A Survey: Comparative Analysis of Classifier Algorithms for DOS Attack Detection

Jatin Patel[1] Vijay Katkar [2] Aditya Kumar Sinha[3]
[1]Department of Computer Engineering ,GTU PG School, Ahmedabad, India
[2]Department of Information Technology ,PCCOE(PUNE),India
[3]Principal Technical Officer C-DAC ACTS, Pune, India

***Abstract*—**In today's interconnected world, one of pervasive issue is how to protect system from intrusion based security attacks. It is an important issue to detect the intrusion attacks for the security of network communication.Denial of Service (DoS) attacks is evolving continuously. These attacks make network resources unavailable for legitimate users which results in massive loss of data, resources and money.Significance of Intrusion detection system (IDS) in computer network security well proven. Intrusion Detection Systems (IDSs) have become an efficient defense tool against network attacks since they allow network administrator to detect policy violations. Mining approach can play very important role in developing intrusion detection system. Classification is identified as an important technique of data mining. This paper evaluates performance of well known classification algorithms for attack classification. The key ideas are to use data mining techniques efficiently for intrusion attack classification. To implement and measure the performance of our system we used the KDD99 benchmark dataset and obtained reasonable detection rate.

Keywords- -DoS Attack, Intrusion Detection System, Classification of Intrusion Detection, Signature-Based IDS, Anomaly-based -Based IDS, Data Mining, Classification Algorithm, KDD Cup 1999 Dataset.

I. INTRODUCTION

Now a day, intrusion detection is one of the high priority tasks for network administrators and security professionals.
As network based computer systems play increasingly vital roles in modern society, they have become intrusion detection systems provide following three essential security functions:

- *Data confidentiality*: Information that is being transferred through the network should be accessible only to those that have been properly authorized.
- *Data integrity:* Information should maintain their integrity from the moment they are transmitted to the moment they are actually received. No corruption or data loss is accepted either from the random events or malicious activity.
- *Data availability:* The network or a system resource that ensures that it is accessible and usable upon demand by an authorized system user.

DoS attack is attempt by attacker to prevent Internet site or Server from functioning efficiently or properly. There are several ways of launching DoS attacks against a server. Every attack uses any one of the following technique:

1) *Consume Server resources*
2) *Consume network bandwidth*
3) *Crash the server using vulnerability present in the server*
4) *Spoofing packets*

Even though there are different ways to launch attack but every attack makes server either nonresponsive or extremely slow. iv. Spoofing packets Even though there are different ways to launch attack but every attack makes server either nonresponsive or extremely slow. Any intrusion detection system has some inherent requirements. Its prime purpose is to detect as many attacks as possible with minimum number of false alarms, i.e. the system must be accurate in detecting attacks. Data mining techniques like data reduction, data classification, features selection techniques play an important role in IDS.

This work is a survey of data mining classification algorithm that have been applied to IDSs and is organized as follows: Section 2 presents IDs terminology and taxonomy. Section 3 mentions the drawbacks of standard IDs. Section 4 gives brief introduction about data mining .Section 5 illustrates how data mining can be used to enhance IDSs. Section 6 describes the various data mining approaches that have been employed in IDSs by various researchers. Section 7 provides misuse and anomaly detection using data mining techniques. Section 8 describes various data mining algorithms to implement IDs and also compares various data mining algorithms that are being used to implement IDs. Section 9 provides experimental study on weka environment. Section 10 focuses on current research challenges and finally section 10 concludes the work.

II. INTRUSION DETECTION SYSTEM

Intrusion Detection System (IDS) can detect, prevent and more than that IDS react to the attack. Therefore, the main objective of IDS is to at first detect all intrusions at first effectively. This leads to the use of an intelligence technique known as data mining/machine learning. These techniques are used as an alternative to expensive and strenuous human input. IDS can provide guidelines that assist you in the vital step of establishing a security policy for your computing assets.

A. *Classification of Intrusion Detection*

Intrusions Detection can be classified into two main categories. They are as follow:

Host Based Intrusion Detection: HIDSs evaluate information found on a single or multiple host systems, including contents of operating systems, system and application files [1].

Network Based Intrusion Detection: NIDSs evaluate information captured from network communications, analyzing the stream of packets which travel across the network [1].

