ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

Performance Evaluation of Intrusion Detection System Using Classification Algorithms

Dr C Manju

Assistant Professor, Department of Computer Science, Kanchi Mamunivar Centre for Post Graduate Studies, Puducherry, India

ABSTRACT: Communication systems are vital for effective transfer of information over a network. However, there are lot of threats in maintaining confidentiality and promoting high level security of information to increase the quality of network communication and increase the trust of the people involved in it. The main aim of this paper is to analyze performance of Intrusion Detection System using various classification approaches. The objective of this paper is to analyze and predict the network attacks by classifying them as normal and anomaly and to implement and measure the performance of our system here we the standard KDD99 benchmark dataset. An implementation of an Intrusion Detection System is proposed using Statistical methods, Probabilistic method, Ensemble methods and Rule based classifiers. The experimental results clearly shows that system achieve high accuracy rate in identifying whether the records are normal or anomaly ones and lower false alarm with reduced number of features using rule based classifiers.

KEYWORDS: Intrusion Detection System, KDD99 dataset, Classification Algorithms, Statistical method, Ensemble Method, Probabilistic Method, Rule based Classifiers

I. INTRODUCTION

An intrusion detection system is a system that can be used to detect hostile activities in network. The main aim of such a system is to prevent and detect the activities that affect the security of the system. It analyses and monitors various events happening in computer network and tries to recognize the abnormal activities. It's like acting against a normal activity internally and externally in the network.

The system does this process by monitoring traffic and activities in the network thereby finding malicious activities that occur in the network. This can be done in different ways. It can be network based, host based Signature based or anomaly based. The network based IDS monitor the traffic to and from all nodes in the network to monitor abnormal activities taking place in the network. The host based system run on each device or node in the network. It focuses on activities taking place in each node. The signature based system monitor packet in the network and malicious packet is fond by validation of signature. In anomaly based system is used to find anomalous activities that are taking place in the network[1][2]. From the categorization of different type of Intrusion detection system we can find that each system focuses on different field and a separate analysis and study is required for each field .This paper focuses on anomaly based intrusion and it focuses on the prediction of Intrusion in Network congestion and exaggerates the presence of intrusion in a network with the use of research tools that can predict the presence of intrusion in a host or network based architecture. The calculations are made based on the statistical analysis of data in the network dataset KDD'99 and results are predicted using rule based classification algorithms, statistical algorithms Probabilistic algorithms and Ensemble methods.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

The paper discuses various detection methods in section II .In section III and IV discuses about description of dataset and design for evaluation is done .Fine section IV describes evaluation of performance between rule based classification algorithms and other type of algorithms is done .

II. VARIOUS APPROACHES FOR ANOMALY BASED IDS

(i) Statically based intrusion detection [1][2] system relies on statically methods for finding normal or anomalous behaviour of network. It is defined by taking various samples over a period of time that is in a dataset. It identifies and track patterns and analyses network data and then assigns values according to the algorithm used. Based on the value it finds out percentage of data in normal or abnormal form.
Statistical anomaly detection falls into two categories. Threshold detection and profile based system [1]. Threshold detection involves counting the number of occurrences of a specific event type over specific interval of time. Profile based focuses on characterizing the past behaviour of individual users or related groups of users and detecting significant deviations. Various techniques include principal component analysis, Chi-Square distribution, Gaussian distribution system, Logistic regression, RBF network method, etc.

(ii) Rule based detection uses set of rules to detect and decide the behaviour of the various intruders in the system. Rules are applied to various events which ultimately give a decision regarding occurrence of suspicious activities in the system [3]. As in the case of statistical approaches here also records are analysed, instead of acquiring knowledge of security vulnerabilities in the system, this method generates rules that describe events in the system.

(iii)Signature based detection is like a pattern matching approach[1][2] and it make use of log files for tracking and finding anomalies, It requires updating of databases for pattern which will be vulnerable when network traffic is high ,in turn result in performance break down[4].

III. DATA SET DESCRIPTION

KDD Cup ’99 intrusion detection datasets which are based on DARPA ’98[9] dataset provides labelled data for researcher working in the field of intrusion detection and is the only labelled dataset publicly available. The KDD dataset is generated using a simulation of a military network consisting of three target machines running various operating systems and traffic. Finally, there is a sniffer that records all network traffic using the Tcpdump format. The total simulated period is seven weeks. Normal connections are created to profile that expected in a military network and attacks [5]
Network Intrusion Detection dataset is available at UCI machine learning data repository contains 41 fields with one class attribute. It includes both numeric and nominal attributes. The class shows whether the networks with intrusion are normal or anomaly. The intention of the dataset is to forecast the presence or absence of network attack to test the result. The network intrusion dataset contain 125973 samples belonging to two different target classes.
This data set consists of 41 feature and separate feature (42nd feature) that labels the connection as ‘normal’ or a type of attack [5].

