Unsupervised Machine Learning Models to Predict Anomalous Data Quality Periods

JOSEPH C HARDIN\textsuperscript{1}, NITIN BHARADWAJ\textsuperscript{1}, MAHANTESH HALAPPANAVAR\textsuperscript{1}, ADAM THEISEN

\textsuperscript{1} Pacific Northwest National Laboratory \textsuperscript{2} University of Oklahoma

ARM/ASR Science Meeting, 2018
Problem Statement

- ARM produces a large amount of data (>1PB).
  - More than can be looked at by hand
- ARM data quality is a key priority
- Machine learning is a promising approach to tackle the problem
- Supervised machine learning has challenges with training data for detecting instrument malfunctions.
- Unsupervised learning potentially sidesteps this problem.
  - Exploit statistical relations between parameters in the data.
- This talk will discuss our recently proposed approach to address data quality using machine learning.
Machine Learning

- Machine learning:
  - solve problems by analyzing data without explicitly programming in solutions – often referred to as learning from the data

- Broadly split into 2 categories (Supervised and Unsupervised):
  - Supervised learning fits a model to relate input data, to labeled output data
    - Given y, x, fit y=f(x)
    - This requires creating a labeled training set relating the input and the outputs.
    - This can be very expensive and time consuming.

- Unsupervised learning
  - Fit y=f(x) given only x.
Unsupervised Machine Learning

- We plan to utilize a variation on unsupervised clustering.
- Break data up into $N$ statistically different groups
  - Not predefined, but data driven
- Clusters represent statistical modes of operational returns.
- Use in cluster fits to detect anomalies.

- One of the largest challenges in unsupervised clustering:
  - You can’t force certain clusters.
  - You can always find $N$ clusters. Doesn’t mean they are statistically independent.
AMF2 MAGIC KAZR Toy Example

![AMF2 MAGIC KAZR Toy Example](image)

AMF2 MAGIC KAZR 23-SEP-2013 12:00 UTC

**Height (km)**

0 1 2 3 4 5

**Time (hrs)**

0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1

**dBz**

-60 -55 -50 -45 -40 -35 -30 -25 -20 -15 -10 -5 0 5 10 15 20 25 30 35 40 45 50 55 60 65 70 75

- Cloud

- No Cloud
Figure 5: Classification Surface as a function of three input variables.
Proposed Method

- Unsupervised clustering to detect statistically independent clusters.
  - “typical operating regimes”
- Data Clustering for initial pointwise classification
  - Clustering on a graph/b-matching
- Region based aggregation
  - Convert point estimates into time periods.
- Human-in-loop review to tweak hyperparameters and verify.
- Envisioned as a way to make data quality review more effective – focus on likely problematic times.
- Test set will use the Oliktok KAZR radar
Timeline

- Interviews for the position have concluded.
- **September 2018**: Preliminary implementation completed.
- **December 2018**: Evaluation of performance, and DQ table completed for testing on OLI KAZR. ADI integration if requested.
- **May 2019**: Work with ARM staff to transition code to infrastructure. Preparation of technical report.
Questions?
Deliverables

- The source code required to run the analysis set up on ARM’s Stratus system.
- Results of running model on a period of Oliktok KAZR data. This will be in the form of an evaluation dataset released to the ARM ADC.
- A technical report describing and assessing the implemented algorithm.