International Journal of Informatics and Communication Technology (IJ-ICT)
Vol.3, No.1, February 2014, pp. 23~29
ISSN: 2252-8776

The Establishment of Decision Tree Model in Network Traffic Incident Using C4.5 Method

Made Sudarma *, Dandy Pramana Hostiadi **
*Computer System and Informatics, Department of Electrical Engineering,
Faculty of Engineering Udayana University, Bali, Indonesia
** Department of Information System, School of Information Technology and Computer
(STMIK STIKOM BALI), Bali, Indonesia

Article Info

Article history:

Received Nov 15th, 2013
Revised Dec 10th, 2014
Accepted Jan 30th, 2014

Keyword:

Network traffic
Network traffic incident
Network capture

ABSTRACT

Computer network traffic is computer activity information in a network which explains interconnection between communication processes. The complexity of network traffic itself depends on the number of communication models used. The utilization of classification analyzing of network incident traffic is one of utilization forms of network traffic. The accuracy of classification model itself depends heavily on the formation of data training. The usage of C4.5 method in the establishment of decision tree in data training toward network incident traffic is with the basis to see the efficiency of decision tree formation. Wireshark tool is used as network capture tool to perform data acquisition in the formation of data training.

Corresponding Author:

Jl. Tukad Yeh Aya No. 46, Denpasar 80225
Bali - Indonesia
Telp./Fax.: +623617956800 / +62361257055
Email: sudarma@ee.unud.ac.id

1. INTRODUCTION

Knowledge development regarding Information Technology and Communication is growing rapidly. Example of development is from security side. Computer security means that an action of protection from computer user's attack or irresponsible network access user[1]. The security of computer network can be associated also with self prevention and detection for unknown intruder's action in computer system[2]. According to ID-CERT data presents that data regarding the development of computer's crimes occurred in Indonesia, which the number of reports received in the year 2012 were at 783.457 reports. The largest report was Network Incident: 780.318 reports consisted of Brute Force attack (80%), Open Proxy (15%) and DDoS (5%). Security handling model of computer network, one of the ways that can be done is by conducting analysis in the form of classification process. In classification process, the establishment of accurate classification depends on the formation of trained data or often called as data training.

Based on the role of trained data formation as the basis for classification process then it is suggested an establishment of decision tree model for network incident traffic by using the derivation of Decision Tree namely C4.5. The using of method above as a formation of decision tree model is on the basis that C4.5 method is well fit for solving decision with discrete value. The formation of branching from a root node is used as decision alternative when the branch node is not fulfilling a condition in the rules of a decision tree.

The establishment of class label data which being used is a network captured data. Wireshark application is used to capture traffic from network attack. Wireshark is a reliable application in terms of network traffic capture[3]. The result of network incident traffic captured is made a class label data in the formation of C4.5 decision tree.

2. THEORETICAL PLATFORM

A. C.45

C4.5 algorithm is one of data classification algorithms with decision technique which is popular and favored due to its advantages. The advantages for examples: can process numeric data (continuously) and discrete, can handle a lost attribute value, generates the rules that easy to be interpreted and fastest among algorithms which using main memory in the computer[4]. C4.5 algorithm itself is an algorithm which is developed from ID3 algorithm where ID3 algorithm itself was found by J. Ross Quinlan since 1986. C4.5 algorithm is often called as decision tree algorithm which popular among decision tree algorithm group. Basically C4.5 algorithm has a similarity with ID3 from the establishment of decision tree model.

The main differences between C4.5 and ID3 are:

a. C4.5 can handle continuous and discrete attribute.
b. C4.5 can handle training data with missing value.
c. The outcome of C4.5 decision tree will be trimmed after established.
d. Attribute selection is done by using Gain ratio.

The referred gain ratio in C4.5 is to overcome the occurring bias in ID3. It is kind of normalization form to get information using "split information", which defined as follows:

$$SplitInfoA(D) = \sum_{j=0}^{v} \frac{|Dj|}{|D|} \times log2\left(\frac{|Dj|}{|D|}\right) \quad (1)$$

Where:

D = sample space (data) used for training.
Dj = the amount of samples for attribute i.

