Optimization Algorithms For Intrusion Detection Systems

Kunal Kwatra, 19BCE1784
B.TECH - CSE
VIT CHENNAI
Tamil Nadu, India

Hasnain Sikora, 19BAI1072
B.TECH - CSE
VIT CHENNAI
Tamil Nadu, India

Abstract

In current times, a vast number of people are vulnerable to security breaches but do not know how they can avoid them. Currently, there are various Intrusion Detection System and Intrusion Prevention System known, used, and actively being built. For any big company, common user, or vendor, a problem arises when they want to decide upon which IDS/IPS should be implemented on their network or host as per their requirements and the features of the IDS/IPS. These attacks are done using the network of bots (networked devices infected with malware) called the botnet. Once the attacker gets the access to the botnet, he can flood the network of target using the IP address of the systems in botnet, resulting in interruption in the working of the victim system/network/server. And since each bot has a legitimate IP address, it becomes difficult todifferentiate between an attack and a normal access. This project addresses this issue and proposes potential solutions to implement security systems and guide users through the procedure involved. It aims at recognition of the attack in the network flow by analyzing various attributes of the incoming network. For the classification ofthe network various classification algorithms are being used and their resultsare analyzed on the basis of the difference in their accuracy. This project covers some well known algorithms and some of the optimization algorithms used in the field of intrusion detection in past ten years such as XGBoost, LightBGM, Genetic Algorithm (GA), and Particle Swarm Optimization (PSO), Random Forest, Decision Tree, Naive Bayes, Ensemble Algorithm, Conditional Variational Autoencoder .

***Keywords*—**Intrusion Detection, Algorithms, KDD-99

1. INTRODUCTION

This project aims to provide useful insights about the intrusion detection literature and is a good source for anyone interested in applying one of the used optimization algorithms in the field of intrusion detection. The dataset used for the experiments carried out in this paper is a standard benchmark dataset known as KDD Cup 99. It was investigated and utilized to study the effectiveness of the proposed method in this problem domain. We also look at and analyse the other algorithms mentioned and compare the results obtained. This process also should facilitate any study in the field of Optimization Algorithms, not just for Intrusion Detection but also any other neural network or machine learning model in general. Here, we first look at Genetic Algorithm (GA) as a tool capable of identifying malicious connections in a computer network. Genetic Algorithms (GA) are search algorithms which function based on the principles of natural selection and genetics. GA develops a population of initial individuals to a population of high quality individuals, where each individual signifies a solution of the problem to be solved. Particle swarm optimization (PSO) is one of the bio-inspired algorithms and it is a simple one to search for an optimal solution in the solution space. It is different from other optimization algorithms in such a way that only the objective function is needed and it is not dependent on the gradient or any differential form of the objective. It also has very few hyperparameters.

Before we proceed to understand optimization techniques that the IDS may use, we need to understand what type of breaches we will be dealing with. An intrusion can be defined as any set of actions attempting to compromise the integrity, confidentiality, or availability of a resource. We have established the importance of having an IDS/IPS, but many people and organizations still do not seem to be taking any active measures in this field. Many organizations dealing in e-business and online modes of transacting require to have a tight security system as any downtime caused by intrusions can cost an of revenue. But there are many companies, both large scale or small scale, that do not see security as an important issue. This is due to the expense of implementing a security system, as well as a general lack of awareness. Subsequently, there is a need to understand and compare different security systems available by summarizing the advantages, disadvantages, and requirements.

An Intrusion Detection System (IDS) is a network security technology originally built for detecting vulnerability exploits against a target application or computer.

1) Network Intrusion Detection System (NIDS): Network intrusion detection systems (NIDS) are set up at a planned point within the network to examine traffic from all devices on the network.

2) Host Intrusion Detection System (HIDS): Host intrusion detection systems (HIDS) run on independent hosts or devices on the network. An example of HIDS usage can be seen on mission-critical machines, which are not expected to change their layout.

3) Protocol-based Intrusion Detection System (PIDS): Protocol-based intrusion detection system (PIDS) comprises a system or agent that would consistently reside at the front end of a server, controlling and interpreting the protocol between a user/device and the server.

4) Application Protocol-based Intrusion Detection System (APIDS): An application Protocol-based Intrusion Detection System (APIDS) is a system or agent that generally resides within a group of servers.

5) Hybrid Intrusion Detection System: A hybrid intrusion detection system is made by the combination of two or more approaches to the intrusion detection system.

Intrusion Prevention Systems (IPS) extended IDS solutions by adding the ability to block threats in addition to detecting them and has become the dominant deployment option for IDS/IPS technologies.

