Feature Analysis with Information Gain for Anomaly Detection

Diana Pham
Department of Computer Science
University of Southern California
dianadph@usc.edu

Abstract

This paper 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 [1]. 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 pre-processing, feature selection using information gain, classification with four algorithms (Random Forest, Random Tree, Naive Bayes, and XGBoost), 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), Random Tree (RT), Naive Bayes (NB), and XGBoost (XGB), is evaluated using metrics like Accuracy, F1 Score, and execution time.

This experiment faced several challenges, with the major issue being the different between Python and Weka, the software that the original authors used to conduct their analysis. In addition, balancing feature selection to avoid over fitting or under fitting 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.

1 Introduction

In the realm of cybersecurity, detecting zero-day attacks remains a critical challenge, prompting the utilization of anomaly-based intrusion detection techniques. Despite the proliferation of such methods, they grapple with several key obstacles, notably the high dimensionality of data, computational complexity, and time constraints.

Addressing the dimensionality issue, researchers have turned to feature selection techniques. This strategy, pivotal for understanding data dynamics, not only trims computational demands but also mitigates the adverse effects of dimensionality, enhancing predictive machine performance. Feature selection, integral to dimensional reduction, aims to identify an optimal feature subset that encapsulates the dataset's essence.

Numerous studies have explored feature selection's role in refining anomaly detection accuracy, often employing datasets like KDD Cup 99 and methodologies such as Chi-Square and Support Vector Machine. However, these approaches have predominantly been tested on datasets with limited features, overlooking the challenges posed by larger datasets. Particularly concerning is the potential for over fitting and escalating computational costs associated with analyzing datasets with extensive features, an aspect scantily addressed in existing research.

In contrast, Information Gain emerges as a prevalent tool for discerning pertinent features, and while Information Gain has been extensively applied to datasets with constrained features, its efficacy on more intricate datasets like CIC-IDS2017 remains under explored. Information gain quantifies how much information a feature provides about the class label. It's the reduction in entropy or uncertainty achieved by splitting the dataset on a particular feature. When selecting features for a model, features with high information gain are considered more important as they contribute the most to reducing the uncertainty in predicting the target variable.

$$Entropy\,(S) = \sum_{i}^{c} -P_i log_2 P_i$$

$$Gain\,(S, A) = Entropy\,(s) - \sum_{Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

2 The Dataset

The dataset that was used for this paper was the CIC-IDS2017 Dataset from the Canadian Institute for Cybersecurity. It's a comprehensive dataset with about 53GB of data designed for evaluating intrusion detection systems with realistic network traffic containing both benign and most up-to-date common attacks.

Previous network intrusion detection systems datasets, such as KDD Cup 99, were out of date and unreliable. Some of those datasets suffered from lack of traffic diversity, volume, and did not cover the variety of known attacks today. But with this dataset, it contains a wide range of network intrusion scenarios, including various types of attacks such as Denial-of-Service, Distributed Denial-of-Service, Brute-force attacks, and Malware infections.

Each instance in the dataset is labeled with information about whether it represents normal network traffic or an attack. And his labeled data is crucial for training and evaluating machine learning models for intrusion detection. In addition to the PCAP files, the CIC also provided CSVs with extracted features and this is what was used for this paper.

3 Methodology

Data Selection and Pre-processing
The first step of the experiment was to randomly select 20 percent of the data, remove redundant features, and relabel observations into updated attack categories. Since the dataset had hundred of thousands of rows in the CSV files, there were about 60,000 rows after randomly selecting 20 percent

of the data. For data pre-processing, redundant features such as Fwd Header Length were removed to reduce the number of features that will be processed in future tasks. What was done differently from the original authors of the paper was including the down sample of the cases for Normal traffic by 20 percent and removing Infiltration attacks since there were only single digit cases for that attack each time a random selection of the data was selected. In addition, the original labels were relabeled to the following updated class attacks:

Updated Class Label	Original Class Label
Normal	Benign
Bot	Bot
Brute Force	FTP-Patator, SSH-Patator
DoS/DdoS	DDoS, DoS, DoS GoldenEye, DoS Hulk, DoS Slowhttptest, DoS slowloris, Heartbleed
Infiltration	Infiltration
Portscan	PortScan
Web Attack	Web Attack - Brute Force, Web Attack - Sql Injection, Web Attack - XSS

Feature Selection with Information Gain

Feature selection was conducted on the training dataset by using information gain as the criteria. To do so, two functions were created that calculated entropy and information gain for each variable in the dataset. After storing the information gains associated with each variable within a Python dictionary, they were sorted in descending order and feature groups were created according to their weights.

