A.I. APPLICATIONS: INTELLIGENT FIREWALLS

GAGNIUC MIRCEA-BOGDAN

University Politehnica of Bucharest – Faculty of Automatic Control and Computers

Abstract: This paper will analyze the possibility to create firewalls with AI algorithms to implement prediction methods for setting firewall rules of data traffic acceptance. Using ID3 decision tree algorithm, created by J. Ross Quinlan will be demonstrating a possible prediction possibility to implement automat decision for firewall questions. The paper will use a simple example, calculating a simple decision tree based on ID3 algorithm.

Keywords: Internet, Security, Firewall, ID3, Decision, Tree, Quinlan, Entropy, Gain

1 .COMPUTER SECURITY – FIREWALLS

Internet Services represents a gate to the whole human knowledge, and, more than that, an entire lot of commercial, educational and entertainment possibilities. A gate is a double sense way, some persons can go out through that gate, other persons can go in. An internet user opens a gate, and through this gate he start browse in Internet virtual world. Sometimes, the Internet user requests some services, some information, and, as an answer of his request, through this gate come in computer requested information. Even simple browsing services represent information requested by the user and question is being represented by the Internet Address who is written in browser software.

The problem of using Internet Services, is the possibility to use the gates opened by Internet user, by the software, applications or services *unrequested by user*. Often, these unrequested enters through our gate, represent malware applications, capable to destroy information kept in computers, or even the computer themselves. Let us remember that personal information represent persons, with an social status. The worst case situations are represented by the materialization of risc to destroy the personal reputation, as a social person.

In our days, Internet represents a revolutionary way to access and share information. But, unfortunately, it is a revolutionary way to destroy the information, also. Even we are talking about a home computer, which is used for entertainment purposes, the user must respect some rules for maintain the functionality of the computer. One of the base rule, is using the protection software. The most used protecting software is an antivirus and an firewall software.

The functionality of antivirus software is relative simple: based on virus signature, when is recognized this signature, the antivirus software is already programmed to disinfect affected software. The user intervention is requested only in the situation when is not programmed in antivirus algorithm.

Firewall functionality is totally different: similar to a door keeper in front of an open gate, the firewall software let information in or out only conform user instructions. The difference between those applications is user intervention. While antivirus can operate without user intervention, the firewalls needs user intervention to let information inside or outside of computer connected at Internet. The problem with user decisions is that sometimes, a message of type “Svchost.exe process needs to connect at 10.54.345.222 address using TCP/IP

protocol at port 1456" can put in difficulty many computer users. Sometimes, the common three options offered by the usual firewalls (Allow, Deny and Create Rule) are chosen randomly by a lot of users. Even worst, to be sure for capability of accessing any Internet service, many users choose option "Allow" for every firewall request.

If we think at gate example, the firewall is a gate keeper, who needs to be instructed for every access request, in or out. And, is sufficient only one unattended rule created by the user, and the gate keeper will proceed exactly by rule conditions. Or, that unattended rule, can permit computer to access, for example, an Internet page, where, an html script, can access personal information, or can open another gate in computer, which can open somewhere, or to someone, personal information.

An intelligent gate keeper, who can create the own rules, based on „lanlord" decisions, can eliminate the risk of unprofessional user response. But, when we talk about an intelligent firewall, we are talking about intelligent decisions – an artificial intelligence which can assist usual computer user, for firewall decisions.

2. HOW FIREWALL WORKS

Firewall is a network component (or a set of components) that restricts access between a protected network and the Internet, or between other components of network. Access restriction is based on rules created by computer user. Restrictions are referred to outcoming and to incoming network traffic.

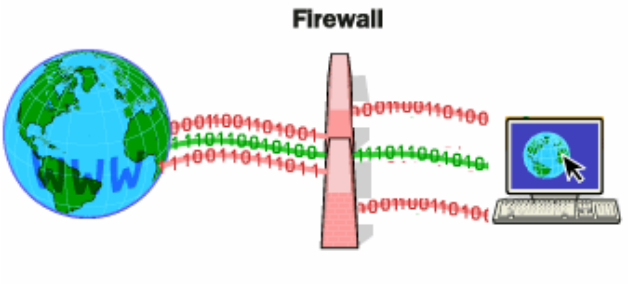

Figure 1 - Firewall Schematic Function

It is easy to understand that bad decisions can determine firewall to block desired information traffic. In this situation, user application can not send information over the (Internet) network or, information requested by user can not be received by computer applications.

Network traffic control is made by firewalls through packet filtering. The computer traffic network consists by packets of bits – packet being fundamental unit of computer network communication. Every packet has a set of headers containing certain information. The main information used by firewalls, is

- IP source address
- IP destination address
- Protocol (most usual protocols being TCP, UDP or ICMP)
- TCP or UDP source port
- TCP or UDP destination port
- ICMP message type
- Packet size.

