Anomaly Detection

Kassandra Munguia, Daniel Ratner, Danielle Sanzone, Nora Norwell

[1]Intern

[2]Linac Coherent Light Source, SLAC National Accelerator Laboratory, 2575 Sand Hill Road, Menlo Park, CA 94025, USA.

[+]Contact: dratner@slac.stanford.edu

Introduction

Our group's main goals were to detect anomalies for data analysis and potentially create a real-time alert system to notify or any potential problems.. We are trying to get a better look at anomaly detection at SLAC.

Definitions:
Anomaly- a point(s) that deviate from the standard
Breakout- a change from one steady standard to another standard

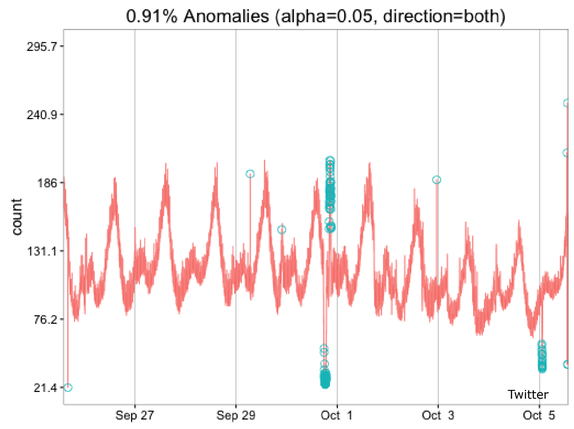


Fig 1. Example of anomalies

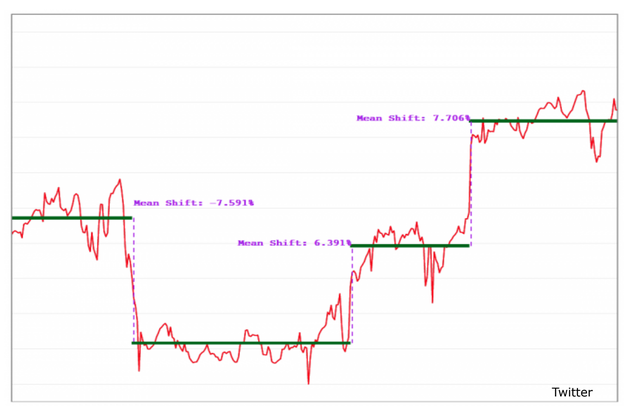
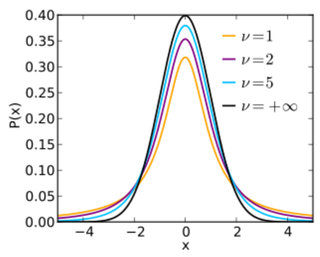


Fig 2. Example of a breakout

Research

In order to understand the Twitter algorithm, we examined the Student's T-Distribution, Extreme Studentized Deviate (ESD), and the Generalized ESD.



Student's T-distribution

A t distribution uses the sample standard deviation instead of the population's standard deviation. This equation is the equation of the colored lines you see in the graph. The black line is the normal distribution. As you increase the degrees of freedom, your t distribution will look more and more like normal distribution.

$$G = \frac{|Y_i - \bar{Y}|}{s}$$

$$G > \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

Extreme Studentized Deviate(ESD)

ESD is used to find outliers in univariate data assuming that the data is normally distributed. It detects one outlier at a time. Once an outlier is detected it is removed from the data. You compare the G value to the critical value and if the above conditional is true, then the point is an outlier.

$$\lambda_i = \frac{(n-i)t_{p,n-i-1}}{\sqrt{(n-i+1+t^2_{p,n-i-1})(n-i+1)}}$$

Generalized ESD

It is a modified ESD calculation. You find G and find the critical value and compare it to each other; however, the critical value and the mean updates as the data set changes each iteration.

Seasonal Hybrid ESD

Twitter modified the generalized ESD to account for seasonality. It uses time series decomposition and robust statistical metrics. Two methods of finding anomalies:

- AnomalyTS- takes in a 2 column data frame: time series + count
- AnomalyVec- takes in a single data frame of counts

Function:

AnomalyDetectionVec = function(x, max_anoms=0.10, direction='pos', alpha=0.05, period=NULL, only_last=F, threshold='None', e_value=F, longterm_period=NULL, plot=F, y_log=F, xlabel='', ylabel='count', title=NULL, verbose=FALSE)

The algorithm produces a list of the timestamp/index where the anomaly is present and the numeric value of the anomaly. Along with it, there is a visual that graphs the data in context.

LCLS data:

The parameter that I changed to produce the following graphs is max_ anoms . By increasing the max_anoms, which indicates the max number of anomalies that S-H-ESD will detect as a percentage of the data, the algorithm was able to catch more anomalies and highlight them in the plot.

25-8 Beam volt

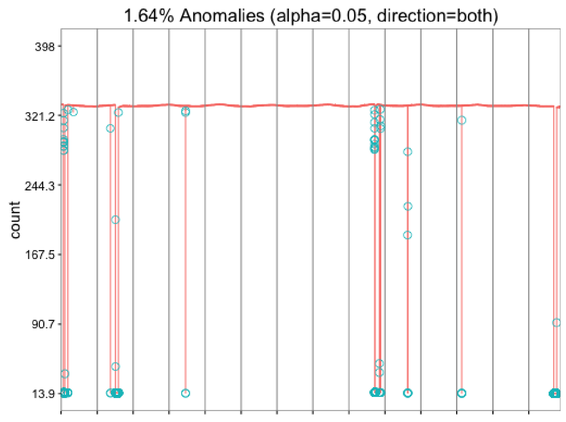


Fig 3. Anomalies on plot is for the modulator high voltage for the of one of the RF stations

It found 1644 anomalies. This detection is useful because if implemented for all RF stations, see which one trips more often in order to see which ones need more work.

Dump Vacuum Gage 365 pressure

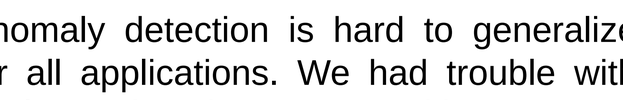


Fig 4. Anomalies on plot is for the modulator high voltage for the of one of the RF stations

It found 30 anomalies. Things that can cause these anomalies are pressure increasing if air gets in the pipe, valves are opening or closing, an RF changes power supply, or the beam hits the wall.

Conclusion

Anomaly detection is hard to generalize for all applications. We had trouble with implementing timestamps with our count and using the timestamp decomposition method. However, these results are promising because it is the basis for a real time notification system and for a key data analysis tool at SLAC.

Acknowledgments

Use of the Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515.

Date: 08/11/2016