CLASSIFICATION TECHNIQUES APPLIED FOR INTRUSION DETECTION

Saranya.V[1]
[1]M.Phil Research Scholar,
PSGR Krishnammal College for Women,
Coimbatore-641004.
E-mail Id:[1] saranya.dhilipkumar@gmail.com

Amsaveni.R[2]
[2]Assistant Professor, Dept of Computer science
PSGR Krishnammal College for Women,
Coimbatore-641004.
E-mail Id:[2] amsaveni@psgrkc.com

Abstract:

Intrusion detection system (IDS) was designed to monitor the network activity and it identifies the normal and abnormal behavioral pattern in the network. If there was any abnormal pattern, it indicates the system is in attack by compromising the confidentiality, availability or integrity of the computer system. IDS perform three functions namely monitoring, detecting and responding for malicious activity. Experiment is based on kdd99 dataset to categorize normal and abnormal pattern. Goal of this paper is to compare three classification techniques by considering two classifiers from each technique and find out the best one based on the true positive, false positive and average accuracy.

Keywords: intrusion, data mining, classification algorithm.

1 INTRODUCTION

Internet plays an important role in our day-to-day life but providing security to the Internet was a great challenge ahead. In this paper applying data mining classification technique to intrusion detection was discussed. IDS based on the assumption that the behaviors of intruders are different from normal user**.**

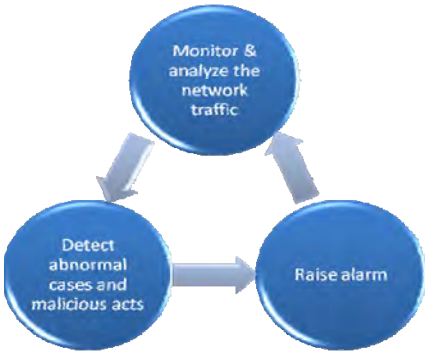
[1] Data mining was one of the sophisticated data analysis tool to discover previously unknown pattern and valid pattern from the large data set. The tool includes mathematical algorithm, statistical tool and machine learning methods. Data mining functionality can be classified into two different ways descriptive mining and predictive mining. The descriptive mining techniques are clustering, association and sequential matching. The predictive mining techniques are classification and regression etc.

1.1 ROLE OF IDS

The role of IDS is to detect abnormal traffic pattern from normal traffic where IDS monitors incoming and outgoing traffic. Common approach for implementing IDS is anomaly detection and misuse/signature detection.

Anomaly detection:-It is based on the normal behavioral pattern of the user when the user deviates from normal behavior then it is termed as intrusive.

Misuse/signature detection:- It is based on already known attack, if the attack matches already known attack then it is termed as intrusive.



1.2 COMPONENTS OF INTRUSION DETECTION SYSTEM

IDS consists of several component following are some of the most important components

Agent or Sensor: - The term Agent is typically used in the host and the term sensor is used in network. It generates security events, monitors and analyzes activity in network.

Console: - It is a program that provides an interface between the IDS user and administrator.

Engine: - Records the event logged by the sensor in a database and uses a system of rules to generate alerts from security events received.

1.3 TYPES OF ATTACKS

Denial of Service (DOS): -

In this type of attack it slow down the system or shut down the system so it disrupt the service and deny the legitimate authorized user.

User to Root Attack (U2R): -

In this type of attack at first attacker starts to access normal user account on the system e.g: by taking down the password, dictionary attack. At last attacker achieves root to access the system

Remote to User Attack (R2U): -

In this type of attack an attacker who has the capability to send packet to a machine over a network but does not have an account on that machine, make use of some vulnerability to achieve local access as a user of that machine.

Probes: -

In this type of attack examines a network to collect information or discover well-known vulnerabilities. These network investigations are reasonably valuable for an attacker who is staying an attack in future. An attacker who has a record, of which machines and services are accessible on a given network, can make use of this information to look for fragile points.

1.4CLASSIFICATION TECHNIQUES

The goal of classification technique is to assign objects (intrusions) to classes based on the values of the objects features. Classification algorithm can be used for both misuse and anomaly detection. In misuse detection, network traffic data are collected and labeled as “normal” or “intrusion”. This labeled dataset is used as a training data to learn classifiers of different types, which can be used to detect known intrusions. In anomaly detection, the normal behavior model is learned from the training dataset that are known to be “normal” using learning algorithms. Classification can be applied to detect intrusions in data streams; a predefined collection of historical data with their observed nature helps in determining the nature of newly arriving data stream and hence will be useful in classification of the new data stream and detect the intrusion. In this paper we compared three classification techniques bayes, function based classifier and rule base classier. In each technique we took two algorithms and compared.

