Anomaly Detection and Pattern discovery for Predictive Maintenance

**Paul Boniol**
Paris Descartes University

**Edouard Mehlman**
Paris Descartes University

**Themis Palpanas**
Institut Universitaire de France
Paris Descartes University

**Niklas Boers**
Paris Descartes University

PGMO Days, November 2017
Motivation

- data series from sensors monitoring operation health of various equipment
- most sensed values are normal
- wish to identify anomalies in observed values and trends
  - these can then be used to predict abnormal behavior
  - perform predictive maintenance
Problem

• develop anomaly detection techniques based on sequences (data series), not on individual values
  ○ individual values can be normal, but their sequence can be abnormal!
Problem

● develop anomaly detection techniques based on sequences (data series), not on individual values

○ individual values can be normal, but their sequence can be abnormal!

150 points in a sequence $S$

values are not outside critical thresholds

values are normal
Problem

● develop anomaly detection techniques based on sequences (data series), not on individual values
  ○ individual values can be normal, but their sequence can be abnormal!
Context : What data?

Target Sensor

Parameter sensors (PRE, PUI, TEMP, VIT, etc)
First approach: detection of predefined anomaly types
Specific type anomaly detection techniques

● Define a score function for each anomaly type specified by the domain expert
  ○ **Trend** (up and down): gradual value increase/decrease
  ○ **Step**: sudden value increase/decrease
  ○ **Spike**: sudden value peak in short period of time
  ○ **Oscillation**: high frequency and amplitude value oscillation in short period of time

● Then go through the dataset to identify these anomalies
Trend Anomaly

Slope of the linear regression on a specific window length
Step Anomaly

\[ \forall i \in [0, \text{length}(s) - 1] \quad \text{score}[t_i] = |\text{median}(s_{j \in [i+1, i+w]}) - \text{median}(s_{j \in [i-w, i-1]})| \]
Spike Anomaly

\[ \forall i \in [0, \text{length}(s) - 1] \quad \text{score}[t_i] = |s_{i+1} - s_i| \]
Oscillation Anomaly

\[ \forall i \in [0, \text{length}(s) - 1] \quad \text{score}[t_i] = \frac{\sum_{j=i-w}^{i+w} \text{var}(s_{j-v:j+v})}{2 \times w} \]
Anomalies of varying durations

We compute each score for different window lengths:

- 1 day, 7 days, 30 days, 180 days for the trends
- 6h, 1 day, 7 days for the steps
- 10 minutes for the spike
- 3 hours for the oscillation
Threshold $T$

Most important anomalies
Shortcomings of previous method

- Hard-coded formulas: do not generalize to other applications
  - Need to define score function and threshold for each anomaly type
- Need to go through all the data for each anomaly, and for each length
- Supervised operation
Second approach: detection of abnormal sequences
Core idea

- define normal behavior
- identify subsequences that look different from normal
- propose scalable algorithm
  - based on Matrix Profile technique
What is matrix profile?

The matrix profile is a meta-data structure that gives information about:

- Discords
- Recurrent patterns
What is matrix profile?

The matrix profile is a meta-data structure that gives information about:

- Discords
- Recurrent patterns
What is matrix profile?

The matrix profile is a meta-data structure that gives information about:

- Discords
- Recurrent patterns
Our anomaly detection approach

- we extract subsequences of normal behavior, called Ground Truth (GT)
  - can be extracted automatically and/or with the help of domain experts

- we use the matrix profile to efficiently identify:
  - subsequences that are not similar to GT
  - and that may repeat (approximately the same)
Ground Truth (GT) example
$\text{GT} \otimes_{m} \text{mA}$

GT (length $m$)

series $S$ (length $M$)

sequences in $S$ that are very different from GT
$S \otimes_m S$

series $S$ (length $M$)

sequences in $S$ that repeat
anomalous subsequences
anomalous subsequences
anomalous subsequences
Clustering of anomalies

subsequences grouped according to their shape
Summary

- benefits:
  - unsupervised method
  - fast execution time
  - identification of anomalies of all types
Summary

- benefits:
  - unsupervised method
  - fast execution time (10min/length)
  - identification of anomalies of all types
  - detection of non-defined anomalies!
  - and more!
Ongoing work
anomaly prediction
Predicting using sequence of events

Goal: *Predict the type of anomaly that will occur next*

Anomaly feature:
- target sensor: cluster 1

Anomaly of type 3
Predicting using sequence of events

Goal: *Predict the type of anomaly that will occur next*
Predicting using sequence of events

Goal: *Predict the type of anomaly that will occur next*

Anomaly feature:
- target sensor: cluster 1

Anomaly feature:
- target sensor: cluster 2

Anomaly feature:
- target sensor: cluster 3
Goal: Predict the type of anomaly that will occur next
Analyze prefix and suffix of patterns

Pattern of length 2500 point (~ 2 weeks)

Pattern of length 11000 point (~ 2 months)
Analyze prefix and suffix of patterns
Analyze prefix and suffix of patterns

Suffix / consequence

Prefix 1

Prefix 2
Conclusions
Conclusions

● operational monitoring becomes increasingly important
  ○ leads to very large collections of sensor data series

● these data have to be analyzed in their full detail
  ○ analyze raw data, not summaries
  ○ analyze sequences, not individual values
Conclusions

- operational monitoring becomes increasingly important
  - leads to very large collections of sensor data series

- these data have to be analyzed in their full detail
  - analyze raw data, not summaries
  - analyze sequences, not individual values

- proposed solutions for data series anomaly detection
  - easy to use: unsupervised
  - general: detect anomalies of different types
  - effective: produce results interesting to experts (not identified before)
  - efficient: fast execution time

- these results can be used for predictive maintenance
Thank you!
Questions?