Analysis of Ensemble and Hybrid Approaches for Intrusion Detection

An Intermediate project report submitted in partial fulfillment of the requirements for B.Tech. Project

B.Tech.

by

Abhinav Jaiswal (2016IPG-004)
Mohit Kumar (2016IPG-053)
Surendra Singh Gangwar (2016IPG-107)

विश्वजीवनामृतं ज्ञानम्

ABV INDIAN INSTITUTE OF INFORMATION TECHNOLOGY AND MANAGEMENT
GWALIOR-474 010

2019

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Background

In recent years, there has been a revolutionary change in the field of networking, use of the internet is growing day by day and so is the risk of intrusion. Nowadays, it is very crucial to have a secure network because of increasing dependability on the internet and to achieve high-level security [5]. One of the possible ways is through the intrusion detection. Intrusion detection is an efficient technique, which aims to identify various network attacks and further classify these network attacks into several categories. Presently, the web is broadly utilized in various organizations and the business prerequisites have caused enterprises to convey their very own data frameworks on the web [4]. Due to large dependability on an information system, it became a major bottleneck, and thus the primary target for the intruders.

Nowadays, with the evolution of new advances like cloud computing, Big data, Internet of things, block chain, etc., which generates the enormous amount of network traffic, it becomes almost impossible to analyze the data and detect the intrusion in time. With the advancement in the machine learning (ML) domain, ML techniques enabled an intrusion detection system (IDS) results in less error rate and more accurate to classify various attacks in less possible time by eliminating the data which are of no use [1].

In this work, we investigate and evaluate two different hybrid approaches, ensemble methods and mixture of experts, which take advantage of various machine learners with the aim to give higher and accurate prediction performance that can not be achieved by using a single classifier [2].

1.2 Motivation

The methods used for intrusion detection have proven to be advantageous but classification of different intrusion attack type efficiently is still a major concern.Previous work performed by Chia-Ying Lin in year 2009 [2] includes classification of intrusions using different machine learning algorithms.

The ability of different machine learning algorithms to accurately detect a particular attack type,drives a motivation to use them to build a hybrid model which is cable of predicting all attack type precisely with low bias and low variance.

The methodology proposed by S. Reza [11] in 2014 describes how Mixture-of-experts technique can be used with various experts that can out-stand many single machine learning algorithm for multiclass classification.

1.3 Literature Survey

1.3.1 Analysis of Machine learning approaches for intrusion detection

Many different types of machine learning approaches has been implemented to classify the intrusion, some has used single learning techniques such as Support Vector Machine(SVM) ,Logistic Regression(LR) , K-nearest Neighbors(KNN), Artificial Neural Network (ANN) [2] and some uses multilevel techniques in which they first use some nature inspired algorithm and genetic algorithm to do feature selection [1] and then used single classifier to predict the result,enormous works has been done in the area of classification of intrusion detection but most of them are only able to binary classify the attacks into normal or abnormal. Only few work has been done to classify the attack type with less accuracy rate.

It is very important to classify the attack type so that proper measure can been taken to built a defence against it.It is important to quickly identify the intrusion at runtime which can only be possible if we identify the most important features of dataset and reduce it's dimensions to quickly train and predict from your model [8].

1.3.2 Ensemble Approaches

Ensemble learning is an effective technique to increase classification and prediction accuracy.It combines several machine learning algorithms (stacking) or uses many algorithms of same type(bosting ,bagging) into one model in order to reduce loss and variance.Author Jie Gu [3] proposed a framework for intrusion detection which uses SVM ensemble with feature enhancement that gives the robust performance with high precision and accuracy than any existing model only for binary classification.

1.3.3 Mixture of Experts

Mixture of experts works on the principle that every expert is specialized in a particular domain of a input data space by imitating a gating network which is liable for learning the combined weight of the specialized experts for a particular input [11] ,by this method input space is dynamically divided and conquered by the gating network and the experts.

