Talagala_notes

September 15, 2018

Thank you for inviting me to share my thoughts on this publication. I will attempt to talk through the paper and identify the key points, making use of the figures to illustrate these as much as possible but if you have specific questions please feel free to chime in at any point.

Ok so lets begin. This article is all about studying efficient and robust methods for outlier detection (or anomaly detection) in time series data

1 Background

1.1 Nomenclature and types of time series anomalies

- **Time series data**: anything where a quantity is measured repeatedly at a series of different times. A collection of such data is known as a time-series. These are very useful for and have applications in a huge range of fields, weather forecasting, software development, monitoring hospital patients, jet engines and manufacturing processes. In astrophysics these are known as light curves and I use them to measure the masses of black holes in distant galaxies. Also can be used to hunt for planets around other stars and look for signs of extra-terrestrial life.
- **Anomaly:** Anything that doesnt fit with the rest of the time series. A more mathematical definition is any point in the time-series that is very unlikely assuming some background distribution (Figure 2a for example).

This paper consider three types of outliers (Figure 2).

1. **Contextual Anomalies within a given time-series:** A few rogue points amongst many examples of ordinary behavior
2. **Anomalous sub-sequences within a given time series:** Have a period of anomalous behavior rather than just a few points.
3. **Anomalous series within a space of collection of series:** Rather than rogue points within a time series, have an entire anomalous time-series relative to normally behaved time-series.

Having identified three types of outliers, the paper now discusses two sources of time series data and challenges of dealing with each.

1. **Batch processing:** Here the entire time series is available and outliers can be identified after the observations.

2. **Data Streams:** Here the time series is continuously evolving and the outlier detection algorithm must be able to adaapt to new 'typical' behaviour as it scans for outliers. Problems include 'concept drift' where the background model might eveolve with time (non stationarity).

Section 2.3: Extreme value theory for anomaly detection

This paper proposes a new method for anomaly detection in data streams (where observations are continously evolving with new data). The model requires some understandin of extreme value theory. A key concept here is that of the Fisher-Tippett theorem.

Fisher-Tippett theorem: If you take M subsamples from any parent distribution, calculate the maximum value in the subsample and repeat. The distribution of maximum values will tend toward one of three distributions

1. **Frechet Distribution:** $\frac{\alpha}{s}\left(\frac{x-m}{s}\right)^{-1-\alpha} e^{-\left(\frac{x-m}{s}\right)^{-\alpha}}$
2. **Weibull Distribution:** $\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}$
3. **Gumbell Distribution:** $\frac{1}{\beta}e^{-\left(\frac{x-\mu}{\beta}+e^{-\frac{x-\mu}{\beta}}\right)}$

build a new disribution out of these maximum values