Comparison of Probabilistic and Decision Tree Approach of NIDS

M. Sadiq Ali Khan*

Department of Computer Science, University of Karachi, Karachi, Pakistan

Abstract: Network Intrusion Detection System is non-trivial requirement in appropriate monitoring of ever increasing use of computer network system. Different approaches including probabilistic, statistical and hardware-based solutions are used for early warning system for system intrusion. In this paper decision tree and Bayesian network methods have been compared. Although both methods are found almost equally effective however decision tree based system gives comparatively better accuracy in detection of less frequent attack types.

Key Words: NIDS, Decision Tree, Bayesian Network.

INTRODUCTION

Unauthorized access of computer network and databases is a critical concern with increasing use of network and online databases. Hackers are using novel methods for unauthenticated access to resources. A different method for monitoring of network intrusion system has been developed including probabilistic, statistical and hardware based solutions. Burney et al [1] has showed that Bayesian Network and Naïve Bayes classifier can be used efficiently to detect network intrusion. Zhang et al [2] Mulay et al [3] has showed using support vector machine and decision tree for intrusion detection.

NETWORK INTRUSION DETECTION SYSTEM

The log patterns collected at the network levels which are used by the intrusion detector [4]. These have limited intrusion detection ability because it is difficult to deduce the related information directly from the network log patterns.

Large log patterns at the network level may also affect the accuracy of the intrusion detection because without analyzing many log patterns are allowed to pass through the network and only the summary statistics at different interval of times are given. These statistical data may include total number of connections, amount of incoming and outgoing traffic. Network based intrusion detection system applies predefined attack signatures to identify hostile traffic of each frame. The management console is notified if it finds a match against any signature. NIDS are operating system independently and its sensors can detect attacks which host-based sensors fail to detect.

IDS monitors traffic on a real time and can detect malicious activity as they occur. The attacker cannot remove evidence of attack by using network IDS because it uses live network traffic and does real time intrusion detection.

In the domain of network security IDS play a vital role. Security is usually implemented as a multi layer infrastructure and different security approaches can be categorized into the following major areas [5]:

Avoidance of Attacks: In order to prevent the launching of an attack we should try to increase the amount of apparent danger of harmful consequences for the intruder. A strong ID system is required for the attack avoidance. However, it requires well-built verification against the attacker in case an attack is launched. Methods used in this area are discussed in which may effectively mark out the actual source of attack. Mechanism for attack avoidance by using the cryptography is discussed in [6].

Prevention of Attacks: Before reaching the targeted machine it aims to prevent an attack by blocking. Practically it is very hard to prevent all kinds of attacks due to incomplete knowledge for the attacks and allow normal activities.

Deflection of Attacks: It refers to trapping the intruder by the system and the attacker intentionally made to reveal the attack. For instance honey pots discussed in [7].

Detection of Attack: It refers to the process of detecting an attack when it is still in progress or to detect such kind of attacks which occurred in the past. It is more significant in order to system recovery and to take preventive measures for occurrence of similar kinds of attacks in future.

Attack Recovery: On the detection of attack, current system must perform the recovery procedures as per the security policy. Performing attack detection followed by reaction and recovery are known as the intrusion detection systems.

The main types are host-based (HIDS) and network-based (NIDS) systems [8]. IDS look like a defense tool which is the need of every organization. Efficient detection of attacks is the major concern of IDS. However we should develop such a system which basically detects the intrusion

*Address correspondence to this author at the Department of Computer Science, University of Karachi, msakhan@uok.edu.pk

0250-5363/11

pattern earlier in order to minimize its negative impact. [9]. The main challenges for IDS are discussed below:

- False alarm rate should be low as much as possible. As the system decreases the false alarm rate it efficiency increases. It is very necessary that ratio of false alarm should be low in a real network with huge amount of traffic. Detection of different types of attacks with reduced false-alarm is the main points for building ideal IDS.
- IDS notifies the administrator incase detection of malicious activity and to access control list in order to stop a malicious connection. But it is still very vital to check the IDS logs frequently to stay on top of the occurrence of events. Checking the logs on a daily basis is essential to analyze the kind of malicious activities detected by the IDS over a period of time.
- IDS should be capable of operating in a real environment by initiating a positive response quickly as attack is detected. IDS should have the capability to handle large amount of data set with better performance.
- Proper response should be generated by the IDS. A system would basically link a response generated by the detector module to the main security event is enviable. It is equally important for IDS to identify the attack type as it detects the abnormality in the system.
- The task is to build a scalable detection system which detects all possible types of attacks and have the capability to incorporate different set of methods/ mechanism in order to provide efficiency and detection accuracy.

