American J. of Engineering and Applied Sciences 2 (4): 721-725, 2009
ISSN 1941-7020

Intrusion Preventing System using Intrusion Detection System Decision Tree Data Mining

[1]Syurahbil, [1]Noraziah Ahmad, [1]M. Fadly Zolkipli and [2]Ahmed N. Abdalla
[1]Faculty of Computer System and Software Engineering, University Malaysia Pahang, Malaysia
[2]Faculty of Elec and Elect Engineering, University Malaysia Pahang, Pahang, Malaysia

Abstract: Problem statement: To distinguish the activities of the network traffic that the intrusion and normal is very difficult and to need much time consuming. An analyst must review all the data that large and wide to find the sequence of intrusion on the network connection. Therefore, it needs a way that can detect network intrusion to reflect the current network traffics. **Approach:** In this study, a novel method to find intrusion characteristic for IDS using decision tree machine learning of data mining technique was proposed. Method used to generate of rules is classification by ID3 algorithm of decision tree. **Results:** These rules can determine of intrusion characteristics then to implement in the firewall policy rules as prevention. **Conclusion:** Combination of IDS and firewall so-called the IPS, so that besides detecting the existence of intrusion also can execute by doing deny of intrusion as prevention.

Key words: Firewall rules, network security, intrusion detection, network traffics, decision tree

INTRODUCTION

With the global Internet connection, network security has gained significant attention in research and industrial communities. Due to the increasing threat of network attacks, firewalls have become important elements of the security policy is generally[7]. Firewall can be allow or deny access network packet, but firewall cannot detect intrusion or attack, so to need intrusion detection and then implemented to firewall is access control systems as prevention. Intrusion detection are also considered as a complementary solution to firewall technology by recognizing attacks against the network that are missed by the firewall[10]. Firewall and IDS represent an old stuff terminology in the field of IT security. Firewall is good for protection a system and network and can minimization risk of attack to network. IDS can detect existence intrusion or attack. The joining ability of IDS and firewalls, that is so-called IPS. That is a functioning tool to detect intrusion and then denying by firewall for prevention.

For each type of network traffics, there are one or more different rules. Every network packet, which arrives at firewall, must be check against defined rules until a matching rule found[1,10]. The packet will be then allow or banned access to the network, depending on the action specified in the matching rule. Each rule identifies specific type of network traffic[4,12]. Characteristics to reflect the current of network traffics can observe from network traffic logs[4] as human pattern recognize[9].

This Study focus on some methods to prevention from attempt intrusion to find intrusion characteristics in the network traffic as IDS then implementation to firewall policy rules as prevention. To find rules of intrusion characteristics using decision tree machine learning data mining. Method used to generate of rules is classification by ID3 algorithm of decision tree. It is an efficient and optimized to make the rules filtering in firewall.

Theoretical background:

Intrusion Detection System (IDS): Intrusion detection can be performed manually or automatically[8]. Manual intrusion detection might take place by examining log files or other evidence for signs of intrusions, including network traffic. A system that performs automated intrusion detection is called an Intrusion Detection System (IDS). IDS play a vital role in ensuring the security of modern computer installations. Such systems are need in order to detect hostile activity and to respond appropriately. As networks continue to expand and become more exposed to a diversity of sources, both hostile and benign, IDS need to be able to deal with a large and ever-increasing flow of alerts and events. Therefore, automatic procedures for detecting and responding to intrusion are becoming increasingly essential[5].

Corresponding Author: Syurahbil, Faculty of Computer System and Software Engineering, University Malaysia Pahang, Malaysia

```
-A INPUT –s 203.130.206.5 –p tcp –d 10.10.15.7 --
dport 80 –j DROP
```

Fig. 1: Firewall rules

Firewall rules: A firewall security policy is a list of ordered filtering rules that define the actions performed on packets that satisfy specific conditions. Before to develop rules filtering by using packet filter, anything have to be considered beforehand how far demarcation which will be applied, because more and more demarcation applied hence increases the search time and space requirements of the packet filtering process[1] and consequences to make downhill performance progressively[11]. This matter because every incoming network packet and go out the network checked beforehand by rules alternately until matching rule found in firewall[12].

Firewall rules can limit to access the connection of pursuant to parameter: source IP, destination IP, source port, destination port, protocol and others[8,10].

