Lecture Notes in Electrical Engineering 1211

Koushlendra Kumar Singh
Sangeeta Singh
Subodh Srivastava
Manish Kumar Bajpai *Editors*

Machine Vision and Augmented Intelligence

Select Proceedings of MAI 2023

Springer

Editors
Koushlendra Kumar Singh
Department of Computer Science
and Engineering
National Institute of Technology
Jamshedpur
Jamshedpur, Jharkhand, India

Subodh Srivastava
Department of Electronics
and Communication Engineering
National Institute of Technology Patna
Patna, Bihar, India

Sangeeta Singh
Department of Electronics
and Communication Engineering
National Institute of Technology Patna
Patna, Bihar, India

Manish Kumar Bajpai
Department of Computer Science
and Engineering
National Institute of Technology Warangal
Warangal, Telangana, India

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-97-4358-2 ISBN 978-981-97-4359-9 (eBook)
https://doi.org/10.1007/978-981-97-4359-9

Contents

Feature Engineering Using Machine Learning Techniques on CIC-IOT-2023 Dataset

Komal Jakotiya, Vishal Shirsath, and Raj Gaurav Mishra

1 Introduction

Today, in this constantly changing world of the Internet of Things (IOT), data-driven understanding is critical to increasing device efficiency and safety, as well as overall performance. Analyzing this data presents significance potential for insights and applications across industries, from predictive maintenance in manufacturing to smart city initiatives. However, to unlock this potential, it's critical to employ effective feature engineering techniques, along with machine learning, on IOT datasets like the CIC-IOT-2023 dataset.

The CIC-IOT-2023 dataset is a comprehensive dataset specifically curated for IoT-related research and analysis. It encompasses a wide range of IoT devices, network protocols, and traffic patterns. This dataset offers useful insights into the behavior, security, and performance of IOT devices and networks. IOT assaults Dataset Facilitating Advancing Security Analytics Applications in Real-Time IoT Environments. To achieve this objective, we executed 33 distinct attacks within an IoT network consisting of 105 interconnected devices. These attacks have been categorized into seven distinct types: DDoS DoS, Recon, web-based and brute force attacks as well as spoofing and Mirai. Second, most significantly all of these attacks have been launched by malicious IOT devices just to attack other elements on the same network. This data set includes diverse attacks which have not been seen in other IOT datasets, giving a chance to IOT professionals for pioneering new security analytics innovation. Moreover, feature engineering forms the basis of many effective machine

K. Jakotiya (✉) · V. Shirsath · R. G. Mishra
School of Engineering, ADYPU, Pune, India
e-mail: Komal.jakotiya@adypu.edu.in

V. Shirsath
e-mail: Vishal.shirsath@adypu.edu.in

R. G. Mishra
e-mail: raj.mishra@adypu.edu.in

K. Kumar Singh et al. (eds.), *Machine Vision and Augmented Intelligence*, Lecture Notes in Electrical Engineering 1211, https://doi.org/10.1007/978-981-97-4359-9_64

learning applications. Models through it can learn from data efficiently and easily understand the nature of what they are processing. In this way, we hope to give a general theory for feature engineering on the CIC IOT Dataset 2023. Our goal is to provide researchers, data scientists and practitioners with the right tools and techniques for extracting valuable insight from IoT data so that its potential can be realized in applications ranging from predictive maintenance to anomaly detection.

2 Related Work

A method of feature selection in Intrusion Detection Systems (IDS) is presented by Nimbalkar et al. [1], which focuses on detection of DoS and DDoS attacks. Their method involved insertion and union operations on feature subsets selected from the top 50% of Information Gain values, or the bottom 50% of gain ratio values. When assessed on the IoT-BoT and KDD Cup 1999 datasets with the JRip classifier, their approach demonstrated better performance compared to the initial feature set and conventional IDSs, using just 16 and 19 characteristics, respectively.

In the domain of feature selection and anomaly detection in the context of the Internet of Things (IoT), Li et al. [2] introduced an innovative model architecture that merges deep learning with security breach detection. Their study introduces a technique for examining learning models to aid in migration while choosing pertinent system features. They performed experiments using the KDD CUP 99 dataset as their experimental dataset.

