Power Grid Anomaly Detection

CMPT 318 - Cyber Security

Sajandeep Toor, Steven Xia

Overview

- Threat of cyber attacks are increasing
- Critical infrastructure such as power grids are under attack
- Use Hidden Markov Models to detect anomalies in power grid stream data

Introduction

- Advanced Persistent Threats (APTs)
- Supervisory Control and Data Acquisition (SCADA)
- Hidden Markov Models

Threats to Cybersecurity

- Number of threats growing rapidly
- Threats themselves are evolving
- Rise of IoT devices
- Advanced Persistent Threats

Anomaly Detection

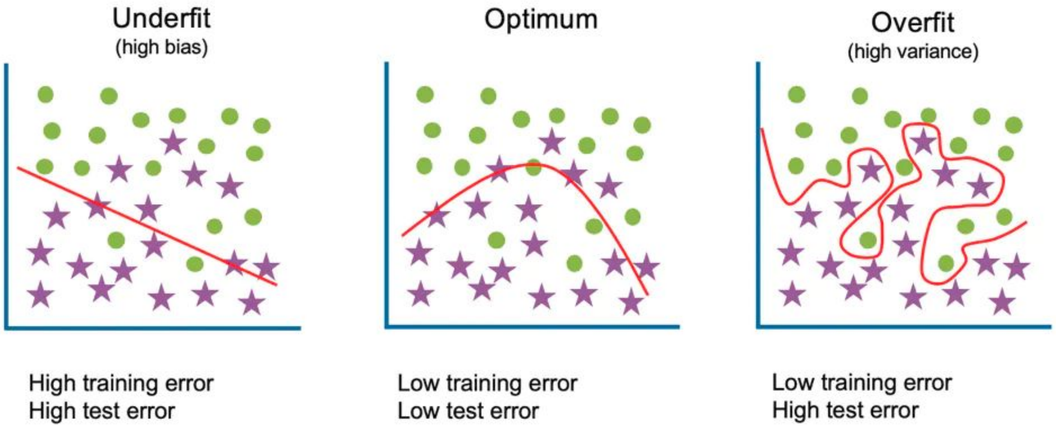
- What is an anomaly?
- Types of anomalies
  - Point Anomalies
  - Contextual Anomalies
  - Collective Anomalies

Use of Hidden Markov Models

- Modeling normal behavior
- Detecting non-normal behaviour

Fitting of a Model

- Overfitting
- Underfitting



IBM Cloud Education. (2021, March 23). *What is underfitting?* IBM. Retrieved December 4, 2021, from https://www.ibm.com/cloud/learn/underfitting.

Methodology

Data Analysis

- Power grid time series data December 16th, 2006 to December 1th 2009
- Features:
  - Global Active Power
  - Global Reactive Power
  - Global Intensity
  - Voltage
  - Sub metering 1 - 3

Data Cleaning and Scaling

- Problem: Data had a lot of missing values
- Solution: Processing and cleaning the data
  - Linear Interpolation
- Scaling
  - Scaled using standardization

Feature Selection

Feature Selection

- Too few features is going to underfitting
- Too many features will cause overfitting
- Choice: Global_active_power and Global_intensity