Based on these elements, a network firewall filter fundamental network elements – packets – and let in or out, only packets accepted by user. Then, starting from user decision, firewalls respect the same rule for the same events, asking the applicable rule only for new events. It is obvious that an incorrect rule, will determine an application or system malfunction, because necessary traffic will be blocked. Or, even worst, unnecessary traffic can be allowed, and malware applications or software can enter in system.

Let us imagine a real situation. A computer is protected by a software firewall, installed over his operating system. That computer, will have an IP address, and his computers neighbors exists in a local area network, some computers having IP addresses in the same class with our computer, other, in different class (another local network).

Every time when an application needs information exchange over the network, firewall software will ask the user about the rule applicable for that application. When the question is displayed, the user is informed about application or process who needs network information interchange, about IP remote address, about used protocol, about data direction (in or out). With this information, user must decide if traffic will be allowed or not. In the same time, user can set the future response of the firewall for similar future events. Based of user settings, the firewall may apply the established rule, or will ask the user again for a new decision.

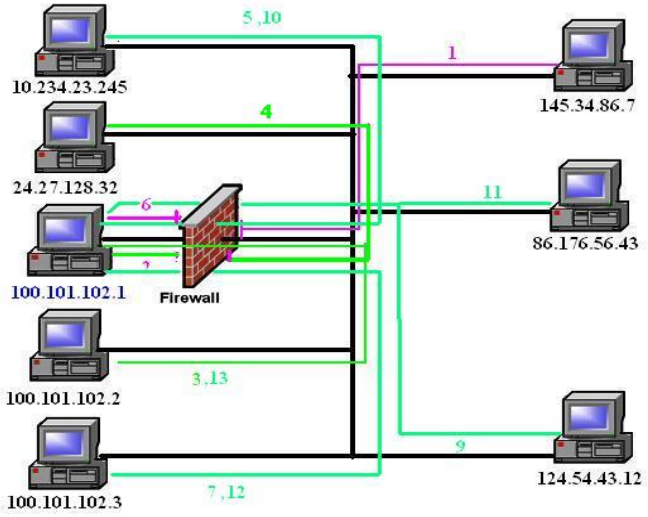

Figure 2 – Firewall decision in some situations.

As it can be seen in Figure 2, for example of firewall function, let presume that user computer has IP address 100.101.102.1. Our user has a software firewall installed on his computer. User has connection with another four computers (with following IP addresses: 101.101.102.2, 101.101.102.3, 24.27.128.32, 10.234.23.245), inside a local area network. In some cases, from network outside (another network, or Internet), some stations (with IP 145.34.86.7, 86.176.56.43, 124.54.43.12) will try to connect to user station.

When user starts the computer, after operating system boot is finished, and with him the firewall is started, the following events and user decisions will intervene:
1. The computer with IP address 147.34.86.7 sends an ICMP packet (an echo request, for example). Firewall blocks the traffic and informs the user, waiting for his decision. Because the user was not starting yet an application and he did not recognize the process, or the IP address, he will create the rule “Block Traffic from that IP address and for that protocol”.
2. The computer is started; the user has no run any application, but an unknown user process send a ping command to IP address 114.234.43.1. The user does not have any information about the process or about the remote IP, the decision (and rule created) is to maintain blocking network traffic.
3. Another ICMP packet is send by network, with IP destination 100.101.102.1, our computer IP address. The sender has IP address 100.101.102.2. The user has no idea about the process describe by firewall, about what mean ICMP echo request address, but based on his own IP address, he recognize one neighbor system, and unblock network traffic, creating a rule for the future, with this decision.
4. Using a application (process) unknown by user, the station from remote IP 24.27.128.32 wants to send a TCP packet. The user setting is to allow that traffic, because he knows what user is behind that IP address and he wants to change information’s with that computer.

5. The user starts an application. Suddenly, firewall informs that the applications just started, wants to receive an UDP packet from 10.234.23.245. Because user knows application, and because that application probably needs incoming information's, the rule will be "allow' traffic, for that application, from that remote IP, using UDP protocol.
6. A process (known by user) wants to transmit an UDP packet to remote IP 67.42.7.98.234. Because user knows that IP is outside his neighbor computers, and because he did not want to send information's to unknown people, the user choose will be "block traffic".
7. An application running on user computer wants to transmit an UDP packet to another machine, with IP 100.101.102.3. Because IP remote address is known by user, and application is under user control, decision (and rule from now one for this application and protocol) is to "allow" traffic.
8. A new event, an unknown process wants to receive a TCP packet from 67.43.23.65.78. User gives the correct answer, blocking the traffic.
9. A user application, wants to receive an UDP packet from 124.54.43.12. User allows traffic, for correct function of his application.
10. Another application known by user, need to receive a TCP packet from 10.234.23.245. When firewall inform about this event, user will accept that kind off traffic, based on known application.
11. When another application ran by computer user needs to transmit a TCP packet to 86.176.56.43, user will answer to firewall question allowing traffic for that combination of application, protocol and remote IP.
12. One unknown process needs to transmit a TCP packet to 101.101.102.3. Based a known IP address, user will allow that traffic, too.
13. A user application needs to receive a ICMP packet from IP address 100.101.102.2.In this case, user know the all parameters (IP remote address, application), so the traffic will be allowed.
14. In the last example, an unknown process needs to transmit a TCP packet to a IP remote address about user know that is outside his network. For this reason, the user decision will be to block that traffic.