This value constitutes a potential information resulted by separating data training set, D, to be v partition, fit with the result v from the test on attribute A. To search for gain ratio value, is defined as follows:

$$GainRation(A) = \frac{Gain(A)}{SplitInfo(A)} \quad (2)$$

B. Network traffic

The measurement and analysis of network traffic is important to do in order to get a knowledge regarding network traffic characteristic. In general a network traffic data has information such as:

- IP Address
  IP address is frequently known as computer address. This computer address serves as computer identity in a network communication. This address is divided into two parts namely as source and destination identity. IP Source Address is a source address which can be identified with the sender in occurring data communication process. Meanwhile IP Destination Address is a destination address which can be identified with the data receiver in data communication process.
- Protocol
  Protocol is a rule that being applied in data communication process which its process is identified based on its service type. Each protocol running will be named according to the process conducted in communication process of computer network. The examples of communication protocols including tcp, udp, http, ftp, icmp protocol, etc.
- Length
  Length is the size of data quantity running in computer network. The size commonly used in network traffic is in byte.

Network traffic itself can be displayed in the form of raw data (data in the form of traffic record such as the outcome) or in completed form (in the form of graphic).

C. Wireshark

Wireshark is one of so many Network Analyzer tools which frequently used by Network administrators to analyze their network performances. Wireshark is preferred by many due to its interface which using Graphical User Interface (GUI) or graphic display. Wireshark is used for network troubleshooting, software analysis and communication protocol development, and education. Wireshark is widely used by network administrators to analyze their network performances. Wireshark is capable to capture data/information passing over a network which we observe in the form of network traffic. The benefits of using Wireshark application are as follows:

- Capturing information or packet data which being sent and received in a computer network.
- Discovering the activity occurring in a computer network.

- Find out and analyze our computer network performances such as access speed/data share and network connection to the internet.
- Observing the security of our computer networks.

Several information which can be captured by wireshark tool as network traffic information among others is time elapse (the time recorded in certain period), source address (source address from the data sender, can be IP address or mac address), destination address (destination address of transmitted data, can be IP address or mac address), protocol (service which running in computer network), length (the size of data transmitted), and info (additional information of each service running in computer network).

3. RESEARCH METHOD

A. Network Incident Traffic Capturing

Taking of network incident traffic is using wireshark application. Network traffic capture is conducted by capturing traffic. The taking of network incident traffic results in approximately up to tens of millions of traffic records. But the number of records yielded each day is not the same. The inequality of the number of traffic records is due to inequality of communication model in computer network performed by the user.

B. Data Filtering

Data filtering is conducted by selecting the data which will be used as a transform data process. Network incident traffic data will be given a label class in accordance with the mode of attacking type obtained from network incident traffic capture. The amount of label classes used is of 33 data with attacking type is of 8 attacks namely the types of arp spoof attack, arp sniffing attack, SSH brute force attack, HTTP DOS attack, HTTP SQL Injection, Mac Flooding, ICMP Flooding, HTTP brute force attack.

C. The establishment calculation of Tree C4.5 Model

After conducting labeling process, in the calculation of C4.5 is started with the calculation of entropy, gain information, split info and gain ratio. Example of calculation with C4.5 calculation is:

a. *Entropy*

Entropy calculation is:

$$E(S) = \sum_{i}^{N} -P_i * \log_2(P_i)$$

E(total) =

$$\left(\left(\frac{-2}{33}\right)\log_2\frac{2}{33}\right) + \left(\left(\frac{-3}{33}\right)\log_2\frac{3}{33}\right) + \left(\left(\frac{-11}{33}\right)\log_2\frac{11}{33}\right) + \left(\left(\frac{-2}{33}\right)\log_2\frac{2}{33}\right) + \left(\left(\frac{-3}{33}\right)\log_2\frac{3}{33}\right) + \left(\left(\frac{-1}{33}\right)\log_2\frac{1}{33}\right) + \left(\left(\frac{-2}{33}\right)\log_2\frac{2}{33}\right) + \left(\left(\frac{-9}{33}\right)\log_2\frac{9}{33}\right)$$