Following example of firewall rules in Fig. 1. Firewall rule of above explaining to enhance the order by the end of chain (A) for the traffic of incoming to firewall (INPUT) by source IP address (-s) 203.230.206.5 with the type protocol (-p) tcp to destination IP address (-d) 10.10.15.7 and destination port (--dport) 80 hence done by action (-j) dropped (DROP) by firewall.

Log files: Log files can give an idea about what the different parts of system are doing. Logs can show what is going right and what is going wrong. Log files can provide a useful profile activity. From a security standpoint, it is crucial to be able to distinguish normal activity from the activity of someone to attack server or network[3].

Log files are useful for three reasons[11]:

- Log files help with troubleshooting system problems and understanding what is happening on the system
- Logs serve as an early warning for both system and security events
- Logs can be indispensable in reconstructing events, whether determined an intrusion has occurred and performing the follow-up forensic investigation or just profiling normal activity

Following some example from log files in Fig. 2 and 3.

```
[**] INFO - Possible Squid Scan [**]
04/20-14:06:49.953376 192.168.0.33:1040 ->
192.168.0.1:3128 TCP TTL:128 TOS:0x0 ID:393
IpLen:20 DgmLen:48 DF
******S* Seq: 0x60591B9  Ack: 0x0  Win: 0x4000
TcpLen: 28
TCP Options (4) => MSS: 1460 NOP NOP SackOK
```

Fig. 2: Snort log files

```
Feb 17 21:02:13 (none) sshd[14938]: Failed
password for albi from 172.18.64.26 port 3419
ssh2
```

Fig. 3: Syslog files

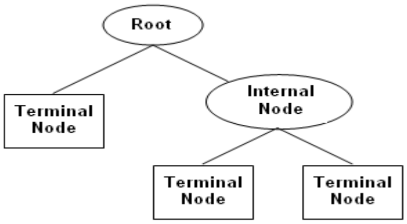

Fig. 4: Decision tree structure

The example in Fig. 2 using Snort log files (/var/log/snort), there is effort for the scan of existence of Squid proxy server at 192.168.0.1 port 3128 from workstation 192.168.0.33 port 1040. Workstation 192.168.0.33 delivering package of Synchronization TCP seen from ****** S*.

Follow the example in Fig. 3 using syslog files (/var/log/syslog), there is someone trying albi user login and the failed password from the IP address 172.18.64.26 port 3419 passing ssh service (port 22 protocol tcp).

Decision tree of data mining: Decision tree is a technique in classification method of data mining for learning patterns from data and using these patterns for classification. Decision tree are structures used to classify anddata andwith andcommon andattributes andas andshown andin Fig. 4. Each decision tree represents a rule, which categorizes data according to these attributes[4,6].

Where each node (nonleaf node) denotes a test on an attribute, each branch represent an outcome of the test and each leaf node or terminal node holds a class label. The topmost node in a tree is the root node[2].

A decision tree classifier is one of the most widely need supervised learning methods used for data exploration. It is easy to interpret and can be re-represented as if-then-else rules. This classifier works well on noisy data. A decision tree aids in data exploration in the following manner:

- It reduces a volume of data by transformation into a more compact form that preserves the essential characteristics and provides an accurate summary
- It discovers whether the data contains well-separated classes of objects, such that the classes can be interpreted meaningfully in the context of a substantive theory
- It maps data in the form the leaves to its root. This may used to predict the outcome for a new data or query[6]

MATERIALS AND METHODS

This research using decision tree a technique of data mining machine learning to find the intrusion characteristics for intrusion detection. Algorithm is used ID3 to construct Decision tree. Network traffic logs as data training that describes the human behavior in network traffics as intrusive activities and normal activities. The results of decision tree training will get rules of intrusion characteristics then these rules to implement in the firewall rules as prevention.

Determining occurrence of intrusion or normal activities at network traffic log can be conducted with two way of that is:

- Observe manually activities network traffic in log files. Example, application software of log files is syslog, syslog_ng, tcpdump and others. Pattern found to see intrusion through log seen modestly, for example there are some times trying to access using login or password failed, trying port scan, abundant ping, delivery of abundant package by repeat
- Using software as a means of assists functioning as Network Intrusion Detection System (NIDS) able to determine intrusion activities or normal activities, for example snort software

RESULTS

Collect and extract log files of intrusive activities and normal activities become five of parameter as attributes and belongs to a class ‘Yes’ or ‘No’ of intrusive for the data training of decision tree. The parameter is IP address source, IP address destination, port source, port destination and protocol as shown in Table 1.

