SIR CHHOTU RAM INSTITUTE OF ENGINEERING & TECHNOLOGY

CHAUDHARY CHARAN SINGH UNIVERSITY MEERUT

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

A Project Synopsis

On

ANOMALY BASED INTRUSION DETECTION

Submitted in partial fulfillment of the requirement for the award of the degree of

BACHELOR OF TECHNOLOGY

In

COMPUTER SCIENCE ENGINEERING

By:

ANMOL SINGH (100180110)

SHRISTI (100180149)

SHIKSHA UPADHYAY (100180147)

UNDER THE GUIDANCE

Dr. AMIT SHARMA

(Asst. Professor , Computer Science and engineering)

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

Signature	Signature	Signature
Anmol Singh	Shristi	Shiksha Upadhyay
(100180110)	(100180149)	(100180147)

Date:

Sir Chhotu Ram Institute Of Engg. & Tech.

Approved by A.I.C.T.E., New Delhi Chaudhary Charan Singh

University Meerut Campus

CERTIFICATE

This is to certify that this project synopsis entitled "ANOMALY BASED INTRUSION DETECTION" which is submitted by Anmol Singh (100180110), Shristi (100180149), Shiksha Upadhyay (100180147) in the partial fulfilment ofthe requirement for the award of degree of Bachelor of Technology in Computer Science Engineering from thiscollege is a record of candidate own work carried out by him/her under our supervision.

The matter embodied in this Project Synopsis has not been submitted earlier for award of any degree ordiploma in any university/institution to the best of our knowledge & belief.

Dr. Amit sharma
Project Guide

Er. Ritu sharma
Project Coordinator

Er. Millind singh
HOD, Department. Of CS

ACKNOWLEDGEMENT

It gives me a great sense of pleasure to present the report of the Project Work undertaken during B. Tech. Final Year. I owe special debt of gratitude to my Project Coordinator Er. Ritu Sharma, Department of Computer Science Engineering, SCRIET C.C.S.University, Meerut for her constant support and guidance throughout the course of my work. It is only her/his cognizant efforts that my endeavors have seen light of the day.

My deepest thanks to my Project Guide **Dr, Amit Sharma,** SCRIET, CCS University Meerut the Guide of the project for guiding and correcting various documents of mine with attention and care.

I also take the opportunity to acknowledge the contribution of **Prof. Niraj Singhal, Director,** SCRIET CCS University, Meerut and **Er. Millind Singh, Assistant Professor & Co-ordinator** Department of Computer Science Engineering, SCRIET, Meerut for his full support and assistance during the development of the project.

I also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of my project.

Last but not the least, I acknowledge my friends for their contribution in the completion of the project.

TABLE OF CONTENTS

1. ABSTRACT

Abstract— In the network communications, network intrusion is the most important concern nowadays. The booming contingency of network attacks is a devastating problem for network services. Various research works are already conducted to find an effective and efficient solution to prevent intrusion in the network in order to ensure network security and privacy. Machine learning is an effective analysis tool to detect any suspicious events occurred in the network traffic flow. In this paper, we developed a classifier model based on SVM and Random Forest based algorithms for network intrusion detection. The NSL-KDD dataset, a much improved version of the original KDDCUP'99 dataset, was used to evaluate the performance of our algorithm. The main task of our detection algorithm was to classify whether the incoming network traffics are normal or an attack, based on 41 features describing every pattern of network traffic. The detection accuracy more than 95 % was achieved using SVM and Random algorithms. The results of two algorithms compared and it is observed that Random Forest algorithm is more effective than Support Vector Machine.