For IDS technology there is a lot of enhancement, so the organizations need to define their goals. IDS have not so far reached the level where it can offers past analysis of the intrusions detected over a period of time. This is still a manual activity. The result of an IDS implementation depends to a great extent on how it has been deployed. A set of plans is required in the design as well as the completion phase.

SIGNATURE BASED MONITORING

The process of comparison of signatures against the observed events to identify possible abnormality is Signature based detection. Signature based detection works efficiently on known threats as in Fig. (**1**). This simplest method compares the current unit of activity to a list of signatures using string comparison operations.

When known attack contains in the set A, the system tries to detect signature based or misuse based system [10]. This applies pattern matching techniques to detect the known attacks. It prevents attacks that comprise multiple events that do not contain a clear indication of an attack.

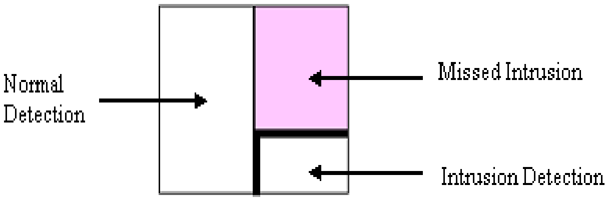

Fig. (1). Signature Based Systems.

False alarm rate is low but they have limited detection capability, as they cannot find the unseen intrusion. They totally depend upon the information stored in the set A. i.e. Information contains $A_{\text{R-intrusion}}$. To make it more reliable and effective, such systems need full information set of intrusions/attacks i.e. $A_{\text{R-intrusion}}$ should be equivalent to $S_{\text{R-intrusion}}$, which not always possible.

This type of NIDS has shortcoming, as it is not efficient in detecting new kind of attack [11]. In this paper signature based IDS is studied.

DECISION TREES

In order to get the information for the decision making process, decision trees are used [12]. Its start with a originating node on which it is for users to take actions. According to Decision tree learning algorithm, each node is split recursively form that originating node. In last a decision tree built in which each branch represents a probable situation of conclusion and its ending, it is based on inductive learning.

In order to define information gain entropy is defined first. Lets assume, that the resulting decision tree classifies instances into two categories, as P(positive) and N(negative). Given a set S, containing these positive and negative targets, the entropy of S related to this Boolean classification is given as:

$$\text{Entropy}(S)= - P(\text{positive})\log_2 P(\text{positive}) - P(\text{negative})\log_2 P(\text{negative})$$

Where entropy is a measure of the impurity in a collection of training sets. In S we need to select the optimal attribute for splitting the tree node, which we can easily imply that the attribute with the most entropy reduction is the best choice. We define information gain as the expected reduction of entropy related to specified attribute when splitting a decision tree node. The information gain, Gain(S,A) of an attribute A,

Gain(S,A)= Entropy(S) - Sum for v from 1 to n of (|Sv|/|S|) * Entropy(Sv)

This notion of gain can be used to rank attributes and to build decision trees where at each node is located the attribute with greatest gain among the attributes not yet considered in the path from the root and successively at each step next attribute of max information gain is selected until the last attribute is reached.