To automate the identification of potential security breaches within IoT traffic generated using the MQTT (Message Queue Telemetry Transport) protocol, an analysis is conducted [3]. This analysis encompasses five distinct categories of cyber-attacks: brute force, SlowITe, flooding, denial of service (DoS), and malformed data attacks, in addition to legitimate traffic. The study involves the training of five well-known machine learning (ML) models, namely MLP, AdaBoost, Light GBM, Decision Tree and Random Forest Classifiers, for the purpose of predicting cyber-attacks. The anticipation of malevolent activities within network systems has been a well-established field of study, dating back to the 1990s. As machine learning (ML) technologies have evolved, numerous research endeavors have emerged, delving into the potential of ML models for forecasting cyber threats [4]. Within this domain, a prevalent choice for research is the utilization of the KDDCUP99 dataset [5], constructed from TCP/IP (Transmission Control Protocol or Internet Protocol) traffic in IP networks. However, it's worth noting that this dataset falls short in presenting a genuine depiction of malicious traffic patterns within IoT networks.

Typically, IoT networks are confined to private domains, characterized by an overseeing entity that manages and monitors the activities of interconnected devices. In the context of IoT networks commonly found in intelligent home control systems, there is a preference for the MQTT protocol, which operates atop the TCP communication protocol. This choice is driven by MQTT's simplicity and minimal resource demands [6]. The MQTT protocol delineates three primary components: the MQTT

Broker, which serves as the Master; the MQTT Publisher, responsible for communication with clients; and the MQTT Subscriber, functioning as a client. These three entities collectively constitute the IoT network, facilitating the exchange of information via the MQTT protocol.

While MQTT stands as a favored protocol for IoT applications, there is a notable scarcity of research dedicated to scrutinizing MQTT traffic behavior for the prediction of cyber-attacks [7]. A substantial portion of research employing MQTT traffic analysis encounters two significant challenges: the absence of publicly accessible datasets and a lack of focus on smart home IoT networks. Consequently, replicating the outcomes documented in the existing literature for comparative assessment becomes a formidable task. Additionally, existing studies primarily concentrate on the detection of two prevalent cyber-attacks: security breach and denial of service (DoS), largely overlooking other potential threats.

One of the novel threats that can potentially impact IoT networks is the Slow DoS against Internet of Things Environments (SlowITe) attack [8]. SlowITe falls into the category of denial-of-service attacks and has the capability to target MQTT servers operating within a TCP environment. Additionally, IoT networks are susceptible to a range of other security breaches, including brute force attacks, flooding attacks, traditional denial of service (DoS) attacks, as well as attacks involving malformed data. These threats collectively pose significant challenges to the security and stability of IoT systems.

Supervised machine learning classifiers involve training models using labeled data to predict known outcomes. Among the popular and state-of-the-art supervised learning classifiers are Linear Discriminant Analysis, Naïve Bayes, Neural Networks, Decision Trees, Stochastic Gradient Descent, SVM (Support Vector Machines) and Ensemble Models. These classifiers play a pivotal role in a wide range of machine learning applications, from classification to regression tasks.

In the next stages of the proposed Network Traffic Log identification framework, the Feature Extraction and Scalable Hypothesis (FRESH) algorithm, as described in [9], is employed. Its primary objective is to extract and meticulously choose the most relevant data characteristics from the provided dataset. After this process of feature extraction and selection, the model is then trained using the CatBoost. This combination of FRESH and CatBoost contributes to the effectiveness and accuracy of the NTL detection framework.

Ndayishimiyepas et al. [10] Various methods have been employed to predict missing data records in datasets, and now there's an opportunity to introduce an advanced approach. XGBoost, short for Extreme Gradient Boosting, represents a significant advancement in the field. Its core development objective was to enhance both model performance and computational efficiency. Essentially, XGBoost is realization of the Gradient Boosting techniques Machine, designed to significantly amplify the computational capabilities of boosted trees algorithms.