Keywords— Network Intrusion, Support Vector Machine, Random Forest, accuracy

2. INTRODUCTION

Network Security maintenance is one of the major safety concerns for neutralizing any unwanted activities. It is not only for protecting data and network privacy issues but also for avoiding any hazardous situations. For decades, Network security is one of the major issues and different types of developed systems are being implemented. Network intrusion is an unauthorized activity over the network that steals any important and classified data. Also sometimes it's the reason of unavailability of network services. The unexpected anomaly occurs frequently and a great loss to internet cyber world in terms of data security, the safety of potential information's etc. Therefore, the security system has to be robust, dependable and well configured. Traditionally, network intrusion detection systems (NIDS) are broadly classified based on the style of detection they are using: systems relying on misuse-detection monitor activity with precise descriptions of known malicious behavior , while anomaly-detection systems have a notion of normal activity and flag deviations from that profile. Signature based detection system involves analyzing network traffic for a series of bytes or packet sequences known to be an anomaly Signature based type detection also has some disadvantages. A signature needs to be created for each attack and they are able to detect only those attacks. They are unable to detect any other novel attacks as their signatures are unknown to the detection scheme. Anomaly based NIDS operate based on the idea that the ambient traffic in a network collected over a period of time reflects the nature of the traffic that may be expected in the immediate future. Anomaly intrusion detection identifies deviations from the normal usage behavior patterns to identify the intrusion. The normal usage patterns are constructed from the statistical measures of the system features, for example, the CPU and I/O activities by a particular user or program. The behavior of the user is observed and any deviation from the

constructed normal behavior is detected as intrusion.

Now a days, Machine learning techniques are heavily being adapted and developed in intrusion detection to enhance the efficacy of the systems and in other applications as well . Suthaharan in his work stated that due to the large size and redundant data in the datasets the computation cost of the machine learning methods increases drastically. They proposed ellipsoid-based technique which detects anomalies and side by side cleans the dataset. The research of deals with intrusion detection technique which is a combination of k means clustering, neuro-fuzzy and radial basis support vector machine. In their technique, firstly k-means clustering is used to spawn the training subsets, on them various neuro fuzzy models are trained, after that a vector used by svm classification is generated and finally classification task is carried by radial SVM technique.

We propose a method that is based on the classification algorithm named as random forests and use it to detect the intrusions. Random forest is based on ensemble approach and is closely related to decision trees and nearest neighbor methods that are widely used in the task of intrusion detection. Random forest initiates with decision tree, which can be said to be a weak learner approach. A random forest creates a strong learner by combining trees which were stated as weak learners. Random forest works better than decision trees when the number of samples is more [5]. In random forests features are selected arbitrarily after each split, this ensures a higher classification power and greater efficiency. Moreover, this method overcomes the problem of over fitting and also it not only pertains the qualities possessed by decision trees, but by utilizing its paging mechanism and voting scheme it produces better results than decision trees mostly [6]. In this paper, we present a model that we implemented an intrusion detection system for classification of intrusion types which outperforms the support vector machine method and the nearest centroid classification method in terms of accuracy, the detection rate

and false alarm. An analysis has been performed for each type of attack mentioned in the dataset that has been utilized for this study.

2. NSL-KDD Dataset

=The dataset to be used in this research is the NSL-KDD dataset [7] which is a new dataset for the evaluation of researches in network intrusion detection system. It consists of selected records of the complete KDD 99 dataset. NSLKDD dataset solve the issues of KDD 99 benchmark and connection record contains 41 features. Among the 41 features, 34 features are numeric and 7 features are symbolic or discrete. The NSL-KDD training set contains a total of 22 training attack types; with an additional 17 types in the testing set only

3.TECHNOLOGY USED

3.1. Python

Python is a high-level programming language that is commonly used. Guido van Rossum, a Dutch programmer, created it in 1991. The Python Software Foundation continued to grow it. It was created with the aim of making the code as readable as possible.

Python's syntax allows programmers to express their thoughts in fewer lines of code. Python is a programming language that, when compared to other programming languages, allows you to operate faster and integrate systems more effectively in less time.

Features of Python Programming Language-

1. Easy to learn: It is very simple to read.

2. Simple to Learn: Python is simple to learn because it is an intuitive and compact language. This means that it is simple to understand and therefore learn.

3. Cross Platform: Python is cross-platform, meaning it can run on a variety of operating systems. As a result, it's a cross-platform and highly portable language.