Feature Selection

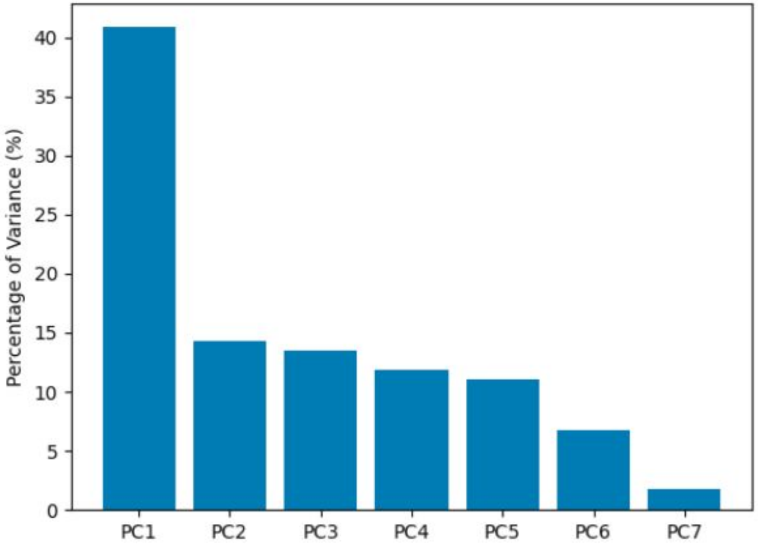


Figure 1: Principal Component Analysis Variance of Project Data

Feature Selection

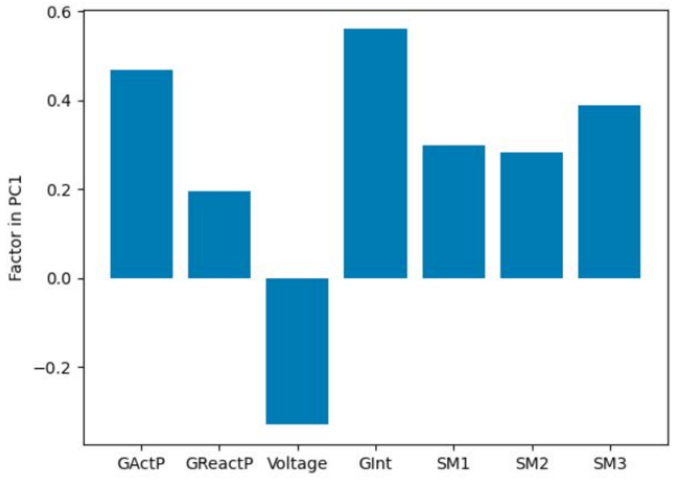


Figure 2: Variance of each feature in PC1

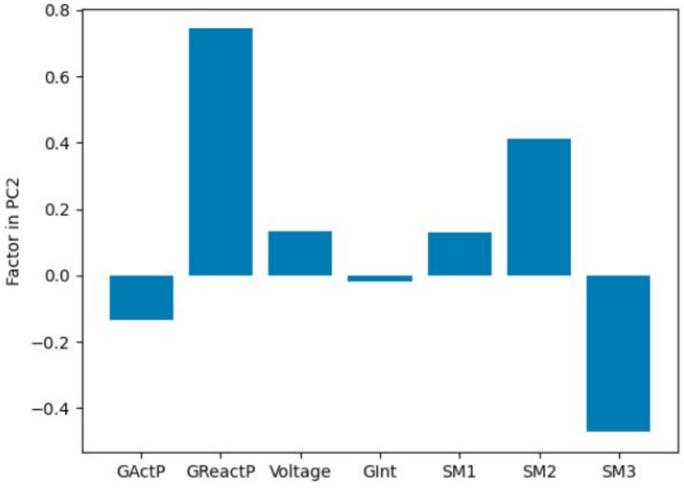


Figure 3: Variance of each feature in PC2

Feature Selection

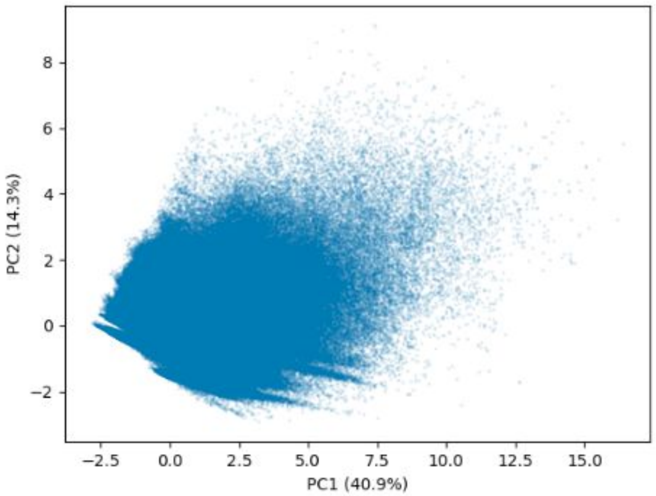


Figure 4: PCA Graph: PC1 vs PC2

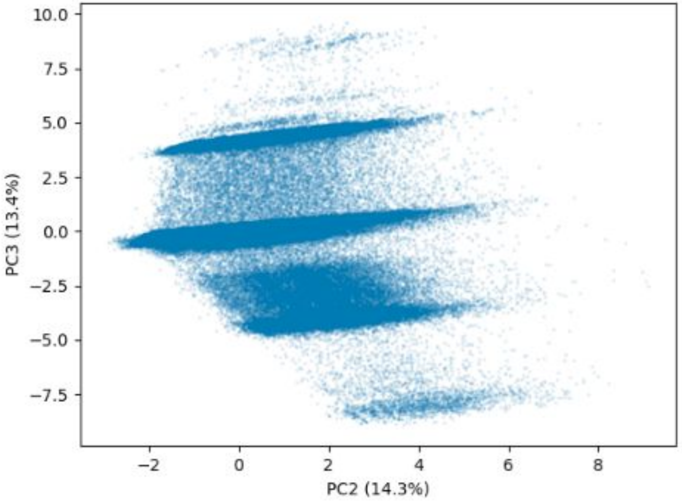


Figure 5: PCA Graph: PC2 vs PC3

Training Models

Training Models

- Use a third of the data for testing
- Train 21 models to determine the best number of states

