To establish baseline for threat detection

DISSERTATION

Submitted in partial fulfillment of the requirements of the

Degree : MTech in Data Science and Engineering

By

Goyal Taruchit Tarun Chitra
2022DC04496

Under the supervision of

Prathibha Panduranga Rao
Vice President

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE
Pilani (Rajasthan) INDIA

(December, 2024)

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
FIRST SEMESTER 2024-25

DSECLZG628T **DISSERTATION**

Dissertation Title :_To establish baseline for threat detection

Name of Supervisor : Prathibha Panduranga Rao

Name of Student : Goyal Taruchit Tarun Chitra

ID No. of Student : 2022DC04496

Courses Relevant for the Project & Corresponding Semester:
1. Introduction to Data Science (Semester 1)
2. Machine Learning (Semester 2)
3. Artificial and Computational Intelligence (Semester 2)
4. Data Visualization and Interpretation (Semester 2)

Abstract

Cybersecurity is key for any organization; attackers keep evolving and learning new ways to evade cyber-attack detection deployed by organizations. By analyzing the events, security operations center (SOC) can detect threats and make existing detections more effective. While analyzing network dataset for detecting cyber-attacks, the volume of records and dimensionality of records generated are very high. As the result, building automated analysis and detection of potential threats can lead to noisy outcomes. In most of the historical research, the power of advanced graph or deep learning models are leveraged for handling high-dimensional dataset. But it comes at the cost of extensive tuning, computation power and time. Thus, the dissertation aims to leverage optimization algorithms for feature selection which enables to handle high dimensional dataset efficiently and effectively, allowing to identify the most optimal set of features for training the models, improving model's overall performance. Most often in network dataset, the features are not linearly correlated, thus, for handling non-linearity of features, optimization algorithms are useful. The features are then used to train two models: the first model performs binary classification to differentiate an attack from a normal event. The second model performs a multi-class classification to identify the type of attack. This enables to handle both the models independently and make choices which allow to get optimal results for the specific objectives of each model. Finally, the project evaluates each model based on the corresponding subset of optimal features obtained from each optimization algorithm and rank the outcomes. Thus, the projects demonstrate mitigating dependence on advance and complex models for higher accuracy, and rather use existing optimization algorithms with Machine Learning algorithms to achieve the same. This also allows to define baseline of results using Machine Learning algorithms, which can be later used as a benchmark for more advanced models.

Key Words: Classifier model, Supervised Learning, Feature selection, Optimization algorithms

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI
I SEMESTER 24-25
DSECLZG628T **DISSERTATION**
Dissertation Outline (Abstract)

BITS ID No. 2022DC04496 **Name of Student:** Goyal Taruchit Tarun Chitra

Name of Supervisor: Prathibha Panduranga Rao

Designation of Supervisor: Vice President

Qualification and Experience: M.Sc. I.T.

Official E- mail ID of Supervisor: ppandurangarao@statestreet.com

Topic of Dissertation: Establish baseline for threat detection

Taruchit
(Signature of Student)
Date:09 Dec. 2024

Prathibha
(Signature of Supervisor)
Date:09/12/2024

Project Work Title: To establish baseline for threat detection.

Discussion on the chosen topic

1. **The purpose of the work and expected outcome of the work**

    a. To use an intrusion detection dataset for building a binary model and multi-class model.

    b. Binary model will enable to determine whether a previously unseen event in the network is normal or attack.

    c. Multi-class model: If the event was identified as an attack, the model will determine the type of attack.

    d. The project will use different optimization algorithms for feature selection and compute the results on common Machine learning model. This will enable us to compare and determine how different optimization approaches in feature selection impacts the results in cyber security use case and compare the outcomes.

