Performance Evaluation of Classification Techniques for Intrusion Detection in Noisy Datasets

Prince Vijay
Computer Science & Technology
Central university of punjab
Princevijay6002@gmail.com

Abstract:- Data mining provides a useful environment and set of tools for processing large datasets such as Intrusion Detection Systems` (IDS) logs. Researchers improve existing IDS models by comparing the performance of various algorithms on these datasets. It is very important to keep in mind that an IDS often has to work in a noisy network environment. Network noise is one of the most challenging issues for efficient threat detection and classification. In this study, normal and noisy datasets for network IDS domain are used and various classification algorithms are evaluated. The results show that an evaluation of algorithms without noise is misleading for IDSs since algorithms that perform best without noise do not necessarily achieve the same in a realistic noisy environment. Moreover refined NSL KDD dataset allows a more realistic evaluation of various algorithms than the original KDD 99 dataset.

Keywords—*Intrusion detectionSystem; Security; Anomaly Detection;NSL,KDD,Noisy datsets; Classifiers,data mining*

*****

I. INTRODUCTION

Security concerns are inevitable for information technology (IT) because of the value of data. Three pillars of information security are confidentiality (only authorized data disclosure); integrity (accuracy and consistency) and availability (accessibility when it is required). Several threats jeopardize security of IT assets in cyber space. Intrusion is "any set of actions that attempt to compromise the integrity, confidentiality and availability of a resource. Intrusion detection system(IDS) is a major countermeasure against intrusions and malicious attempts. IDSs monitor events at the endpoints or in the network depending on their deployment for harmful actions which are likely to cause violations. It is an increasing trend as more assets connect to internet and/or intranet every single day. Since several services are delivered through IT networks, they are valuable targets. Public institutions, universities and private companies provide (inter) net-enabled services and the health of these networks is critical. The value of data is very high and it is risky to ignore even a slightest threat onthe security. IDSs are one of the most significant network boundary and system protection mechanisms against malicious parties with unauthorized access intentions to others' information. The efficiency of IDS is heavily affected by network noise which is an interference that can trigger a false alarm. IDSs may throw a lot of false alarms because of the noisy environment it has to work in so noisecan reduce the reliability of IDS. Current IDSs often suffer from the noise in real networks. IDSs may have a high "false-alarm threshold" that causes certain attacks to be ignored in order to alleviate credibility Classification of Intrusion Detection Systems are Intrusion Detection Systems can be classified based upon Data Collection & Storage and Data analysis & processing**.**

Data collection & storage based Intrusion Detection Systems: Classification of IDS based on data collection & storage is further divided into two types which are as under:

Network Based System Intrusion Detection Systems (NIDS): NIDS examines all traffic on the entire network. It can detect intrusions that cross a specific network segment. Administrators sometimes place IDS sensor units inside and as well as outside of the firewall. NIDS are not able to see processes running and data stored in the memory of computers.

Host Based Intrusion Detection Systems (HIDS): Host Intrusion Detection Systems are deployed on individual hosts or devices on the network. A HIDS monitors all the traffic and activity for a particular machine or a host and will alert the user or administrator of suspicious activity is detected.

Data analysis & processing based Intrusion Detection Systems: Classification of IDS based on data analysis and

processing is further divided into two types:

Signature based Intrusion Detection Systems: A signature based IDS 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. The issue is that there will be a lag between a new threat being discovered in the field and the signature for detecting that threat being applied to your IDS. During that lag time your IDS would be unable to detect the new threats.

Anomaly based Intrusion Detection Systems: An anomaly based IDS detects the intrusions on the basis of comparing the active behavior of the network with its normal behavior and generates an alert for an intrusion if behavior differs from normal behavior. It can detect new types of attacks but requires more overhead and processing and may generate many false positives

This paper is composed as follows: in Section 2, related works are described, and in Section 3, we propose an intrusion detection system and explain its whole structure and each detection model. In Section 4, we describe an experiment using actual packets of NSL network and its results, and the last section, Section 5, is a conclusion.

II. RELATED WORKS

Intrusion detection techniques have been published in various studies at home and abroad based on whitelist, communication pattern, traffic and machine learning, etc.