BAYESIAN NETWORK

A Bayes net is specific graphical model. It reflects the states of some state of a world that is being modeled and it describes how those states are related by probabilities. A Bayesian network consists of a directed acyclic graph (DAG) and a set of local distributions. The random variable are represented by a node in the graph while random variable denotes an attribute, feature, or hypothesis about which we wish to draw inference. Each random variable has a set of mutually exclusive and collectively exhaustive possible outcomes. The graph represents direct qualitative dependence relationships; the local distributions represent quantitative information about the strength of those dependencies. The graph and the local distributions together represent a joint distribution over the random variables denoted by the nodes of the graph. Let X_1, $X_2 \ldots X_i$ represent random variable each reprinting node in DAG then the joint probability is defined as the probability that a series of events will happen concurrently. The joint probability of several variables can be calculated from the product of individual probabilities of the nodes based on Bayes Theorem can be given as:

$$\mathrm{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} \mathrm{P}(X_i \mid \mathrm{parents}(X_i))$$

The Bayesian Network determined is depicted in Fig. (**2**), which reflects the interdependence of nodes. In Bayesian network also count and source byte is found to be important feature.

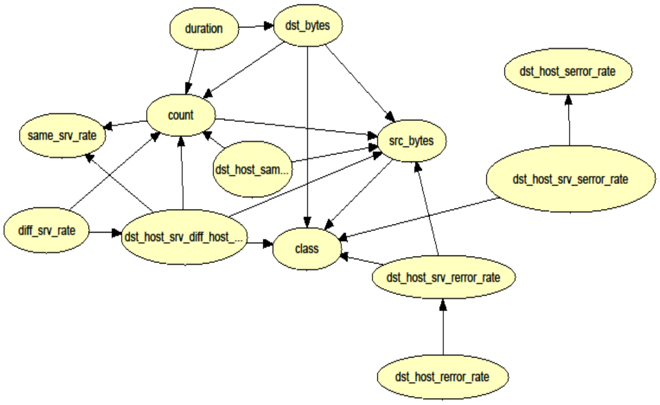

Fig. (2). Bayesian Network Model IDS.

EXPERIMENT

In this paper reduced feature data of KDD data set is used for analysis. 15 parameters mentioned in Table **1** are selected using PCA. 8000 instances are selected using stratified random sample proportional to attack type. The data set is split in to two parts for training (66%) and testing (33%). Open source Weka software used for running the experiment

Table 1. Selected Features from Dataset

S#	Feature
1	Duration
2	Land
3	src_bytes
4	dst_bytes
5	su_attempted
6	Count
7	diff_srv_rate
8	same_srv_rate
9	dst_host_same_src_port_rate
10	dst_host_srv_diff_host_rate
11	dst_host_serror_rate
12	st_host_srv_serror_rate
13	dst_host_rerror_rate
14	dst_host_srv_rerror_rate
15	Class

RESULT & DISCUSSION

The Decision Tree approach C45 is compared with the Bayesian method although both methods are equally effective in attack classification however C45 method is slightly better in determining certain type of attack. The precision of determining normal data packet Bayesian network (99.5%) is slightly better than C45 (98.8) as mentioned in Table **2**.

Table 2. Comparison

	Bayes Network	C45
Correctly Classified Instances	97.5	98.925
Incorrectly Classified Instances	2.5	1.075
Kappa statistic	0.9589	0.9822

However, some attack type (Imap, buffer overflow, Phf, Multihop) which are less frequent are difficult to identify as shown in Table **3**. However if larger data set is considered then accuracy may be improved. C45 method resulted in pruned tree of size 141 and leaves density of 71 src_bytes is found to be the feature with maximum information gain followed by count feature.

Based on two methods a model is proposed as depicted in Fig. (**3**). The process of NIDS is mainly depending upon appropriate selection of parameter which has been carried out through component analysis. Secondly the identification of parameters which gives maximum information of the data patched. The model accommodates both of these features.

