International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 03 | Mar-2016 www.irjet.net p-ISSN: 2395-0072

INTRUSION DETECTION MODEL IN DATA MINING BASED ON ENSEMBLE APPROACH

VIKAS SANNADY [1], POONAM GUPTA [2]

[1]*Asst.Professor, Department of Computer Science, GTBCPTE, Bilaspur, chhattisgarh, India*
[2]*Asst.Professor, Department of Computer Science, GTBCPTE, Bilaspur, chhattishgarh, India*

Abstract - *Nowadays the need for secure networks is tremendously increased so Security of computers and the networks that connect them is increasingly becoming of great significance. Computer attack has become very common. Although there are many existing mechanisms for Intrusion detection, but the major issues is the security and accuracy of the system. In data mining based intrusion detection system we should have thorough knowledge about the particular domain in relation to intrusion detection so as to efficiently extract relative rule from huge amounts of records. In this paper we investigate and evaluate the ensemble bagging data mining techniques as an intrusion detection mechanism. Our research shows that bagging decision trees gives better overall performance.*

Key Words: **Bagging ensemble approach, intrusion detection System.**

1.INTRODUCTION

An intrusion detection system (IDS) monitors network traffic and monitors for suspicious activity and alerts the system or network administrator. In some cases the IDS may also respond to anomalous or malicious traffic by taking action such as blocking the user or source IP address from accessing the network. There are various types of intrusion detection techniques -

NIDS: Network Intrusion Detection Systems are placed at a strategic point or points within the network to monitor traffic to and from all devices on the network.

HIDS: Host Intrusion Detection Systems are run on individual hosts or devices on the network.

Signature Based: Signature based IDS will monitor packets on the network and compare them against a database of signatures or attributes from known malicious threats. This is similar to the way most antivirus software detects malware.

Anomaly Based: An IDS which is anomaly based will monitor network traffic and compare it against an established baseline. The baseline will identify what is normal for that network [1].

This technique aims to design and develop security architecture for computer networks. We build the model to improve the classification rate for known attacks with minimum number of false alarm rate. We train and test our proposed model on the normal and the known attacks in NSL-KDD datasets. The proposed system should have an adaptive capability.

LITERATURE SURVEY

In this portion we discuss the literature survey on Classification techniques applied on the IDS data. There are many research papers published regarding the classifiers in order to detect the intrusions in the network dataset. The following are the some of the related works described below.

Manikandan R et.al [2] proposed a new
Ensemble boosted decision tree approach for intrusion detection system.

Amit Kumar et.al [3] proposed the purpose of an intrusion detection system is to detect attacks. However, it is equally important to detect attacks at an early stage in order to minimize their impact. I have used Dataset and Classifier to refine Intruders in Networks.There are a variety of approaches of intrusion detection, such as Pattern Matching, Machine Learning, Data Mining, and Measure Based Methods. This paper aims towards the proper survey of IDS so that researchers can make use of it and find the new techniques towards intrusions.

Zibusiso Dewa et.al [4] this article gives an overview of existing Intrusion Detection Systems (IDS) along with their main principles. Also this article argues whether data mining and its core feature which is knowledge discovery can help in creating Data mining based IDSs that can achieve higher accuracy to novel types of intrusion and demonstrate more robust behaviour compared to traditional IDS.

Poonam Gupta et.al [5] in this paper they investigate and evaluate the decision tree data mining techniques as an intrusion detection mechanism. Our research shows that Decision trees gives better overall performance.

Vivek Nandan Tiwari et.al[6] This constraint has guide researchers in the IDS society to not only extend better detection algorithms and signature tuning methods, tuning methods, but to also focus on determining a variety of relations between individual alerts, formally known as alert correlation.

Priya U. Kadam et.al [7] proposes effectiveness and accuracy of an approach to generate rules for different types of anomalous connection. The KDDCUP99 training and testing dataset is used to generate effective new rules by adopting reasonable detection rate.

Sandhya Peddabachigari et.al [8] In this paper we investigate and evaluate the decision tree data mining techniques as an intrusion detection mechanism and we Compare it with Support Vector Machines (SVM). Intrusion detection with Decision Trees and SVM were tested with benchmark 1998 DARPA Intrusion Detection dataset. Our research shows that Decision trees gives better overall performance than the SVM.