A firewall is a network security device that monitors incoming and outgoing network traffic and decides whether to allow or block specific traffic based on a defined set of security rules. Firewalls have been the first line of defense in network security for over 25 years.Malware is an umbrella term that describes all forms of malicious software designed to wreak havoc on a computer. Common forms include viruses, trojans, worms, and ransomware.A virus is a type of malware aimed to corrupt, erase or modify information on a computer before spreading to others. However, in more recent years, viruses like Stuxnet have caused physical damage.Trojan horse is a piece of malware that often allows a hacker to gain remote access to a computer through a "back door".

A type of software application or script that performs tasks on command, allowing an attacker to take complete control remotely of an affected computer. A collection of these infected computers is known as a "botnet" and is controlled by the hacker or "bot-herder".An acronym that stands for distributed denial of service – a form of cyber attack. This attack aims to make a service such as a website unusable by "flooding" it with malicious traffic or data from multiple sources (often botnets).

In context of Information Security, intrusions can be defined as activities that break and violate the security policy of that system, intrusions can be identified by intrusion detection. Since the advent of the internet, intrusion detection systems (IDS) are one of the most important types of the security software that have been used to deal with the intrusions. Intrusion detection systems are among the necessary systems within the information security system.

IDS are hardware or software that observe the processes of the computer network, waiting for any violation of network management policies, or monitors any change such as modification, files addition, or files deletion on the host device. Intrusion is indicated as any types of unauthorized activity that results in corrupted to the information system. Whereas, any attack that poses a potential threat to the integrity and confidentiality of the information is considered intrusion. For example, When computer services do not respond to legitimate users. In the past and to this day, Cyber criminals have focused on stealing from banks and credit card customers or robbing bank account. Therefore, it is of vital importance to have IDS for detecting various attacks. The goal of IDS is identifying all sorts of different attacks as soon as possible, which a traditional firewall cannot be achieved. Besides, to distinguished between the system activities which are normal and behaviors which be classified as suspicious or intrusions.

2. Literature Review

In [1], The authors proposed a methodology for the improvement of a generalized ML-based model for the detection of DDoS attacks. After exploring a range of attributes of the dataset chosen for this study, they suggested an integrated feature selection (IFS) technique which consists of three ranges and integration of two unique methods, that is, filter and embedded techniques to choose features that especially make contributions to the detection of a range of kinds of DDoS attacks. They used a light gradient boosting machine (LGBM) algorithm for training the model for the classification of benign and malicious flows.

In [2], Saikat Das, Deepak Venugopal and Sajjan Shiva suggested ensemble unsupervised ML method to implement an intrusion detection system which has the good accuracy to discover DDoS attacks. The aim of this study is to expand the DDoS attack detection accuracy whilst reducing the false positive rate. The NSL-KDD dataset and twelve characteristic features from current research are used for experimentation to evaluate our ensemble results with those of our individual and different present models.

Authors in [3], addresses the prediction of application-layer DDoS attacks in real-time with exceptional machine mastering models. They utilized the two machine learning strategies Random Forest (RF) and Multi-Layer Perceptron (MLP) thru the Scikit ML library and big data framework for the detection of Distributed Denial of Service (DDoS) attacks. In addition to the detection of DDoS attacks, They optimized the overall performance of the model by way of minimizing the prediction time as in contrast with different available processes using the big data framework. They accomplished a mean accuracy of 99.5% of the models both with and without big data approaches.

In [4], Authors recommend DDoSNet, an intrusion detection system towards DDoS attacks in SDN environments. Their technique is based totally on Deep Learning (DL) technique, combining the Recurrent Neural Network (RNN) with autoencoder. They considered the model using dataset CICDDoS2019, which includes a complete range of DDoS attacks and addresses the gaps of the present contemporary datasets. They obtained a considerable enhancement in attack detection, as in contrast to different benchmarking methods. Hence, the model offers notable confidence in securing these networks.

In [5], authors proposed a set of new entropy-based features that assist to become aware of attacks precisely and brought a novel multi classifier system based totally on the proposed set of multiple entropy-based aspects and machine learning classifiers to extend the generality and accuracy of detecting low-intensity and high-intensity DDoS attacks. Experiment results confirmed that approach accomplished greater precision and greater recall values in contrast to countless latest approaches.

3. Algorithms Used

3.1 Naive Bayes

It's a classification method that uses Bayes' Theorem and assumes predictor independence. A Naive Bayes classifier, to put it simply, assumes that the existence of one feature in a class is independent to the presence of any other feature. It's a classification method that uses Bayes' Theorem and assumes predictor independence. A Naive Bayes classifier, to put it simply, assumes that the existence of one feature in a class is independent to the presence of any other feature.