1.4 Research gaps

The past work done in the field of intrusion detection predominantly centers around binary classification,detecting a specific intrusion type effectively is as yet a noteworthy concern.Additionally, their is a need of consistent dataset so that their wouldn't be any baisness in order to predict any attack type.

The proposed work uses different machine learning algorithm to detect a particular attack type in which they give the best result,which can be used as hybrid method using mixture of expert, and also using different feature selection algorithm to reduce the features which are of less significance for the predicating the different attack type can be useful for increasing the efficiency of our model.

1.5 Problem statement

Comparative Analysis of previously built intrusion detection algorithms with hybrid Mixture of expert algorithm which combines the expertise of different machine learning models to built a meta-model that is able to provide a result which would out-perform many intrusion algorithms previously developed.

1.6 Thesis Objective

1. Evaluating the performance of many different machine learning algorithms for the intrusion detection.

2. Implementation of mixture of experts technique for intrusion detection.

3. To perform a comparative analysis of ensemble approaches(Bagging,Boosting) with hybrid approach(Mixture of experts).

CHAPTER 2

Methodology

Methodology discusses about the work flow during the entire process of developing a model for intrusion detection.

1. Data cleaning and preprocessing of the given dataset.

2. Division of different types of attacks into four major categories, User to Root attack(U2R), Denial of Service attacks (DoS), Remote to Local Attack(R2L), and Probing Attack.[6]

   (i) Denial of Service Attack (DoS): In dos attack, attacker sends many requests to a server to exhausts all the resource of the server so that legitimate users are unable to access the service.

   (ii) User to Root Attack (U2R): By this type of attack intruders first access the system with normal privileges later on try to access the user with administrator privileges.

   (iii) Remote to Local Attack (R2L): In this type of attack intruders try to gain unauthorized access to the system over the network by sending the packets to that system.

   (iv) Probing Attack: It is an attempt from the intruders to gain access the information and vulnerabilities of the system[12], later on this information can be used to generate an attack[7].

3. Evaluation of the performance of different base learners with the given dataset to know the best base learners for specific attack types.

4. Implementation of ensembles (bagging, boosting) and hybrid model (mixture of experts) with selected base learners.

5. Comparative analysis of bagging, boosting and mixture of experts on the basis of accuracy score, precision, recall and F1-score.