4. Open Source: It is a programming language that is free to use.

5. Large Stabdard Library: Python comes with a wide standard data library that contains useful codes and functions that we can use when coding in Python.

6. It's free: It's completely free.

7. Supports exception handling: Since Python supports exception handling, we canwrite less error-prone code and evaluate different situations that might result in an exception later.

8. Advanced features: Generators and list comprehensions are supported.

9. Automatic memory management: Python has built-in support for automatic memory management, which guarantees that memory is optimised at all times.

3.2. Jupyter Notebook-

The Jupyter Notebook is a free, open-source web application that allows programmers to build and share files containing equations, visualisations, code, and text. The people behind Project Jupyter are in charge of keeping this website up to date.

Jupyter Notebook is a fork of the IPython project, which once had its own IPython Notebook project. The name Jupyter comes from the three main programming languages that it supports: Julia, Python, and R. IPython kernel is included with Jupyter, allowing you to write Python code. However, we now have access to more than 100 additional kernels.

It's useful for data cleansing and transformation, computational simulation, mathematical modelling, data visualisation, machine learning, and a variety of other tasks.

Here are listed the main libraries used:

- Numpy: standard library for math operations
- Scipy: used to compute test statistics and distributions
- Pandas: used to manipulate data inside dataframes and for basic computations
- Sklearn: used to apply different ML models to the data
- Pyplot to plot visualizations
- Seaborn built on top of pyplot (nicer visualizations)Other libraries:

- random: used to generate random numbers
- HTML and matplotlib.animation: used for the animations

3.3 Machine Learning

Machine learning (ML) is a type of artificial intelligence (AI) that allows software applications to become more accurate at predicting outcomes without being explicitly programmed to do so. Machine learning algorithms use historical data as input to predict new output values. Recommendation engines are a common use case for machine learning. Other popular uses include fraud detection, spam filtering, malware threat detection, business process automation (BPA) and predictive maintenance. Classical machine learning is often categorized by how an algorithm learns to become more accurate in its predictions. There are four basic approaches: supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. The type of algorithm data scientists choose to use depends on what type of data they want to predict. Primarily there are 4

1. Logistic Regression

It is a statistical method for analyzing a data set in which there are one or more independent variables that determine an outcome. The outcome is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent

(predictor or explanatory) variables. Logistic regression is a Machine Learning classification algorithm that is used to predict the probability of a categorical dependent variable. In logistic regression, the dependent variable is a binary variable that contains data coded as 1 (yes, success, etc.) or 0 (no, failure, etc.).

2. Decision Tree

It is one of the most powerful and popular algorithm. Decision-tree algorithm fallsunder the category of supervised learning algorithms. It works for both continuous as well as categorical output variables. Decision tree builds classification or regression models in the form of a tree structure. It breaks

down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. A decision node has two ormore branches and a leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

Decision tree builds classification or regression models in the form of a tree structure. It utilizes an if-then rule set which is mutually exclusive and exhaustive for classification. The rules are learned sequentially using the training data one at

a time. Each time a rule is learned, the tuples covered by the rules are removed. This process is continued on the training set until meeting a termination condition. It is constructed in a top-down recursive divide-and-conquer manner. All the attributes should be categorical. Otherwise, they should be discretized in advance. Attributes in the top ofthe tree have more impact towards in the classification and they are identified using the information gain concept. A decision tree can be easily over-fitted generating too many branches and may reflect anomalies
due to noise or outliers.

3. SVM (Support Vector Machine)

Support Vector Machine(SVM) is a supervised machine learning algorithm used for both classification and regression. Though we say regression problems as well its best suited for classification. The objective of SVM algorithm is to find a hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features. If the number of input features is two, then the hyperplane is just a line. If the number of input features is three, then the hyperplane becomes a 2-D plane. It becomes difficult to imagine when the number of features exceeds three.

