original master thesis is protected by a confidentiality clause, very little can be said about its background or the origin of the two time series shown here.

The aim of the thesis was to compare different machine learning approaches and find which worked best on the data at hand. A large section of the thesis is dedicated to preprocessing the time series data and preparing the two time series to be used in model training. In the end, the results of the evaluation looked to be quite promising, however there is still a lot of work left to do until the findings of this thesis can be applied in practice.

Fig. 1: Comparison between regular and anomalous time series



(a) Regular



(b) Anomaly