Contemporary IoT networks are comprised of small-scale sensing devices characterized by constrained computational capabilities and limited data storage. Therefore, when formulating machine learning models for the purpose of monitoring and identifying cyber-attacks, it becomes imperative to account for these resource constraints

Table 1 Different machine learning models comparative study

Machine learning model	Strength	Limitations
Decision Tree classifier	Interpretable and can handle both categorical and numerical data	Are prone to overfitting with complex trees may not capture intricate relationships in the data
CatBoost	Effective in handling categorical features and has strong predictive performance	Longer training times compared to some other algorithms
MLP classifier	Capture complex non-linear patterns in data	Require a larger amount of data and may be prone to overfitting
Random Forest	Resilient, adept at managing high dimensional data, and furnishes feature significance scores	May over fit on noisy data and can be computationally expensive for large datasets
Light GBM	Highly efficient for large dataset excellent predictive performance	Require fine tuning sensitive to overfitting for limited data
AdaBoost	Combines feeble learners to construct a robust classifier and is less prone to overfitting	Be sensitive to noisy data and outliers
XgBoost	Excels in predictive accuracy and computational efficiency	Tuning hyper parameters can be challenging more memory for large datasets

and the feasibility of deploying such models within the IoT network infrastructure. The selection of models outlined in Table 1 is a direct response to these precise limitations inherent to IoT networks.

3 Methodology

3.1 Dataset

The study utilized the CIC-IOT-2023 benchmark dataset as its primary data source. In the initial phase of data preprocessing, columns containing null values were eliminated, resulting in a refined dataset. This cleaned dataset was then employed for feature selection and feature scaling. During this stage, various techniques, including Random Forest, XGBoost, and CatBoost, were employed, and the data was balanced to ensure a comprehensive analysis of the dataset's characteristics and performance (Figs. 1 and 2).

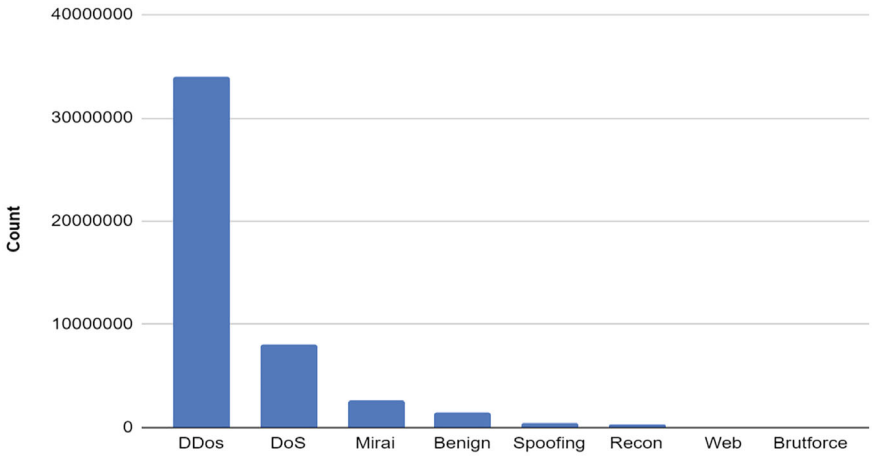

Fig. 1 Types of attacks

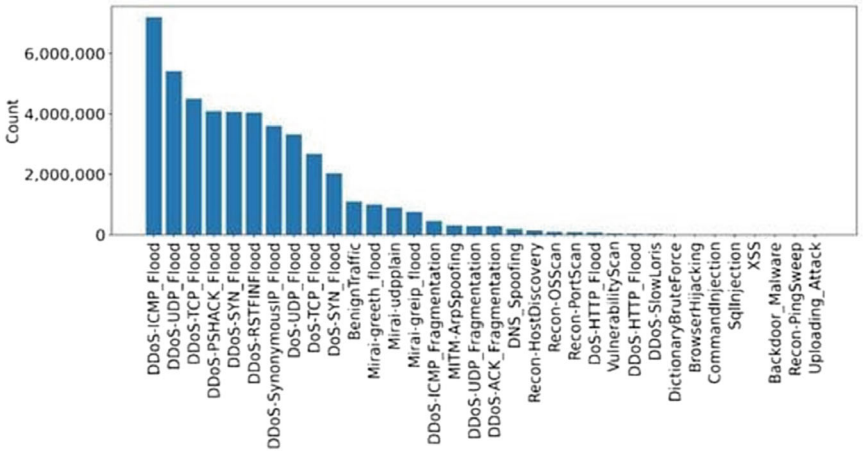

Fig. 2 Different number of records for each subtype of attacks

3.2 Proposed Model

In this research, we explore the following fundamental aspects of feature engineering on the CIC IoT Dataset 2023.

