Tracking of Intruder in Local Area Network Using Decision Tree Learning Algorithms

K. Raja[1], and M. Lilly Florence[2]

[1]Research Scholars
Bharathiar University, Coimbatore, Tamil Nadu, India
raja.ktkm@gmail.com

[2]Professor,
Adhiyamaan College of Engineering, Hosur, Tamil Nadu, India
lilly_swamy@yahoo.co.in

ABSTRACT--- ***This paper is mainly deals with Intrusion Detection System which is used to monitor the activities in network or an application and also used to track and filter the unauthorized data using Decision Tree and Entropy techniques. Entropy and information Gain are the two Mathematical Concept and Computational techniques, to find the Solution for particular dataset using ID3(Iterative Dichotomiser 3) Techniques. Entropy is mainly focused on Iterative steps base on three rules less than, greater than and equal to whereas Information Gain is used the leaf node from the decision tree. This paper is to track and filter the unauthorized data using Intrusion Detection System (IDS) in appropriate LAN network in an organization.***

Keywords--- Intrusion Detection System, Iterative Dichotomiser 3, Information Gain, Entropy, Decision Tree Algorithm

---

1. INTRODUCTION

The Intrusion Detection System is a system that inspects all network activities and identifies suspicious pattern in LAN network in an organization**.** This paper deals with intrusion inside Local Area Network with appropriate port and protocols. Here we are using three types of protocol as sample data, with appropriate port number. If Port number or protocol mismatch, then intrusion will occur. Intrusion is also called as an act of intruding or the state of act being intruded. According to law, intrusion means an illegal act of entering, seizing, or taking possession of another's property. An application which is used to monitor the content in the Network which may be used in LAN or WAN or MAN and protecting from the intruder[7]. Since today we are living in the cyber world in daily life, there is possibility of hacking the data from an application in various types of network. There are two types of Intrusion Detection System, Which are Network Based Intrusion Detection System(NIDS) and Host Based Intrusion Detection System(HIDS) which is used to analyze the packet of data entering or leaving from the Network. These NIDS are not detected based on real time but if it is configured correctly are true real times. These HIDS and it is act as agents which is used to detect the unauthorized data to and fro in the network. Decision Tree is a nonlinear data structure which is used to represent each attribute of data and elements in a diagrammatic representation. This diagram representation of decision tree consists of three parts root, parent node and leaf node[2]. Each Root node and Parents node denotes different types of attributes of particular elements. Each Leaf Node denotes the output values corresponding to attributes given from root to leaf. A tree can be learned into splitting the attributes of set into subset based on the values of the attributes. The main objective of this tree is used to classify the data set and based on the classification, possible to estimate the Information Gain. This Decision Tree is combination of Mathematics and computational technology, according to Data Mining Techniques[1]. The algorithm which is used to construct a decision tree from the given data set is called Iterative Dichotomiser 3(ID3) invented by Ross Quintan in the year of 1997. This ID3 Algorithm is classified into two parts. In the First part we will find out Entropy and Second part we will retrieve Information Gain from the Entropy[4].

2. ENTROPY

Let us Consider S be the Dataset $S=\{s_1,s_2,s_3,\ldots s_n\}$, where $s_1,s_2,s_3,..s_n$ be the different types of attributes of the Set S. On Each Iteration, it iterates through every subset $\mathbf{s_i}$.

Estimate Entropy H(S) and Information Gain IG(S), H(S) which is used to select the attribute si with smallest entropy and IG(S) which is used to estimate and select the largest information value, Continuous Recursive or iterative on each subset will estimate Information Gain, where H represents Entropy and IG represents Information Gain. Termination of iterative process may be in dead state when its satisfies any one of these cases[5].

Case 1: Each node will be labeled as leaf when every element in the subset belongs to class of same type.
Case 2: Each node will be labeled as leaf when every element in subset belongs to most common class of same type.
Case 3: Each node will be labeled as most common class of the super Set or Parent Set when no attributes found to be matched on the specific set of value related to appropriate attributes.

This Decision Tree (ID3) constructed by Terminal Node and Non Terminal Node. Each Node is labeled as final set then it is called Terminal Node. Each Node on which data can be Split based on the selected attributes is called Non Terminal Node[6].

3. DECISION TREE LEARNING ALGORITHMS

The main objective of this algorithm is to detect the intrusion in the cyber world. Using Tree Learning algorithm, we can able to detect the intrusion with appropriate category.

The Process of Decision Tree Learning Algorithms as follows.

Let S ,R be the Set which consist of the subset of $s_1,s_2,s_3,\ldots s_i$ where $s_1,s_2,s_3,\ldots s_i$ be the set of the attributes for a particular dataset.