Applying Decision Tree to Find Intrusion Characteristic: Suppose train a decision tree using the example in Table 1.

Table1: Network Traffics

No.	Source IP	Destination IP	Source port	Destination port	Protocol	Intrusion
1	122.306.13.100	10.10.1.2	1360	22	TCP	Yes
2	122.306.13.100	10.10.1.2	1425	22	TCP	Yes
3	122.306.13.100	10.10.1.2	1488	22	TCP	Yes
4	122.306.13.100	10.10.1.5	1559	22	TCP	Yes
5	122.306.13.100	10.10.1.5	1620	80	TCP	Yes
6	122.306.13.100	10.10.1.3	2156	80	TCP	Yes
7	203.306.14.20	10.10.1.5	2158	22	TCP	Yes
8	203.306.14.20	10.10.1.5	1624	22	TCP	Yes
9	203.306.14.20	10.10.1.3	4207	22	TCP	No
10	203.306.14.20	10.10.1.3	4607	80	TCP	Yes
11	206.145.206.4	10.10.1.2	4690	21	TCP	Yes
12	206.145.206.4	10.10.1.2	1552	80	TCP	Yes
13	206.145.206.4	10.10.1.3	1572	80	TCP	no
14	206.145.206.4	10.10.1.5	1430	80	TCP	Yes
15	206.145.206.4	10.10.1.5	1630	80	TCP	No

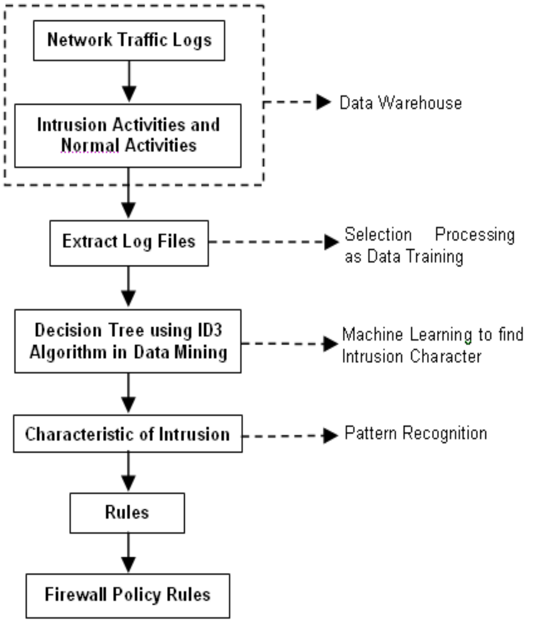
Network Traffic Logs
Intrusion Activities and Normal Activities
Data Warehouse
Extract Log Files
Selection Processing as Data Training
Decision Tree using ID3 Algorithm in Data Mining
Machine Learning to find Intrusion Character
Characteristic of Intrusion
Pattern Recognition
Rules
Firewall Policy Rules

Fig. 5: Research process

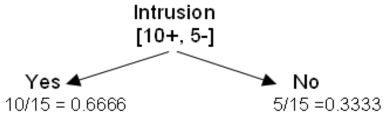

Fig. 6: Set of 15 examples are 10 ‘Yes’ and 5 ‘No’

First, work out which attribute will be put into the node at the top of tree as root node as shown in Fig.5.

S is a set of 15 examples are 10 ‘Yes’ and 5 ‘No’, show in Fig. 6.

Second, Calculate entropy of S:

$$E\ (S) = -p_{+} \log_2(p_{+}) - p_{-} \log_2(p_{-})$$
$$E\ (S) = -0.6666 \log_2 0.6666 - 0.3333 \log_2 0.3333$$
$$= 0.9182$$

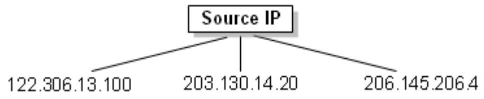

Fig. 7: Source IP as root node

Table 2: Source IP

	Intrusion		
Source IP	Yes	No	Total
122.306.13.100	6	0	6
203.130.14.20	1	3	4
206.145.206.4	3	2	5
	Total		15