2. **Literature review done in connection with the work**

    a. Graph based feature extraction and normalization, with LSTM based neural network architecture were used for classification.

    b. Single-tree based learning algorithm was used for classification of network events.

    c. Deep learning approach with multiple neural network designs and hidden layers were used to analyze and classify network events.

    d. Features are categorized into mandatory and non-mandatory features. On non-mandatory features, feature selection process was carried out using Univariate and ANOVA. The best subset of non-mandatory features was selected and merged with mandatory features to train the model.

    e. Naïve Bayes, Decision Trees and Neural Networks were used to build models and compare their results.

    f. Feature selection techniques: Information Gain and Gini importance were used for fetching the subset of features for training model.

g. Pre-processing techniques: variance thresholding and one-hot encoding were used to clean the dataset. For feature selection, filters, wrapper and embedded method were used.

h. Both supervised and unsupervised machine learning algorithms were used to train the model. Graph based feature extraction technique was used to extract features from dataset.

i. Autoencoders were trained on normal data to learn the underlying patterns. Thus, they were used for dimensionality reduction.

j. Ant Colony optimization was used for feature selection by running multiple iterations to determine feature importance.

k. Ant Colony optimization was used to extract label correlation which was combined with heuristic factor as label weights, enabling to eliminate redundant and irrelevant features.

l. Particle Swarm Optimization was used to fetch best subset of features, and Support Vector Machine was used to build the model.

m. Two types of attacks were selected from the dataset: Smurf and Neptune. Then, Particle Swarm optimization was used for fetching best subset of features. Finally, Artificial Neural Network was used to train the model.

n. Ant Colony Optimization was used for dimensionality reduction. Support Vector Machine was used to train the model.

o. Simulated Annealing algorithm was used for feature selection in different classification problems and their results were compared with feature selection done without Simulated Annealing algorithm.

3. Brief discussion on existing processes and its limitations

a. Most of the previous research in analyzing network dataset is based on Graphs and Deep Learning, which requires extensive computation time and memory.

b. For analyzing a new environment and building models for threat detection, using Graphs and Deep Learning may not be feasible due to lack of suitable volume of data available for processing.

c. It will also get complicated to identify which features were ultimately used to train the model.

4. Justification for selecting a particular methodology for completing the tasks

a. We want to study how different optimization algorithms help in high dimensional network dataset.

b. Thus, it will help to leverage their efficiency for feature selection which often gets overlooked when alternative and highly complex model building processes are used which handle the features themselves.

c. The method will also help us compare how the optimal subset of features over-lap and differ in each algorithm.

d. Using optimization algorithms for feature selection will enable us to train a less complex model, which will result in less overfitting.

e. Thus, in a relatively new or unknown network environment, this approach will enable to build an accurate and efficient Machine Learning model for high dimensional dataset, which can later be used as a baseline for more advanced and complex models.

5. Brief discussion on the Project Work methodology

a. Data Collection: -

   i. Identifying different open-source environments where network datasets which are labelled can be fetched.

   ii. Initial analysis of dataset to understand its features and usage for the project.

   iii. Finalization of the dataset having required attributes for the project.

b. Data Pre-processing: -

   i. Handling missing values in each field.

   ii. Handling of categorical features.

   iii. Scaling and normalization of the data.

c. Exploratory Data Analysis

   i. Data visualization for understanding different features.

   ii. Understanding distribution of target feature labels.

   iii. Identifying relation between different independent features and target feature.

d. Feature Engineering

   i. Creating new features (if required) that may help to train the model.

   ii. Handling the imbalanced nature of dataset with suitable approaches.

e. Feature selection

   i. Using different optimization approaches for selecting best subset of features in the dataset for training two models: -

      1. Binary model: To differentiate an event between normal and attack.

      2. Multi-class model: To identify the type of attack based on characteristic of the event.

   ii. Storing the list of features from each approach.

f. Training the model

   i. Training two models: Binary model and Multi-class classification model based on subset of features obtained from feature selection.

g. Evaluation of the model

   i. Using the following metrics to compute the results of each model: -
      1. Accuracy
      2. Precision
      3. Recall
      4. F1-Score
      5. Confusion matrix

h. Comparison of results

   i. Compare the outcomes of different optimization approaches and rank them.

ii. Fetch the subset of features obtained from each optimization approach and note the similarities and differences.