1.4.1 Bayes classification

Bayes networks are one of the most widely used graphical models to represent and handle uncertain information. In bayes classification two classifiers are handled for our experiment.

1.4.1.a. Bayes net:

Bayes net learns from Bayesian network under the assumption of nominal attribute and no missing values. There are two different parts for estimating the conditional probability tables of the network. For our experiment we used BayesNet with the SimpleEstimator and k2 search algorithm without using ADT tree

1.4.1.b. Naïve Bayes:

Naïve bayes classifier is simple approach it represents learning probability knowledge. It relies on two important simplifying assumes that the predictive attributes and conditionally independent given the class and it posits that no hidden or latent attributes influence the prediction process.

1.4.2 Functional based classifier:

In function based classifier two classifiers are used for evaluation

1.4.2.a. MLP:

Multi Layer Perceptron (MLP) is commonly used in neural network classification algorithms. This architecture consists of three layer feed forward neural network: one input, hidden layer and output layer. Parameters selected for the model are learning rate=0.3; momentum=0.2; randomseed=0; validationthreshold=20;

1.4.2.b. SMO:

Sequential Minimal Optimization used for training the support vector classifier using polynomial or gausian kernel. SMO parameters are c=1.0; epsilon=1.0E-12; kernel=PolyKernel; num-folds=-1; randomspeed= 1.

1.3.3 Rules

Rule based classifier consist of certain rules for evaluating two classifier are used in this intrusion detection system

1.3.3.a. JRip:

RIPPER (JRip) is one of the most popular algorithm in rule based classifier. Classes are examined based on increasing size and an initial set of rules for the class is generated using incremental reduced error pruning. RIPPER is evaluated through JRip parameters are used is folds=3; minNo=2; optimization=2; seed=1; usePruning=true.

1.3.3.b. OneR

OneR is another basic algorithm for rule based model. It generates a one-level decision tree expressed in the form of set of rules that all test one particular attribute it consists of simple rules which are characterized as structure of data.

2 RELATED WORKS

In 1980 the concept of intrusion detection was first found by James Anderson [1]. He defined an intrusion model or threat classification model that develops a security monitoring inspection system based on detecting anomalies in user behavior.In 1987 [2] Dorthy Denning introduced solution for the problem of intrusion detection. Intrusion behavior modeled as abnormal behavior pattern. He used rule based pattern matching system model for commercial IDS development are based on statistics markov chains and time series etc.

In the late of 1990's the concept of data mining is integrated with network intrusion detection system.In 1999 Jake Ryan et al [3] has applied neural network to the intrusion detection system. It is used to learn the behavior pattern of the user and it is stored if the administrator finds some abnormal behavior from the user then the administrator gets alert. Back propagation technique is used for this process.

In [4], author proposed a framework for Network intrusion detection based on the Naïve bayes classifier. Experiment resulted in KDDCUP'99 dataset using naïve bayes classifier. It is observed that the proposed technique perform better in terms of false positive rate, cost and computation time when applied to KDDCUP'99 data sets compared to back propagation neural network based approach. In [5], author compared naïve bayes with the decision tree. Experimental study is done on the KDD'99 intrusion data set. By considering

three level of attacks, whole attack (all attack classes presented by KDD dataset in addition to the normal situation) five classes (categorizing four attacks Denial-Of-Service, Remote to Local, User to Root and probing) and Two classes (normal and abnormal by grouping all attacks in the same class i.e. Abnormal). For gathering attacks in the five class and two class cases two strategies are followed based on gathering before classification and gathering after classification. The idea of gathering before classification is modifying the dataset by grouping attacks belonging to the same attack category (i.e. DOS, R2L, U2R or probing) grouping them into a abnormal connection.

In gathering after classification type in five classes of training sets remains unchanged. Fuzzy logic [7] is integrated with the intrusion detection. It supports both anomaly detection and misuse detection components at both system level and network level both fuzzy and non-fuzzy are supported within the system. Genetic algorithm also used to tune the fuzzy variable used by the system and select the most effective. By combining Naïve bayes and decision tree [8], it analyzes large volume of network data and considers the complex properties of attack behavior to improve the performance of detection speed and detection accuracy. It minimizes the false positive and maximize balance detection rate on the 5 classes of KDD99 dataset.

Author [9] applied genetic algorithm to network intrusion detection. The genetic algorithm starts with a population that has randomly selected rules. Using the crossover and mutation operators can do the population. Due to the effectiveness of the evaluation function, the succeeding population is biased toward rules that match intrusion connections. Ultimately algorithm stops, rules are selected and added into the IDS rule base.