4. Classification Model

In general, the category of problems which contains data as well as the additional attributes that we want to predict comes under supervised learning approach. Under supervised learning approach the classification problem comes into account when the instances belong to two or more classes and our intention is to forecast the class of the unlabeled instances. Under the category of supervised learning methods, a technique known as Support vector machines (SVM) holds its place for classification. This method is effective for high dimensional spaces, is memory efficient since it utilizes subset of training data points in the decision function called as support vectors, also it is adroit as for the decision function various kinds of kernel functions can be stated . If the count of features is bigger than the count of samples this technique is liable to give mediocre performance.

A) Support Vector Machine:

The SVM uses a portion of the data to train the system, finding several support vectors that represent the training data. These support vectors will form a SVM model. A basic input data format and output data domains are listed as follows

(Xi, Yi)……Xn, Yn)

Where

$X \in R^m$ and $Y \in \{0, 1\}$

(Xi, Yi)……. (Xn, Yn) is training data records, n is the numbers of samples m is the inputs vector, and y belongs to category of class ‘0’ or class ’1’ respectively. On the problem of linear, a hyper plane can be divided into the two categories as shown in Figure.

The hyper plan formula is: $(w \cdot x) + b = 0$

The category formula is: $(w. x) + b \geq 0$ if $Yi = 1$ $(w. x) + b \leq 0$ if $Yi = 0$

A classification task usually involves with training and testing data which consist of

some data instances. Each instance in the training set contains one “target value" (class labels: Normal or Attack) and several “attributes" (features).The goal of SVM is to produce a model which predicts target value of data instance in the testing set which is given only attributes. To attain this goal there are four different kernel functions.in this experiment RBF kernel function is used The Formula for RBF Kernel Optimization function .

B) Random Forest Classification Model:

The suggested classification model is comprised of random forest algorithm. In Random Forest, the method to build an ensemble of classifiers is to change the training set using the same strategy as bagging (Breiman, 1996). Bagging creates new training sets by resampling from the original data set n times, n being the number of samples in the original training set, randomly with replacement. This means the sample just being chosen will not be removed from the data set in the next draw. Hence, some of the training samples will be chosen more than once while some others will not be chosen at all in a new set. Bagging helps classification accuracy by decreasing the variance of the classification errors. In another words, it taps on the instability of a classifier. ‘Instability’ of a classifier means that a small change in the training samples will result in comparatively big changes in accuracy. The classifiers are combined by a majority vote and the vote of each classifier carries the same weight. In the case of a tie, the decision can be made randomly or by prescribed rules. Random Forest creates multiple trees using the impurity gini index (Breiman et al., 1984). However, when constructing a tree, Random Forest searches for only a random subset of the input features (bands) at each splitting node and the tree is allowed to grow fully without pruning. Since only a portion of the input features is used and no pruning, the computational load of Random Forest is comparatively light. In addition, in case a separate test set is not available, an outof-bag method can be used. For each new training set that is generated, one-third of the samples are randomly left out, called the out-of-bag (OOB) samples. The remaining

(in-thebag) samples are used for building a tree. For accuracy estimation, votes for each sample are counted every time when it belongs to OOB samples. A majority vote will determine the final label. Only approximately one-third of the trees built will vote for each case. These OOB error estimates are unbiased in many tests (Breiman, 2001). The number of features for each split has to be defined by the user, but it is insensitive to the algorithm. Majority vote is used to combine the decisions of the ensemble classifiers. The Algorithm: The random forests algorithm (for both classification and regression) is as follows:

1. Draw $ntree$ bootstrap samples from the original data.

2. For each of the bootstrap samples, grow an unpruned classification or regression tree, with the following modification: at each node, rather than choosing the best split among all predictors, randomly sample $mtry$ of the predictors and choose the best split from among those variables. (Bagging can be thought of as the special case of random forests obtained when $mtry$ = p, the number of predictors.)

3. Predict new data by aggregating the predictions of the $ntree$ trees (i.e., majority votes for classification, average for regression).