6. Benefits derivable from the work

a. By using optimization algorithms for feature selection, we can improve the performance of machine learning models.

b. Network datasets have high dimensionality, thus, with optimization algorithms, we can reduce the dimensionality for faster training of the models.

c. This enables us to build efficient models in unknown environment with restricted computation resources and time.

d. By building an effective model to differentiate between events and identify attacks, we can contribute to improve the process of handing cybersecurity issues.

e. With effectively trained model, we can reduce false positives, which is often the biggest concern in real-world process, leading to noisy and false alerts.

f. By building on a large static network dataset, we can understand how to effectively build models over large enterprise networks.

1. **Broad Area of Work**

    a. AI and threat intelligence are important components of Cybersecurity.

    b. User Behavior Analytics (UBA) has become vital for analyzing events and identifying potential threats.

    c. In usual network, around 10% of the events may be anomalous, but only 0.01% of events potential threats. Thus, it becomes extremely important to have an efficient system which helps to get actual alerts and mitigates noisy and false alerts.

    d. The output of Machine Learning models helps to contribute in advanced analytics and add value on top of existing security tools.

2. **Objectives**

    a. To build effective and efficient Machine Learning models on network datasets for predicting if the given event is an attack or normal activity, if it is an attack, then identifying the type of attack.

    b. To use optimization algorithms for feature selection.

3. **Scope of Work**

    a. The dataset used will be a network dataset with labelled target features.

    b. For feature selection, optimization algorithms will be used.

    c. Two types of Machine Learning models will be trained: -

        i. To differentiate an event between normal and attack.
        ii. To identify the type of attack among the ones observed in the dataset.

Detailed Plan of Work (for 16 weeks)

Serial Number of Task	**Tasks or subtasks to be done**	**Start Date - End Date**	**Planned duration in weeks**	**Specific deliverable in terms of the project**
1	Research about cybersecurity use cases which align with work of current employer.	25 Nov 2024 - 08 Dec 2024	2	1. Identifying different use cases and problem statements.

				2. Fetching prior research papers, whitepapers and understanding the work done around those use cases. 3. Determining the limitations of existing research and lesser explored areas, which can be addressed as part of the dissertation.
2	Searching for different open-source datasets and analyzing them to see if they can be used for the project. Since the project will be built on personal system, its important to analyze the dataset and its properties to understand the feasibility of using it.	09 Dec 2024 - 15 Dec 2024	1	1. Fetching list of different open-source datasets that are applicable for the project's use case. 2. Preliminary analysis of datasets such as: - 2.1 File size 2.2 Number of records 2.3 Number of files 2.4 Format of files 3. Finalization of the dataset to be used for dissertation.
3	Data preprocessing and Data visualization	16 Dec 2024 - 22 Dec 2024	1	1. Clean the dataset by handling missing values. 2. Convert all categorical columns into numerical. 3. Analyze the features using visuals and derive important inferences from the dataset. 4. Determine and implement if we need

				to perform oversampling, under-sampling or any other method to handle imbalanced nature of target variable in the dataset.
4	Feature engineering	23 Dec 2024 - 29 Dec 2024	1	Extract new features from existing ones which may be useful to train the model.
5	Research about usage of optimization algorithms for feature selection	30 Dec 2024 - 05 Jan 2025	1	1. Document different optimization algorithms and their usage. 2. Understand how the optimization algorithms are used for feature selection.
6	Research different classification algorithms	06 Jan 2025 - 12 Jan 2025	1	1. We need to build two types of models: - 1.1 Binary classification model 1.2 Multi-class classification model Thus, we need to determine suitable algorithms for each type of model and the reason for selecting the algorithm.
7	Feature selection	13 Jan 2025 - 26 Jan 2025	2	1. Use optimization algorithms to fetch optimal subset of features from each algorithm.
8	Training the models	27 Jan 2025 - 09 Feb 2025	2	1. Build Machine Learning models by using different subset of features
9	Evaluation of models	10 Feb 2025 - 16 Feb 2025	1	1. Evaluate each model using metrics such as: - 1.1 Accuracy 1.2 Precision

				1.3 Recall 1.4 F1-Score 1.5 Confusion matrix
10	Documentation of results	17 Feb 2025 - 02 Mar 2025	2	1. Compare the results obtained by models built with different optimization algorithms used for feature selection. 2. Compare the subset of features obtained by different optimization algorithms. 3. Based on results observed from prior research in literature survey, compare the results of our project
11	Review and corrections (if needed)	03 Mar 2025 - 15 Mar 2025	2	This is buffer time which can be utilized if required for any extra activities that need to be carried out or to complete the tasks that we may have not been able to complete as per the timeline.

