An Efficient Algorithm for Network Intrusion Detection System

V. Jaiganesh
Doctoral Research Scholar,
Manonmaniam Sundaranar University, Tirunelveli, India

P. Rutravigneshwaran
M.Phil. Scholar,
Department of Computer Science, Dr. N.G.P. Arts and Science College, Coimbatore, India

P. Sumathi, Ph.D
Doctoral Research Supervisor,
PG & Research Department of Computer Science,
Assistant Professor,
Government Arts College, Coimbatore, India

ABSTRACT

Nowadays it is a vital role to provide a high level security to protect highly sensitive and private data. A lot of work is required to be done in the field of application of AI techniques to the field of network security. Firewall alone is just not enough to protect a corporate network from all type of internal and external threats. It is beyond the limits of conventional technologies to cope up with new and new attack patterns and exploitations of system vulnerabilities. Intrusion Detection System is a software or hardware device which is the essential technology in Network Security. Intelligent IDSs have to be developed. Today researchers have interested on intrusion detection system using Data mining techniques as an artful skill. In this paper, implement a model based on ID3 algorithm and proposed Exclusive ID3 algorithm which detects unknown attacks with the help of optimized decision tree from available set of data also follow predefined rules for accurate decision making for system Administrator.

Keywords

IDS, Decision Tree, Network Intrusion Detection, ID3 algorithm, Time-To-Live Value, Intrusion Attacks

1. INTRODUCTION

Each computer have risk for unauthorized and intrusion, however, with sensitive and private data are at a higher risk. Even the most secure systems are vulnerable to insider attacks. New intrusions continually emerge and new techniques are needed to defend against them. Since there are always new intrusions that cannot be stopped, IDS have introduced to notice possible abuses of a security policy by monitoring response and entire system activities. Intrusion Detection is a key technique in Information Security plays an important role detecting different types of attacks and secures the network system. IDSs are rightly called the second line of defense, since IDS comes into the picture after an intrusion has been occurred. Once we detect the attack it comes into the network, a response should be initiated to stop or minimize the harm to the system. It also assists prevention techniques develop by providing information about intrusion techniques.

IDSs are divided into two broad categories: host-based (HIDS) and network-based (NIDS). A host-based ID requires small programs to be installed on individual systems to be supervised. A network-based Intrusion Detection System is placed on a network segment or boundary and monitors all traffic on that segment. The current trend in intrusion detection is to combine both host based and network based information to develop hybrid systems that have more efficient.

There are currently a variety of approaches being utilized to accomplish the desirable elements of an intrusion detection system. Specially, there are two general approaches to intrusion detection:

- Anomaly detection
- Misuse detection

The methods used by intruders can often contain any one, or even combinations of distributed denial of service, Spoofing, Network port scans, Viruses and Worms, Trojan horse and Buffer overflow. There is a need to enhance the security of computers and networks for protecting the critical infrastructure from threats. Accompanied by the rise of electronic crime, the design of safe-guarding information infrastructure such as the intrusion detection system (IDS) for preventing and detecting incidents becomes increasingly challenging. The development of new IDS is motivated by the following factors because, Most existing systems have security was that render them susceptible to intrusions, and finding and fixing all these deficiencies are not feasible. The Prevention techniques cannot be sufficient. This technique is almost impossible to have an absolutely secure system. Most of the modern Intrusion Detection Systems are poised of four parts. A packet capturing mechanism, a classifier, a database of known attack patterns and an optional user interface. The packet capturing mechanism captures network traffic from an identified network segment and passes it on to the classifier. Figure 1 shows the implemented system having one module of classifier. The classifier classifies the incoming traffic as "innocent" or "Suspicious" based on the results of comparison with attack Patterns already there in the database. ID3 is the decision tree algorithm developed by Ross Quinlan in 1979. ID3 builds a decision tree from a fixed set of examples and the resulting tree is used to classify future samples. The example has many attributes and belongs to a class (like yes or no). In the decision tree, the leaf nodes contain the class name whereas a non-leaf node is a decision node defined.