B. *Components of Intrusion Detection System*

An intrusion detection system normally consists of three functional components [2]. The first component of an intrusion detection system, also known as the event generator, is a data source. Data sources can be categorized into four categories namely Host-based monitors, Network-based monitors, Application-based monitors and Target-based monitors. The second component of an intrusion detection system is known as the analysis engine. This component takes information from the data source and examines the data for symptoms of attacks or other policy violations. The analysis engine can use one or both of the following analysis approaches:

1) *Signature-Based/ Misuse-based detection IDS*

Misuse-based detection [2] is named knowledge-based detection too. Knowledge-based detection is equipped with a database that contains a number of signatures about known attacks. The audit data collected by the IDS is compared with the content of the database and, if a match is found, an alert is generated. Events that do not match any of the attack models are considered as a part of legitimate activities. The main advantage of misuse-based systems is that they usually produce very few false positives. But this approach has drawbacks. It cannot detect previously unknown attacks, and sometimes it even cannot detect the variations of known attacks.

2) *Anomaly-based -Based IDS*

Anomaly-based detection [2] is a behavior-based detection method. It is based on the assumption that all anomalous activities are malicious and all the attacks are subset of anomaly activities. By building a model of the normal behavior of the system, then it looks for anomalous activities that do not conform to the established model. Data mining techniques can be used for intrusion detection efficiently.

III. DATA MINING

Data mining [3] is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. Data mining can be used for solving the problem of network intrusion based security attack. It has Ability to process large amount of data and reduce data and by extracting specific data, With this Easy data summarization and visualization that help the security analysis. It is a fairly recent topic in computer science but utilizes many older computational techniques from statistics, information retrieval, machine learning and pattern recognition.

Here are a few specific things that data mining might contribute to an intrusion detection project:

- Remove normal activity from alarm data to allow analysts to focus on real attacks
- Identify false alarm generators and ”bad” sensor signatures
- Find anomalous activity that uncovers a real attack
- Identify long, ongoing patterns (different IP address, same activity)

To accomplish these tasks, data miners employ one or more of the following techniques:

- Data summarization with statistics, including finding outliers
- Visualization: presenting a graphical summary of the data
- Clustering of the data into natural categories
- Association rule discovery: defining normal activity and enabling the discovery of anomalies
- Classification: predicting the category to which a particular record belongs

A. *Data Mining Classification Algorithms*

The central theme of our approach is to apply data mining Classification algorithms for intrusion detection System for detecting DoS Attack. Data mining generally refers to the process of (automatically) extracting models from large stores of data [8]. The recent rapid development in data mining has made available a wide variety of algorithms, drawn from the fields of statistics, pattern recognition, machine learning, and database. Several types of algorithms are particularly relevant to our research:

Classification [3] [4] data mining technique Classification maps a data item into one of several pre-defined categories. These algorithms normally output "classifiers", for example, in the form of decision trees or rules. An ideal application in intrusion detection will be to gather sufficient "normal" and "abnormal" audit data for a user or a program, then apply a classification algorithm to learn a classifier that will determine (future) audit data as belonging to the normal class or the abnormal class. there are many types of classifiers are available like tree, bayes, function ,rule . Basic aim of classifier is predict the appropriate class.

1) *Decision Tree*

Decision tree [5] is an important method for data mining, which is mainly used for model classification and prediction. This predictive machine-learning model that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The internal nodes of a decision tree denote the different attributes; the branches between the nodes tell us the possible values that these attributes can have in the observed samples, while the terminal nodes tell us the final value (classification) of the dependent variable.

a) *J48 Algorithm*

The J48 [5] is a Decision tree classifier algorithm. In this algorithm for classification of new item, it first needs to create a decision tree based on the attribute values of the available training data. It discriminate the various instances and identify the attribute for the same. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then we terminate that branch and assign to it the target value that we have obtained.

b) Classification and Regression Trees (CART)

CART algorithm was developed by Brieman, Friedman, Olshen, and Stone in 1984. CART creates trees that have binary splits on nominal or interval inputs for a nominal, ordinal, or interval target. The CART algorithm does not require binning; data is handled in its raw state. The CART algorithm uses the gini comparing classification algorithms in data mining 25 index to measure impurity at the node. For a binary class the GINI measure of impurity is given by GINI (t) =1-∑ [p (t/j)] $^{2, \text{Where}}$ *p (j | t)* is the relative frequency of class j at node t. When a node p is split into x partitions, the quality of split is given by GINIsplit =∑ (ni/n) GINI (t) Where, =number of records at child i n i n = number of records at node p CART also supports the towing splitting criterion which can be used for multi-class problems. It uses the minimal cost complexity pruning to remove features from the classifier that are not significant. CART algorithm automatically balances the class variable, can handle missing values, and allows for cost-sensitive learning and probability tree estimation.