Table 3. True Positive & Precision Table

	True Positive		Precision	
Attack Type	**Bayes**	**C45**	**Bayes**	**C45**
Neptune	0.993	0.998	0.993	0.999
Normal	0.976	0.995	0.988	0.988
Ipsweep	0.929	0.964	0.941	0.977
Nmap	0.844	0.969	0.818	0.939
Rootkit	0	0	0	0
Satan	0.922	0.959	0.957	0.981
Smurf	1	0.97	0.977	0.982
Teardrop	0.983	1	0.831	0.984
Portsweep	0.968	0.989	0.726	0.974
Warezclient	0.877	0.842	0.833	0.96
Back	0.984	0.968	0.984	0.952
Warezmaster	0	0.75	0	0.75
Imap	0	0	0	0
buffer_overflow	0	0	0	0
Pod	0.667	0.333	0.889	0.571
guess_passwd	0.667	0	0.889	0
Phf	0	0	0	0
Multihop	0	0	0	0

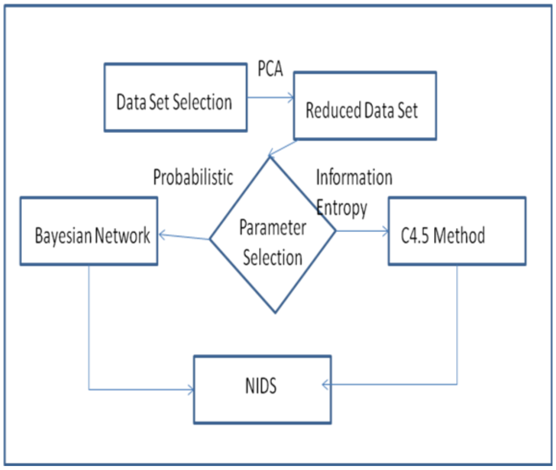

Fig. (3). Proposed Network Intrusion Detection Model.

REFERENCES

[1] Burney, S. M. Aqil, Sadiq Ali Khan, Jawed Naseem, (2010) “Efficient Probabilistic Classification Methods for NIDS” IJCSIS Vol 8 No 8 pp 168-172

[2] Zhang, Yongli, Yanwei Zhu “Application of Improved Support Vector Machines in Intrusion Detection” 978-1-4244-5895- (2010) IEEE

[3] Mulay, Snehal A. P.R. Devale G. V. Garje, 2010, Intrusion Detection System using Support Vector Machine and Decision Tree International Journal of Computer Applications (0975 - 8887). Volume 3 - No.3, June (2010)

[4] Carol Taylor and Jim Alves-Foss; “An Empirical Analysis of NATE: Network Analysis of Anomalous Traffic Events”, Proceedings of the 2002 Workshop on New Security Paradigms, pages 18–26. ACM, (2002).

[5] Christopher Kruegel, Fredrik Valeur, and Giovanni Vigna. Intrusion Detection and Correlation: Challenges and Solutions. Springer, (2005).

[6] Rodica Tirtea, Geert Deconinck; Fault Detection Mechanisms for Fault Analysis Attacks Resistant Cryptographic Architecture; Third International Conference on Systems, Signals & Devices; March 21-24, Sousse, Tunisia, (2005).

[7] Lokesh D. Pathak and Ben Soh; “Incorporating Data Mining Tools into a New Hybrid-IDS to Detect Known and Unknown attacks; Ubiquitous intelligence and computing; Lecture Notes in Computer Science, Volume 4159/2006, 826-834, DOI: 10.1007/11833529, (2006).

[8] Elvis Tombini, Herve Debar, Ludovic Me, and Mireille Ducasse. A Serial Combination of Anomaly and Misuse IDSes Applied to HTTP Traffic. In Proceedings of the 20th Annual Computer Security Applications Conference, pages 428-437. IEEE, (2004).

[9] Saira Beg et.al, 2010; “ Feasibility of Intrusion Detection system with High Performance Computing: A survey”, International Journal of Advances in Computer Science, ISSN-2218-6638, Vol 1 Issue 1 Dec (2010).

[10] Rebecca Bace and Peter Mell. Intrusion Detection Systems. Gaithersburg, MD : Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology, (2001).

[11] Karthikeyan K.R.; “Intrusion Detection Tools and Techniques: A Survey”, International journal of Computer Theory and Engineering Vol 2, No 6; 1793-8201, (2010).

[12] Chih-Fong Tsai; Jung-Hsiang Tsai; “Performance Evaluation of the Judicial System in Taiwan Using Data Envelopment Analysis and Decision Trees”, Second International Conference on Computer Engineering and Applications (ICCEA), (2010) Vol 2 , Page(s): 290 – 294.