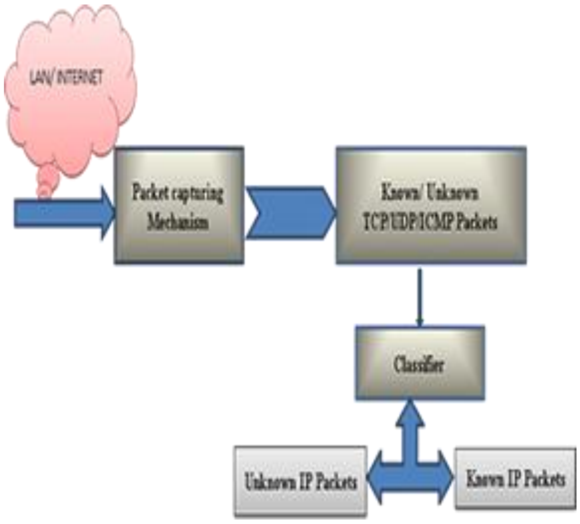
LAN/ INTERNET
Packet capturing Mechanism
Known/ Unknown TCP/UDP/ICMP Packets
Classifier
Unknown IP Packets
Known IP Packets

Figure 1.1: Contemporary IDS with classifier

2. RELATED WORKS

At first a model to embed primitive intelligence in the network intrusion detection system have implemented by Anant R. More, Vikas N. Nandgaonkar, Dr. ManojNagmode, Pramod P. Patil (2013) [1] using C#. In [2] the authors has studied that Intrusion detection has become a critical component of network administration due to the vast number of attacks persistently threaten our computers in 2011. In [3] the author has discussed the Efficiency is one of the major issues in intrusion detection. Reference [4] deployed sequential pattern mining to identify attack patterns that hackers frequently submit, and classified the *modus operandi* that suspects used in the commission of crimes into predefined crime types.

Reference [5] used a pattern extraction technique to identify particular crime data, such as segmenting and extracting a suspect from a picture on a security video. In [6] the paper has discussed about the security purpose in information system. In this proposed research there are five types of classifiers used. They are Feed Forward Neural Network (FFNN), Elman Neural Network (ENN), Generalized Regression Neural Network (GRNN), Probabilistic Neural Network (PNN) and Radial Basis Neural Network (RBNN). Finally it is clear that Probabilistic Neural Network has better accuracy than rest of other neural networks. In [7] paper has studied that one of the central areas in network intrusion detection is how to build effective systems that are able to distinguish normal from intrusive traffic. Reference [8] presented that with the tremendous growth in information technology field, network security is the important challenging issue and so as Intrusion Detection system (IDS). In [9] the paper has said that the Intrusion Detection system is an active and driving secure technology to compromise the integrity, confidentiality, availability, or to bypass the security mechanisms of a network.

3. METHODOLOGY

3.1 ID3 Algorithm

ID3 algorithm is a typical decision tree algorithm. Algorithm ID3 uses information gain (or entropy) to help it decide which attribute belongs into a decision node.

The following steps involving in ID3 algorithm:

1. Evaluate the entropy (or information gain) of every attribute using the data set
2. Partition the set into subsets using the attribute for which entropy is minimum (or, equivalently, information gain is maximum)
3. Build a decision tree node containing that attribute
4. Recourse on partitioned subsets using the remaining attributes

Traditional ID3 algorithm chooses attributes, and often tends to choose the attributes that get more values, because the weighted sum method makes the classification of examples set tend to the metadata group that discarding small data group, but the attribute has more properties is not always optimal one. The attributes in the learning model building process include the knowledge level of originally subject in learning ability database, the multiple factors of learning mode in learning mode database, and the learning motivation classification in learning motivation database. The final decision tree classification results are not certainly consistent with the actual situation according to the traditional ID3 classification because there are many types of attributes based on Entropy.

Suppose that a training example set is X, the purpose is to divide the training examples into n classes, recorded as C=(X1, X2, …Xn). On the assumption that the number of ith training examples is | Xi | = Ci, the probability that an example belongs to this training examples is P (Xi).