4. Literature References

a. Kapil Sinha, Arun Viswanathan, Julian Bunn “Tracking temporal evolution of network activity for botnet detection”, in arXiv:1908.03443, 2019

b. Akinyemi Moruff Oyelakin, Rasheed Gbenga Jimoh “Tree-Based Learning models for botnet malware classification in real world sub-sample dataset”, in Innovative Computing Review, 2024

c. Abdulghani Ali Ahmed, Waheb A Jabbar, Ali Safaa Sadiq “Deep Learning-Based Classification Model for Botnet attack detection”, in Journal of Ambient Intelligence and Humanized Computing, 2020

d. Dandy Pramana Hostiadi, Tohari Ahmad, Muhammad Aidiel Rachhman Putra, Gede Angga Pradipta, Putu Desiana Wulaning Ayu, Made Liandana “A new approach of botnet activity detection models using combination of Univariate and ANOVA Feature Selection Techniques”, in International Journal of Intelligent Engineering & Systems, 2024

e. Songhui Ryu “Comparison of Machine Learning algorithms and their ensembles for Botnet Detection”, in Prude University, 2018

f. Javier Velasco-Mata, Victor Gonzalez-Castro, Eduardo Fidalgo Fernandez, Enrique Alegre “Efficient Detection of Botnet Traffic by Features Selection and Decision Trees”, in IEEE, 2021

g. Anand Ravindra Vishwakrma “Network Traffic based botnet detection using machine learning”, in SJSU ScholarWorks, 2020

h. V. Krishna Sahithi, R. Jyothika, S. Preethi, Dr. A.R, Siva Kumaran “BotChase: Integrated Unsupervised Learning with Decision Tree classifier for Graph-Based Bot detection”, in Turkish Journal of Computer and Mathematics Education, 2023

i. Rany ElHousieny “Leveraging Autoencoders for Anomaly Detection: A Case Study with the KDD Cup 1999 dataset” in Level Up Coding, 2024

j. Duygu Yilmaz Eroghu, Unmut Akcan “An Adapted Ant Colony Optimization for Feature Selection”, in Applied Artificial Intelligence, 2024

k. Ting Caim Chun Ye, Zhiwei Ye, Ziyuan Chen Mangqing Mei, Haichao Zhang, Wanfang Bai, Peng Zhang “Multi-label Feature Selection on Improved Ant Colony Optimization Algorithm with Dynamic Redundancy and Label Dependence”, in Computers, Materials & Continua, 2024

l. Harshit Saxena, Vineet Richaariya “Intrusion Detection in KDD99 Dataset using SVM-PSO and Feature Reduction with Information Gain”, in International Journal of Computer Applications, 2014

m. S. Norwahidayah, Noraniah N. Farahah, Ainal Amirah, N. Liyana, N. Suhana “Performance of Artificial Neural Network (ANN) and Particle Swarm Optimization (PSO) using KDD Cup 99 Dataset in Intrusion Detection System (IDS)”, in Journal of Physics, Conference Series, 2021

n. Ahmed Abdullah Alqarni “Towards support-vector machine-based ant colony optimization algorithms for intrusion detection”, in Springer Nature Journal, 2023

o. S Francisca Rosario, Dr. K. Thangadurai “Simulated Annealing Algorithm for Feature Selection”, in International Journal of Computer & Technology, 2015

Supervisor's Rating of the Technical Quality of this Dissertation Outline

EXCELLENT / GOOD / FAIR / POOR (Please specify): Excellent

Supervisor's suggestions and remarks about the outline (if applicable).

Date 09/12/2024

Prathibha
(Signature of Supervisor)

Name of the supervisor: Prathibha Panduranga Rao

Email Id of supervisor: ppandurangarao@statestreet.com

Mobile # of supervisor: 9620202200