In The task for the classifier learning contest organized in conjunction with the KDD'99 conference was to learn a predictive model (i.e. a classifier) capable of distinguishing between legitimate and illegitimate connections in a computer network[1]. A study by suggested dataset of Lincon laboratory dataset for military purpose intrusion detection system the Lincoln Laboratory of MIT conducted a comparative evaluation of Intrusion Detection Systems developed under DARPA funding[2] proposed a So**me methodologies used in the evaluation are questionable and may have biased its results** Categories and Subject Descriptors: K.6.5 [Management of Computing and Information Systems] Security and Protection—Invasive software (e.g., viruses, worms, Trojan horses) General Terms: Security Additional Key Words and Phrases: Computer security, intrusion detection, receiver operating curves (ROC), software evaluation[3]tried to use a different approach than the previous works here author describe usefulness of DARPA dataset for intrusion detection system evaluation using java concept that detect lincon dataset 1998 the military dataset for calculation of intrusion in week based data.[4]. All the experiments were performed using WEKA Tool on equally selected instances of five class category of attacks (2016) to prevent oversampling of normal classes over minority attack categories using 10 folds Cross validation[5]. Author kajal rai An IDS can bebroadly classified as Signature based IDS and Anomaly based IDS. In our proposed work, the decision treealgorithm is developed based on C4.5 decision tree approach.[6]. The authors **(Tavallaee & et.al., 2009)** proposed another idea about how anomaly detection coverage signature based system for this purpose KDDCUP 99 dataset used to evaluate performance.

[7]. **Gandhi & kumaravel Appavoo** describe some features in addition to algorithm for performance of classifier the era of information society, computer networks and their related applications are the technologies. In this paper, we evaluate the performance of a set classifier algorithms of rules (JRIP, Decision Table, PART, and OneR) and trees (J48,RandomForest, REPTree, NBTree).In **Nikolaos Avouris & Daskalaki**paper This study concludes to a framework that provides the “best ” classifiers, identifies the performance measures that should be used as the decision criterion and suggests the “best ” class distribution based on the value of the relative gain from correct classification in the positive class.[8], a detection method based on network flow and periodicity was proposed based on the idea that a number of cyber attacks targeting control systems cause the change of traffics such as the periodicity, size and noise of network data. It has a limitation that it can be applied only part of attack types. In A **Sabri & kamaruzzaman seman,**Denial of Service (DOS), Probes and User to Root (U2R) attacks[9], a flow-based abnormal behavior detection method was proposed, which measured average packet size and average inter-arrival time in a specific time interval.

 ISSN: 2321-8169
1011 – 1016

III. PROPOSED INTRUSION DETECTION SYSTEM FOR UNDER NOISY DATSETS

A. *Data Preprocessing*

Classification Algorithms

This section describes the Classification methods used to perform experiments in this research work. Classification is a data mining technique that is used for predicting the class or group membership among the instances in dataset.

Lazy Classifier

k-Nearest Neighbors algorithm(k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space.

Neural network Algorithm

In more practical terms neural networks are non-linear statistical data modeling tools. They can be used to model complex relationships between inputs and outputs or to find patterns in data. Using neural networks as a tool, data warehousing firms are harvesting information from datasets in the process known as data mining **Multilayer Perceptron** In a multilayer neural network, there are three kinds of layers. Each layer contains a set of neurons. The first layer, called input layer, sets the activation of its neurons according to the provided pattern in question. The output layer provides the answer of the network.

Fuzzy Algorithms

In Fuzzy logic algorithm there is fuzzy thresholds or boundaries that are defined for each category. This is an advantage over rule-based system as these rule-based systems involve sharp cut-offs for continuous attributes. Each category in fuzzy algorithm then represents a Fuzzy set. **Decision tree** is a recursive and tree like structure for expressing classification rules. It uses divide and conquer method for splitting according to attribute values. Classification of the data proceeds from root node to leaf node, where each node represents the attribute and its value & each leaf node represent class label of data. **Random Forest** is first introduced by Lippett et.al. and it is ensemble classification technique which consists of two or more decision trees. In Random Forest, every tree is prepared by randomly select the data from dataset.