If we choose the attribute A to test, with a set of properties a1, a2, a3,...ai, the number of examples that belonged to the ith category when A = a j is Cij

$$P(Xi : A = aj) = Cij / | X | \qquad (1)$$

The value of P(Xi:A=aj) is the probability that the test attribute A belongs to the ith category. Yj is the examples set when A= aj, then the degree of uncertainty to the decision tree classification is the entropy of the training examples set to attributes A:

$$H(Yj)= -\_P(Xi|A=aj)log2P(Xi|A=aj) \quad (2)$$

We increase the user interest _when calculating the taxonomic information entropy of each leaf node= aj extended from attribute A, and then strengthen the label of important attribute, and reduce the label of non-important attribute. The formula as follows:

$$H(X|A)=\_[P(A=aj)+\_]H(Xj) \qquad (3)$$

The information provided by attributes A for classification (the information gain of attribute A) is:

$$I(X: A) =H(X) -H (X|A) \qquad (4)$$

By doing all the steps which was discussed a new decision tree will be created with the possibility to identify known and unknown incoming packets.

3.2 Proposed Exclusive ID3 Algorithm

It is assumed that IP packets with strange TTL (Time-To-Live) values are malicious and detecting malicious packets using TTL values. A computer sets the TTL at the initial value when it sends an IP packet. When a packet with abnormal TTL is observed, then there are two possible reasons applicable why TTL is abnormal.

First, the packet actually came through more than 30 routers. However, a packet rarely hops more than 30 routers. *Second*, a sender modifies the initial TTL value. So, this is considered that packets with an abnormal TTL were sent with a malicious intention. With the basis of this assumption, a novel method for distinguishing malicious packets from legitimate ones is proposed.

It is possible to estimate the number of routers (hop count) along a path from the sender host to the destination host by initial value. The estimation is based on the following facts.

1. The TTL initial values of popular OSs are well separated it is easy to judge the initial value. It is not necessary to identify the OS. Here only need the initial TTL value.

2. The maximum count of routers along a path (hop count) is around 30.

If a host receives an IP packet with TTL equal to value t, the initial value of the TTL, t0, can be assumed to be the minimum value that is larger than t in Table 1.

The hop count can be calculated for the received packet as follows: (hop count) = t0 − t. For example, when a host receives a packet with a TTL value of 120 (t = 120), the minimum number in Table 1 that is larger than t is 128 (t0 = 128). Therefore, the hop count is 8 (128 − 120 = 8).

Table 1: Initial TTL values of standard operating systems

OS	Protocol	Initial TTL
Linux 2.4 kernel	ICMP	255
BSDI BSD/OS 3.1 and 4.0	ICMP	255
Windows Server 2008	TCP, UDP, ICMP	128
Windows7	TCP, UDP, ICMP	128
Windows XP	TCP, UDP, ICMP	128
Linux RedHat 9	TCP, ICMP	64
FreeBSD5	ICMP	64
MacOS X (10.5.6)	TCP, UDP, ICMP	64
AIX	TCP	60

First, the tree must be created and trained. The tree must be offered with sufficient labeled samples that are used to create the tree itself. It is done this by dividing the samples into subsets based on features. Here the sets of samples at the leaves of the tree define a classification. The tree is created in such a way that the simplest and smallest tree, then that is consistent with the training samples, which is the best predictor.

Our methodology is to capture the packet data at the gateway and the proposed Exclusive ID3 algorithm involves the following steps after the successful creation of training data set.

1. Formulate a table from the incoming data packet
2. Identify class attribute and classes
3. Use ID3 algorithm for creating decision tree
4. Identify useful attributes for classification (relevance analysis)
5. Learn Exclusive ID3 algorithm using training examples in training set
6. Use Exclusive ID3 algorithm to classify the unknown data samples

4. EXPERIMENTAL RESULTS

The proposed IDS is experimented using the Waikato Environment for Knowledge Analysis (WEKA) and the dataset used is KDD Cup99 dataset. The tool WEKA is a complete set of Java class libraries that execute several state-of-the-art machine learning and data mining approaches.