IV. SYSTEM DESIGN

After Careful analysis and Review of Literature, the existing system and the system was formulated based on the efficiency of the algorithm. The overall system design is as follows
(i) Selection of data set

As prescribed in the above section KDDCUP data set is used [5].

.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

(ii) Applying feature selection on dataset

Feature Selection or attribute selection [6] is a process by which you search for the best subset of attributes in our dataset. They may be assessed by building a model and evaluating the accuracy of the model. Three key benefits of performing feature selection of our data are

- **Reduces Over fitting:** Less redundant data means less opportunity to make decision based on noise.
- **Improves Accuracy:** Less misleading data means modelling accuracy
- **Reduces Training time**: Less data means that algorithm train faster.

The feature selection [6] of the data is done using Correlation-based Feature Selection. This algorithm relies on a heuristic for evaluating the worth or merit of a subset of feature. This heuristic takes into account the usefulness of individual features for predicting the class label along with the level of inter correlation among them. A feature is said to be redundant if one or more of the other features are highly correlated with it.

(iii)Evaluation using Classifiers

In order to analyse the system and finding the attacks in the system, Data Mining techniques have been used here to classify the system to be normal or some mischief happened in it. This is done various classification approaches as discussed below.

(i) Probabilistic classifiers

- **Navie Bayes Classifier** [9]: It is a probabilistic classification technique based on Bayes' Theorem with an assumption of independence among predictors. . It is based on probability models that incorporate strong independence assumptions. The independence assumptions often do not have an impact on reality. Therefore they are considered as naive .In simple terms, a Naive Bayes classifier [9] assumes that the presence of a particular feature in a class is unrelated to the presence of any other feature. Bayes theorem provides a way of calculating the posterior probability, P(c|x), from P(c), P(x), and P(x|c). Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence.

  P(c/x) = P(x/c)*P(c)/ P(x) Where,

  P(c|x) is the posterior probability of class (target) given predictor (attribute).

  P(c) is the prior probability of class.

  P(x|c) is the likelihood which is the probability of predictor given class.

  P(x) is the prior probability of predictor [9].

The Navie Bayes is applied on the dataset of after feature selection dataset and the confusion matrix is generated for class intrusion having two possible values i.e. normal and anomaly.

Confusion Matrix is as follows

a	b	
65863	1480	a=normal
21336	37294	b=anomaly

It is inferred that the after training and testing the Naïve Bayes Classification Model has a prediction accuracy of 81.8% and the false positive is 19.2%.

(ii) Ensemble Classification Algorithm

Various classification algorithm can be combined to improve performance. Such algorithm is called as ensemble method [10]. Stacking is a parallel combination of classifiers in which different classifiers are executed in parallel and execution take place in meta level.

Stacking takes place in two phases.

- In the first phase each of the base level classifiers takes part in the j- fold cross validation training.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

- In the second phase this input is given for the Meta learning algorithm which adjusts the errors in such a way that the classification of the combined model is optimized [10].

This process is repeated for k-fold cross validation to get the final stacked generalization model [10]. The algorithm is applied on the dataset of after feature selection dataset and the confusion matrix is generated for class intrusion having two possible values i.e. normal and anomaly. Confusion Matrix is as follows,

a	b	
67343	0	| a=normal
58630	0	| b=anomaly

It is inferred that the after training and testing the Stacking Classification Model has a prediction accuracy of 53.45% and the false positive is 53.5%.

(iii). Statistical Approach

- Logistic regression model

This classifier is a statistical method[13] and is a product of combining the tree structure and the function of logistic regression[13] to construct a single decision tree that has the function of logistic regression in the leaves , that is providing a model of piecewise linear regression which is a real valued .

The Confusion Matrix is

a	b	<-- classified as
63429	3914	| a = normal
11592	47038	| b = anomaly

It is inferred that the after training and testing the Logistic regression model has a prediction accuracy of 87.6% and the false positive is 13.3%.

- RBFnetwork method

A Radial Basis Function (RBF)[14] network classification by measuring the similarity of input to data in the dataset. It is an artificial neural network that uses radial basis functions as activation functions. It is a linear combination of radial basis functions. They are used in function approximation, time series prediction, and control. It have three layers: an input layer, a hidden layer with a non-linear RBF activation function and a linear output layer. The output of the network is thus $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ [14] Where N is the number of neurons in the hidden layer, is the centre vector for neuron i, and are the weights of the linear output neuron.

