Adaptive Thresholds and Scan Statistics

Nathen byford

Contents

Adaptive Thresholds

Motivation

- Communications networks have many components that report statistics about their health on a regular basis, daily, hourly, minutely, or even more frequent.
- How do we measure extreme counts when that data is being collected every second?
- When setting these manually by treating each time period and region separately for each metric:
  - Leads to far too many thresholds to be set by hand, and
  - Far too many false alarms to be investigated.

> "The goal is to detect events with each incoming count, without looking at the past raw counts or being distracted by strong cyclical patterns, trends high background variability, or stretches of missing data."

Previous methods

- Batch the data into intervals that are sufficiently short in so counts can be considered identically distributed.
  - Problem is that these intervals make it difficult to detect in real time because all analysis is done after the interval.
- Others have assumed that daily patterns repeat so that homogeneity holds not just within short intervals, but also across days. Specifically counts at the same times (Say 9:00am to 9:05am).

New Method Requirements

- For a model that is monitoring extremes, the model must have accurate tails. (Negative-binomial distribution)
- Smooth continuously and and tracks both cyclical and long-term trends.
- Be able to capture both extreme value spikes and persistent low-level degradation.[1]
- Threshold on a severity metric S_t. The severity metric could capture both the magnitude and diration of an event.

1. Formulate the problem as statistical process control

Adaptive count thresholding

With the problem in process control terms, we can standardize each count and then use an exponentially weighted moving average (EWMA) process to get an adaptive threshold.

Adaptive count thresholding consists of four basic steps:

1. Interpolate the stored grid values to obtain the estimated parameters for the reference negative binomial distribution F_t in effect at time t
2. Score X_t by computing its p value, p_t, under its reference distribution F_t and its normal score $Z_t = \Phi^{-1}(p_t)$.
3. Threshold the updated severity metric, $S_t = (1 - w) \times S_{t-1} + wZ_t$, against a constant threshold
4. Update stored grid values with a count X_t, or with a random draw from F_t if X_t is missing, or with a random draw from the tail of F_t if X_t is an outline.

Adaptive count thresholding pros

- Each step is quick to compute
- The p values provide a natural way to monitor the performance
- The negative binomial distribution will provide a roughly uniform distribution on the p values

The procedure

The following procedure is proposed to calculate thresholds for a time stamp t:

1. *Index:* obtain time stamp t from the minute, hour, and day information
2. *Compute reference parameters:*
3. *Validate:*
4. *Threshold:*
5. *Outliers and missing data:*
6. *Update reference distribution:*
7. *Update grid values:*

Simulation and results

Discussed further research

Scan Statistcs

Motivation

Adversaries will sometimes get into a computer network and then need to move through the network of computers to reach their intended target. We can identify anomalies in communications between two computers that do not often communicate to find these adversaries.

Method

- Use scan statistics on the network of computer systems and communications to identify anomalies.

Scan statistics

Results

Discussed further research

Questions or Ideas?

References

Lambert, Diane, and Chuanhai Liu. 2006. “Adaptive Thresholds: Monitoring Streams of Network Counts.” *Journal of the American Statistical Association* 101 (473): 78–88. https://doi.org/10.1198/016214505000000943.

Neil, Joshua, Curtis Hash, Alexander Brugh, Mike Fisk, and Curtis B. Storlie. 2013. “Scan Statistics for the Online Detection of Locally Anomalous Subgraphs.” *Technometrics* 55 (4): 403–14. https://doi.org/10.1080/00401706.2013.822830.