b. *Gain Information*

Gain information calculation is:

$$G(S,A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} E(S_v)$$

Gain Information(total, time) =

$$2{,}557 \left(\left(\frac{33}{33}\right)2{,}557\right)$$

$$= 0{,}000$$

c. *Splitinfo*

Splitinfo calculation is:

$$SplitInfoA(D) = \sum_{j=0}^{v} \frac{|Dj|}{|D|} \times log2\left(\frac{|Dj|}{|D|}\right)$$

Splitinfo (Time)=

$$-\left(\frac{33}{33}\right)\log_2\frac{33}{33}$$

$$= 0.000$$

d. *Gain Ratio*

Gain Ratio calculation is:

$$GainRation(A) = \frac{Gain(A)}{SplitInfo(A)}$$

Gain Ratio (Protocol) =
1,508 / 2,2537 = 0,66911

In the first iteration after performing calculation of Gain Ratio will generate a table mapping having the same label classes with filter based on the highest gain ratio value. Initial calculation will be a root node. If there are different label classes in one table, then it will be performed a recalculation of entropy calculation, Gain information, split info and Gain ratio up to producing table with same label classes.

D. The establishment of Tree C4.5 Model

To facilitate the calculation in the establishment of decision tree with C4.5 method, the labeling process is performed. From Table 1, the sample of labeling result generates to the following Table..2:

Table 1. Attack Labeling Table

Protocol Name	Label
arp_spoof attact	A
arp_sniffing attact	B
SSH brute force attack	C
HTTP DOS Attack	D
HTTP SQL Injection	E
Mac Flooding	F
ICMP Flooding	G
HTTP brute force attack	H

Tabel 2. Result Generates Labeling Table

Attribute		Total cases	Type							
			A	**B**	**C**	**D**	**E**	**F**	**G**	**H**
Total		33	2	3	11	2	3	1	2	9
Time	49	33	2	3	11	2	3	1	2	9
Support_Source_address (%)	60	33	2	3	11	2	3	1	2	9
Protocol	ARP	4	2	2	0	0	0	0	0	0
	NBNS	1	0	1	0	0	0	0	0	0
	TCP	11	0	0	5	1	0	1	0	4
	SSHv2	6	0	0	6	0	0	0	0	0
	HTTP	9	0	0	0	1	3	0	0	5
	ICMP	2	0	0	0	0	0	0	2	0

4. RESULTS AND ANALYSIS

In the calculation conducted on the previous stage, will generate several tables in each iteration. First iteration generates an Attribute Protocol as a node of first root. However on first iteration it does not present a uniform label class value from root node protocol. Therefore recalculation is performed until producing table or branch node which generating a uniform label class[5]. The sample of calculation t having the same label class and unnecessarily to conduct a recalculation is as follows:

Tabel 3. The Uniform Label Class Table

Time (menit)	Support_Source_Address	Protocol	Length	Info	Class_label
49	60 %	ICMP	98	Echo (ping) request	ICMP Flooding
49	60	ICMP	98	Echo (ping) reply	ICMP Flooding

For the table which has different Label Class should be performed a calculation until generating the same label class. The table instance with different label class is showed on table 4.

Tabel 4. Non Uniform Label Class Table

Time (menit)	**Support_Source_Address**	**Protocol**	**Length**	**Info**	**Class_label**
49	60%	TCP	74	ssh [SYN]	SSH brute force attack
49	60%	TCP	66	ssh [ACK]	SSH brute force attack
49	60%	TCP	66	ssh [FIN	SSH brute force attack
49	60%	TCP	66	ssh > …..[ACK]	SSH brute force attack
49	60%	TCP	66	ssh > …..[FIN	SSH brute force attack
49	60%	TCP	294	[TCP segment of a reassembled PDU]	HTTP DOS Attack
49	60%	TCP	54	[Malformed Packet]	Mac Flooding
49	60%	TCP	74	http [SYN]	HTTP brute force attack
49	60%	TCP	74	http > ….. [SYN	HTTP brute force attack
49	60%	TCP	66	http [ACK]	HTTP brute force attack
49	60%	TCP	66	http > ….. [FIN	HTTP brute force attack

