Network Anomaly Detection

Amit Hasan	**170104055**
Tunazzin Rahman Topu	**170104066**
Almas Shahriar Ador	**170104074**

Project Report

Course ID: CSE 4214

Course Name: Pattern Recognition Lab

Semester: Fall 2020

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

September 2021

Network Anomaly Detection

Submitted by

Amit Hasan	**170104055**
Tunazzin Rahman Topu	**170104066**
Almas Shahriar Ador	**170104074**

Submitted To

Faisal Muhammad Shah, Associate Professor

Farzad Ahmed, Lecturer

Md. Tanvir Rouf Shawon, Lecturer

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Department of Computer Science and Engineering

Ahsanullah University of Science and Technology

Dhaka, Bangladesh

September 2021

ABSTRACT

Regarding to the advancement in network information access, the need of its security also arises. To ensure the security of the system, the challenge is to identify malicious activity and intrusion across the network using anomaly-based approaches. Feature Selection (FS) is the important technique, which gives the issue for enhancing the performance of detection. The objective of feature selection is to remove irrelevant and redundant attributes from the dataset to improve the predictive power of a classification algorithm. A number of solutions have been proposed, but further investigation is required to enhance efficiency. In this project, we introduce a filter-based feature selection model for anomaly-based intrusion detection systems. Many existing feature selection methods mainly focus on the relationship among features. However, in network data sets, feature having correlations with class label is vital for selecting useful features. Our approach scores feature pairs based on their correlation with the predicting class. To classify selected features, we apply some basic machine learning algorithms.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Over the last decade, the exponential increase in computer networks and developments, as well as the availability of the internet, has brought many benefits, but it has also become a source of security threats. These cyber-attacks become more attractive and potentially more disastrous as our dependence on information technology increases. Our main objective is to detect anomalies or attacks in the network traffic through classification. With access to more data than ever before, it is even more important to analyze it and interpret it correctly. When it comes to security, finding the outliers and determining if the outlier is a security threat is a must. Anomaly detection is the process of finding patterns in data that don't conform to a model of normal behavior. It is an important tool for detecting fraud, network intrusion, and other rare events that may have great significance but are hard to find. Anomaly detection techniques rely on machine learning which can be used to learn the characteristics of a system from observed data, helping to enhance the speed of detection. When building a machine learning model, it's almost rare that all the variables in the dataset are useful to build a model. Adding redundant variables reduces the generalization capability of the model and may also reduce the overall accuracy of a classifier. Furthermore, adding more and more variables to a model increases the overall complexity of the model. That is why the idea of feature selection is considered. Feature selection finds a subset of features to improve classification accuracy. Feature selection is a technique that allows us to choose those features in any data that contribute most to the target variable Instead, features are selected on the basis of the principal characteristics of the training data, distance measure, correlation measures, consistency measures and information measure. In this project, we are using the most up to date IDS dataset, the CICIDS2018 which simulates a real-world environment and is explained in detail further on. In our proposed methodology, Filter methods are used as a preprocessing step to remove constant and redundant features and univariate feature selection is used to select the best features based on univariate statistical tests. We have applied the SelectKBest algorithm to select the highest k scoring features.

Chapter 2

Literature Reviews

2.1 K- Nearest Neighbors.

K- Nearest Neighbor or KNN is one of the simplest classification algorithms which is based on feature similarity. Despite its simplicity it has proven to be incredibly effective at certain tasks in world of machine learning. It can also be used for regression problems. But it is more widely used in classification problem. It is a supervised learning algorithm because it is used for labeling an unknown data point in existing labeled data. K in KNN is a parameter that refers to the number of nearest neighbors to include in the voting process. For implementing the algorithm, we need to follow the steps. Firstly, the algorithm Select the number K of the neighbors then calculate the Euclidean distance of k number of neighbors. The next step is to take the K nearest neighbors as per the calculated Euclidean distance. Among these k neighbors, count the number of the data points in each category. Now assign the new data points to that category for which the number of the neighbor is maximum. The Euclidean distance is measured by the following formula:

d(p,q) = ($\sum_{i=1}^{n}$)*(qi-pi)*(qi-pi)