Training Models

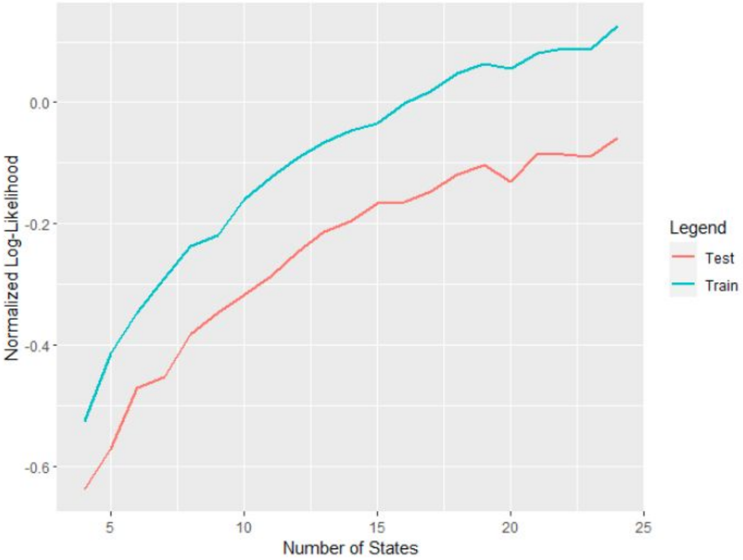


Figure 6: Log Likelihood of Training and Testing Data

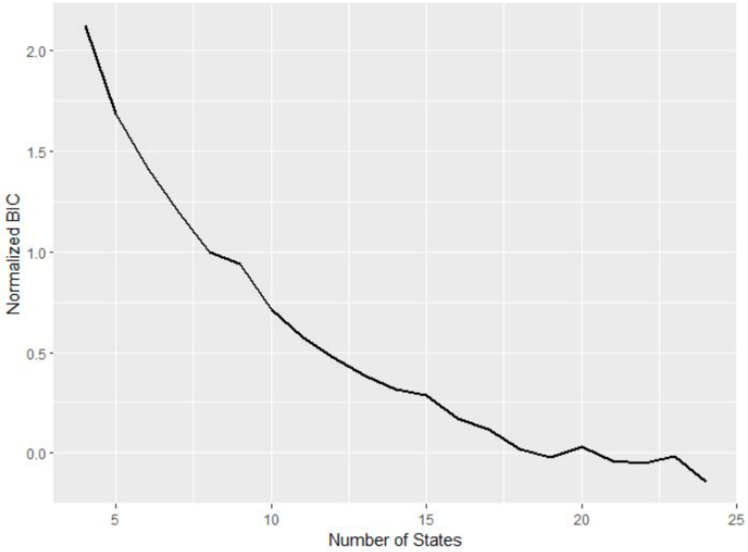


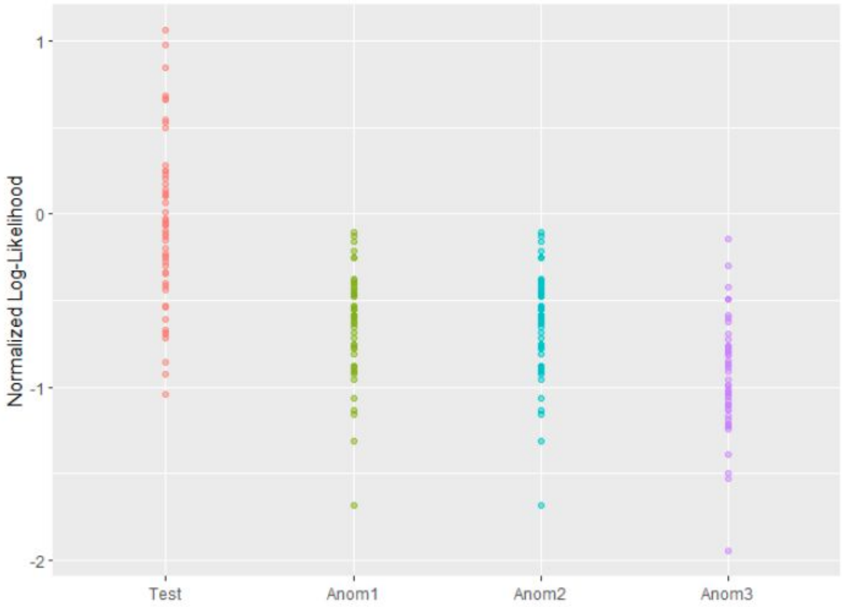
Figure 7: Normalized BIC of Training

Anomaly Detection

Anomaly Detection

- Run the trained model on data with injected anomalies
- Choose threshold for detecting anomalies
    - Depends on properties of false positives to false negatives

Anomaly Detection



Point Anomaly Detection

- Used a different approach than HMMs
  - Took the average of each minute of each feature in the interval
    - This represents normal behaviour of the power grid for the interval
  - Find thresholds using testing data
- Results:
  - Anomaly Dataset 1: 47
  - Anomaly Dataset 2: 47
  - Anomaly Dataset 3: 2,530

Conclusion

Accomplishments

- An HMM that represents normal behavior in power grids
  - Can be used to find anomalies
- Adjustment of log-likelihood thresholds for anomalies

- For threshold –0.342, we get 25.5% of normal data as false positive and 9.3% of anomalous data as false negative.

Challenges

- Substantial amount of overlap between the log-likelihoods of individual intervals from the testing data set and the anomaly data set

Lessons Learned

- Feature selection is important!
- How we split the dataset into training and testing is also important!

Thank you for listening!

Any questions?

References

[1] Canadian Centre for Cyber Security. National Cyber Threat Assessment 2020. Communications Security Establishment, Government of Canada, 2020.

[2] Oracle. *What is the internet of things (IOT)?* — Retrieved December 3, 2021, from https://www.oracle.com/ca-en/internet-of-things/what-is-iot/.