6. Building a web based application with the resultant model to classify dynamically generated attacks.

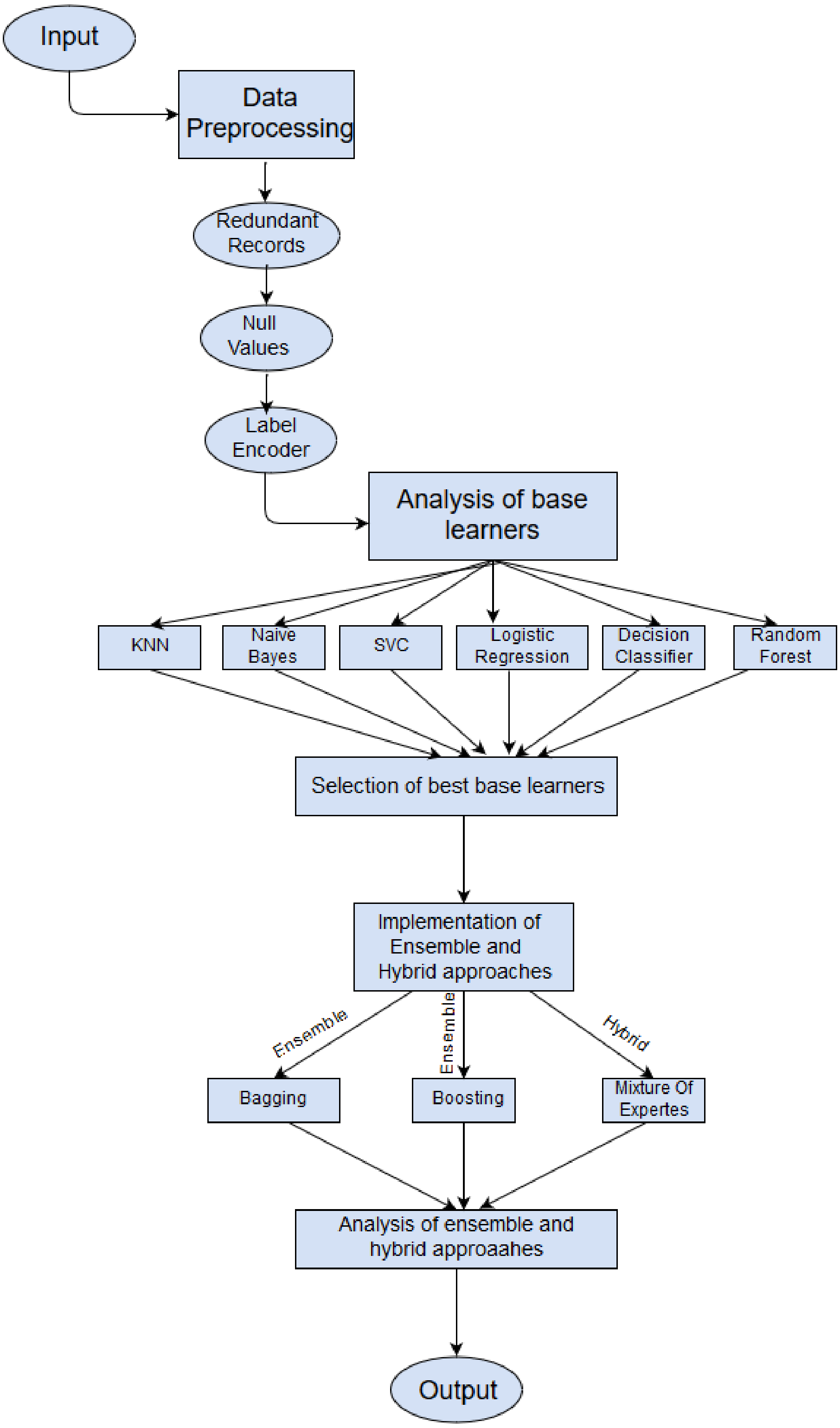

Figure 2.1: Work Flow

2.1 Hybrid Approaches

2.1.1 Mixture of Experts:

The mixture-of-expert method based on the divide and conquer approach technique that subdivides the problem based on input regions and trains different models with specific input regions of the problem. So that each model has its own contribution in the final output of the problem .

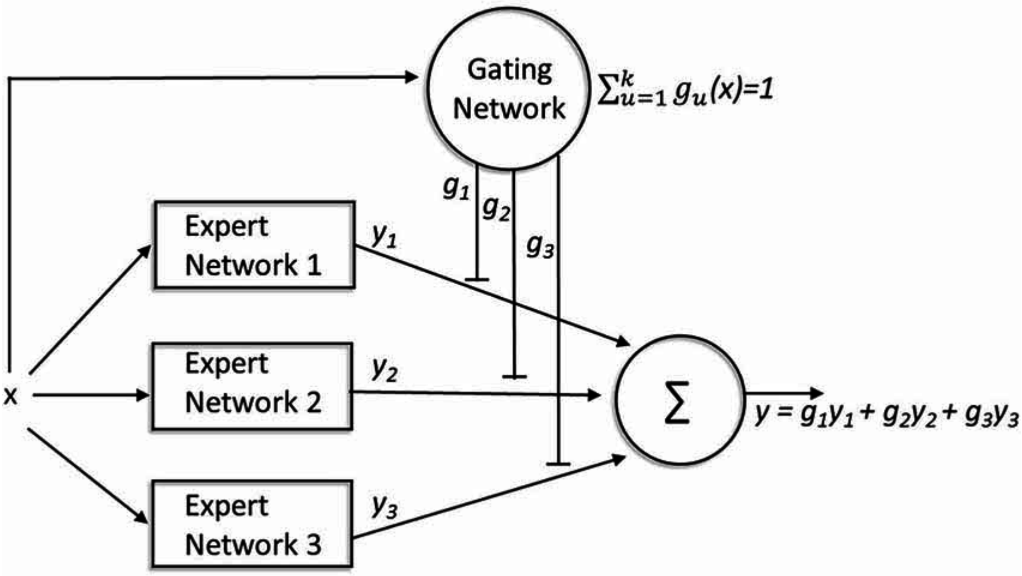

Figure 2.2: Mixture of experts
[10]