The branch of root node yielded is depicted in each iteration as follows:

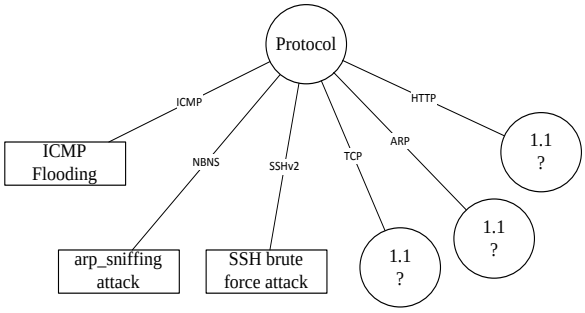

Figure 1. Root Node Protocol

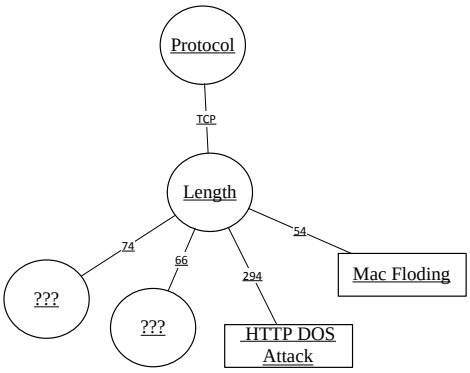

Figure 2. TCP Branch Root Node

In Figure 1, it is seen that the Protocol becomes first root node in the establishment of Decision Tree C4.5. In the branch node protocol of TCP, ARP and HTTP is not yet having the same label class. So it performs recalculation of Gain Ratio value which having Attribute Protocol value of TCP, ARP and HTTP. The result is depicted in figure 2, 3, and figure 4. In Figure 2 for branch node length of 74 and 66 it is necessary to perform calculation to get the uniform label class.

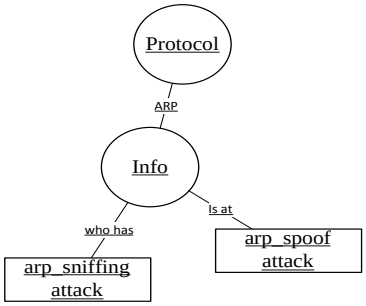

Figure 3. ARP Branch Root Node

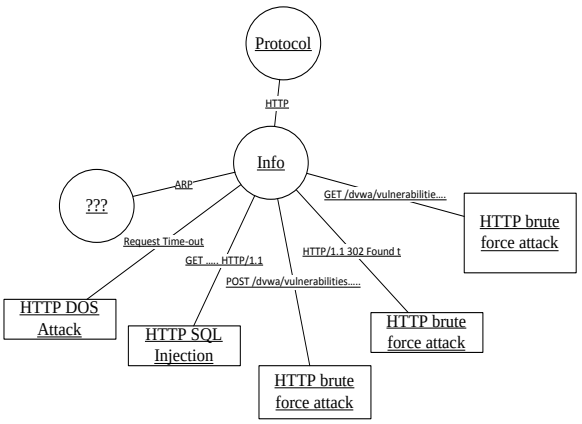

Figure 4. HTTP Branch Root Node

In Figure 3 root node info is already having the same branch node. The final result of decision tree states the attack decision of arp sniffing attack and arp spoofing attack. The final result of the establishment of decision tree with C4.5 method is like in Figure 5. In Figure 5 is the result of establishment of decision tree from attacking label class which is obtained from network incident traffic capture. The symbol of circle shape is a symbol which representing root node, meanwhile straight line representing branch node. So when later it is tested with a trial data for classification process, then clarification decision flow is based on established decision tree.