2) Naïve Bayes

The naïve Bayes [3] classifier works on a simple but intuitive concept. It is based on Bayes rule of conditional probability. Naïve Bayes assumes that all attributes of the dataset are independent of each other given the context of the class. The assumption of the conditional probability may be expressed as (Larose, 2005, p. 216)

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_m = x_m \mid \theta) = \prod_{i=1}^{m} p(X_i = x_i | \theta)$$

The naïve Bayes classification is therefore given as (Larose, 2005, p. 216):

$$\theta_{NB} = \arg_\theta \max \prod_{i=1}^{m} p(X_i = x_i | \theta) p(\theta)$$

Because of this assumption, the parameters for each attribute can be learned separately, and this greatly simplifies the learning, especially when the attributes are very large (McCallum & Nigam, 1998). Also naïve Bayes model has shown itself to be more consistently robust to violation of the conditional independence assumption. Naïve Bayes uses a single scan of the data set to estimate the components.

3) Support Vector Machine

Support Vector Machines [3] have been proposed as a novel technique for intrusion detection. An SVM maps input (real-valued) feature vectors into a higher-dimensional feature space through some nonlinear mapping. SVMs are developed on the principle of structural risk minimization [11]. Structural risk minimization seeks to find a hypothesis h for which one can find lowest probability of error whereas the traditional learning techniques for pattern recognition are based on the minimization of the empirical risk, which attempt to optimize the performance of the learning set. Computing the hyper plane to separate the data points i.e. training an SVM leads to a quadratic optimization problem. SVM uses a linear separating hyper plane to create a classifier but all the problems cannot be separated linearly in the original input space. SVM uses a feature called kernel to solve this problem. The Kernel transforms linear algorithms into nonlinear ones via a map into feature spaces. There are many kernel functions; including polynomial, radial basis functions, two layer sigmoid neural nets etc. The user may provide one of these functions at the time of training the classifier, which selects support vectors along the surface of this function. SVMs classify data by using these support vectors, which are members of the set of training inputs that outline a hyper plane in feature space. Computing the hyper plane to separate the data points i.e. training a SVM leads to quadratic optimization problem. SVM uses a feature called kernel to solve this problem. Kernel transforms linear algorithms into nonlinear ones via a map into feature spaces. There are many kernel functions; some of them are Polynomial, radial basis functions, two layer sigmoid neural nets etc. The user may provide one of these functions at the time of training classifier, which selects support vectors along the surface of this function. SVMs classify data by using these support vectors, which are members of the set of training inputs that outline a hyper plane in feature space. The implementation of SVM intrusion detection system has two phases: training and testing. SVMs can learn a larger set of patterns and be able to scale better, because the classification complexity does not depend on the dimensionality of the feature space. SVMs also have the ability to update the training patterns dynamically whenever there is a new pattern during classification. The main disadvantage is SVM can only handle binary-class classification whereas intrusion detection requires multi-class classification

IV. KDD CUP'99 DATA SET

The data set used to perform the experiment is taken from KDD Cup '99[6], which is widely accepted as a benchmark dataset and referred by many researchers. "10% of KDD Cup'99" from KDD Cup '99 data set was chosen to evaluate rules and testing data sets to detect intrusion. The entire KDD Cup '99 data set contains 41 features. Connections are labeled as normal or attacks fall into 4 main categories.

1. DOS: - Denial Of Service
2. Probe: - e.g. port scanning
3. U2R:- unauthorized access to root privileges,
4. R2L:- unauthorized remote login to machine.

In this dataset there are 3 groups of features: Basic, content based, time based features.

- Training set consists 5 million connections.
- 10% training set - 494,021 connections
- Test set have - 311,029 connections
- Test data has attack types that are not present in the training data .Problem is more realistic
- Train set contains 22 attack types

- Test data contains additional 17 new attack types that belong to one of four main categories.

V. EXPERIMENTAL SETUP

To assess the effectiveness of the algorithms for proposed intrusion detection, the series of experiments were performed in Weka. The java heap size was set to 1024 MB for weka-3-7. KDD99 IDS evaluation dataset is used in this paper.KDD99 contains training dataset and testing dataset, its training dataset contains normal and attack connection events. This paper chooses a training file kddcup.data_10_percent.gz as the training dataset. And chooses testing file corrected.gz that contains connect flags as the testing dataset. We are used fuzzy logic for preprocessing training and testing KDD 99 dataset for input to the weka. For fuzzification, we are used triangular membership function in matlab and fuzzily both training and testing KDD99 dataset. Here we are created 3 to 17 intervals of training and testing dataset .We are applied 3 to 17 interval training and testing data on the Weka collection of classification algorithms.