Random Tree as its name indicating it's a tree build by picking

Attack in datasets	**Attack Types (37)**
Dos	Back,Land,Neptune,Pod,Smurf,Teardrop,Maildrop,Processtable,UdpStrom,Apache2,Worm
Probe	Satan,IPsweep,Nmap,Portsweep,Mscan,Saint
R2L	Guess_password,Ftp_write,Imap,phf,Multihop,Warezmaster,Xlock,Xsnoop,Snmpguess,Httptunnel, Sendmail,Named
U2R	Buffer_overflow,Loadmodule,Rootkit,Perl,Sqlattack,Xterm,Ps

Fig Attacks in network system

Random branches from a possible set of trees. Each tree has an equal probability of being get sampled in this algorithm or we can say the trees are distributed in a uniform way.

Rule based classifiers

PART a Indirect Method ¾ Extract rules from other classification models (e.g. decision trees, etc.). PART of z Combines the divide-and-conquer strategy with separate-and conquer strategy of rule learning.

Meta Classifiers

Random committee These types of classifiers find the optimal set of attributes. After finding the best possible setup the meta-classifiers then train an instance of the base classifier with these parameters and use it for subsequent predictions (Kamari, 2013).

B. *Measuring Diversity under noisy datasets*

KDD datasets Dataset was produced from DARPA IDS Dataset 1998 which was then widely used in several studies and researches. Its effectiveness seems to be controversial and it is criticized as "harmful, results based on it are questioned due to using synthetic simulated normal data with scripted attack data" and its further use is discouraged in (McHugh, 2000)

NSL KDD data

Further investigation about the usability of the original dataset revealed a new study that was performed on a modified

(refined) version which was named as NSL KDD(Gandhi & kumaravel Appavoo, 2010) It is a more recent dataset which claims solving problems like no validation against real world data, low data rate, traffic irregularities and huge number of records by removing duplicates and radically reducing the redundancies.

The experiments wereperformed on full training data set having 125976 records and test data set having 42831 records. The KDD99 dataset consists of 42 features and one class tribute. The class attribute has 42 classes that fall under four types of attacks: Probe attacks,User to Root (U2R) attacks, Remote to Local (R2L) attacksand Denial of Service (DoS) attacks

C.Problem statement

In networking environment network noise is challenging issues for efficent threat detection and classification due to this quality of intrusion detection system degrades.The algorithm which perform without noise not give proper idea about accuracy of Intrusion detection system and unable to do best in realistic noisy environment.There is need to evaluate performance of different classification algorithms with network log dataset

IV. EXPERIMENTAL RESULT

*A. **Training Dataset and Experimental Environment***

Rsearch work will be carried out at Software Lab/Research Lab Nof Centre for Computer Science and Technology, Central University of Punjab.

-WEKA (Waikato Environment for Knowledge Analysis) environment using 43 attributes

-Dataset :NSL,KDD Datasets

-Training:Around 94,000 instances from complete NSL KDD dataset
-Testing:48,000 instances
-False positive (FP): It defines the number of activities classified as an attack while the activities are actually normal activities.

-False negative (FN) The IDS classify an intrusive activity as normal one.

-True positive (TP): It is situation when IDS triggers alarm in response to an attack.

-True negative (TN): An event when no attack has taken place and no detection is made.

Accuracy: It is the percentage of correct predictions. On the basis of Confusion Matrix it is calculated by using the formula below:

$Accuracy = TP+TN/n$ Here n is total number of instances.

Mean Absolute Error: It is the mean of overall error made by classification algorithm. Least the error and best will be the classifier.

TPR: True Positive Rate is same as accuracy so we have not considered this metrics.

