COMP3217 Security of Cyber-Physical Systems 23/24 Coursework 2

Apishan Thayananthan (at16g21)

April 18, 2024

Google Colab Link: `https://colab.research.google.com/drive/1lSmRlXJdarYY_KwJN5ZPbCGQ__4aiTnf?usp=sharing`

Part A

Introduction

Problem Description

In this coursework, we are presented with a cyber-physical challenge within a power system grid. The grid contains two power generators (G1 and G2) and Intelligent Electronic Devices (IEDs) responsible for controlling breakers labelled BR1 through BR4. The IEDs (R1 through R4) each govern a breaker, leveraging a distance protection scheme to detect faults. Additionally, a manual override exists for maintenance purposes. The dataset consists of 6,000 system traces with 128 features each, derived from the measurements taken by 4 PMUs. These features are augmented with control panel logs and other system alerts. Each trace is labelled as normal or indicative of a data injection attack. The task involves using machine learning to classify these traces accurately, thereby detecting potential cyber-attacks on the system. The ultimate goal is to compute labels for an unlabeled test set of system traces, replicating the format of the labelled training data. This classification will enhance the detection and subsequent response to real-world power system attacks.

Dataset Characterisation

The dataset used for Part A is well-balanced, containing 3,000 instances of normal operational events and an equal number of data injection attack events. This

equilibrium facilitates a fair training environment for machine learning models, enabling them to learn and distinguish between the two event types without inherent bias toward one class. Each trace comprises 128 features, including measurements from four PMUs, which measure electrical waveforms for synchronisation purposes. The features are identified by an index referencing the PMU and type of measurement, totalling 116 PMU measurement columns. The dataset includes 12 columns about control panel logs, snort alerts, and relay logs containing operational and security-related information. The labelling scheme distinguishes between normal events, labelled 0, and data injection attacks, labelled 1, aiming to train the machine learning model on a balanced representation of the two event classes. This dataset's robust and multi-dimensional nature is crucial for developing an effective classification model to identify cyber threats within the grid accurately.

Methodology

The initial step involved standardising features to ensure a uniform scale across all inputs. This is crucial for models relying on distance metrics, preventing any feature from disproportionately influencing the model due to scale variances.

Then, hyperparameter tuning was conducted using GridSearchCV to optimise parameters such as 'n_estimators'. This balance between model complexity and learning capability enhances the models' robustness and generalisability. 'n_estimators' is mainly targeted because it directly influences the model's capacity to learn from data without underfitting or overfitting.

StratifiedKFold cross-validation was utilised to maintain a representative proportion of each class in every fold, which is critical for achieving consistent model accuracy across imbalanced datasets. This method ensures reliable performance across different data subsets.

Various metrics were used for model evaluation. Accuracy metrics directly assessed effectiveness, confusion matrices highlighted classification errors, and ROC curves evaluated sensitivity and specificity trade-offs. These evaluations guided model adjustments to enhance accuracy and reliability. Finally, simplicity in model design was prioritised to prevent overfitting and ensure models generalise well to new data. This approach is essential for deploying models that perform reliably and accurately under varied real-world conditions, making them suitable for practical applications in power system security.

Results

The cross-validation accuracy showcased scores ranging from approximately 0.978 to 0.986. This high degree of accuracy indicates a robust predictive capability. The

gradual decrease in training loss over 100 iterations, from 0.48853 to 0.00490, signifies a significant reduction in training error, highlighting effective model learning and generalisation without overfitting.

The confusion matrix and classification report corroborate the model's precision and recall, scoring around 0.99, demonstrating an excellent balance between sensitivity and specificity. The ROC curve's AUC, which is 0.99, displays outstanding model performance with an excellent trade-off between the true and false positive rates.

Overall, these results suggest the model is highly effective in distinguishing between normal and attack events, which is pivotal for deploying a reliable security measure within power system grids.