E(Intrusion, Source IP) = $6/15\ (6/6\ \log2\ 6/6–0/6\ \log2\ 0/6)+4/15\ (1/4\ \log_2\ 1/4–3/4\ \log_2\ 3/4)+5/15\ (3/5\ \log_2\ 3/5–2/5\ \log_2\ 2/5)$
= 0.3783
G(Intrusion, Source IP) = 0.9182–0.5399
= 0.3783

G(Intrusion, Dest IP) = 0.2516
G(Intrusion, Dest Port) = 0.0421

Table.2 the highest of Gain is Source IP. As a note, ignore protocol to the calculation, because only one value of protocol attributes that is TCP, but for each path from the root to a leaf node assume there is a TCP protocol. Meanwhile, Source Port attributes have large numbers of values called to super attributes.

Third, GainRatio can be use for attribute selection between Source IP and Source Port.

Split(Intrusion, Source IP) = $((6/15\ \log_2\ 6/15)+(4/15\ \log_2\ 4/15)+(5/15\ \log2\ 5/15))$
= 1.3675
Split(Intrusion, Source Port) = 3.9060
Gain Ratio (Intrusion, Source IP) = 0.9182/1.3675 = 0.6714
Gain Ratio (Intrusion, Source Port) = 0.9182/3.906 = 0.2350

Source IP has the highest gain ratio, therefore, it is used as the decision node, show in Fig. 7.

This process goes on until all data classified perfectly or run out of attributes. The complete of tree show in Fig. 8.

Decision tree can simplified by pruning all connections are assumed normal and not classified as intrusions as shown in Fig. 9.

DISCUSSION

Rule extraction and characteristic of intrusion: The knowledge represented in decision trees can be extracted and represented in the form of IF-THEN rules.

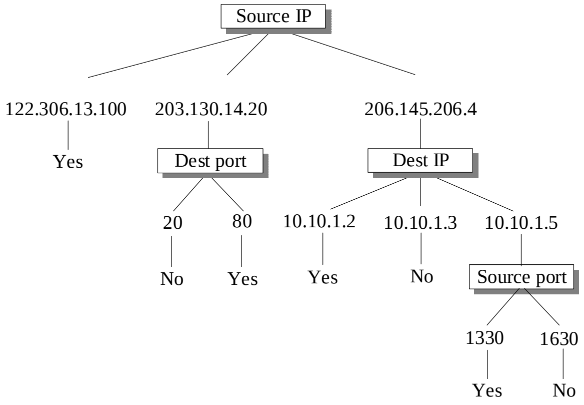

Fig. 8: Decision tree network traffic for IDS

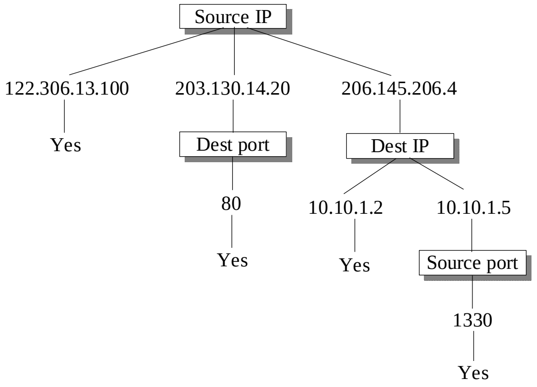

Fig. 9: Pruning Decision tree network traffic for IDS

R1 : IF (Source IP =123.202.72.140) THEN Intrusion = Yes
R2 : IF (Source IP = 203.130.14.20) AND (Dest Port = 80) THEN Intrusion = Yes
R3 : IF (Source IP = 206.145.206.4) AND (Dest IP=10.10.1.2) THEN Intrusion= Yes
R3 : IF (Source IP = 206.145.206.4) AND (Dest IP=10.10.1.5) AND (Source IP=1630) THEN Intrusion= Yes

Fig. 10: Rule extraction to IF-THEN

One rule can be created for each path from the root to a leaf node. Each attribute-value pair along a given path forms a conjunction in the rule antecedent (“IF” part). The leaf node holds the class prediction, forming the rule consequent (“THEN” part). The IF-THEN rules may be easier for humans to understand[2,6] is shown Fig. 10.