FPR: False Positive Rate is calculated by using the formula:

$FPR=FP/TN+FP$

Recall: It is the proportion of instances belonging to the positive class that are correctly predicted as positive. $Recall=TP/TP+FN$

Precision: It is a measure which estimates the probability that a positive prediction is correct $Precision=TP/TP+FP$

Data Preprocessing

Iin evaluating classifiers KDD give less accuracy as compared NSL dataset. Here accuracy in NSL (%) shows KDD have less accuracy as compared to NSL dataset. For e.g. Random forest classifieds NSL dataset with 99.94% accuracy whereas only achieved 99.90% for KDD

Table 1 Accuracy under KDD dataset

Classifiers	KDD	NSL
KNN IB1	**99.67**	**99.80**
Neural MLP(Multilayer perception)	98.18	99.71
Fuzzy logic Random Forest	99.90	99.94
Fuzzy logic Random Tree	99.71	99.84
Meta Random Committee	99.90	99.91
Fuzzy logic PART	99.80	99.92

Beside table depicts NSL have TPR rate 97% for multilayer perceptron whereas 98% approximate which also shows NSL have greater intrusion detection capability as compared to

Table 2 Accuracy under NSL dataset

Classifiers	TPR	FPR	PRC	Recall	Correctly classified	Incorrectly classified
KNN IB1	0.998	0.000	0.998	0.998	42749	82

 ISSN: 2321-8169
1011 – 1016

Neural MLP(Multilayer perception)	0.997	0.000	0.996	0.997	42707	124
RF Random Forest	0.997	0.000	0.999	0.999	42808	23
RT Random Tree	0.998	0.000	0.998	0.998	42765	66
RC Random Committee	0.999	0.000	0.999	0.999	42794	37
PART	0.999	0.000	0.999	0.999	42799	32

KDD.For incorrectly classified instance as compared to KDD .

For incorrectly instance we can find in NSL it is only 124 whereas in the KDD it is only 124.whereas in KDD it is 777 which is too large

The table depicts accuracy between KDD and NSL which shows NSL Have less redundant data as compared to the KDD

Datasets. The NSL have capability of handle more class types of attacks . After KDD have less attacks categorization like normal and anomaly.

Table3 : Accuracy detection in KDD & NSL datasets(in %)

Classifiers	KDD	NSL
KNN IB1	**99.67**	**99.80**
Neural MLP(Multilayer perception)	**98.18**	**99.71**
Fuzzy logic Random Forest	**99.90**	**99.94**
Fuzzy logic Random Tree	**99.71**	**99.84**
Meta Random Committee	**99.90**	**99.91**
Fuzzy logic PART	**99.80**	**99.92**

Table4 Accuracy under NSL noisy dataset

Classifiers	NSL	NSL Noise10 %	NSLNoise20 %	NSLNoise30 %
KNN IB1	99.80	99.75	99.74	99.75
Neural MLP(Multilayer perception)	99.71	99.73	99.73	99.72
RF Random Forest	99.94	99.94	99.94	99.94
RT Random Tree	99.84	99/.76	99.75	99.57
RC Random Committee	99.91	99.89	99.91	99.91
PART	99.92	99.61	99.91	99.91

NSLVs NSL Noise 10%,20%,30%(individual classifiers are characterize basis of their response toNSLnoise dataset).Consider the case of Random tree classifiers accuracy is 99.97% as compared to NSL 99.84% it shows when noise increases the NSLNoise10% reduces accuracy and if noise again increase by 20% or 30% the accuracy significantly reduces.

Table5 Accuracy under NSL 10% dataset

	TPR	FPR	PRC	Recall	Correctly classified	Incorrectly classified
KNN IB1	**0.998**	**0.001**	**0.997**	**0.998**	**42724**	**107**
Neural MLP (Multilayer perception)	**0.997**	**0.000**	**0.996**	**0.997**	**42717**	**114**
RF Random Forest	**0.999**	**0.000**	**0.999**	**0.999**	**42809**	**22**
RT Random Tree	**0.998**	**0.000**	**0.998**	**0.998**	**42729**	**102**
RC Random Committee	**0.999**	**0.000**	**0.999**	**0.999**	**42787**	**44**
PART	**0.999**	**0.000**	**0.998**	**0.999**	**42746**	**35**

Table 6 Accuracy under NSL 20% dataset

Classifiers	TPR	FPR	PRC	Recall	Correctly classified	Incorrectly classified
KNN IB1	0.998	0.001	0.997	0.998	42723	108
Neural MLP(Multilayer perception)	0.997	0.000	0.996	0.997	42717	114
RF Random Forest	0.999	0.000	0.999	0.999	42809	22
RT Random Tree	0.998	0.000	0.998	0.998	42726	105
RC Random Committee	0.999	0.000	0.999	0.999	42796	35
PART	0.999	0.000	0.999	0.999	42793	38

In NSL 20% Noise apart from PART values are changes it is 35 to 38.whereas minor changes in KNN .random tree are for when noise increase by 20% it is 105.