The Experimental data comes from KDD CUP 1999 dataset [13], which from DARPA 98 Intrusion Detection Evaluation handled by Lincoln laboratory at MIT. It is test set widely used in Intrusion detection field. It includes about 4.9 million simulative attack records and 22 types of attack. Because the entire data set is too large. Here, we have used subset of 10% of KDD Cup 99 dataset for training and testing. At first, the training dataset is classified into five subsets so that, four types of attacks DoS (Denial of Service), R2L (Remote to Local), U2R (User to Root), Probe) and normal data are separated.

Particularly, the four broad classes of attack type defined in IDS as: *DoS, Probe, R2L and U2R*. KDD Cup99 is an audited set of standard dataset which includes training and testing set. Data has the following four major groups of attacks:

- ➢ Denial-of-Service (DoS) like neptune, smurf, land etc.
- ➢ Remote-to-Local (R2L) like sendmail, guess_passwd, spy etc.
- ➢ User to Root (U2R) like ps, perl, xtern and so on.
- ➢ Probing like nmap, saint etc.

- *Denial-of-Service (DoS)*: These are attacks designed to make some service accessible through the network unavailable to legitimate users.
- *Probe*: A Probe is reconnaissance attack designed to uncover information about the network that can be exploited by another attack.
- *Remote-to-Local (R2L)*: This is where an attacker with no privileges to access a private network attempts to gain access to that network from outside, for e.g. over the internet.

- *User-to-Root (U2R)*: The attacker has a legitimate user account on the target network. Though, the attack is designed to escalate his privileges so that one can perform unauthorized actions on the network.

Initially pre-processing is taken out; at this phase the data is segmented into training and testing. In this process data that is collected from the IDS or IPS sensors needs to be put into some canonical format or a structured database format based on the preprocessing. Then applied proposed system to the dataset.

Detection of attack be able to calculated by using the following metrics:

- False positive (FP): Or false alarm, which Equals to the number of detected attacks but it is in fact normal.
- False negative (FN): Equals to the number of detected normal instances but it is actually attack, in other words these attacks are the target of intrusion detection systems.
- True positive (TP): Equals to the number of detected attacks and it is in fact attack.
- True negative (TN): Equals to the number of detected normal instances and it is actually normal.

A. *Performance Measures*

The performance measures are used to evaluate the proposed EID3 agianst ID3 is

- Detection rate and
- False-alarm rate

The accuracy of an IDS is computed based on the detection rate and false alarm rate.

B. *Detection Rate Comparison*

Detection rate indicates the percentage of detected attack among all attack data, that is given as,

$$Detection\ Rate = \frac{TP}{TP + TN} \times 100 \qquad (1)$$

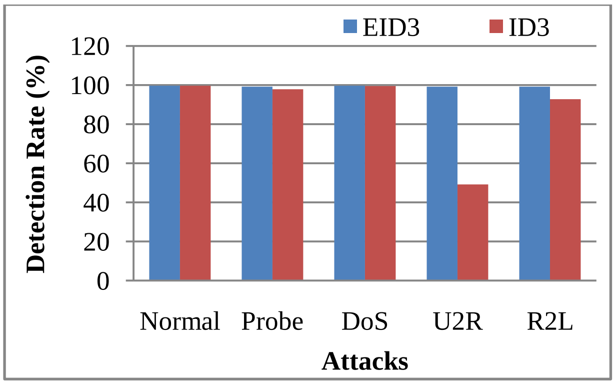

Figure 1.2: Comparison of Detection Rate on Four Attacks

The results of detection rate for different type of attacks are shown in figure 1.2. From the results it is observed that in case of U2R attacks, detection rate for ID3 and proposed EID3 is 49.21 and 99.2, respectively. Similarly the detection rate for EID3 is better than ID3 in all other attacks.

