Log Anomaly Detection

Aamir Khan, Conrad Yeung

Steps

1. Log Collection
2. Log Parsing
3. Feature Extraction
4. Anomaly Detection

Log Collection

What logs we will be using:

- Zookeeper **(Zoo)**
- Hadoop Distributed File System **(HDFS)**

https://github.com/logpai/loghub

TABLE II: Parsing Accuracy of Log Parsing Methods (Raw/Preprocessed)

	BGL	HPC	HDFS	Zookeeer	Proxifier
SLCT	**0.61/0.94**	0.81/0.86	0.86/0.93	0.92/0.92	0.89/-
IPLoM	0.99/0.99	0.64/0.64	0.99/1.00	0.94/0.90	0.90/-
LKE	0.67/0.70	0.17/0.17	**0.57/0.96**	0.78/0.82	0.81/-
LogSig	**0.26/0.98**	0.77/0.87	0.91/0.93	0.96/0.99	0.84/-

An Evaluation Study on Log Parsing and Its Use in Log Mining, He et al.
https://pinjiahe.github.io/papers/DSN16.pdf

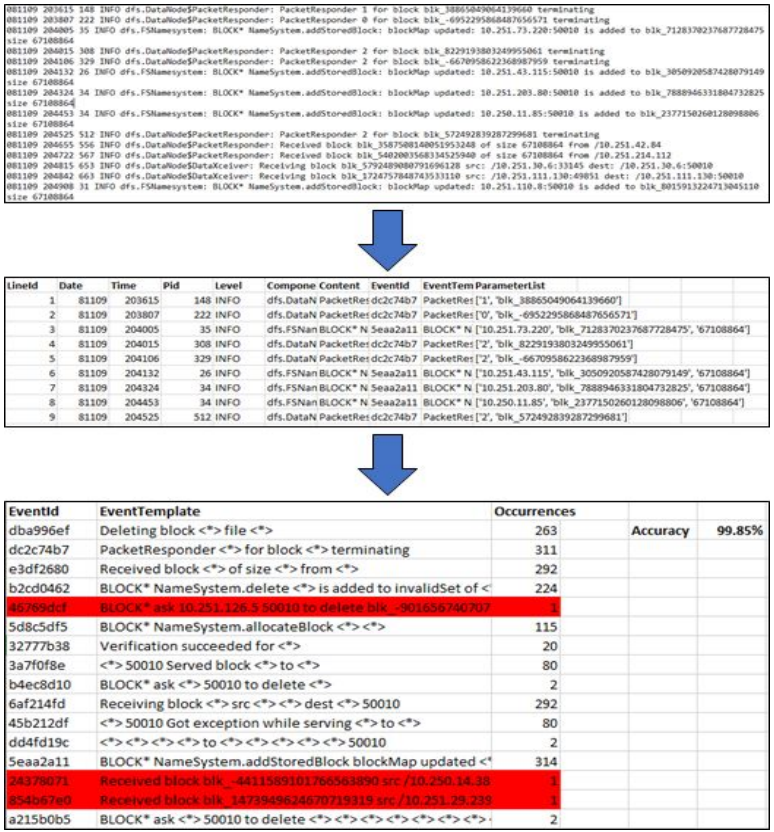
Log Parsing

What parsing method will we use:

- Iterative Partitioning Log Mining **(IPLoM)**

Reason:

- Deterministic
- High parsing accuracy
- No need for pre-processing
- One optional parameter



https://github.com/logpai/logparser

Feature Extraction

- Fixed Window: **Zoo & HDFS**
- Sliding Window: **Zoo & HDFS**
- Session Window: **HDFS** (blk_id)

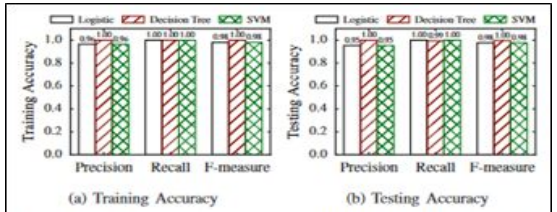


Figure 5: Accuracy of supervised methods on HDFS data with session windows

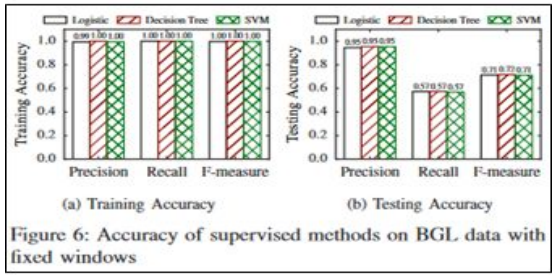


Figure 6: Accuracy of supervised methods on BGL data with fixed windows

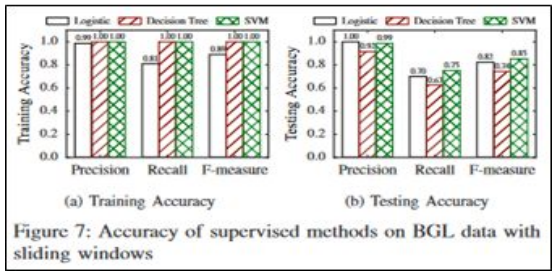


Figure 7: Accuracy of supervised methods on BGL data with sliding windows

Experience Report: System Log Analysis for Anomaly Detection, He et al.
https://netman.aiops.org/

Model Selection

- Several possible directions
- Supervised learning
  - Use “anomaly_label.csv” to add labels column
  - Possible models to explore: Logistic regression or Random forest
- Unsupervised learning
  - Use the log data directly
  - Possibly use PCA to find anomalous patterns.
  - K-means clustering to gather alike events.

Model Selection

- Use *normal distribution* to detect anomaly
- The more an event occurs away from the distribution curve, more chances that it is an anomaly
- We can add level of certainty to our decisions.

References

- https://netman.aiops.org/~peidan/ANM2018Fall/6.LogAnomalyDetection/LectureCoverage/2016ISSRE_System%20Log%20Analysis%20for%20Anomaly%20Detection.pdf
- https://pinjiahe.github.io/papers/DSN16.pdf
- https://www.kaggle.com/shelars1985/anomaly-detection-using-gaussian-distribution