A total of seven groups were created for features with: 1) Information gain greater than 0.6, 2) Information gain greater than 0.5, 3) Information gain greater than 0.4, 4) Information gain greater than 0.3, 5) Information gain greater than 0.2, 6) Information gain greater than 0.1, and 7) All features. These groups are not mutually exclusive, so if a feature was greater than 0.6, then it was also included in Group 1 as well as the other six groups. As we can see from Figure 1 below, the most important features for identifying different network attacks were Flow bytes per second, Average packet size, Flow packets per second, Packet length standard deviation, Packet length variance, and Forward packets per second since they had the highest information gain values.

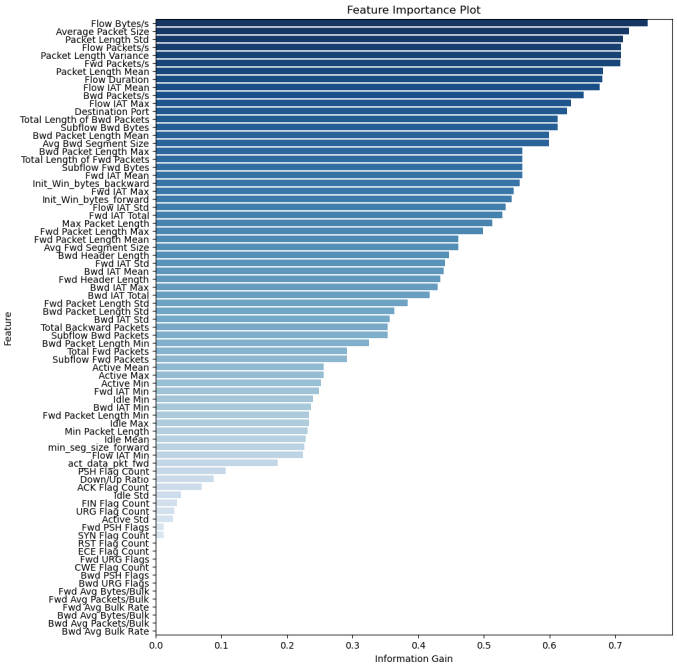

Figure 1: Feature Importance using Information Gain

Classification
With the seven feature groups that were created during feature selection, four classification models were built: Random Forest (RF), Random Tree (RT), Bayesian Network (BN), and XGBoost (XGB) with 10-fold cross-validation. The authors of the research paper briefly mentioned that these classifiers were chosen based on their research of previous works and found that they had good performance in terms of accuracy, learning ability, scalability, and speed. The original authors built a Bayesian Network but since they didn't describe the structure of the network, it was difficult to recreate it since the kernel would crash during every run due to memory constraints. So, after further research, it was decided to replace the Bayesian Network with XGBoost for this paper since it can handle large and imbalanced datasets.

4 Results

To evaluate the classification models that were built, confusion matrices were constructed to calculate each model's Accuracy, F1 Score, and Execution Time. The execution time is measured during the training time (the time measured between the classification process starting and after predictions are made).

Accuracy
In Figure 2 below, the x-axis is the number of features in each feature group and the y-axis is the accuracy. The bar chart rounds the values to two decimal places so it's difficult to see the difference in performance, so a table with the accuracies was also included as Figure 3. By observing these figures, they show that Random Forest performed the best out of the 4 models with at least 99 percent accuracy until it was tested on the last feature group with all 77 features.

It's also notable that Naive Bayes performed poorly and had the same accuracy for each feature group. This may be due to its assumption that all features are conditionally independent given the class label, which may not hold in real-world data, and it struggles with class imbalance. For this experiment, it predicted almost every observation as a Brute-Force attack.