Let R be the final Given Set which consist of two values whether it satisfies or not.

The Following Step to Estimate Entropy of H(S) and Information Gain IG(S).

Step 1: Calculate the probability from the given Set R which Classified into two category Pyes (Probability with yes attributes) and Pno(Probability with no attributes) .
Entropy (S)= -[(Pyes)Log2(Pyes)+(Pno)Log2(Pno)] ----------(A)

Step 2: Calculate the Iterative process of all given set of attributes from $s_1,s_2,s_3,\ldots s_i$ such that E (Si) for all Subset Si from 1 to n attributes from the given set with Pyes and Pno ----------(B)

Step 3 : Calculate the Information Gain IG(S) for the Given Set such that difference between equation (A)-(B).

Step 4: Create Matrix with MXN such that records are in the form of the subset elements and Columns are classified into two categories. One is entropy value E(S) and Second One is Information Gain IG(S).

Step 5: Decision can be taken with appropriate record based on Maximum value of the Information Gain.

Step 6: Stop the Process.

4. ESTIMATION OF SAMPLE DATA

The Mathematical Concept and Computational Methodologies are used to estimate Information Gain IG(S) from the decision tree using sample data set. Consider the Sample Data which Consist of Class A, Class B IP addresses, protocol, and port number which used to transfer file within an organization.

Sl. No	IP Address	Protocol	Port Number`	Intruder
1	0.x.x.x (Class A)	SMTP	25	No
2	128.x.x.x (Class B)	POP3	1886	Yes
3	56.x.x.x (Class A)	UDP	1078	Yes
4	190.x.x.x (Class B)	UDP	110	No
5	45.x.x.x (Class A)	SMTP	25	No
6	146.x.x.x (Class B)	SMTP	2013	Yes
7	75.x.x.x (Class A)	POP3	1886	Yes
8	147.x.x.x (Class B)	UDP	1093	Yes
9	44.x.x.x (Class A)	POP3	68	No
10	76.x.x.x (Class A)	SMTP	25	No
11	172.x.x.x (Class B)	POP3	1886	Yes
12	98.x.x.x (Class A)	UDP	110	No
13	165.x.x.x (Class B)	SMTP	1856	Yes
14	166.x.x.x (Class B)	POP3	68	No

Table: Sample data for an organization

STEP 1: Calculate Combination of intruder based on Yes and No

Pyes(Probability of Yes)=7/13

Pno(Probability of No)=7/13

E(Start) =-[(Pyes)log2(Pyes)-(Pno)log2(Pno)] ----------(C)

=- [(7/13)log2(7/13)-(7/13)log2(7/13)]

= -0.17456 ---------- (1)

STEP 2: Consider the Specialty with different value port number 25,68,110 and other port number with combination of Pyes and Pno respectively

E(P25)

Pyes(Probability of Yes)=3/5

Pno(Probability of No)=2/5

By Applying Equation (c)

= - [(3/5)log2(3/5)+(2/5)log2(2/5)]

= -0.15653 ---------- (2)

E(P68)

Pyes(Probability of Yes)=2/5

Pno(Probability of No)=3/5

By Applying Equation (c)

= - [(2/5)log2(2/5)+(3/5)log2(3/5)]

= -0.15653 ----------(3)

E(P110)

Pyes(Probability of Yes)=2/4

Pno(Probability of No)=2/4

By Applying Equation (c)

= - [(2/4)log2(2/4)+(2/4)log2(2/4)]

= -0.1505 ----------(4)

E(Pother)

Pyes(Probability of Yes)=7/14

Pno(Probability of No)=7/14

By Applying Equation (c)

= - [(7/14)log2(7/14)+(7/14)log2(7/14)]

= -0.1505 ----------(5)

To Calculate the Specialty By applying Equation 1,2,3,4,5 we get Entropy of Specialty

E(New) = (3/14)*E(P25) + (2/14)*E(P68) + (2/14)*E(P110) + (7/14)*E(Pother)

= (3/14)*(-0.15653)+(2/14)*(-0.15653)+(2/14)*(-0.1505)+(7/14)*(-0.1505)

= -0.15265 ----------(6)

E(Port) = E(Start)-E(New)

= -(0.17456+0.15265)

= -0.32825 ----------(7)

STEP 2: Consider the protocol such as SMTP, POP3 and UDP Protocol with combination of Probability of Yes and Probability of No respectively

E(SMTP)

Pyes(Probability of Yes)=3/5

Pno(Probability of No)=2/5

By Applying Equation (c)