The Confusion Matrix is as follows

a	b	<-- classified as
47320	20023	| a = normal
2110	56520	| b = anomaly

It is inferred that the after training and testing the RBFnetwork method has a prediction accuracy of 82.4% and the false positive is 17.5%.

(iv) Rule based classifiers [7]

Various classification algorithms used here for evaluation are rule based systems[7]. It uses set of rules to make decisions about the data in the intrusion detection system. Set of predefined rules are induced to the collection of facts that are given by system or administrator. Facts are condition representing certain situation and rules represent action to be executed in case of condition.This classification is done by generating set of rules for classifying dataset [7].

- Ripple down rule leaner(Ridor)

It generates a default rule first and then exceptions for default rule with least (weighted error) rate are found [10]. The incremental reduced error pruning is used to find exceptions with smallest error rate, finding the best exceptions .It generates a tree like expansions of exceptions.

Confusion Matrix is as follows

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

```
    a        b   <-- classified as
66975      368 |    a = normal
  608    58022 |    b = anomaly
```

For it is inferred that the after training and testing the Ripple down rule leaner Model has a prediction accuracy of 99.2% and the false positive is 0.8%.

- Repeated Incremental Pruning to Produce Error Reduction(Ripper)

It is an interfaced and rule based learner used to classify elements with proportional rules [11]. This is a direct method that is used to extract the rules directly from data. The algorithm progresses through 4 phases such as growth, pruning, optimization and selection.

Confusion Matrix is as follows

```
    a        b   <-- classified as
67184      159  |    a = normal
 1442    57188 |    b = anomaly
```

It is inferred that the after training and testing the Ripper Model has a prediction accuracy of 98.5% and the false positive is 1.5%.

- Genetic Algorithm

Genetic Algorithm (GA) is a search-based optimization [12] technique based on the principles of Genetics and Natural Selection. It is frequently used to find optimal or near-optimal solutions to difficult problems which otherwise would take a lifetime to solve. It is frequently used to solve optimization problems, in research, and in machine learning. Optimization is the process of making something better. Confusion Matrix is

```
    a        b
41971    10752    | a= anomaly
  287    60366    | b= normal
```

It is inferred that the after training and testing the Genetic Algorithm Model has a prediction accuracy of is 91.09% and the false positive is 8. 31%.

V. EXPERIMENTAL RESULTS

After careful experimentation, various new results are identified and presented as input to the system. The results were identified from WEKA for generation of reports. The following results were identified for the algorithms based on the following parameters:
The table below show the accuracy and precision of various classifiers for Intrusion Detection System. However, the rule based system showed tremendous improvement for other methods such as statistical, embedded methods. They show high accuracy with low false positive [15].
The table also show that the rate of accuracy and precision along with other factors like Recall, F-Measure, Kappa and FP Rate increases with the rule

Classification	Accuracy	Precision	Recall	F-measure	Kappa	FP-Rate
Logistics	87.7	88.2	87.7	87.6	75.5	13.3
RBF network	82.4	85.5	82.4	82.2	65.38	17.5
NaiveBayes	81.89	85.1	81.19	81.2	62.7	20.5
Stacking	53	28.6	53.5	37.2	50	53
Genetic Algorithm	91.09	98.38	98.43	98.41	94.7	8.31
Ripper	98.5	98.7	98.5	98.5	97	1.5
Ridor	99.2	99.6	98.8	99.6	99	0.8

Fig. 1. Table showing performance of classifiers seven classifiers

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

The graphical representation is as follows

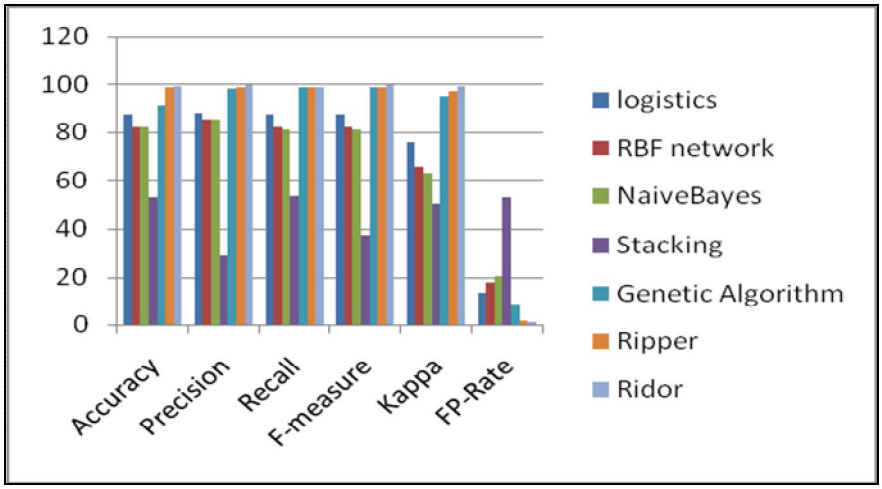

Fig. 2 Graphical representation of performance of classifiers on various features