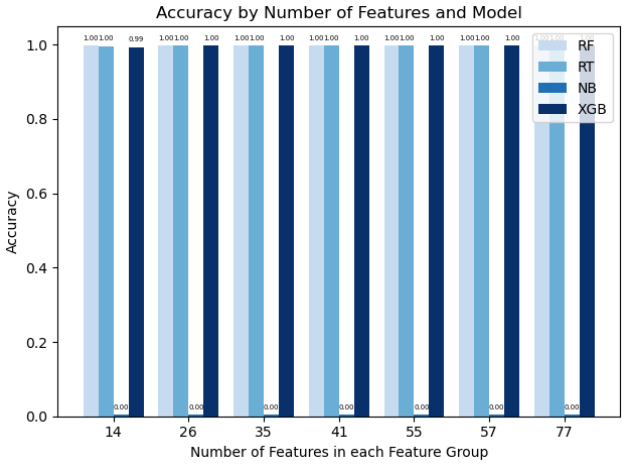

Figure 2: Model Accuracy Bar Chart

Accuracy				
	RF	RT	NB	XGB
14 Features	0.99721	0.99617	0.00495	0.99383
26 Features	0.99862	0.99779	0.00495	0.99781
35 Features	0.99865	0.99788	0.00495	0.99761
41 Features	0.99866	0.99781	0.00495	0.99776
55 Features	0.99866	0.99778	0.00495	0.99828
57 Features	0.99865	0.99797	0.00495	0.99843
77 Features	0.80346	0.99782	0.00495	0.99831

Figure 3: Model Accuracy Table

F1 Score

Similar to Accuracy, a bar chart and table were also created to showcase each model's F1 Score. The findings from Figures 4 and 5 are similar to the ones discussed for Accuracy, except Random Forest performed the best out of the 4 models with at least 99 percent accuracy for all feature groups here.

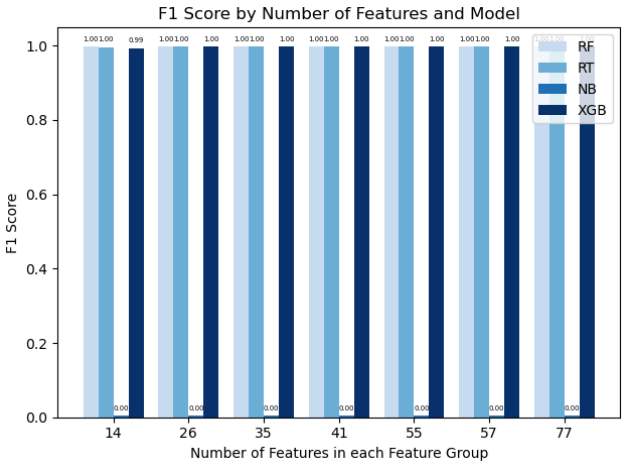

Figure 4: Model F1 Score Bar Chart

F1 Score				
	RF	RT	NB	XGB
14 Features	0.99297	0.98415	0.01135	0.98855
26 Features	0.99936	0.99044	0.01135	0.99744
35 Features	0.99936	0.98728	0.11345	0.99744
41 Features	0.99936	0.98536	0.01135	0.99744
55 Features	0.99936	0.99361	0.01135	0.98861
57 Features	0.99936	0.98784	0.01135	0.99681
77 Features	0.99872	0.98473	0.01135	0.99681

Figure 5: Model F1 Score Table

Execution Time

Figure 6 shows the models' execution times for each feature group. Random Tree and Naive Bayes were relatively quick, but the number of selected features has a significant impact on Random Forest and XGBoost when looking at the first feature group with 14 features versus all 77 features. This is a key insight because in the real world, detection systems need to be able to identify attacks not only accurately, but also quickly.