2.1.2 Bagging:

It consists of two approaches, first it randomly selects bootstrapped samples from the given dataset and build a classifier for each bootstrap sample and then combines the results from all classifiers.

2.1.3 Boosting:

Boosting is a sequential learning algorithm which trains weak learners to converts them into the strong learner. In each iteration, it tries to increase the weights of poorly predicted instances[9].

2.2 Scikit-Learn

During this experiment, we are using Scikit-learn which is a machine learning library developed in python. All of the models are implemented in scikit-learn library. Scikit-learn is a simple and efficient tool for data analysis.

(i) Accessible to everybody

(ii) Built on NumPy, SciPy, and matplotlib

2.3 Programming setup:

Windows 10 is used as an operating system. The models are trained on Nvidia Quadro P2000 GPU with 5GB of dedicated graphics memory. We have implemented our model on jupyter notebook. All code has been written in python 3.0 and training of the model is done in the notebook.

CHAPTER 3

Dataset Description

3.1 Dataset

The proposed work uses KDD 1999 data set. It was used for the Third International Knowledge Discovery and Data Mining Tools Competition. Dataset has following features:

- It contains 48 Lakh tuples and 41 features.
- It contains 22 attack types which are further categorized into 4 major classes (DOS,U2R,R2L,Probing attack).
- It includes 3 types of protocol http,tcp and udp.

The percentage of attack type in the datatset has been pictorially shown by the pie chart.