= - [(3/5)log2(3/5)+(2/5)log2(2/5)]

= -0.15653 ---------- (8)

E(POP3)

Pyes(Probability of Yes)=2/5

Pno(Probability of No)=3/5

By Applying Equation (c)

= - [(2/5)log2(2/5)+(3/5)log2(3/5)]

= -0.15653 ----------(9)

E(UDP)

Pyes(Probability of Yes)=2/4
Pno(Probability of No)=2/4

By Applying Equation (c)

= - [(2/4)log2(2/4)+(2/4)log2(2/4)]
= -0.1505 ----------(10)

E(Fault)

Pyes(Probability of Yes)=7/14
Pno(Probability of No)=7/14

By Applying Equation (c)

= - [(7/14)log2(7/14)+(7/14)log2(7/14)]
= -0.1505 ----------(11)

To Calculate the Specialty By applying Equation 8,9,10,11 we get Entropy of Specialty

E(New) = (3/14)*E(SMTP) + (2/14)*E(POP3) + (2/14)*E(UDP) + (7/14)*E(Fault)
= (3/14)*(-0.15653)+(2/14)*(-0.15653)+(2/14)*(-0.1505)+(7/14)*(-0.1505)
= -0.15265 ----------(12)

E(Protocol) = E(Start)-E(New)
= -(0.17456+0.15265)
= -0.32825 ----------(13)

STEP 3: Consider the Class A and Class B IP Address with combination of Pyes and Pno respectively

Specialty based on transfer

E(Class A)

Pyes(Probability of Yes)=2/7
Pno(Probability of No)=5/7

By Applying Equation (c)

= - [(2/7)log2(2/7) + (5/7)log2(5/7)]
= -0.17816 ----------(14)

E(Class B)

Pyes(Probability of Yes)=5/7
Pno(Probability of No)=2/7

By Applying Equation (c)

= - [(5/7) log2 (5/7) + (2/7) log2 (2/7)]
= -0.17816 ----------(15)

To Calculate Entropy E(class) by equation 14,15

E(transfer) = (7/14)E(Class A) + (7/14)E(Class B)
= (7/14)*(-0.17816) + (7/14)*(-0.17816)
= -0.17816 ----------(16)

E(Class) = E(Start)-E(New)
= -(0.17456+0.17816)
= -0.35272 ----------(17)

5. EXPECTED OUTCOME

To Calculate Information Gain, Subtract the equation 07, 13 and 17 from equation 1 we will get the following table with the appropriate values

	Entropy	Information gain
IP address	-0.17816	-0.35272
Protocol	-0.15265	**-0.32825**
Port Number	-0.15265	**-0.32825**

From the Table we conclude that Information Gain should be Maximum to be taken into account for our decision tree. So According to our sample table, decision should be taken only port number or protocol rather than IP address.

6. CONCLUSION

In this paper we concluded that, to detect the intruder in Local Area Network of a particular organization we can use techniques like Entropy, ID3 with Decision tree learning algorithm. In this paper we worked with real time data which we have collected from a LAN in particular Organization. We would like to conclude that the intruder can be detected with minimum of three parameters like IP address, Protocol, and Port number. In future this work can be extend to any type of network. This paper which leads to track and filter the intruder in Local Area Network of an Organizations or an industry.

7. REFERENCES

1. Dr. S.Vijayarani1 and Ms. Maria Sylviaa.S," Intrusion Detection System-A Study", International Journal of Security, Privacy and Trust Management (IJSPTM) Vol 4, No 1, February 2015.
2. Singh Vijendra, Hemjyotsana Parashar and Nisha Vasudeva," A New Method for Classification of Datasets for Data Mining", 3rd International Conference on Machine Learning and Computing V5, pp 551-554, 2011.
3. Shilpi Gupta, Roopal Mamtora, "Intrusion Detection System Using Wireshark", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 11, November 2012.
4. Ankita Agarwal, Sherish Johri, Ankur Agarwal, Vikas Tyagi, Atul Kumar, "Multi Agent Based Approach For Network Intrusion Detection Using Data Mining Concept", Journal of Global Research in Computer Science, Volume 3, No. 3, March 2012.
5. 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, pg. 145-153, October- 2013
6. Archana D Wankhade, Dr P.N.Chatur," Comparison of Firewall and Intrusion Detection System", International Journal of Computer Science and Information Technologies, Vol. 5 (1) 674-678, 2014.
7. Dr S. Vijayarani and Ms Maria Sylviaa S, "Intrusion Detection System –A Study", International Journal of Security, Privacy and Trust Management, Vol. No. $, Issue. No. 1, February 2015.