Implementation to firewall rules: The examples of extract rule of tree decision is shown in Fig. 10. representing characteristic of intrusion earn implementation into firewall rules is shown in Fig. 11. Do not forget to every rule there is a TCP protocol.

```
-A INPUT -p tcp -s 123.202.72.140 -j DROP
-A INPUT -p tcp -s 203.130.14.20 --dport 80 -j DROP
-A INPUT -p tcp -s 206.145.206.4 --d 10.10.1.2 -j DROP
-A INPUT -p tcp -s 206.145.206.4 --sport 1630 -d 10.10.1.5 -j DROP
```

Fig. 11: Intrusion characteristic implementation into firewall rules

Firewall policy rules above representing preventive action, where every network packet with criteria like rules firewall above will DROP.

CONCLUSION

Network traffic logs to describe patterns of behavior in network traffic accident with intrusive or normal activity. Decision tree technique is good for the intrusion characteristic of the network traffic logs for IDS and implemented in the firewall as prevention. The both of this combination is called IPS. The other hand, this technique is also good efficiency and optimize rule for the firewall rules such as avoid redundancy.

REFERENCES

1. Al-Shaer, E.S. and H.H. Hamed, 2004. Discovery of policy anomalies in distributed firewalls. Proceeding of the 23rd Annual Joint Conference on Computer and Communications Societies, Mar. 7-11, IEEE Computer Society, Washington, DC., USA., pp: 2605-2616. http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1354680.
2. Han, J. and M. Kamber, 2006. Data Mining Concepts and Techniques. 2nd Edn., Morgan Kaufmann Publishers, Elsevier Inc., USA., ISBN: 10: 1558609016, pp:800.
3. Terpstra, J.H., P. Love, R.P. Reck and T. Scanlon, 2004. Hardening Linux. 1st Edn., McGraw-Hill/Osborne, Apress, California, ISBN: 13: 978-1590594445, pp: 584.
4. Golnabi, K., R.K. Min, L. Khan and E. Al-Shaer, 2006. Analysis of firewall policy rules using data mining techniques. Proceeding of the 10th IEEE/IFIP Symposium on Network Operation and Management, Apr. 3-7, IEEE Computer Society, Washington, DC., USA., pp: 305-315. DOI: 10.1109/NOMS.2006.1687561.
5. Kantardzic, M.M. and J. Zurada, 2005. Next Generation of Data-Mining Applications. 1st Edn., Jhon Wiley and Son, Inc., Hoboken, New Jersey, ISBN: 10: 0471656054, pp: 696.
6. Michael W. Berry and M. Browne, 2006. Lecture Notes in Data Mining. Word Scientific Publishing Co. Pte. Ltd. ISBN:978-981-256-802-1.
7. Benelbahri, M.A. and A. Bouhoula, 2007. Tuple based approach for anomalies detection within firewall filtering rules. Proceeding of the 12th IEEE Symposium on Computers and Communications, July 1-4, IEEE Computer Society, Washington, DC., USA., pp: 63-70. DOI: 10.1109/ISCC.2007.4381486.
8. Ye, N. and X. Li, 2001. A scalable clustering technique for intrusion signature recognition. Proceedings of the IEEE Workshop on Information Assurance and Security United States Military Academy, June 5-6, United States Military Academy, West Point, New York, USA., pp: 1-4. http://www.itoc.usma.edu/Workshop/2001/Authors/Submitted_Abstracts/paperT1A1(01).pdf.
9. Winding, R., T. Wright and M. Chapple, 2006. System anomaly detection: Mining firewall logs. Proceeding of the Securecomm and Workshops, Aug. 28-Sept. 1, IEEE Computer Society, Washington, DC., USA., pp: 1-5. DOI: 10.1109/SECCOMW.2006.359572.
10. **Steve** Suehring, and R.L. Ziegler, 2006. Linux Firewalls. 3d Edn., Pearson Education, Inc., Novell Press. ISBN-13: 9780672327711.
11. Abbes, T., A. Bouhoula and M. Rusinowitch, 2004. Protocol analysis in intrusion detection using decision tree. Proceeding of the International Conference on Information Technology, Apr. 5-7, IEEE Computer Society, Washington, DC., USA., pp: 404-404. http://portal.acm.org/citation.cfm?id=978412.
12. Pale, T.K.P., 2007. Optimization of firewall rules. Proceedings of the ITI 29th International Conference on Information Technology Interfaces, June 25-28, IEEE Computer Society, Washington, DC., USA., pp: 685-690. DOI: 10.1109/ITI.2007.4283854.