

Method of multi-feature fusion based on SVM and D-S evidence theory in Trojan detection

Shengli Liu^{*}, Xiang Gao, Pan Xu, Long Liu

State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou, 450002, China

Received 1 March 2014, www.tsi.lv

Abstract

According to the low accuracy and low stability of the single feature-based method for Trojan detection, a multi-feature fusion method based on SVM and DS evidence theory is proposed. First, three types of flow features such as session, upload data of session/download data of session, distribution of data packet size are extracted from the data stream. Then the SVM classification results of each single feature are used as evidences to construct the basic probability assigned (BPA). Finally, we use DS combination rule of evidence to achieve the decision fusion and give the final detection results by fusion results. The experimental results showed that the accuracy of multi-feature fusion method was 97.48% which has good performance on accuracy and stability compared with the single feature method in Trojan detection.

Keywords: Trojan detection, support vector machine, DS evidence theory, multi-feature fusion

1 Introduction

With the fast development of information technology, the rapid growth of data has become a serious challenge as well as a good opportunity in many industries [1]. In the era of big data, Trojan technology based on data stream communication has become the primary means of attackers to steal confidential information, which poses a serious threat to network security. Therefore, the research of Trojans based on the data stream communication brooks no delay.

At present, many scholars, at home or abroad, have made a lot of research and exploration about Trojan detection methods, mainly divided into two host-side and network-side [2, 3]. Among them, only the host-side software protection can't achieve effective detection for Trojan, and the Trojan detection method based on data flow analysis of network traffic gradually becomes a hot topic, attracting more and more attention. However, the existing work, mostly focus on the researches of a single feature[4-6], typically by establishing and maintaining a pre-defined characteristics database, using the database to match the characteristic information with the network data flow. If the match is successful, it gives an alarm. In addition, the comprehensive utilization of multi-feature also has made some achievements [7-9], but these studies are simply integrated multiple features without effective integration, resulting in the low rate of Trojan detection, and prone to false negatives and false positives, thus affecting the accuracy and validity of the detection system.

In order to further improve the accuracy of Trojan detection, on the basis of feature extraction, we propose a

combining multi-feature fusion Trojan detection method based on SVM and DS evidence theory. Using the DS evidence theory's advantage [10-11] of dealing with uncertainty information and the better classification capabilities of SVM in small sample, we combine multiple single features of Trojan detection information and get the final test results according to the decision rules. It's important to note that this paper is to study the Trojans of information stealing. After such Trojans are implanted into the target system, they will record or collect all kinds of important information of the target system, such as stealing all kinds of user names and Passwords, and send the information to the attacker through a particular way.

2 Feature extraction of Trojan

Due to the limit of firewall and NAT networks, the controlled terminal initiates a connection request to the control terminal in general Trojans. After that the connection is established, once the controlled terminal takes orders of the control terminal, they generally open a new session to execute, which will be ended when the session is finished. Figure 1 shows the timing sequence.

Based on the flow characteristics of the Trojans' communication, this paper detects Trojans by analysing the roles of both Trojans communication sides (controlled terminal and control terminal) playing respectively and the inconsistency of performance in the communication and the roles that they play. By analysing the communication theory of many Trojans and a great amount of communication features, we select three types of effective communication behavioural characteristics to

^{*}Corresponding author e-mail: liushengl2013@163.com

distinguish between normal network communication behaviours and Trojan communication behaviours.