Data Preprocessing. We delve into the critical phase of data preprocessing, addressing issues such as missing values, data normalization, and handling class imbalances. Through these techniques, we ensure that the dataset is clean, balanced, and ready for analysis.

Feature Extraction. Drawing upon domain knowledge, we extract relevant features from the raw network traffic data. These features capture the essence of IoT device behavior, including communication patterns, payload characteristics, and protocol-specific attributes.

Feature Selection. To mitigate the challenges of high dimensionality, we explore techniques for feature selection, identifying the most informative attributes while discarding redundant or noisy ones. This step streamlines model training and enhances interpretability.

Feature Transformation. Techniques for reducing dimensionality, including Principal Component Analysis (PCA) and manifold learning techniques, are employed to visualize and transform the dataset into a more manageable and discriminative form.

Advanced Feature Engineering. We push the boundaries of feature engineering by exploring advanced strategies such as feature generation through deep learning architectures, auto encoders, and transfer learning, tailored to the intricacies of IoT data.

Model Selection and Evaluation. We leverage a wide array of machine learning algorithms, including classical models, ensemble techniques, and deep learning frameworks, to train and evaluate predictive models. Rigorous evaluation metrics are employed to assess model performance.

Interpretability. In addition to predictive accuracy, we emphasize the interpretability of models, making the insights gained from feature-engineered attributes comprehensible to users and stakeholders.

3.3 Imbalance Ratio About the Dataset

These datasets exhibit significant imbalances, which necessitate addressing in order to accurately assess the system's efficiency. The metric used to quantify this imbalance is the Imbalance Ratio, denoted as ρ and calculated as shown in Eq. 1:

$$\text{Imbalance Ratio}\,(\rho) = \text{maxi}\{\text{Ci}\}/\text{mini}\{\text{Ci}\} \tag{1}$$

Here, Ci represents the size of data in class i. In essence, the imbalance rate is characterized by the disparity between the count of occurrences in the majority (maximum) class and the minority (minimum) class.

For the CIC-IOT-2023 dataset, the data imbalance rate is notably high, at 7,000,000:10,000. Such substantial differences between data classes significantly impact the system's effectiveness. Additionally, it's important to highlight that skilled hackers frequently concentrate on exploiting the less common data types to accomplish their goals.

4 Machine Learning Algorithms

4.1 SVM

Support Vector Machines (SVMs) are primarily used as classification algorithms, but they can indirectly contribute to feature engineering through feature selection and dimensionality reduction techniques. SVMs inherently identify the most discriminative features during the training process. SVMs can indirectly aid in feature engineering through these techniques, they are not typically used for explicit feature creation or modification. For feature engineering involving feature creation, transformation, or generation, other techniques like polynomial features, feature engineering based on domain knowledge, and autoencoders may be more appropriate. SVMs are generally employed for their ability to handle high-dimensional data and learn complex decision boundaries when feature selection or dimensionality reduction is the primary goal.

4.2 Random Forest

The Random Forest, a versatile ensemble learning algorithm, indirectly contributes to feature engineering by offering insights into feature significance and selection. Besides being a powerful predictive modeling tool, it also ranks highest among feature importance. In Random Forest, the importance of each feature is determined by how much it affects the model's abilities to predict. Gini impurity and Mean Decrease in Accuracy are two important criteria for this assessment.

In particular, popular algorithms such as Random Forest can not only rank features in terms of feature importance but also select informative features. These capacities enhance model explanatories and forecasting accuracy, particularly when working with massive data sets or intricate feature relationships.

4.3 CatBoost

CatBoost is a gradient boosting procedure that deals well with categorical features. With built-in handling of categorical data, which does not require much preprocessing or one-hot encoding.