Table7 Accuracy under NSL30% dataset

Classifiers	TPR	FPR	PRC	Recall	Correctly classified	Incorrectly classified
KNN IB1	0.998	0.001	0.997	0.998	42693	107
Neural MLP (Multilayer perception)	0.997	0.000	0.996	0.997	42054	118
RF Random Forest	0.999	0.000	0.999	0.999	42789	25
RT Random Tree	0.998	0.000	0.998	0.998	42711	180
RC Random Committee	0.999	0.000	0.999	0.999	42790	35
PART	0.999	0.000	0.999	0.999	42746	36

V. **Conclusion**

In Intrusion detection system there is possible way of finding intrusive attempt in NSL dataset. Sometimes in real implementation there is noise in the system in that IDS not

give proper accuracy**.** Individual Classifiers are characterize with their response to NSL Noisy datasets. Consider the case of Random tree classifiers accuracy is 99.76%as compared to NSL 99.84%. It shows when noise increases the NSLNoise10% reduces accuracy and if noise again increased by 20% or 30%the accuracy significantly reduces. The Random Forest is noise tolerant which shows it have no noise effect so it is the best classifiers.

References

[1] D. K. (2013). Selection of best classifiers from different dataset using weka. *I IJERT*, 73-77.
[2] 1999, K. C. (1999). *Computer network intrusion detection*.
[3] A detailed analysis of the KDD CUP99 dataset. (2009).
[4] *IEEE on computational intelligence fo rsecurity and defence application*.
[5] Axelsson, s. (2000). Intrusion detetction system
[6] Brugger, e., & S.terry, B. j. (2007). An assessment of DARPA
[7] IDS Evaluation dataset using Snort. *UCDAVIS department of computer science*. (1999). *Cup, KDD* .
[8] D.V, C. S. (2012). Efficient Algorithm for Intrusion Attack classification by analyzing KDD Cup99. *IEEE*.
[9] Gandhi, M., & kumaravel Appavoo, m. g. (2010).
[10] Effectivenetwork intusion detection using classifies decisiontrees and decision rules. *int .J.Advanced network and application*.
[11] Govindrajan. (2016). Evaluation of Ensemble Classifiers for Intrusion. *International Journal of Computer, Electrical, Automation, Control and Information Engineering* .
[12] Ibrahim, L. M., & mahbod, I. L. (2013). A comprison study for intrusion database(Kdd 99,Nsl kdd) based on self organizing map. *Artificial Neural network Journal of Engineering science and technology*, (pp. 107-119).
[13] Kajal Rai, A. G. (2016). Decision Tree Based Algorithm for intrusion detection. *Int. J. Advanced Networking and applications*, 2828-2834.
[14] M.patra, M. A. (2010). discriminative multinomial naive bayes for intrusion detection system. *Information Assurance and security*.
[15] McHugh, J. (2000). The 1998 Lincon laboratory IDS Evaluation(A critique). *recent advances in intrusion detection*, (pp. 141-161). toulouse,france.
[16] .Nikolaos Avouris, D. I., & Daskalaki, s. (2006). evaluation of classifier forban uneven class distribution problem. *applied artificial intelligence*, pp. 381-417.
[17] Sabri, & kamaruzzaman seman, s. N. (2011). identifiying false alarm rate for intusion detection with data mining. *IJCNSinternationaljournal of computer and network security*, 95.
[18] Stefan, A. (2000). *Intrusion detection system : A survey & taxonomy*.
[19] Tavallaee, & et.al. (2009). A detailed analysis of KDD CUP 99 dataset. *IEEE On computational intelligence for security and defence application*.
Vishwas, T. C. (2008). Usefulness of DARPA dataset for intrusion detection system evaluation. *SPIE defence and securit* accuracy value.