Evaluating the performance of genetic algorithm [10]. It is used to obtain the classification rules for intrusion detection. It filters the traffic data and reduces complexity; it classifies the network behavior either normal or abnormal. This approach is applied for KDD99 dataset and obtains high detection rate up to 99.87% as well as low false positive rate 0.003%. Finally the result of this approach is compared with available machine learning techniques.

Genetic algorithm based network intrusion detection [11] generate a rule using the principles of evolution in a GA to classify all types of smurf attack labels in the training data set, false positive rate is quite low at 0.2% and accuracy rate is high as 100% . Both support vector machine and neural network is applied for intrusion detection [12]. It delivers high accuracy (99% and higher) performance with SVM showing slightly better results as compared with neural network.

Comparing Decision tree with support vector machine [13] .Author evaluated performance of decision tree and SVM where decision trees give better overall performance than the SVM. As the decision tree was used as a binary classifier, results indicates decision tree gives better accuracy than SVM for probe, U2R and R2L classes whereas for normal class both gives same accuracy and for DOS class decision tree gives slightly worse accuracy than decision tree. U2R and R2L classes, which have small training data and for decision tree gives better performance than SVM where decision tree works well with small training data. The result shows testing time and training time of the classifiers are slightly better than SVM. Moreover decision tree is capable of multi-class classification, which is not possible with SVM.

3 EXPERIMENTS:

3.1 KDD cup99 Dataset:

KDD cup 99 dataset is based on 1998 DARPA, where dataset is based on four attack categories (Dos, probing, u2r and r2l) it consists of 41 features. 41 features are classified in to three groups.

1. Basic Features:

In this feature attributes are extracted from TCP/IP connection.

2. Traffic Features:

In this feature attributes are classified as same host features and same service features it is monitored based on past 2 seconds.

3. Content Features:

It consists of attribute which it look for suspicious behavior of the data portion

Following list shows 41 features used in dataset

1	Duration	22	Is_guest_login
2	Protocol type	23	Count
3	Service	24	Srv_count
4	Flag	25	Serror_rate
5	Src bytes	26	Srv_serror_rate
6	Dst_bytes	27	Rerror_rate
7	Land	28	Srv_rerror_rate
8	Wrong_fragment	29	Same_srv_rate
9	Urgent	30	Diff_srv_rate
10	Hot	31	Srv_diff_host_rate
11	Num_field_logins	32	Dst_host_count
12	Logged_in	33	Dst_host_srv_cont
13	Num_compromised	34	Dst_hostdst_same_srv_rate
14	Root_shell	35	Dst_host_diff_srv_rate
15	Su_attempted	36	Dst_host_same_src_portrate
16	Num_root	37	Dst_host_srv_diff_host_rate
17	Num_file_creation	38	Dst_hot_serror_rate
18	Num_shells	39	Dst_host_srv_serror_rate
19	Num_access_files	40	Dst_host_rerror_rate
20	Num_outbound_cmds	41	Class label
21	Is_hist_login		

For our experiment 10% kdd dataset is extracted it consist of 391458 records for DOS attack, 4107 instances for probe attack, 52 records for U2R, 1126 records for R2L and for normal connection 97277 records.

3.2 TYPES OF ATTACK AND ITS CATEGORY IN DATASET

DOS attack:

In this type of attack 6 attacks are categorized. Smurf consist of 280790 samples, Neptune consist of 107201 samples, back consist of 2203 samples, teardrop consist of 979 records, pod consist of 264 samples, land consist of 21 samples.

Probe attack:

In this category 4 types of attacks are categorized. Satan consists of 1589 records, ipsweep consist of 1247 records, portsweep consist of 1040 records, nmap consist of 231 records. Totally probe attack consists of 4107 records.

R2L:

Remote to Local attack consist of 8 attacks. Warezclient consist of 1020 samples, gues_passwd consist of 53 samples, warezmaster consist of 20 samples, imap consist of 12 samples, ftp_write consist of 8 sample records, multihop consist of 7 samples, phf consist of 4 records and spy consist of 2 samples totally 1126 record samples for R2L attack.

U2R:

User to Root consists of four attack category. Buffer_overflow consist of 30 samples, root kit consists of 10 record samples, load module consist of 9 records, perl consist of 3 records.

3.3 PERFORMANCE EVALUATION:

Performances are evaluated based on True Positive (TP) and False Positive (FP) and average accuracy of the algorithm.

True Positive:

This alarm corresponds to the number of detected attacks and its fact attack

False positive:

This alarm corresponds to the number of detected attacks but it is in normal

Average Accuracy:

Totally correctly classified instances / total instances

4 RESULTS:

We used an open source tool called weka, it is a machine learning package because it consist of collection of machine learning algorithm for data mining task that contains tools of preprocessing, classification, regression, clustering, association rules and visualization.