Most of the feature engineering for CatBoost is concentrated on how to properly handle categorical features, while at the same time making use of model's ability to capture interactions across different type of variables. With less need for manual encoding and preprocessing, it shortens the process. This means that you have more time to generate interesting features and work on improving model performance. And it turns out that CatBoost's feature importance analysis is an effective tool

for understanding your model and making it more effective. It will help you to concentrate on how and which ones are best picked, what kind of work needs doing in order to prepare them.

5 Evaluation and Results

5.1 Metrics

But assessment criteria for machine learning are an indispensable means of measuring model effectiveness. These criteria provide data analysts, researchers and pros with valuable viewpoints on the power of their model. They can then make decisions informed by reality when it comes to choosing a suitable model or carefully adjusting its parameters. The following are some frequently utilized machine learning evaluation criteria:

Accuracy. Accuracy acts as a fundamental measure of a model's overall correctness. It is the proportion of accurate predicted instances to the total number of instances in the dataset. Although this metric is easy to calculate but not suitable for imbalance data.

Formula: (Correct Positive Predicted instances + Correct Negative Predicted instances)/(Total instances)

Precision: Precision quantifies the ration of true positives (i.e. correctly predicted positive instances) among all instances predicted as positive**.** This metric proves beneficial in scenarios where reducing false positives which are significant.

Formula: Correct Predicted Positives/(Correct Predicted Positives + Incorrect Predicted Positives

Recall (Sensitivity). Recall metric is the fraction of true positives among all actual positive instances. It is valuable utility in situation where the primary goal is to minimize the false negatives.

Formula: True Positives/(True Positives + False Negatives)

F1-Score. The harmonic mean of precision and recall represented by F1-score, a balance can be achieved between these two metrics. While working with various types of datasets that exhibit an uneven class distribution, balance is especially advantageous.

Formula: 2 * Precision * Recall/(Precision + Recall)

These machine learning evaluation metrics evaluate model performance, collectively providing practitioners with tools to refine their models and make them more effective across an ever-broadening scope of applications.

5.2 Results

Oversampling can be used to handle the class imbalance problems found in CIC IOT 2023 data set. It means balancing the number of instances in minority class (representing anomalies or rare events) with those in majority class.

However, an important point is that oversampling can be used to solve the problem of class imbalance but care should be taken. If adjustment occurs over aggressively, it could lead to the problem of overfitting and poor generalization. Also explore techniques such as feature engineering, various resampling methods or anomaly detection (Figs. 3 and 4).

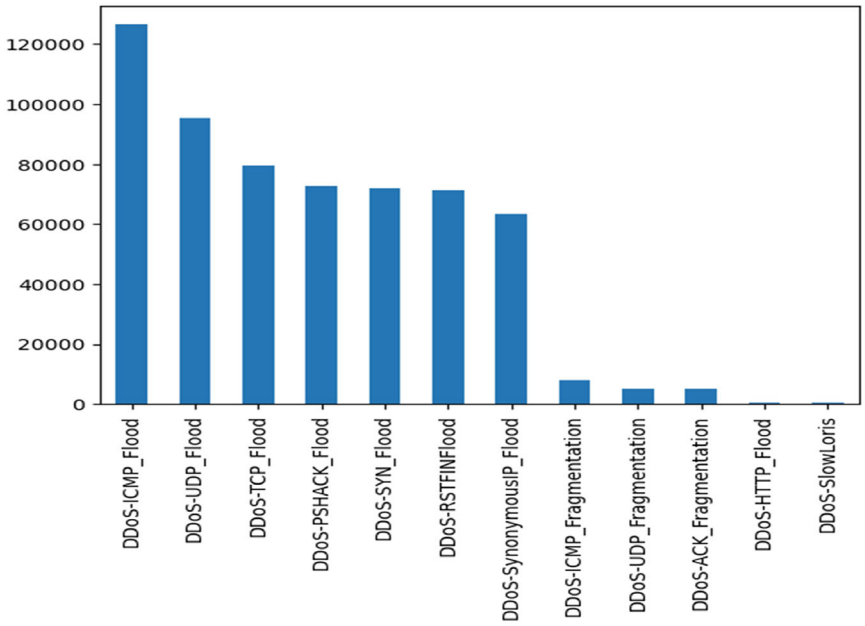

Fig. 3 Imbalanced data (before oversampling)