C. *False Alarm Rate Comparison*

False alarm rate indicates the percentage of normal data which is wrongly recognized as attack that is defined as follows:

$$False\ Alarm\ Rate = \frac{FP}{FP + TN} \times 100 \qquad (2)$$

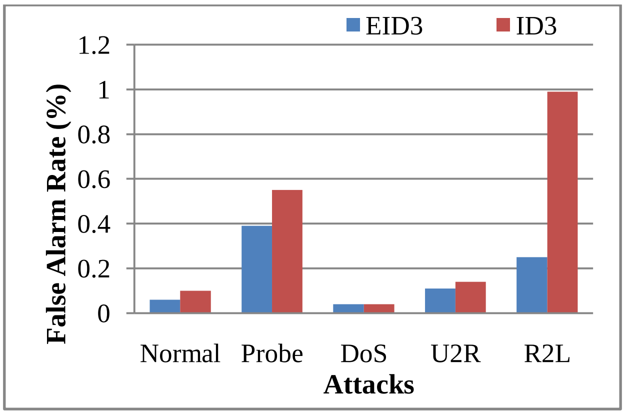

Figure 1.3: Comparison of False Alarm Rate on Four Attacks

The results of False alarm rate for different type of attacks are shown in figure 1.3. From the results it is observed that in case of R2L attacks, detection rate for ID3 and proposed EID3 is 0.99 and 0.25, respectively. Similarly the false alarm rate for EID3 is better than ID3 in all other attacks.

5. CONCLUSION

Internet and local networks have become everywhere. So organizations are increasingly employing various systems that monitor IT security breaches because intrusion events are growing day by day. An IDS has the capability to learn and take inferences. This paper proposed Exclusive ID3, in which the TTL (Time-To-Live) values are tuned. TTL value is utilized to obtain a converged solution to use ID3 algorithm of decision trees and identification trees in the classifier of IDS. The experiment is carried out in WEKA by using KDD Cup 1999 dataset and the results indicate that the proposed system can provide better detection rate and low false alarm rate than the ID3.

6. REFERENCES

[1] Anant R. More, Vikas N. Nandgaonkar, Dr. ManojNagmode, Pramod P. Patil “ID3 Algorithm for Intrusion Detection” International Conference on Recent Trends in engineering & Technology - 2013(ICRTET'2013).

[2] Ahmed Youssef and Ahmed Emam “Network Intrusion Detection using Data Mining and Network. Behavior Analysis” International Journal of Computer Science & Information Technology (IJCSIT) Vol 3, No 6, Dec 2011.

[3] AnazidaZainal, MohdAizainiMaarof and SitiMariyamShamsuddin “Data Reduction and Ensemble Classifiers in Intrusion Detection” in 2008 IEEE.

[4] Chau M., Xu J.J. and Chen H. (2002) National Conference on Digital Government Research, 271-275.

[5] Devaraju .S, Ramakrishnan .S “Detection of Accuracy for Intrusion Detection System using Neural Network

Classifier" International Journal of Emerging Technology and Advanced Engineering(ISSN 2250-2459 (Online), An ISO 9001:2008 Certified Journal, Volume 3, Special Issue 1, January 2013).

[6] Devendrakailashiya, Dr. R.C. Jain "Improve Intrusion Detection Using Decision Tree with Sampling" in IJCTA | MAY-JUNE 2012.

[7] GuangqunZhai, Chunyan Liu "Research and Improvement on ID3 Algorithm in Intrusion Detection System" in 2010 IEEE.

[8] Jorge Blasco, Agustin Orfila, Arturo Ribagorda "Improving Network Intrusion Detection by Means of Domain-Aware Genetic Programming" DOI 10.1109/ARES.2010.53 in IEEE 2010.

[9] Joshi .S.A, VarshaS.Pimprale "Network Intrusion Detection System (NIDS) based on Data Mining"International Journal of Engineering Science and Innovative Technology (IJESIT) Volume 2, Issue 1, January 2013.

[10] Mohd. JunedulHaque, Khalid.W. Magld, NisarHundewale "An Intelligent Approach for Intrusion Detection Based on Data Mining Techniques" in 2012 IEEE.

[11] YacineBouzida, Frederic Cuppens "Neural networks vs. decision trees for intrusion detection" in 2011. SIGMOD Rec-ord, 30 (4), 25-34.

[12] YacineBouzida, Frederic Cuppens "Neural networks vs. decision trees for intrusion detection" in 2011. SIGMOD Rec-ord, 30 (4), 25-34.