Heba Ezzat Ibrahim et.al [9] our experimental results showed that the proposed multi-layer model using C5 decision tree achieves higher classification rate accuracy, using feature selection by Gain Ratio, and less false alarm rate than MLP and naïve Bayes. Using Gain Ratio enhances the accuracy of U2R and R2L for the three machine learning techniques (C5, MLP and Naïve Bayes) significantly. MLP has high classification rate when using the whole 41 features in Dos and Probe layers.

K.Nageswara rao et.al [10] in this paper, we evaluated the influence of attribute pre-selection using Statistical techniques on real-world kddcup99 data set. Experimental result shows that accuracy of the C4.5 classifier could be improved with the robust pre-selection approach when compare to traditional feature selection techniques.

PROPOSED ARCHITECTURE

Proposed Research work introduces a new framework for offline analysis as shown in fig 1. For network intrusion detection. In this framework KDD99cup [11] dataset is given to Preprocessing stage which includes bagging technique.

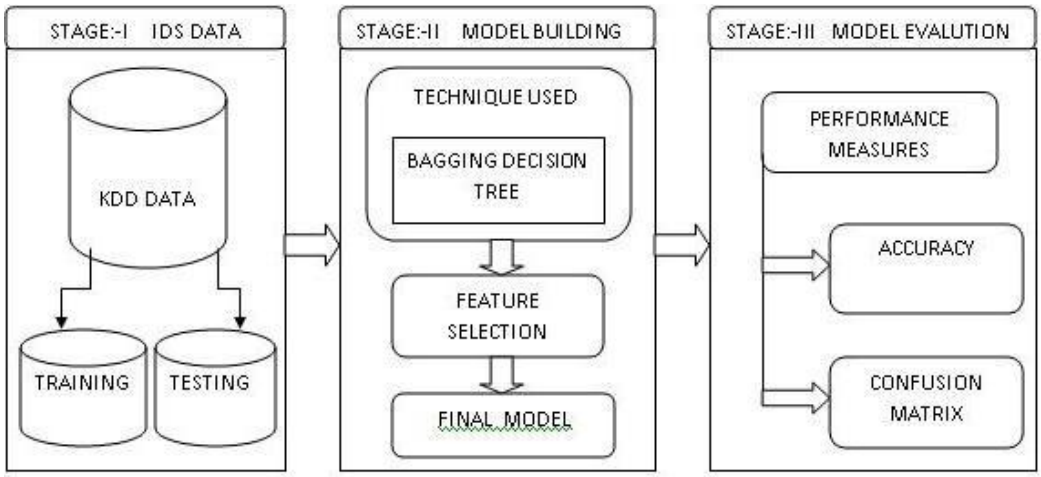


Fig 1: Proposed Architecture for Bagging Technique.

OUR APPROACH

This proposed model uses bagging decision tree i.e. hoeffding tree classification techniques to increase performance of the intrusion detection system. An ensemble model is a combination of two or more models to avoid the drawbacks of individual models and to achieve high accuracy. The two models are combined by using high confidential wins scheme where weights are weighted based on the confidence value of each prediction. Then the weights are summed and the value with highest total is again selected. The confidence for the final selection is the sum of the weights for the winning values divided by the number of models included in the ensemble model. In this work many models with the combination of all the above individual models are tried and finally an ensemble model is selected, because this model produces highest accuracy among all the ensemble models as well as individual models [12].

BAGGING

For each trial t= 1, 2…T. a training set of size N is sampled (with replacement) from the original instances. This training set is the same size as the original data but some instance may not appear in at. While others appears more than once. The learning system generates a classifier C^t

from the sample and the final classifiers C^* is formed by aggregating the T classifier from these trails. To classify an instance x, a vote for class k is recorded by every classifier for which $C^t(x) = k$ and $C^*(x)$ is then the class with the most votes.

Using CART as the learning system, breiman (1996) reports result of bagging on seven moderate size datasets. With the number of the replicates T set at 50 the average error of the bagged classifier C^* ranges from 0.57 to 0.94 of the corresponding error when a single classifier is learned. Breiman introduces the concept of an ordered-correct classifier learning system as one that over many training sets tends to predict the correct class of a test instance more frequently than any other class. An order correct learner may not produce optimal classifier but breiman shows that aggregating classifier produced by an order correct learner result in an optimal classifier. Breiman notes:

"The vital element is the instability of the prediction method if perturbing the learning set can cause significant changes in the predictor constructed then bagging can improve accuracy [13]."