Evaluation

The evaluation of the XGBoost model's performance on the binary classification task reveals a highly accurate system. Cross-validation accuracy rates remain exceptionally high, demonstrating the model's robustness. The descending training loss over numerous training iterations reflects the model's increasing proficiency in accurately predicting the test data while reducing the risk of overfitting. The confusion matrix and classification report attest to the model's exceptional precision and recall across both classes. Moreover, with an AUC near perfect, the ROC curve highlights the model's ability to distinguish between the classes effectively. These results collectively affirm the model's potential as a reliable tool for real-time anomaly detection in power system grids.

Part B

Introduction

Problem Description

Part B of the coursework expands on Part A by introducing an additional dimension to the classification task, now encompassing three distinct classes of events within the power grid. This multi-class scenario distinguishes between normal operations, data injection attacks, and command injection attacks within the system traces. The dataset mirrors the complexity of real-world cybersecurity threats to critical infrastructure. It requires developing a machine-learning model capable of precisely recognising and categorising these nuanced event types. Like Part A, the goal is to classify the provided training data effectively and to generalise this capability to a new, unseen set of system traces.

Dataset Characterisation

The dataset is similar to Part A. The dataset includes 6,000 system traces. However, it now has three labelled events: normal operations, labelled 0; data injection attacks, labelled 1; and command injection attacks, labelled 2. Each event is equally represented by 2000 instances each. Each trace consists of 128 features derived from various measurements and logs similar to Part A's.

Methodology

In Part B, the methodology primarily focused on leveraging XGBoost, which was recognised for its superior performance in Part A. However, other models were also explored to see if they performed better, such as LightGBM and TabNet, for multi-class classification.

The process began with standardising features to maintain consistent input scales, which is crucial for models reliant on distance calculations. Scaling data, even with evenly distributed classes, is essential in preparing the dataset for machine learning models, particularly those like XGBoost. Scaling standardises the range of continuous initial features, ensuring that no single feature dominates the model's learning due to its variance or scale. This uniformity is crucial because models relying on gradient descent methods, as in XGBoost, can converge more quickly and effectively when all features contribute equally. It helps avoid biases toward larger-scale features and improves the algorithm's sensitivity to all input features equally.

Hyperparameter tuning with GridSearchCV, particularly 'n_estimators', optimised XGBoost's learning depth without causing overfitting.

StratifiedKFold cross-validation ensured balanced class representation, enhancing model accuracy on the varied dataset. Training and validation split facilitated robust performance evaluations through accuracy assessments, confusion matrices, and ROC analysis. While LightGBM and TabNet were considered, XGBoost was the primary model due to its effectiveness in handling multi-class challenges, guided by extensive evaluations that confirmed its efficacy and generalisability for deployment in detecting diverse cyber threats within power systems.

Results

The results section for Part B of the coursework reflects the XGBoost model's proficiency in classifying power system events into three distinct categories: normal, data injection attack, and command injection attack. The model achieved an overall accuracy of 0.9667, with the confusion matrix indicating an impressive true positive rate across the classes: 394 for normal events, 373 for data injection at-

tacks, and 393 for command injection attacks. Training loss metrics further detail the model's adept learning. This is shown by the descending multi-class training loss values, starting from an initial 0.90353 and steadily decreasing to a final 0.03304, indicating an optimal fit. These results suggest the model's capability to distinguish subtle patterns and differentiate between events with high accuracy, an essential quality for ensuring the reliability and security of cyber-physical power systems.

Evaluation

When evaluating the results for Part B, the XGBoost model's performance stands out. The evaluation centred on the model's accuracy, precision, and recall, all crucial for a nuanced understanding of its effectiveness. The model showcased robust predictive capabilities with minimal errors. This is shown by the confusion matrix's high true positive rates and low false positives and negatives. The decline in training loss during the training phase implies that the model's predictions became more confident and accurate with each epoch. It suggests that the model was not merely memorising the training data but learning to generalise. Furthermore, the high overall accuracy score of 0.9667 underscores the model's adeptness at distinguishing between normal operations and various types of cyber-attacks. This level of performance denotes a well-tuned machine-learning pipeline, affirming the model's potential as a reliable tool in the cybersecurity of power grids.