Bayes:

a. Bayes Net:

Table 1 shows TP and FP vales for each attack

Type of attack	**TP**	**FP**
DOS	94.6	0.2
Probe	83.8	0.13
U2R	30.3	0.3
R2L	5.2	0.6

b. Naïve Bayes:

Table 2 shows TP and FP vales for each attack

Type of attack	**TP**	**FP**
DOS	79.2	1.7
Probe	94.8	13.3
U2R	12.2	0.9
R2L	0.1	0.3

Function based classifiers:

a. MLP:

Table 3 shows TP and FP vales for each attack

Type of attack	**TP**	**FP**
DOS	96.9	1.4
Probe	74.3	0.1
U2R	20.1	0.1
R2L	0.3	0.5

b. SMO

Table 4 shows TP and FP vales for each attack

Type of attack	**TP**	**FP**
DOS	96.4	0.8
Probe	74.3	0.3
U2R	13.3	0.1
R2L	0.1	0.4

Rule based classifier:

a. JRIP

Table 5 shows TP and FP vales for each attack

Type of attack	**TP**	**FP**
DOS	97.4	0.3
Probe	83.8	0.1
U2R	12.8	0.1
R2L	0.1	0.4

b. OneR

Table 6 shows TP and FP vales for each attack.

Type of attack	**TP**	**FP**
DOS	94.2	6.8
Probe	12.9	0.1
U2R	10.7	2
R2L	10.7	0.1

Average Accuracy

Table 7 shows Average Accuracy of each classifier

Algorithm name	**Classifier**	**AA**
Bayes	Bayes net	90.62
	Naïve bayes	78.32
Function based classifier	MLP	92.03
	SMO	91.65
Rules	JRIP	92.30
	SMO	89.31

5 CONCLUSIONS

In this paper we presented three classification techniques in that we selected two classifier from each technique. Based on the attack classifying for DOS attack JRip perform well and for Probe attack Naïve bayes perform well and U2R attack bayesNet performs well and for oneR performs well. Performance based on Average Accuracy in bayes classifier by comparing BayesNet and Naïve bayes, where BayesNet have high accuracy, in function based classifier by comparing MLP and SMO, where MLP accuracy is high, based on rule based classifier by comparing JRip and SMO, where JRip performs well. By considering overall performance JRip rule based classifier performs well.

6 REFERENCES

[1] Anderson J.P “Computer Security Threat Monitoring & Surveillance”, technical Report, James Anderson.co Fort Washington, Pennyslavia 1980.
[2] Dorthy E.Denning, “An intrusion detection model”, IEEE transaction on Software Engineering”, SE-13(2), 1987 pp222-232.
[3] Jake Ryan Meng-Jang Lin, Risto Miikkulainen “Intrusion detection with Neural Network”
[4] Mrutyunjaya Panda and Manas Ranjan Patra “Network intrusion detection using Naïve Bayes” IJCNS December 2007
[5] Nahla Ben Amor, Salem Benferhat and Zied Eloedi “Naïve Bayes vs Decision trees in Intrusion detection system.
[6] Yacine Bouzida and Frederic Cuppens “Neural Network Vs decision trees for intrusion detection system.

[7] Susan M.Bridges and Rayford B Vaughn “Intrusion detection via Fuzzy data mining” Twelfth Annual Canadian Information Technology
[8] Dewan Md.Farid, Nouria Harbi and Mohammad Zahidur Rahman “Combinind naïve bayes and decision tree for Adaptive Intrusion detection” International Journal of Computer Security and its Application 2010
[9] Wei Li “Using Genetic Algorithm for Network Intrusion Detection” Department of Computer Science Engineering.
[10] B.Abdullah , I .Abd –alghafar , Gounda I.Salama and A.Abd-Alhafez “Performance Evaluation of a Genetic algorithm based approach to Network Intrusion detection”.
[11] Anup Goayal ,Chetan Kumar “GA-NIDS: A Genetic Algorithm Based Network Intrusion Detection”, Northwestern University Illionis.
[12] Srinivas Mukkamala, Guadalupe Janoski, Andrew Sung “Intrusion Detection: Support Vector Machine and Neural Networks”, Department of Computer Science, New Mexico Institute of mining and technology.
[13] Sandhya Peddabachigari, Ajith Abraham, Johnson Thomas “Intrusion Detection System Using Decision Trees and Support Vector Machines”, Department of Computer Science, Oklahoma State University, USA.