Evaluation on KDDCup'99 Data Set

The experiment is carried out on a intrusion detection real data stream which has been use in the Knowledge Discovery and Data Mining (KDD) 1999 Cup competition. In KDD99 dataset the input data flow contains the details of the network connections, such as protocol type, connection duration, login type etc. Each data sample in KDD99 dataset represents attribute value of a class in the network data flow, and each class is labeled either as normal or as an attack with exactly one specific attack type. In total, 41 features have been used in KDD99 dataset and each connection can be categorized into five main classes as one
normal class and four main intrusion classes as DOS, U2R, R2L and Probe. There are 22 different types of attacks that are grouped into the four main types of attacks DOS, U2R, R2L and Probe tabulated in Table I[2].

Table I. Different Types of Attacks

Main Attack Classes	22 Different Attack types
DOS – Denial of service	Back , land, Neptune , pod , smurf , teardrop
U2R –User to Root	Buffer_ overflow , loadmodule , perl , rootkit
R2L – Remote to user	ftp_write , guess_password , imap , multihop , phf , spy.
Probe	Ipsweep , nmap , portsweep

EXPERIMENTAL RESULTS

The experimental setting is for the KDD99 Cup, taking 10% of the whole real raw data
Streams are selected as per proposed algorithm.
Table II: Testing the system by cross validation datasets- However, in this experiment k=10 have highest accuracy & its Confusion Matrix is also given below

Cross Validation Technique								
Datasets used for testing	Correctly Classified instance	Incorrectly Classified instance	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
K=2	99.583	0.416	0.996	0.004	0.996	0.996	0.996	0.999
K=4	99.607	0.393	0.996	0.004	0.996	0.996	0.996	0.999
K=6	99.611	0.389	0.996	0.004	0.996	0.996	0.996	0.999
K=8	99.626	0.373	0.996	0.004	0.996	0.996	0.996	0.999
K=10	99.642	0.357	0.996	0.004	0.996	0.996	0.996	0.999

Confusion Matrix for k=10:-

a	b		Classified as
13414	35	|	a=Normal
55	11688	|	b=Anomaly

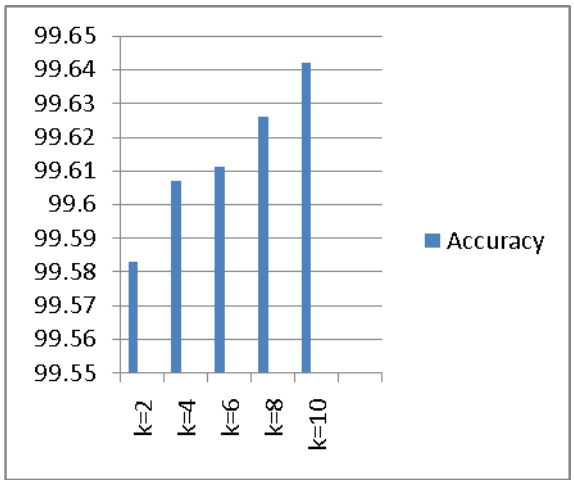
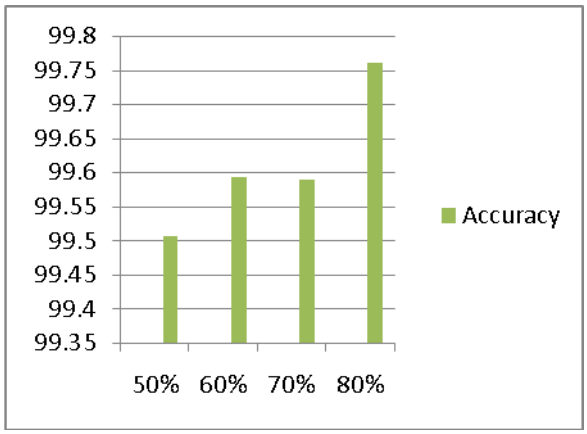