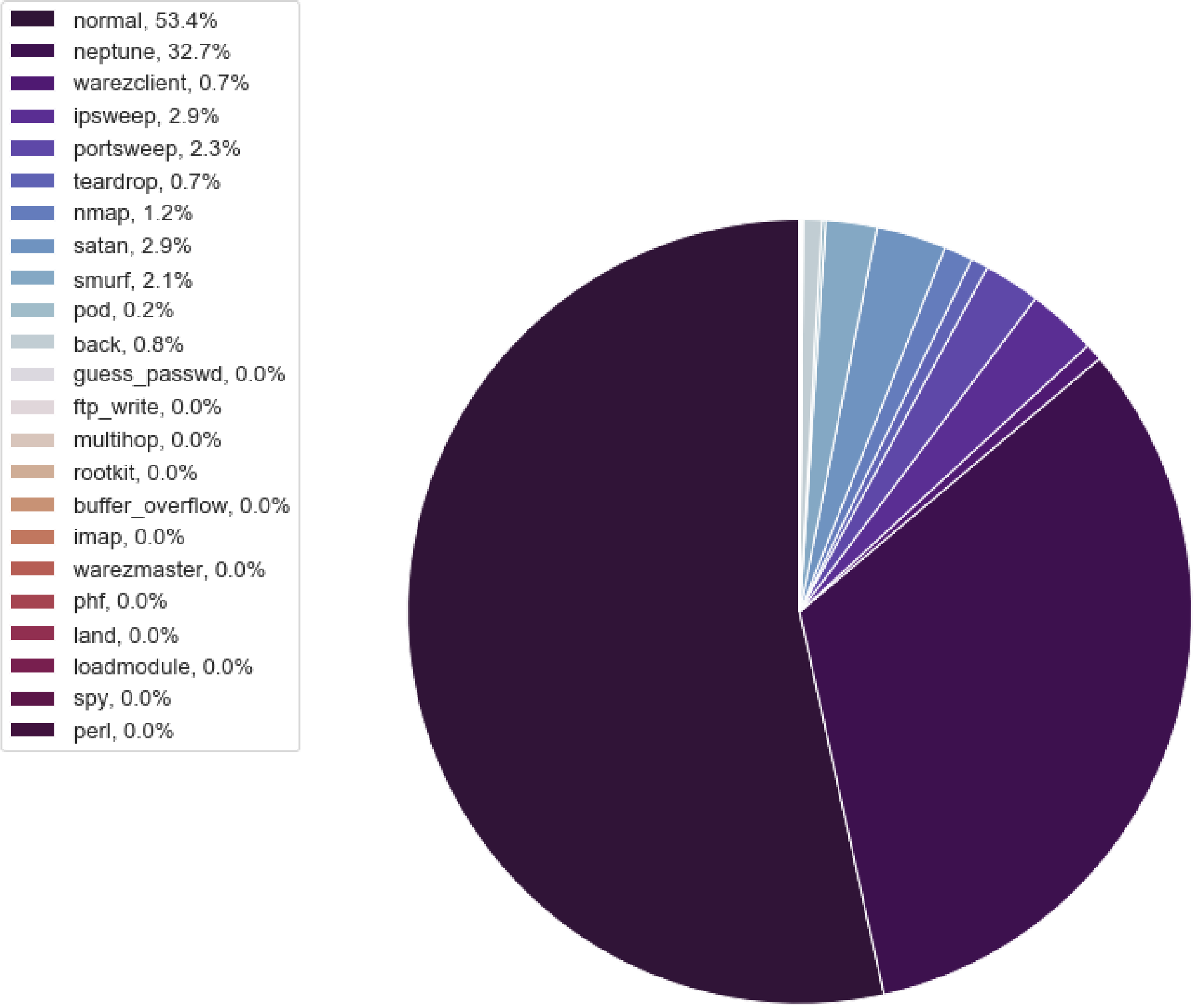

Figure 3.1: Attack Distribution

3.1.1 Redundancy in Dataset

The Previously used dataset was inconsistent.It contains several duplicate tuples, which may cause the machine learning algorithms to be biased towards the frequent records, and so stop them from learning unusual records which are typically a lot of harmful to networks. Also, the presence of frequently repeated records in the test set results in the assessment to be biased by the methods which have better detection accuracy on the frequent records.

CHAPTER 4

Activities Aand Results

4.1 Description of Activities Completed

We have performed the following activities includes literature survey, data prepossessing and also trained the dataset with the various classification algorithms so that we can select different base Lerners which can help our hybrid model to classify each attack type successfully, a brief description of above activities is given below.

4.1.1 Literature Survey

We have reviewed many research papers to get knowledge of various work which has been done in the field of intrusion detection and what are the research gaps which has to be covered.

4.1.2 Data Cleaning and Preprocessing

Data Preprocessing is a technique through which raw data is converted into useful and efficient format.

(a) **Removing the Data Redundancy**
All the redundant records of the dataset are analyzed and removed to maintain data consistency.

(b) **Filling the missing values**
All the missing values of various features of the dataset are searched and replaced by the meaningful values(mean, median and mode value of the corresponding column) to make it trainable for the learners.

(i) **Mean**
The mean of a dataset is calculated by adding all the numbers in the dataset and then dividing by the total number of values in the dataset.

(ii) **Mode**
The mode is the number which has the highest frequency in the dataset.

(iii) **Median**
The median is the middle value when the data set is sorted in the increasing order.

(c) **Encoding the data**
All the categorical data has been encoded with the tools provided by scikit-learn library of python.

(i) **LabelEncoding**
It is used to convert the categorical text data into model-understandable numerical data.

(ii) **OneHotEncoding**
One hot encoding is a process by which categorical variables are converted into a form which is easier for finding the prediction.

(d) **Analysis of different machine learning Algorithms**
We have tested different machine learning algorithms for the selection of base learners for the hybrid Approach trained and tested on the KDD 1999 dataset. Following are the base learners we have used.

(i) **k-nearest neighbors classifer**
the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression.In both cases, the input consists of the k closest training examples in the feature space.The output depends on whether k-NN is used for classification or regression.