All these examples are shown in Figure 2, with green line is represented allowed traffic while with magenta line being represented blocked traffic. Every line has a corresponding number equal to event number.

3. USING ID3 DECISION TREE ALGORITHM FOR FIREWALL DECISSIONS.

Security of computers based on firewalls has two major disadvantages represented by frequently messages that need user intervention, and those interventions need a minimum technical knowledge about computer networks. And, even in that situation, is enough a one unattended answer, and the risk of transmitting confidential data to an unknown user, or to block traffic for a necessary application or process, or to receive in computer a malware software who can crash the operating system, can become real fact.

A firewall that can predict user answer and set its own decisions according to a set of examples can eliminate the disadvantages presented before. Instructions are given by user, based on several attributes of data traffic that needs attention.

Analyzing the situation described in Figure 2, the events can be described with several attributes:
IP address - The computer involved in communication, protected by analyzed firewall, has IP address 100.101.102.1 and is connected to a local area network with another four workstations. Two of them have IP inside the same class with protected computer, and the other have IP address outside computer IP class. Other stations involved in communication with this computer, have IP outside computer class, and more than that, they are outside local area network, being connected to an extern network (which are represented in right part of Figure 2). A usual user will know his computer IP, and he can decide if a network computer is from local area network, or outside his local area network. But, never a computer user will verify an IP address using a WHOIS Service, before allow or deny network traffic. For that reason, the IP address is important not as value, but as member of one instance of this attribute: IP is from extern network, or from LAN having an IP inside IP user class computer or outside IP class.

Data direction is another decision attribute. Direction can have only two instances: in or out. Another attribute essential for decisions, is direction of data traffic, with only two instances, in or out. Sometimes, a user can decide to allow or deny traffic, only based on this information.

Application or process can be a decision attribute, a user can deny network traffic for all unknown application or processes, if he not considers another attributes.

The protocol use in data traffic can be another decision attribute, and in presented example can have three instances: ICMP, TCP or UDP.

Based on these considerations, the exemplified events can be condensed like in Table 1.

Example of Firewall Settings. Table 1

Event	IP remote address	Protocol	Application known	Direction	Allow traffic
E1	Extern Network	ICMP	No	Incoming	No
E2	Extern Network	ICMP	No	Outgoing	No
E3	Inside IP Class	ICMP	No	Incoming	Yes
E4	Outside IP Class	TCP	No	Incoming	Yes
E5	Outside IP Class	UDP	Yes	Incoming	Yes
E6	Outside IP Class	UDP	Yes	Outgoing	No
E7	Inside IP Class	UDP	Yes	Outgoing	Yes
E8	Extern Network	TCP	No	Incoming	No
E9	Extern Network	UDP	Yes	Incoming	Yes
E10	Outside IP Class	TCP	Yes	Incoming	Yes
E11	Extern Network	TCP	Yes	Outgoing	Yes
E12	Inside IP Class	TCP	No	Outgoing	Yes
E13	Inside IP Class	ICMP	Yes	Incoming	Yes
E14	Outside IP Class	TCP	No	Outgoing	No

In Artificial Intelligence theory, decisions based on classifiable events are efficient implemented with decision trees. In presented example, a useful decision tree algorithm is ID3 algorithm developed by J. Ross Quinlan. This algorithm can solve user decision problem.

ID3 builds a decision tree from a fixed set of examples. The resulting tree is used to classify future samples. The example has several attributes and belongs to a class (like yes or no). The leaf nodes of the decision tree contain the class name where as a non-leaf node is a decision node. The decision node is an attribute test with each branch (to another decision tree) being a possible value of the attribute. ID3 uses information gain to help it decide which attribute goes into a decision node. The advantage of learning a decision tree is that a program, rather than a knowledge engineer, elicits knowledge from an expert.

ID3 decide which attribute is the best based on a statistical property, called information gain. Gain measures how well a given attribute separates training examples into targeted classes. The one with the highest information gain (information being the most useful for classification) is selected.

In order to define gain, it is used an idea from information theory called entropy. Entropy measures the amount of information in an attribute.