2.2 Decision tree.

In a decision tree, for predicting the class of the given dataset, the algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree. At first Begin the tree with the root node, says S, which contains the com-

plete dataset. Then, find the best attribute in the dataset using Attribute Selection Measure (ASM). Now divide the S into subsets that contains possible values for the best attributes. After that the decision tree nodes are generated, which contains the best attribute. Finally, make new decision trees using the subsets of the dataset created recursively. Continue this process until a stage is reached where you cannot further classify the nodes and called the final node as a leaf node.

2.3 Network Anomaly Detection and Identification Based on Deep Learning Methods.

Mingyi Zhu et. al. [2] have proposed a deep learning model that detects anomaly using Feed Forward Neural Network (FNN) model and Convolutional Neural Network (CNN). The model has been de- signed to evaluate the performance of FNN and CNN for five-category classification, such as Normal, DoS, R2L, U2R and Probe. The experimental result showed that for FNN with one hidden layer and 100 nodes achieved detection accuracy of 80.34percenr. On the other hand, for two hidden layers with 80 nodes in 1st hidden layer and 60 nodes in 2nd hidden layer got detection accuracy of 80.3 percent. The detection accuracy they have observed for the CNN model is 77.8 percent as enough layers are not designed for CNN due to the hardware limitation.

2.4 A Machine Learning Based Intrusion Detection System for Software Defined 5G Network.

Jiaqi Li et. al. [3] have proposed an intelligent intrusion detection system for Software Defined 5G networks. The model has three layers: Forwarding Layer, Management Control Layer and Data Intelligence layer. It is designed to evaluate the performance for five-category classification, such as normal, DoS, R2L, U2R and Probe. The experimental result showed that classifications with cross validation significantly boosts the accuracy of detecting U2R and R2L attacks as well as gently promoting the rate of other categories. The main types of attacks mentioned above are subdivided into 39 small classes, 17 of which only appear in the model they came up with. So, it shows that the proposed classifier lacks a generalization ability to detect various attacks without previous training.

2.5 CorrCorr: A feature selection method for multivariate correlation network.

anomaly detection techniques. Florian Gottwalt et. al. [4] have proposed a new feature selection technique, CorrCorr, for multi- variate correlation NAD. The method generates feature correlations and stores them in a matrix which is then transformed into a vector and to detect anomalous traffic, the Mahalanobis distance is used. The features selected with CorrCorr delivered significantly better results, with an accuracy close to or above 0.95. Even on the 100k and 150k samples, where the performance of the original features and PCA decreased drastically. CorrCorr with all features, including sttl and ct-statettl and have achieved the best results with an accuracy of 98.65percent.

2.6 Feature Selection Algorithms in Intrusion Detection System: A Survey.

Sofiane Maza et. al. [5] have proposed latest well-known FS algorithms for Intrusion Detection. System (IDS) which are developed to select the best feature subsets. They classify the algorithms into five approaches according to the techniques have been integrated with them which are: Deterministic Algorithms, Intelligent Patterns, Artificial Neural Networks, Fuzzy Rough Set and Swarm Intelligent. Feature selection algorithms which used filter approach has lower time complexity than the wrapper approach which is considered higher, because a filter is based on distance, information, dependency, and consistency measures instead of the wrapper which is based on classifiers error rate. They focus on SVM, DT, BN, k-means, Cuttlefish, Immune Artificial System (IAS) and evolutionary algorithm which is specified on Genetic Algorithms. In these researches, they use reduction techniques with intelligent algorithms to achieve the best features at the minimum time. Among reduction techniques are used with feature selection approaches in this section are: PCA (Principal Component Analysis), LDA (Linear Discriminant Analysis), ICA (Independent Component Analysis) and GPC (Genetic Principal Component).