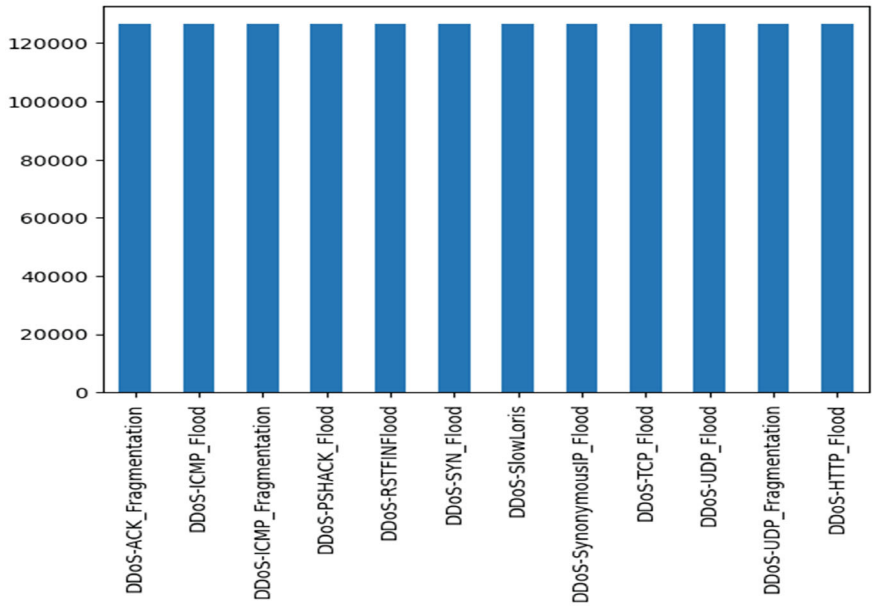

Fig. 4 After over sampling balanced data

6 Conclusion

Our study titled "Feature engineering using Catboost, Random Forest and XGBoost algorithms on CIC IoT Dataset 2023" shows the vital role that feature engineering can play in boosting a machine learning model's predictive performance when it comes to protecting against attacks from adversaries endangering internet of things (IoT) devices. Indeed, Catboost, Random Forest and XGBoost all show their strong points in handling different types of feature sets as well. They also contribute to improving model accuracy. By thoughtfully doing feature engineering on these algorithms, we have already made significant strides in strengthening the analysis capabilities of IoT networks. Yet the choice of the algorithm depends on the specific characteristics of a given dataset and challenge, which once again points to variation as key for security applications in IoT. In general, our research demonstrates the importance of feature engineering and that these advanced algorithms can still support a strengthening in the defenses of IoT networks from cyber threats.

References

1. Nimbalkar P, Kshirsagar D (2021) Feature selection for intrusion detection system in Internet-of-Things (IoT). ICT Express 7(2):177–181

2. Li D et al (2019) IoT data feature extraction and intrusion detection system for smart cities based on deep migration learning. Int J Inf Manag 49:533–545
3. Dissanayake Maheshi B (2021) Feature engineering for cyber-attack detection in internet of things. IJ Wirel Microw Technol 6:46–54
4. Vaccari I et al (2021) Exploiting Internet of Things protocols for malicious data exfiltration activities. IEEE Access 9:104261–104280
5. Imran et al (2023) Improving reliability for detecting anomalies in the MQTT network by applying correlation analysis for feature selection using machine learning techniques. Appl Sci 13(11):6753
6. Al Enany MO, Harb HM, Attiya G (2021) A Comparative analysis of MQTT and IoT application protocols. In: 2021 international conference on electronic engineering (ICEEM). IEEE
7. Ferrag MA et al (2020) Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study. J Inf Secur Appl 50:102419
8. Vaccari I, Aiello M, Cambiaso E (2020) SlowITe, a novel denial of service attack affecting MQTT. Sensors 20(10):2932
9. Hussain S et al (2021) A novel feature engineered-CatBoost-based supervised machine learning framework for electricity theft detection. Energy Rep 7:4425–4436
10. Ndayishimiyepas P, Wilson C, Kimwele M (2022) A hybrid model for predicting missing records in data using XGBoost. In: 2022 IEEE international symposium on product compliance engineering—Asia (ISPCE—ASIA). IEEE