Detecting the Unexpected: An Introduction to Anomaly Detection Methods

Kiri Wagstaff
Jet Propulsion Laboratory, California Institute of Technology
May 20, 2019
KISS Technosignatures Workshop

Image credit: NRAO/AUI/NSF

© 2019, California Institute of Technology. Government sponsorship acknowledged. This work was performed at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with NASA.
What is an Anomaly?
What is an Anomaly?
Technosignatures – What are we looking for?

• “Any sign of technology that is not also a biosignature”
  – SETI Nomenclature
  • “... that modifies its environment in ways that are detectable ...”
    [Tarter, 2007]

• Challenge: human technosignatures pollute the data

Example: Fast Radio Bursts

Parkes Telescope
[Lorimer et al., 2007]

VLBA “V-chirp” found by V-FASTR [2012]
Anomaly Types

- Both major anomaly types are valuable
- Each leads to different follow-up actions

- Error
  - Filter data
  - Improve pipeline

- Scientific interest
  - Announce discovery
  - Follow-up observations
  - Revise textbooks
Anomaly Types

• Both major anomaly types are valuable
• Each leads to different follow-up actions

Error

- Imaging artifact
- Modeling error
- Data gap
- Blended source

Scientific interest

- Filter data
- Improve pipeline
- Announce discovery
- Follow-up observations
- Revise textbooks
Anomaly Types

• Both major anomaly types are valuable
• Each leads to different follow-up actions

Error
- Imaging artifact
- Modeling error
- Data gap
- Blended source

Scientific interest
- Shape
- Timing
- Color
- Pulsar
- Quasar
- Green pea galaxy

- Filter data
- Improve pipeline
- Announce discovery
- Follow-up observations
- Revise textbooks
Automated Methods for Finding Anomalies

1. Easily distinguished events
2. Events that fall outside “normality”
3. Difficult-to-model events
(1) Easily distinguished events

• Statistics/probability
  • Assume Gaussian (or other) generating distribution
  • Outlier: >3 sigma from mean
  • Hypothesis tests: Student’s t-test, $\chi^2$-test

• Distance/density
  • Examples
    • K Nearest Neighbor
    • Local Outlier Factor (LOF): ratio of local density to neighbors’ density
  • Clustering
    • Methods often do not scale well to high dimensional data or large data sets
      (although some efficiency improvements exist)

Figures from Zhao et al. (2019)
SDSS galaxy outliers: Local Outlier Factor (LOF)

Elongated galaxies, multiple objects and colors
(1) Easily distinguished events

- Isolation Forest [Liu et al., 2008]
  - Each decision node picks a random feature and a random value to split data

```
f_i < v
  f_i < v
  f_i < v
  ...
  ...
  ...
  ...
```

...
(1) Easily distinguished events

- **Isolation Forest**
  - Score: ease with which a random forest isolates the event

Figure from Zhao et al. (2019)

Figure from scikit-learn documentation
SDSS galaxy outliers: Isolation Forest

Bright, large, diffuse galaxies with single primary color
(2) Events that fall outside “normality”

- Build model of “normality” and score items by their distance to model
  - One-class SVM [Schölkopf, 2001]: learn a tight boundary around all training data and flag items that fall outside

Figure from Zhao et al. (2019)
SDSS galaxy outliers: One-class SVM

Smaller, variable shape galaxies
(3) Difficult-to-model events

• Build a model of “normality” and score events by content that can’t be modeled (reconstructed)
  • Principal Component Analysis (PCA) or Singular Value Decomposition (SVD)
  • Autoencoder or replicator network
  • Self-organizing map (SOM)

Figure from Valkov (2017)
SDSS galaxy outliers: SVD

Heterogeneity of colors, sizes, and number of objects
Explanations for Anomalies

• Final explanation comes from human interpretation
  • “The search for unexpected behaviour in data is presently not completely automated, since interpreting such signals requires human judgement.” [Wheeler and Kipping, 2019]

• Understanding the anomaly
  1. Why was a given item chosen as an anomaly? (algorithm explanation)
  2. How does it compare to other items in the data set? (contextual explanation)
  • This is an active area of research in the machine learning community

• Gathering more information
  1. What does it look like in other views? (complementary observations)
  2. What does it look like if we observe it again? (follow-up observations)
Reconstruction-based Explanations

- Dark Energy Survey (DES) galaxies (n = 11.9M)
  - Observe with $r$, $i$, $g$, $z$ bands

With Eric Huff and Umaa Rebbapragada
Reconstruction-based Explanations

• Green Bank Telescope (GBT) observations of Kepler field
  • Capture radio pulse detections with SNR > 10 (n = 21,430)
  • Frequency range: 1420 +/- 30 MHz (hydrogen 21-cm line)
Summary

1. Definition of “anomaly” or “outlier” is context-dependent
   • Context = prior knowledge and distribution of data collected

2. Anomaly types
   • Data or measurement artifacts
   • Scientifically interesting – potential new discoveries

3. Existing methods look for:
   • Easily distinguished events
   • Events outside of “normality”
   • Difficult-to-model events

4. After finding anomalies, don’t stop there
   • Methods that provide explanations help guide interpretation and understanding of anomalies
Useful References


• Software: **PyOD: Python Outlier Detection**
  - [https://github.com/yzhao062/pyod](https://github.com/yzhao062/pyod)