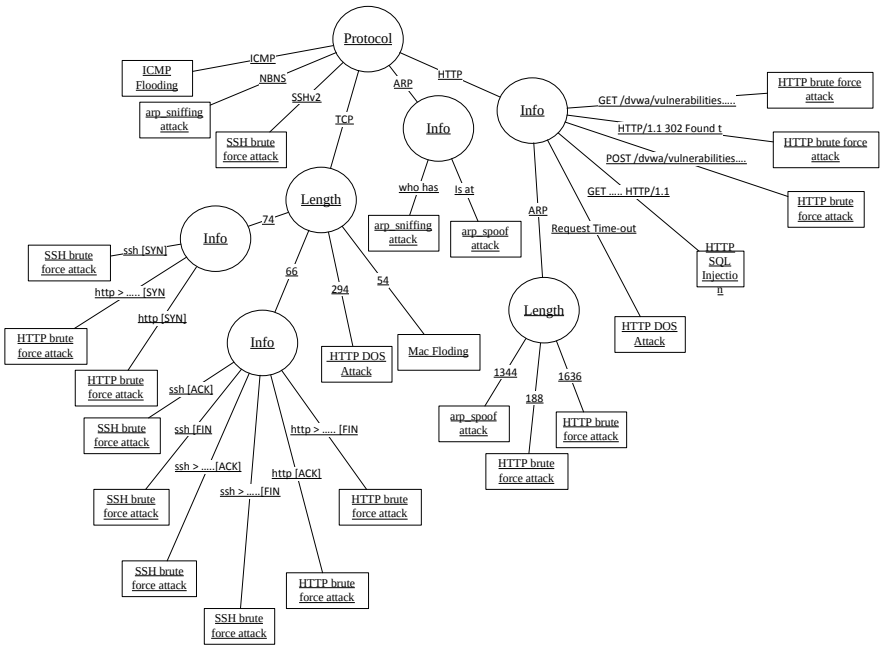

Figure 5. The shape of Decision Tree C4.5

5. CONCLUSION

Based on the establishment of decision tree, then can be concluded that the shape of decision tree C4.5 is capable to provide decision form in classification based on the uniformity value of an attribute. Attribute which possessing uniformity value will be shaped to be a root node and the value possessed in the attribute will become a decision branch node. In the last root node which having the uniformity value is the decision result of classification. The logic of decision tree C4.5 can be applied to the decision which has properties of discrete and continuous.

The development of the research conducted in this study, can be maximized with testing process of trial data which is obtained in a network with network complexity possessing large network traffic, so that it can detect the form of attack occurring in the network.

ACKNOWLEDGEMENTS

A great appreciation goes to promoter, colleague and everybody who has made valuable contributions in this study and their critical comments on this manuscript.

REFERENCES

[1] John D. Horward.1997. An Analysis of Security Incidents on the Internet.Pittsburgh : Pennsylvania USA.
[2] Golman Dieter.2010.Computer Security 3rd edition. ISBN 978-0-470-74115-3
[3] Sec*Tools*.Org: Top 125 Network Security *Tools*, http://sec*tools*.org. Diakses tanggal 24 September 2013
[4] Quinlan. J.R.1993. C4.5 : Progrmas for Machine Learning. San Mateo : Morgan Kaufmann
[5] Jiqing Liu, Jinhua Huang.2010. Broadband *Network traffic* Analysis and Study In Various Types Of Application.IEEE : 978-1-4244-7050-1110.

BIOGRAPHY OF AUTHORS

Dr. Ir. Made Sudarma, M.A.Sc.
Computer System and Informatics
Department of Electrical Engineering
Faculty of Engineering, Udayana University
Bukit Jimbaran Campus, Bali, Indonesia
Tel./Fax : 62361703315
e-mail: sudarma@ee.unud.ac.id; msudarma@baliyoni.co.id

Dandy Pramana Hostiadi, S.Kom.
Department of Information System,
School of Information Technology and Computer
(STMIK STIKOM BALI),
Denpasar, Bali
Email : dan_ganx_cil@yahoo.co.id