From the above figures it can be inferred that after feature selection using rule based classifiers has the highest prediction accuracy. Hence the rules generated by this model can be used for Intrusion Detection System.

VI. CONCLUSION

Anomaly based Intrusion Detection System is to classify the network intrusion whether the intruder is normal or anomaly .Here an analysis of performance is done on statistical classifiers, Ensemble Classifiers, Probabilistic classifiers and Rule based classifiers. , It is found that rule based classifiers provide a high accuracy rate of above 97% after feature selection method with low false positive around 3%. So we can conclude rule based classifiers are suitable for Intrusion Detection System when evaluated with KDD99 dataset.

REFERENCES

[1] Krishna Kant Tiwari , Susheel Tiwari , Sriram Yadav :" Intrusion Detection Using Data Mining Techniques" International Journal of Advanced Computer Technology (IJACT).
[2] Trupti Phutane, Apashabi Pathan :" A Survey of Intrusion Detection System Using Different Data Mining Techniques" International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 2, Issue 11, November 2014.
[3] Subaira.A.S , Anitha.P : "A Survey: Network Intrusion Detection System based on Data Mining Techniques" International Journal of Computer Science and Mobile Computing Vol.2 Issue. 10, October- 2013, pg. 145-153
[4] Jaina Patel , Mr. Krunal Panchal :" Effective Intrusion Detection System using Data Mining Technique" June 2015, Volume 2, Issue 6.
[5] Safaa O. Al-mamory , Firas S. Jassim :"Evaluation of Different Data Mining Algorithms with KDD CUP 99 Data Set" Journal of Babylon University/Pure and Applied Sciences/ No.(8)/ Vol.(21): 2013 .
[6] Jialei Wang, Peilin Zhao, Steven C.H. Hoi, and Rong Jin:" Online Feature Selection and Its Applications" IEEE Transactions On Knowledge And Data Engineering. 2014 - ieeexplore.ieee.org
[7] Asim Das & S.Siva Sathya:" Association Rule Mining For KDD Intrusion Detection Data Set ", International journal for Computer Science and Informatics ISSN :2231-52932 Volume 2, Issue 6.
[8] Safaa O. Al-mamory , Firas S. Jassim :"Evaluation of Different Data Mining Algorithms with KDD CUP 99 Data Set" Journal of Babylon University/Pure and Applied Sciences/ No.(8)/ Vol.(21): 2013
[9] Saidi Ben Boubaker Ourida " A Comparative Study of Spam Detection in Social Networks Using Bayesian Classifier and Correlation Based Feature Subset Selection" International Journal of Computer Sciences and Engineering, Volume-3, Issue-8, E-ISSN: 2347-2693
[10] S S Roy ,P V Krishna" Analyzing intrusion detection system : stacking ensemble approach" ,Signal Processing and Information Technology,2014,IEEE international Symposium , ISSN 2162-7843

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 7, July 2017

[11] http://www.eecs.yorku.ca/tdb/_doc.php/userg/sw/weka/doc/weka/classifiers/rules/package-summary.html
[12] http://www.csee.usf.edu/~lohall/dm/ripper.pdf
[13] https://www.tutorialspoint.com/genetic_algorithms/genetic_algorithms_tutorial.pdf
[14] Deeman Yousif Mahmood , Classification Trees with Logistic Regression Functions for Network Based Intrusion Detection System IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661,
[15] J. Sudhir Kumar, K. Venkata Subbaiah, and P. V. V.Prasada Rao, Prediction of Municipal Solid Waste with RBF Net Work- A Case Study of Eluru, A.P, India " International Journal of Innovation, Management and Technology, Vol. 2, No.
[16] David J. Weller-Fahy, Brett J. Borghetti, and Angela A. Sodemann :" A Survey of Distance and Similarity Measures used within Network Intrusion Anomaly Detection" Citation information: DOI 10.1109/COMST.2014.2336610, IEEE Communications Surveys & Tutorials