(ii) **Random Forest Classifier**
Random Forest Classifier is ensemble algorithm.Ensembled algorithms are those which combines more than one algorithms of same or different kind for classifying objects.Random forest classifier creates a set of decision trees from randomly selected subset of training set.It then aggregates the votes from different decision trees to decide the final class of the test object.

(iii) **Decision Tree Classifier**
Decision Tree Classifier, repetitively divides the working area(plot) into sub part by identifying lines.(repetitively because there may be two distant regions of same class divided by others).

4.2 Intermediate Results

Below are the intermediate results obtained on the kdd dataset. We have performed different base learners on the dataset in order to find the best in its type to classify attacks. The table below describes the findings of different learners on the basis of accuracy, precision, recall, F-score.

Classifer	Accuracy	Precision	recall	F1-score
LinearSVC	0.9221	0.9229	0.9147	0.9187
KNN	0.9690	0.9699	0.9647	0.9674
MultinomialNB	0.4172	0.4161	0.4159	0.3966
RandomForest	0.9688	0.9679	0.9610	0.9645
LogisticRegression	0.8959	0.8912	0.8899	0.8457
DecisionTree	0.9490	0.9474	0.9447	0.9447

Table 4.1 : Intermediate Results

4.3 Future Activities

1. Try to reduce the dimensionality of the dataset with feature selection approaches like PCA (Principal Component Analysis).

2. Implementation of selected base learners with hybrid approaches like a mixture of experts, Bagging and Boosting.

3. Analysis of final results and finding the best method on the basis of accuracy score, precise, recall and F1-Score.

4. Building a web based application to detect the type of attack generated dynamically on the system.

REFERENCES

[1] K.Muneeswaran B.Selvakumar. "firefly algorithm based feature selection for network intrusion detection". *Science Direct,*, 81:148–155, 2018.

[2] Chia-Ying Lin Wei-YangLin Chih-Fong Tsai, Yu-Feng Hsu. "intrusion detection by machine learning: A review". *Science Direct,*, 10:11994–12000, 2009.

[3] Jie Gu, Lihong Wang, Huiwen Wang, and Shanshan Wang. "a novel approach to intrusion detection using svm ensemble with feature augmentation". *Computers Security*, 2019.

[4] Y. P. N. L. S. S. L. C. Guo. "a two-level hybrid approach for intrusion detection". *Research Gate,*, 214:9, 2016.

[5] G. C. Kessler. "defenses against distributed denial of service attacks". *Science Direct,*, 321:12, 2002.

[6] Euntai Kim, Heejin Lee, Minkee Park, and Mignon Park. "a subset feature elimination mechanism for intrusion detection system". *International Journal of Advanced Computer Science and Applications,*, 7:148–157, 2016.

[7] Rohit Kumar Singh Gautam and Amit Doegar. "an ensemble approach for intrusion detection system using machine learning algorithms". *International Conference on Cloud Computing, Data Science Engineering,*, 14:554 – 574, 2003.

[8] W. Lu M. Tavallaee, E. Bagheri and A. Ghorbani. "a detailed analysis of the kdd cup 99 data set". *IEEE international conference on Computational intelligence for security and defense applications,*, 1:53–58, 2009.

[9] Riyad.A.M and M.S Irfan Ahmed. "an ensemble classification approach for intrusion detection". *International Journal of Computer Applications,*, 80:37–42, 2013.

[10] Andrew S Bock, Fine, and Ione. "anatomical and functional plasticity in early blind individuals and the mixture of experts architecture". *Frontiers in human neuroscience*, 8:971, 2014.

[11] A. Nowzari-Dalini M. Ganjtabesh-R. Ebrahimpour. S. R. Kheradpisheh, F. Sharifizadeh. *"Mixture of feature specified experts"*, volume 20. 2014.

[12] Mohit Tiwari, Raj Kumar, Akash Bharti, and Jai Kishan. "intrusion detection system". *International Journal of Technical Research and Applications*, 5:2320–8163, 2017.