Given a collection S of c outcomes

$$Entropy(S) = \sum \; - p(I)\log_2 p(I) \tag{1}$$

Where p(I) is the proportion of S belonging to class I and $\sum$ is over c. Note that S is not an attribute but the entire sample set.

In our case, S is a collection of 14 examples with 9 Yes and 5 No examples, then

$$Entropy(S) = -\frac{9}{14}\log_2(\frac{9}{14}) - \frac{5}{14}\log_2(\frac{5}{14}) = 0.940 \quad (2)$$

Gain(S, A) is information gain of example set S on attribute A and is defined as

$$Gain(S, A) = Entropy(S) - \sum(\frac{|S|}{|S_V|}Entropy(S_V) \quad (3)$$

Where ∑ is each value v of all possible values of attribute A, S_v = subset of S for which attribute A has value v, $|S_v|$ = number of elements in S_v, |S| = number of elements in S.

In presented example, S is a set of 14 examples in which one of the attributes is, for example, Direction. The values of Direction can be *Incoming* or *Outgoing*. The classification of these 14 examples are 9 Yes and 5 No. For attribute Direction, suppose there are 8 occurrences of Direction = Incoming and 6 occurrences of Direction = Outgoing. For Direction = Incoming, 6 of the examples are Yes and 2 are No. For Direction = Outgoing, 3 are Yes and 3 are No. Therefore

$$Gain(S, Direction) = Entropy(S) - \frac{8}{14}Entropy(S_{Incoming}) - \frac{6}{14}Entropy(S_{Outgoing}) = 0.048 \quad (4)$$

For each attribute, the gain is calculated and the highest gain is used in the decision node. In our example, rest of attributes has Gain(S, IP Remote Address)=0.246, Gain(S, Protocol)=0.029 and Gain(S, Application Known)=0.151. "IP remote address" attribute has the highest gain, therefore it is used as the decision attribute in the root node.

Applying iterative ID3 algorithm, will be obtain the following decision tree:

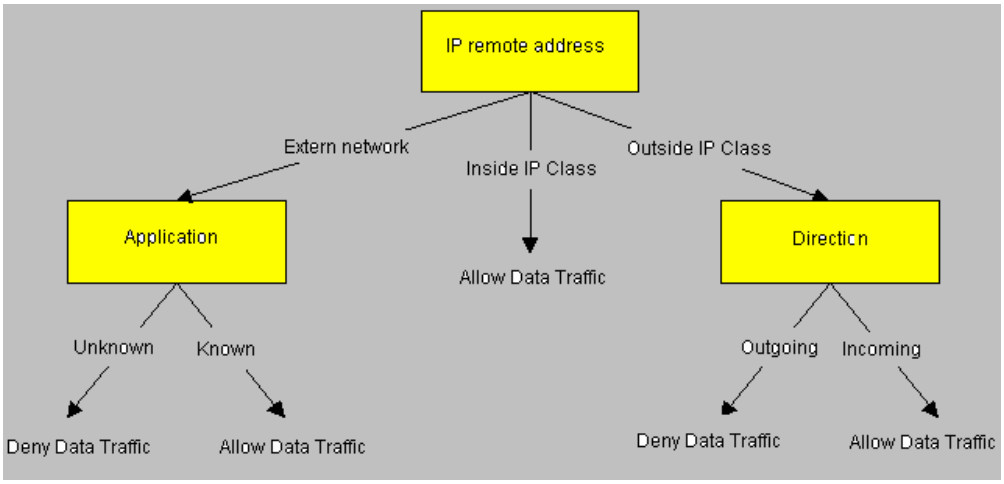

Figure 3 – Firewall Decision Tree

4. CONCLUSIONS

Firewalls represent an important security network component. Like any other security software, simple installing of a firewall system will not offer a total protection and assures network connection. After is installed, a firewall needs a period of settings, when user answer's is the key of entire configuration process. Because the user response is based on several specific communication attributes, firewalls can easily adapted to set automatic configuration, based on preprogrammed examples set. Of course, user decision will not be eliminated, but,

instead of maintain traffic blocked until user set a rule, firewall will predict user answer, based on the decision tree like in Figure 3. When user intervenes with a new rule (for a new event or for an old event changing the old rule) his answer is implemented as a new example, ID# algorithm will be used and a new decision tree will be generated. In that way, a firewall will self adapt and decisions will be predicted in accordance with user desires.

BIBLIOGRAPHY

[1] Elizabeth D. Zwicky & others , “Building Internet Firewalls”, O'Reilly Media, June 2000
[2] Andrew C, “Building Decision Trees with the ID3 Algorithm", Dr. Dobbs Journal, June 1996
[3] Paul E., "Incremental Induction of Decision Trees", Kluwer Academic Publishers, 1989
[4] Tom M., "Machine Learning", McGraw-Hill, 1997