2.7 A Feature Selection Method for Anomaly Detection Based on Improved Genetic Algorithm.

Shi Chen et. al. [6] have proposed an improved genetic algorithm-based feature selection method is proposed to obtain optimal features subset with not only considering the perfor-

mance of classifier but the features generation costs. The cuttlefish algorithm (CFA) was firstly applied as a feature selection method for IDS. After the optimal features subset was obtained, a decision tree (DT) classifier was used as a classifier to train and test the dataset with algorithm (GA) was adopted as a search strategy to obtain an optimal subset of features. GA is known as a perfect algorithm for solving optimization problems. Here applied the crowding mechanism-based niche technology to improve the global optimization ability and convergence speed of GA, and adopt OWNN classifier as an evaluator to improve the classification performance with the selected features subset.

Chapter 3

Data Collection & Processing

3.1 Description of Dataset.

The University of New Brunswick originally generated the CICIDS2018 dataset for the purpose of evaluating attacking data. The final dataset includes some different attack scenarios: Brute-force, Heartbleed, Botnet, DoS, DDoS, Web attacks, and infiltration of the network from inside. In our dataset consists of a total of 80 features. We are working on an imbalanced dataset. It will give biased or overfitted results. So, we want to balance the dataset using smote technique. Then after some further preprocessing, we will get our dataset on which we can apply our machine learning model.

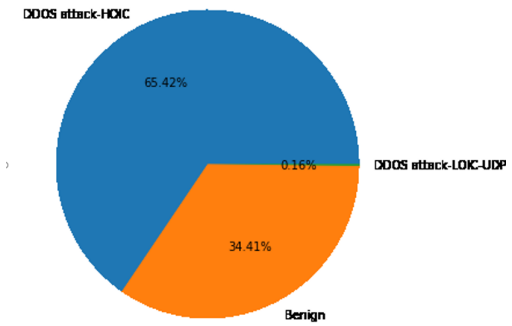

Figure 3.1: Attacks found in Imbalanced Dataset.

The imbalance in the class distribution may vary, but a severe imbalance is more challenging to model and may require specialized techniques. There are many techniques like Use the right evaluation metrics, Re-sample the training set, Use K-fold Cross-Validation in the right way, Ensemble different re-sampled datasets, Re-sample with different ratios, Cluster the abundant class and other so on. We are using Synthetic Minority Oversampling Technique, or SMOTE for short. Oversampling the examples in the minority class is one technique to tackle this problem. This can be accomplished by simply duplicating minority class samples in the training dataset before fitting a model. This can help to balance the class distribution,

but it doesn't give the model any extra information. Synthesizing new instances from the minority class is an improvement over replicating examples from the minority class. This is a sort of data augmentation that works well with tabular data.

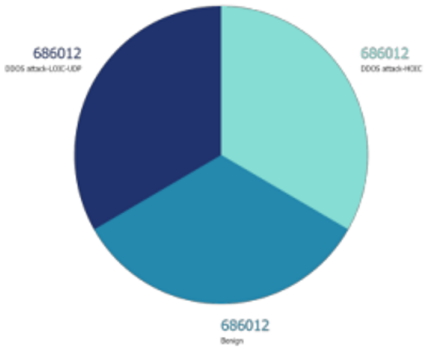

Figure 3.2: Attacks found in Balanced Dataset.

3.2 Preprocessing

The first step of preprocessing is to clean up the dataset, after which the null and infinite values were handled. Then we encoded categorical data using label encoder. The dataset is split into two separate categories: independent data and dependent data. Using Imputer method, we handled missing values. Then we split the dataset into train data (80percent) and test data (20percent). We applied feature scaling to standardize the independent features present in the dataset into a fixed range. Feature scaling is applied to handle highly varying magnitudes or units. We used the standardization technique to scale the features into a limited range. It is a very effective technique which re-scales a feature value so that, it has distribution with 0 mean value and variance equal to 1. fig. 3.3 shows the preprocessing procedure. The first step of preprocessing is to clean up the dataset, after which the null and infinite values were handled. Then we encoded categorical data using label encoder. The dataset is split into two separate categories: independent data and dependent data. Using Imputer method, we handled missing values. Then we split the dataset into train data and test data. We applied feature scaling to standardize the independent features present in the dataset into a fixed range. Feature scaling is applied to handle highly varying magnitudes or units. We used the standardization technique to scale the features into a limited range. It is a very effective technique which re-scales a feature value so that, it has distribution with 0 mean value and variance equal to 1.In addition, features which have constant values have been removed using constant feature removal method and duplicate feature removal method is used to remove the features having duplicate values.