5. RESULT AND DISCUSSION:

The performance of all the classifiers was computed by utilizing a matrix known as confusion matrix. It is a standard metric for benchmarking the effectiveness and robustness of a classification algorithm. Using the confusion matrix, measures like accuracy, detection rate and false alarm rate have been computed which are the generic criteria for evaluating the performance of the IDS. These metrics have been utilized in a number of studies and they ensure a viable means of deciding the efficiency of the model for detecting the intrusions within systems. For a decent level of performance, the intrusion detection system (IDS) needs high accuracy and precision and conversely false alarm rate should be low.

These terms are given by the following formulae:

Accuracy = $TP+TN\ /\ TP+FP+TN+FN$

Precision = $TP/\ TP+FP$ True positive rate (TPR) = $TP/\ TP+FN$

False positive rate (FPR) = $FP\ /\ TN+FP$

True negative rate TNR) = $TN/\ TN+FP$

False negative rate (FNR) = $FN/\ TP+FN$

True positive (TP) indicates the number of instances having the class label of attack and were correctly classified as an attack. True negative (TN) indicates the number of instances having the class label of normal and were correctly classified as normal. False positive indicates the number of instances that have a label of being valid but have been incorrectly classified as intrusion. False negative indicates the number of instances that were having a label of intrusion but were incorrectly classified as normal by the IDS

6. CONCLUSION

In this paper, we have scrutinized some new techniques for intrusion detection and evaluated their performance based on the benchmark KDD Cup 99 Intrusion data. An Intrusion Detection System that was able to assay the dynamic and complex nature of intrusion activities has been built. Random forests classification algorithm surpasses the major classification methods support vector machine. After comparison of models, the proposed model resulted in the highest accuracy and detection rate values as well as the least false rate values. The results specify that the classification ability of the proposed model is inherently superior to the support vector machine model. Anomaly detection methods that are based on artificial intelligence are continuously alluring a lot of attention from the research community. The experimental result shows the efficiency of both Random Forest and Support Vector Machine, which proves that the machine learning techniques can be successfully applied to the Anomaly Intrusion Detection System. The research work can be extended by applying various other soft computing techniques in Anomaly Intrusion Detection.

7. REFERENCES

[1] S. S. Roy, V. M. Viswanathan - Classifying Spam Emails Using Artificial Intelligent Techniques. In International Journal of Engineering Research in Africa, vol. 22, pp. 152-161. Trans Tech Publications, 2016.

[2] S. S. Roy, D. Mittal, A. Basu, A. Abraham - Stock Market Forecasting Using LASSO Linear Regression Model. In Afro European Conference for Industrial Advancement, pp. 371-381. Springer International Publishing, 2015.

[3] S. Suthaharan - An iterative ellipsoid-based anomaly detection technique for intrusion detection systems, In Southeast on, Proceedings of IEEE, pp. 1-6, 2012.

[4] A. Chandrasekhar, K. Raghuveer - Intrusion detection technique by using k-means, fuzzy neural network and svm classifiers, In Computer Communication and Informatics (ICCCI), 2013 International Conference, pp. 1-7.

[5] S. Adusumilli, D. Bhatt, H. Wang, V. Devabhaktuni, P. Bhattacharya - A novel hybrid approach utilizing principal component regression and random forest regression to bridge the eripod of GPS outages, Neurocomputing, 2015.

[6] J. Ali, R. Khan, N. Ahmad, I. Maqsood - Random forests and decision trees, IJCSI International Journal of Computer Science Issues, vol. 9, no. 5, 2012.

[7] NSL-KDD Data set for Network-based Intrusion Detection Systems. Available at: http://nsl.cs.unb.ca/NSL-KDD.

[8] Breiman, L. (1996). Bagging predictors. Machine Learning, 24, 123−140.

[9] Breiman, L. (2001). Random forests. Machine Learning, 45, 5−32.

[10] Breiman, L., Freidman, J. H., Olshen, R. A., & Stone, C. J. (1984). Classification and Regression Trees. Wadsworth.