3.2 Decision Tree

A virus is a type of malware aimed to corrupt, erase or moThe Decision Tree algorithm is part of the supervised learning algorithms family. The decision tree approach, unlike other supervised learning algorithms, may also be utilised to solve regression and classification issues. By learning basic decision rules inferred from past data, the purpose of employing a Decision Tree is to develop a training model that can be used to predict the class or value of the target variable (training data). We start at the root of the tree when using Decision Trees to forecast a class label for a record. The values of the root attribute and the record's attribute are compared. We follow the branch that corresponds to that value and go to the next node based on the comparison.

3.3 Random Forest

Random forest is supervised machine learning algorithm that is commonly used to solve classification and regression issues. It creates decision trees from several samples, using the majority vote for classification and the average for regression. One of the most essential characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous and categorical variables, as in regression and classification. For classification difficulties, it produces superior results

3.4 XGBoost

Extreme Gradient Boosting (XGBoost) is a distributed gradient-boosted decision tree (GBDT) machine learning toolkit that is scalable. It is the top machine learning library for regression, classification, and ranking tasks, and it includes parallel tree boosting.

3.5 Ensemble Algorithm

It is an ensemble algorithm combining Naive Bayes, Decision Tree, Random Forest and Xgboost giving the hierarchy of classification and gradient boost by Xgboost along with Random Forest in the lowest layer of execution path.

3.6 Conditional Variational Autoencoder (CVAE)

The CVAE is a conditional directed graphical model in which the input observations modulate the prior on Gaussian latent variables, which produce the outputs. It has been programmed to optimise conditional marginal log-likelihood.

3.7 LightBGM

It is a gradient boosting framework that makes use of tree based learning algorithms. While other algorithms trees grow horizontally, LightGBM algorithm grows vertically meaning it grows leaf-wise and other algorithms grow level-wise. LightGBM chooses the leaf with large loss to grow. It can lower down more loss than a level wise algorithm when growing the same leaf.

3.8 Particle Swarm Optimizatin (PSO)

Particle swarm optimization (PSO) is one of the bio-inspired algorithms and it is a simple one to search for an optimal solution in the solution space. It is different from other optimization algorithms in such a way that only the objective function is needed and it is not dependent on the gradient or any differential form of the objective. It also has very few hyper-parameters.

3.9 Genetic Algorithm

Genetic Algorithm was developed in order to be applied in a computer network. Genetic algorithms mimic the principles involved in evolutionary sciences, such as the concept of natural selection. This method should be able to identify and estimate the differences between the behavior of the unauthorized connection and normal connection using a proposed fitness (objective) function, the better the fitness valIt begins with generating 100 chromosomes randomly. A chromosome in this context is a set of parameters that define a proposed solution to the problem that the genetic algorithm is trying to solve. Next, an attack recognition between generated chromosomes and training data takes place. This is followed by applying the fitness function to measure fitness value. A fitness function simply defined is a function that takes the solution as input and produces the suitability of the solution as the output.ue, the better the solution, much like how the survival of the fittest works in nature.

4. Dataset and Methodology

3.1 Data Used - KDD99

KDD Cup 99 data set enlists 41 features that represent the variables used in a computer network. Owing to the fact that the process of analyzing these variables is time-consuming and requires large-scale computational steps this research focused on the eight most important features with 6 types of attacks. The basis of selecting the 8 features that were selected was not mentioned, but based on the rest of the paper, I believe that a technique like Recursive Feature Elimination was used to identify the most important features. Elaborating on the dataset used, while the KDD99 dataset is over 20 years old, it is still widely used to demonstrate Intrusion Detection Systems (IDS). KDD99 is in fact also the data set used for The Third International Knowledge Discovery and Data Mining Tools Competition, which was held in conjunction with KDD-99 The Fifth International Conference on Knowledge Discovery and Data Mining.

3.2 Pre-processing

Data is processed using variety of techniques including, removing 6 of the attributes, one-hot encoding, normalizing using min-max scalar, conversion of class labels and SMOTE to handle the imbalanced data. Following that data

Figure3.1: Flow of Implementation

3.3 ML Models

After pre-processing, data is fed to different Machine Learning Algorithms giving a range of results demonstrated in the form Accuracy Percentage.

We trained different models using the given dataset, namely – XGBoost, LightBGM, Genetic Algorithm (GA), and Particle Swarm Optimization (PSO), Random Forest, Decision Tree, Naive Bayes, Ensemble Algorithm, Conditional Variational Autoencoder. In Ensemble, I trained the data using Decision Tree, XGBoost, Naïve Bayesand concatenated the prediction of all the algorithms with the testing data. Then trained the Random Forest Algorithm using the newly formed dataset.

