Feature Analysis with Information Gain for Anomaly Detection

Diana Pham

Executive Summary [250-300 words]:

This project will focus on recreating the results from the research paper, CICIDS-2017 Dataset Feature Analysis with Information Gain for Anomaly Detection by Kurniabudi, D. Stiawan, et al. The research investigates the effectiveness of Information Gain in selecting relevant features for anomaly detection in network traffic. By systematically analyzing feature groups and their impact on detection performance, the study aims to enhance anomaly detection capabilities in complex datasets. Since anomaly detection is vital across various domains for identifying deviations from normal behavior that may indicate security threats or fraud, prioritizing relevant features that contribute most to distinguishing between normal and anomalous instances could lead to improved accuracy and reliability of anomaly detection.

There are four key stages for this research's experiment: data selection and preprocessing, feature selection using information gain, classification with five algorithms (Random Forest, Bayes Net, Random Tree, Naive Bayes, and J48), and evaluating the classifiers. For the feature selection stage, it will be implemented by ranking features based on their relevance and then grouping features together based on their weight scores, resulting in several feature subsets. The performance of feature subsets, classified by Random Forest (RF), Bayes Net (BN), Random Tree (RT), Naive Bayes (NB), and J48, is evaluated using metrics like TPR, FPR, Accuracy, Recall, Precision, and execution time.

However, the project faces several challenges. Balancing feature selection to avoid overfitting or underfitting the model while maintaining interpretability is crucial, and expert intervention is needed to determine the minimum weight value that decides how the features are grouped together. In addition, careful consideration of evaluation metrics is also required to accurately assess the model's effectiveness in anomaly detection. Overall, overcoming these challenges will lead to a more robust anomaly detection system with significant real-world implications.

Tasks

1. Data Selection & Preprocessing
    a. Select what percentage of the CICIDS-2017 data to use
        i. Using a subset so the classifiers can run within a reasonable amount of time so I can fix any errors, if needed
    b. Remove redundant features
    c. Relabel the observations into updated attack categories
    d. Split the preprocessed dataset into train and test sets
2. Feature Selection using Information Gain
    a. Use information gain to select features on the training dataset
    b. Group the features according to their weights
3. Classification with Five algorithms
    a. Each feature subset is classified using Random Forest, Bayes Net, Random Tree, Naive Bayes, and J48
4. Evaluating the Classifiers
    a. Compare and analyze the TPR, FPR, Precision, Recall, Accuracy, Percentage of Incorrectly Classified, and Execution Time of each classifier algorithm
5. Summarize Results into a Deliverable

Metrics of success

The performance of feature subsets, classified by the above five classifiers, will be evaluated using the following metrics: TPR, FPR, Accuracy, Recall, Precision, and execution time.

Milestones

What are the dates/the timeline you will complete the above tasks

Task	Deadline (by 11:59PM)
Data Selection & Preprocessing	Week 8 (2/26/2024)
Feature Selection Using Information Gain	Week 10 (3/20/2024)
Classification with Five Algorithms	Week 12 (4/3/2024)
Evaluating the Classifiers	Week 14 (4/17/2024)
Summarize Results into a Deliverable	Week 15 (4/24/2024)

Deliverables

Each milestone should have two or more deliverables. Make sure to include the following types of deliverables at least once:

- Two paragraphs/bulleted list of the work does towards the task
- Functional program code with a description of how to install and use
- A final report that captures the research project effort
- A blog or Linkedin post or a poster

Task	Deliverables
Data Selection & Preprocessing	☐ Two paragraphs/bulleted list of the work done towards the task ☐ Functional program code with a description of how to install and use
Feature Selection Using Information Gain	☐ Two paragraphs/bulleted list of the work done towards the task ☐ Functional program code with a description of how to install and use
Classification with Five Algorithms	☐ Two paragraphs/bulleted list of the work done towards the task ☐ Functional program code with a description of how to install and use
Evaluating the Classifiers	☐ Two paragraphs/bulleted list of the work done towards the task ☐ Functional program code with a description of how to install and use
Summarize Results into a Deliverable	☐ A final report that captures the research project effort ☐ A blog or poster