Table III: Testing the system by splitting datasets on different percentage

Percentage Split Technique								
Percentage Split	Correctly Classified instance	Incorrectly Classified instance	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC
50%	99.507	0.492	0.995	0.005	0.995	0.995	0.995	0.999
60%	99.593	0.406	0.996	0.004	0.996	0.996	0.996	0.999
70%	99.589	0.410	0.996	0.004	0.996	0.996	0.996	0.999
80%	99.761	0.238	0.998	0.003	0.998	0.998	0.998	0.999

Confusion Matrix for 80%:-

a	b		Classified as
2675	3	|	a=Normal
9	2351	|	b=Anomaly



CONCLUSION

Intrusion detection classification is a crucial and essential task now a days for security. Data mining techniques provides facility to design and develop predictive model for intrusion detection classification. This paper explores various data mining techniques to design an ensemble model for classification of intrusion related security. A testing accuracy of model show the efficiency of ensemble model. Models are also measured in terms of confusion matirix. Result show alternative as a instrusion detection that ensemble model gives higher accuracy than using single model.

REFERENCES

[1]http://netsecurity.about.com/cs/hackertools/a/aa030504_2.htm.
[2] Manikandan R , Oviya P , Hemalatha C , "A New Data Mining Based Network Intrusion Detection Model "Journal of Computer Applications ISSN: 0974 – 1925, Volume-5, Issue EICA2012-1, February 10, 2012 .
[3] Amit Kumar, Harish Chandra Maurya, Rahul Misra , "A Research Paper on Hybrid Intrusion Detection System ",International Journal of Engineering and Advanced

Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-4, April 2013 .
[4] Zibusiso Dewa , Leandros A. Maglaras," Data Mining and Intrusion DetectionSystems ," International Journal of Advanced Computer Science and Applications, Vol. 7, No 1,2016
[5] Poonam Gupta , S. R. Tandan , Rohit Miri , "Decision Tree Applied For Detecting Intrusion ," International Journal of Engineering Research & Technology (IJERT) Vol. 2 Issue 5, May - 2013 ISSN: 2278-0181.
[6] Vivek Nandan Tiwari, Prof. Kailash Patidar,Prof .Satyendra Rathore, Prof. Kumar Yadav, "A Comprehensive Survey of Intrusion Detection Systems ," Computer Engineering and Intelligent Systems ISSN 2222-1719 (Paper) ISSN 2222-2863 Vol.7, No.1, 2016
[7] Priya U. Kadam, P. P. Jadhav," An effective rule generation for Intrusion Detection System using Genetics Algorithm", International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 10, October 2013
[8]Sandhya Peddabachigari, Ajith Abraham*, Johnson Thomas,"Intrusion Detection Systems Using Decision Trees and Support Vector Machines".
[9]]Heba Ezzat Ibrahim,Sherif M. Badr, Mohamed A. Shaheen," Adaptive Layered Approach using Machine Learning Techniques with Gain Ratio for Intrusion Detection Systems", International Journal of Detection Systems Using Decision Trees and Support Vector Machines".
[10]K.Nageswara rao, D.RajyaLakshmi, T.Venkateswara Rao," Robust Statistical Outlier based Feature Selection Technique for Network Intrusion Detection", International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231-2307, Volume-2, Issue-1, March 2012.
[11] "NSL-KDD data set for network-based intrusion detection systems ", Available on: http://nsl.cs.unb.ca/NSL-KDD.
[12] J.R.Quinlan ," Bagging,Boosting, C4.5".
[13] H.S.Hota," Diagnosis of Breast Cancer Using Intelligent Techniques", International Journal of Emerging Science and Engineering (IJESE) ISSN: 2319–6378, Volume-1, Issue-3, January 2013.

BIOGRAPHIES

Vikas Sannady has received his Bachelor degree in Computer Science from Gurughasi Das University, Chhattishgarh, India in 2010. And master Degree from Gurughasi Das Central University, Chhattishgarh, India in 2013. He is currently working as an Asst. Pro. In the Department of Computer Science. His current research interest includes data mining in cyber security.

Poonam Gupta has received her B.E. degree in Computer Science & Engg. from C.S.V.T.U. University, Chhattishgarh, India in 2011. And M.Tech Degree from Dr. C.V.Raman University, Chhattishgarh, India in 2013. She is currently working as an Asst. Pro. in the Department of Computer Science. Her current research interest includes data mining in cyber security.