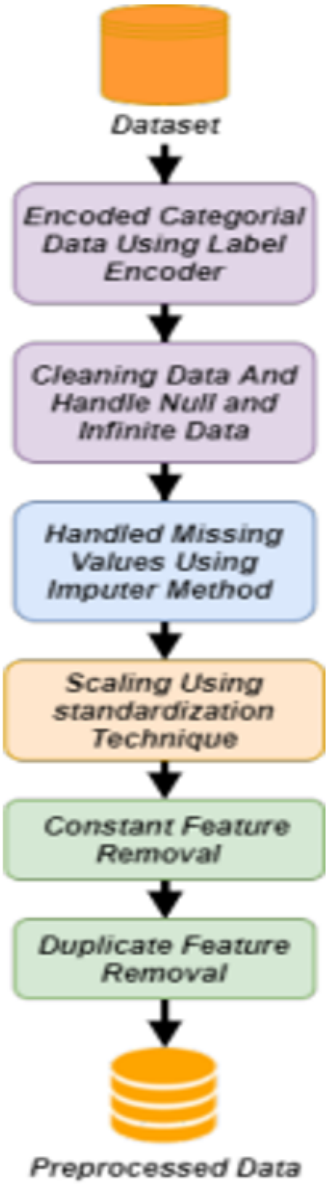

Figure 3.3: Dataset Preprocessing.

Chapter 4

Methodology

4.1 Features Selections

For our project we used some selected features from the datasets. Features selection process followed by some basic preprocessing algorithms.

4.2 SelectKBest Algorithm.

This algorithm has three basic and important benefit in feature selection. Reduces Overfitting: Less redundant data means less possibility of making decisions based on redundant data or noise. Improves Accuracy: Less misleading data means modeling accuracy improves. Reduces Training Time: Less data means that algorithms train faster. SelectKBest is a module that selects k feature that has the highest score. The score is calculated based on univariate statistical analysis, which is an analysis of variables one by one. As per domain standard, the dataset is split into two parts: one used to train the algorithm and the other used to test it. Since the algorithms need a lot of data samples to converge, the split is not usually uniform. The algorithm measures the correlation between the independent variables and the dependent variable using one of the provided metrics.

4.3 fclassif () function.

The scikit-learn machine library provides fclassif () function. This function can be used in a feature selection strategy, such as selecting the top k most relevant features (largest values) via the SelectKBest class. SelectKBest class can be defined to use the fclassif () function and select all features, then transform the train and test sets.

4.4 Training Datasets with ML models.

Two basic machine learning algorithms (Decision Tree and KNearestNeighbor) are trained with 80 perent of data to obtain the model. Furthermore, the accuracy and F1score are calculated from test data. Firstly 61 features are used to fit those models. After using SelectKBest algorithm, features reduced to 20. In both case, accuracy and f1score was the same.