1) Weka

Weka [5] is a collection of machine learning algorithms for data mining tasks. Weka contains tools for data pre processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes. WEKA consists of Explorer, Experimenter, Knowledge flow, Simple Command Line Interface, Java interface.

2) Performance Measurement Terms

To evaluate algorithms' performance several measures have been employed in the thesis.
In general, Positive = identified and negative = rejected. Therefore:

- True positive = correctly identified
- False positive = incorrectly identified
- True negative = correctly rejected
- False negative = incorrectly rejected
- Sensitivity or true positive rate (TPR)=TP/P=TP/(TP+FN)
- False positive rate (FPR)=FP/N=FP/(FP+TN)
- Precision-In the field of information retrieval, precision is the fraction of retrieved documents that are relevant to the search: Precision=TP/(TP+FP) .
- Recall-Recall in information retrieval is the fraction of the documents that are relevant to the query that are successfully retrieved. Recall=TP/(TP+FN)

VI. RESULTS AND DISCUSSION

Our ultimate goal is to evaluate performance of Data mining Classification algorithms on DoS Intrusion Attack. After Applying Data mining Classification algorithm on kdd99 data set on weka tool we get output. For evaluate the output of classification Algorithms for Detecting DoS attack, we are using true positive (TP), false positive (FP) rate, Precision and Recall teams. The detection algorithm maps incoming events to attacks and normal activity. The resulting classification can be used to determine the effectiveness of IDS. Effectiveness is the ability of IDS to maximize the detection rate while minimizing the false alarm rate (false positive rate). In other words, good IDS reports intrusions when they occur, and does not report intrusions when they do not occur.

Algorithm Name-interval of fuzzify training and testing data	TP Rate	FP Rate	Precision	Recall
NAIVE BAYES-3	0.968	0.011	0.948	0.968
J48-15	0.973	0.007	0.952	0.973
SVM-3	0.972	0.007	0.951	0.972
CART-16	0.973	0.006	0.955	0.973

Table 1.1 Comparison of all the algorithms

In This Above Table, we are taking J48, CART, NAÏVE BASED and SVM algorithm with particular fuzzify training and testing data at particular interval that gives maximum detection rate based on evaluation parameter.

In this above table shows that Naïve bayes classification algorithm is give higher detection rate at 3 intervals fuzzify training and testing data. For J48 classification algorithm is give higher detection rate at 15 intervals fuzzify training and testing data. For SVM classification algorithm is give higher detection rate at 3 intervals fuzzify training and testing data. For CART classification algorithm is give higher detection rate at 16 intervals fuzzify training and testing data.

The comparison of all four algorithm shows that the CART algorithm is gives higher TP rate, precision, recall and lowest FP rate.

VII. CONCLUSION

Data mining can improve intrusion based security attacks detection system by adding a new level of surveillance to detection of network data in differences. CART learning algorithm was found to be performing better than other classification algorithms for detecting DoS Attacks in terms of better accuracy and lower error rate. Experiment performed on KDD cup dataset demonstrates that CART algorithm is an efficient algorithm of classification. Accuracy demonstrated helps to improve efficiency of intrusion detection system.

REFERENCES

[1] J. P. Planquart, "Application of Neural Networks to Intrusion Detection", SANS Institute Reading Room.

[2] Mohammad Sazzadul Hoque1, Md. Abdul Mukit2 and Md. Abu Naser Bikas3 "An Implementation Of Intrusion Detection System Using Genetic Algorithm" International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.2, March 2012.

[3] Jiawei Han and Micheline Kamber "Data mining concepts and techniques" Morgan Kaufmann publishers .an imprint of Elsevier .ISBN 978-1-55860-901-3. Indian reprint ISBN 978-81-312-3. 0535-8 .

[4] Stephen Northcutt , Judy Novak "Network Intrusion Detection", Third Edition, New Riders Publishing

[5] N.S.Chandolikar & V.D.Nandavadekar "Comparative Analysis Of Two Algorithms For Intrusion Attack Classification Using Kdd Cup Dataset" International Journal of Computer Science and Engineering (IJCSE) Vol.1, Issue 1 Aug 2012 81-88.

[6] KDD99CUPDataset, http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html.