In Evaluation, testing data was fed to the trained models and the achieved accuracies were compared to find the best classification algorithm. In Result, the comparison of the trained models is presented according to their calculated accuracies.

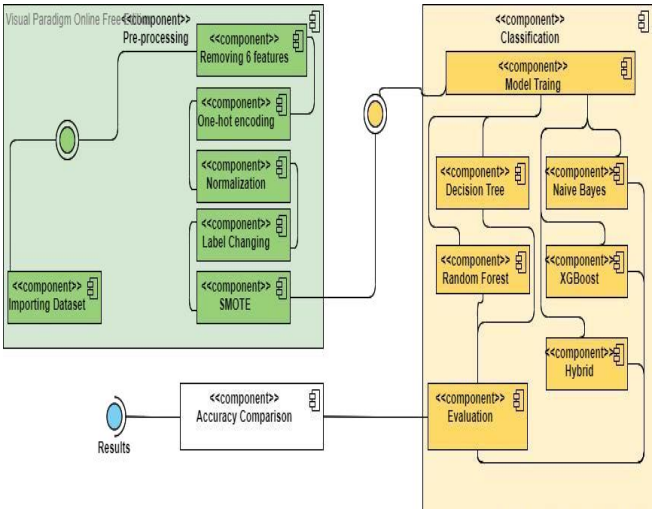


Figure3.2: Component Diagram

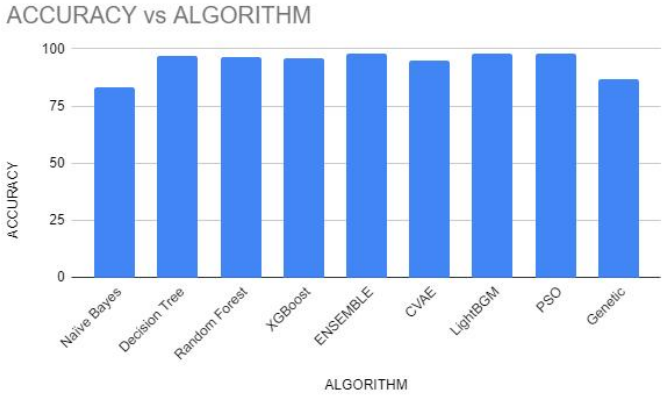
5. **Results and Discussion**

The comparison between the results of the algorithms used is done on the basis of their accuracy score attained after the implementation.The accuracy of the algorithms is as follows:

ALGORITHM	**ACCURACY**
Naïve Bayes	(78 – 83)%
Decision Tree	Approx.(97 %)
Random Forest	Approx.(96.5)%
XGBoost	Approx.(96-97)%
ENSEMBLE	Approx.(98 %)
CVAE	Approx(94-96)%
LightBGM	Approx(98-99)%
PSO	98%
Genetic	87%

As shown in Graph 5.1, Ensemble and LightBGM got rhe highest Accuracy for the used KDD-99 dataset.

Other algorithms like Naive bayes and Genetic got the accuracy better than the most available IDS and along with them CVAE with the average accuracy among the used algorithms also helps in the regeneration of data features providing an extra benefit especially in the case of Data Loss or Data leakage.



GRAPH5.1: ACCURACY- ALGORITHM

6. **References**

[1] Murk Marvi,Asad Arfeen,Riaz Uddin, “A generalized machine learning- based model for the detection of DDoS attacks”, 2020 doi: https://doi.org/10.1002/nem.2152

[2] Saikat Das, Deepak Venugopal, Sajjan Shiva, “A Holistic Approach for Detecting DDoS Attacks by Using Ensemble Unsupervised Machine Learning”, 2020 Advances in Information and Communication, 2020, Volume 1130, doi: https://doi.org/10.1007/978-3-030-39442-4_53

[3] Awan, M.J.; Farooq, U.; Babar, H.M.A.; Yasin, A.; Nobanee, H.; Hussain, M.; Hakeem, O.; Zain, A.M. Real-Time DDoS Attack Detection System Using Big Data Approach.Sustainability2021,13,10743. https://doi.org/10.3390/su131910743

[4] M. S. Elsayed, N. -A. Le-Khac, S. Dev and A. D. Jurcut, "DDoSNet: A Deep- Learning Model for Detecting Network Attacks," 2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM),2020,pp.391-396, doi:10.1109/WoWMoM49955.2020.00072.

[5] A. Koay, A. Chen, I. Welch and W. K. G. Seah, "A new multi classifier system using entropy-based features in DDoS attack detection," 2018 International Conference on Information Networking (ICOIN), 2018, pp. 162-167, doi: 10.1109/ICOIN.2018.8343104.