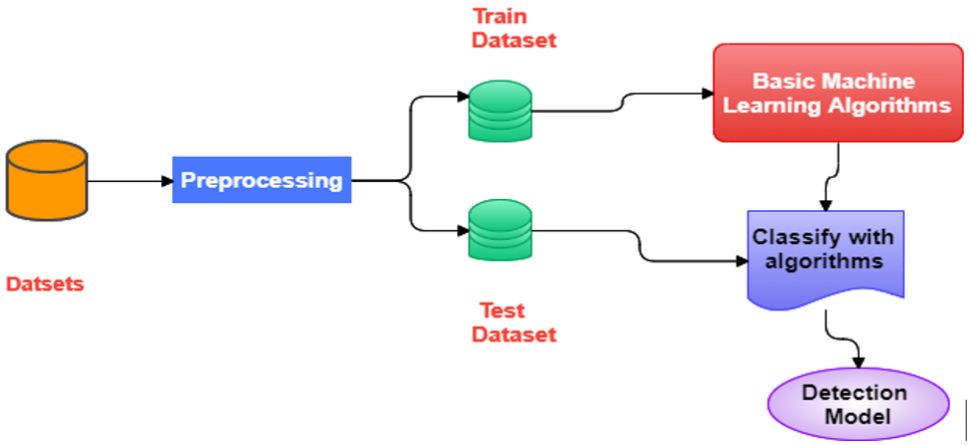

Figure 4.1: Overview of methodology.

Chapter 5

Experiments and Results

In this phase, we will show the progress that has made so far. As, we preprocessed the dataset with some algorithm which refines the dataset by cleaning, deducting nan value, constant features, so our results are good enough and computational time was relatively low. Accuracy is one metric for evaluating classification models. Informally, accuracy is the fraction of predictions our model got right. Formally, accuracy has the following definition: Accuracy=(Number of correct predictions)/(Total number of predictions) For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

Accuracy=(TP+TN)/(TP+TN+FP+FN)

Where TP = True Positives, TN = True Negatives, FP = False Positives, and FN = False Negatives.

And The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is:

F1 = 2 * (precision * recall) / (precision + recall)

Where Precision =TP/(TP+FP) and Recall= TP/(TP+FN).

In the K- Nearest Neighbous Classifier we got ,

Accuracy = 0.96 and F1-Score = 0.97.

In the Decision Tree Classifier we got ,

Accuracy = 0.96 and F1-Score = 0.97.

Chapter 6

Future Work and Conclusion

In this project we use KNN and Decision Tree which are the basic machine learning classification algorithms. In future, we will develop our model using some other machine learning algorithms such as Artificial Neural Network (ANN), Support Vector Machine (SVM) and compare the result with KNN and Decision Tree. Some tuning can be applied in our model for more accuracy. There are lots of datasets preprocessing algorithms for cleaning the datasets. We will use them also to minimize the computational cost by reducing the irrelevant features. Increasingly complex IT infrastructures and applications represent a growing risk for cyber- attacks. To address this risk, extensive research has been performed in cyber security, particularly on network anomaly detection, with a current focus being on multivariate correlation techniques. We think our scope of exploring the project is low as we use the basic machine learning algorithms, though our results are quite good.

Chapter 7

References

[1] CSE-CIC-IDS2018 a collaborative project between the communications security establishment (cse) the canadian institute for cybersecurity (cic).https://www.unb.ca/cic/datasets/ids-2018.html?fbclid=IwAR18Ngt-P9pndGbJMMCziqGZ1X1WrViQ0VAxQvxtn864pP9pM2HbOmyisA,2.

[2] Kejiang Ye Mingyi Zhu and Cheng-Zhong Xu. Network Anomaly Detection and Identification Based on Deep Learning Methods. Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, 2018.

[3] Rongpeng Li Jiaqi Li, Zhifeng Zhao. A Machine Learning Based Intrusion Detection System for Software Defined 5G. The Institution of Engineering and Technology, 2015.

[4] Tharam Dillon Florian Gottwalt, Elizabeth Chang. CorrCorr: A feature selection method for multivariate correlation network anomaly detection techniques. University of New South Wales, Canberra, Australia . La Trobe University, Melbourne, Australia, 2019.

[5] Muneeswaran K Selvakumar B. Firely algorithm based Feature Selection for Network Intrusion Detection. Computers Security., 2018.

[6] Zhen Zuo Xiaojun Guo Shi Chen, Zhiping Huang. A Feature Selection Method for Anomaly Detection Based on Improved Genetic Algorithm. College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha, 410073, China, 2016.

Generated using Undegraduate Thesis LaTeX Template, Version 1.4. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This project report was generated on Sunday 26th September, 2021 at 8:58am.