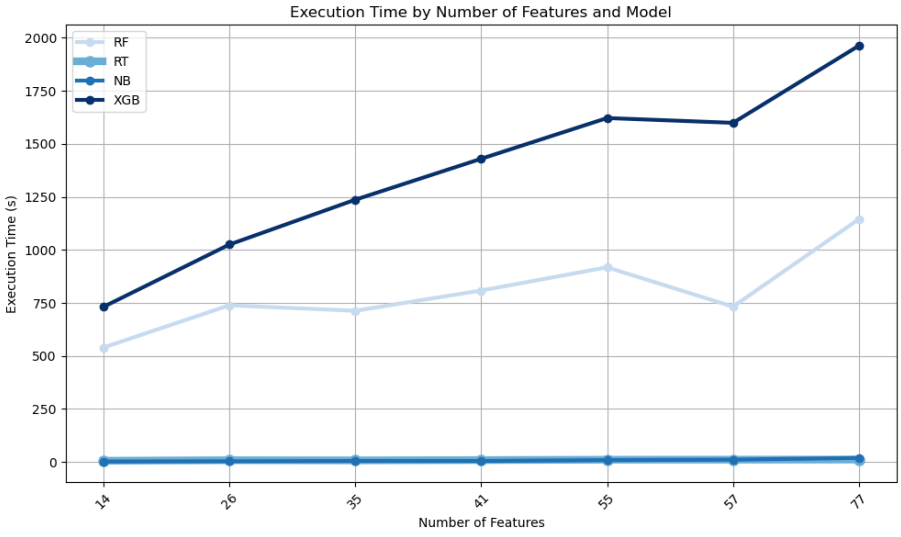

Figure 6: Model Execution Time

5 Conclusions, Limitations, and Future Directions

Conclusion
In conclusion, this study highlights the persistent challenges in zero-day attack detection within cybersecurity, driving the adoption of anomaly-based intrusion detection techniques. Despite their proliferation, key obstacles such as high data dimensionality, computational complexity, and time constraints persist. To address these challenges, researchers have increasingly turned to feature selection techniques, particularly leveraging Information Gain. However, while effective, the application of Information Gain on complex datasets like CIC-IDS2017 remains relatively unexplored.

By conducting experiments on a subset of the CIC dataset and employing Information Gain-based feature selection, this research identifies significant features for detecting various network attacks. Notably, Random Forest emerges as the most robust performer, showcasing the nuanced trade-offs between accuracy, execution time, and the number of selected features. These insights underscore the importance of balancing accuracy and computational efficiency in real-world deployment scenarios, signaling avenues for further research to enhance anomaly-based intrusion detection in evolving cybersecurity landscapes.

Limitations
The paper exhibits several limitations, notably its narrow focus on Information Gain as the primary feature selection technique, overlooking alternative methods that could provide additional insights. Additionally, reliance on a single dataset, CIC-IDS2017, may restrict the generalizability of findings, warranting further evaluation on diverse datasets to enhance validity. Assumptions and simplifications in the experimental setup, such as excluding specific attack categories or downsampling data, raise concerns about representativeness. Lack of real-time evaluation and scalability considerations further limit the practical applicability of the proposed techniques. Addressing these limitations in future research would contribute to a more comprehensive understanding of anomaly-based intrusion detection and facilitate their effective deployment in real-world scenarios.

Future Directions
Future research directions stemming from the findings of the paper include exploring alternative feature selection methods like Mutual Information or Recursive Feature Elimination to further refine anomaly detection on complex datasets like CIC-IDS2017. Additionally, investigating ensemble methods and model fusion techniques could enhance detection accuracy while maintaining computational efficiency, potentially integrating different classifiers such as Random Forest and XGBoost. Dynamic feature selection mechanisms and real-time anomaly detection scenarios offer promising avenues, facilitating adaptation to changing network environments and validation of techniques in dynamic contexts.

Further research could involve evaluating proposed techniques on diverse datasets and network environments beyond CIC-IDS2017 to validate generalizability. Integration of domain knowledge and contextual information, alongside optimization for scalability and efficiency, are critical for deploying intrusion detection systems effectively. By focusing on these areas, researchers can advance anomaly-based intrusion detection and bolster cybersecurity defenses against emerging threats.

References

[1] Kurniabudi, Deris Stiawan, Darmawijoyo, Mohd Yazid Bin Idris, Alwi M. Bamhdi, and Rahmat Budiarto. Cicids-2017 dataset feature analysis with information gain for anomaly detection. *IEEE Access*, 8:132911–132921, 2020.