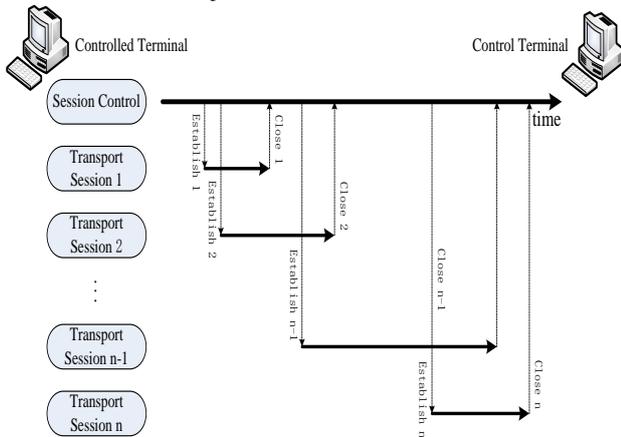


FIGURE 1 The timing sequence of Trojans communication session

2.1 SESSION FEATURE

There is significant difference between Trojans and normal HTTP C/S server. For Trojans, interactive commands, file resource searching and file transfer need artificial waiting and operating time. Therefore a longer time of communication is needed, while the normal HTTP C/S server performance to release the connection after completing the first request, so there is seldom a long connection. We use three parameters to describe the session features. Let T_1 be the number of TCP sessions, T_2 be control session duration and T_3 be IP flow duration.

- 1) The number of TCP sessions T_1 : when the IP connections launch the first TCP session, the value of T_1 is one, before the end of the session, the value of T_1 plus 1 when a TCP session is finished.
- 2) Control session duration T_2 : the control session duration is the duration of the first TCP sessions.
- 3) IP flow duration T_3 : the duration of the data communication between one pair of IP is from the first TCP session between the pair of IP to the end of all sessions.

TABLE 1 The characteristic of data flow

| session feature | | | upload data volume/download data volume features | distribution feature of packet size | |
|-----------------|--------|--------|--|-------------------------------------|-------|
| T_1 | T_2 | T_3 | D | S_1 | S_2 |
| 7.0 | 184.25 | 190.45 | 6.83 | 0.64 | 0.75 |

3 SVM and DS evidence theory

3.1 SVM

SVM (Support-Vector Machines) [12] is an important product of statistical learning theory. SVM is able to make experience risk minimization in a fixed structural

2.2 SESSION UPLOAD DATA VOLUME / DOWNLOAD DATA VOLUME FEATURES

Because of that the data flow of Trojan communication is the upload data flow from the inside out; we extract characteristic D of upload data volume / download data volume in the session. When the controlled terminal initiates the communication, it is presented as the client of connecting C/S model requested resources. After that the connection is established, it will turn into the service side to provide resources, in which cases the amount of data sent is much larger than the amount of data received. Thus this ratio is greater than one for Trojans, while in normal HTTP C/S server, the upload data is far less than the amount of download data, the ratio is less than one.

2.3 THE DISTRIBUTION FEATURE OF PACKET SIZE

For the characteristic of packet size distribution in the communication process, we extract the following characteristics:

- 1) The number of receiving small packets / the number of small packets S_1 , most of the packets received in the controlled terminal from the Trojans are the control commands, which are mainly small packets different from the big packets for sending information. So the value is generally greater than 0.5, and the normal HTTP C/S server is always the case that clients receive resource information from services, such as website information, most of which are big packets.
- 2) The number of big packets sent / the number of large packets S_2 , most of the packets sent in the controlled terminal from the Trojans are the resource files and host information, which rarely receiving a large package, so the value is generally greater than 0.5. The normal HTTP C/S server is the case that the client requests the resource, and only sending the request to the server, receiving server's resources, so the number of big packets sent is low.

To sum up, we extract six features from three types, which are session feature, upload data volume / download data volume features and distribution feature of packet size. Grey pigeons Trojan sample is used as an example, intercepting data flow of a period of time in these samples. Table 1 shows every characteristic of data flow.

risk, because the problem will be converted into a constrained quadratic programming problem in SVM ultimately, so SVM is usually able to get a global optimal solution. SVM is a method that by defining the optimal linear hyper plane and summarizing the search of optimal linear hyper plane algorithm as solving a convex programming problem, and using Mercer nuclear

expansion theorem. The sample space is mapped into a dimensional or even infinite dimensional feature space by nonlinear mapping function T, and making it possible to solve regression and highly nonlinear classification problems in the sample space by using linear learning machine method. The classification function as

$$f(x) = \sum_{x_i \in S_V} a_i y_i k(x, x_i) + b. \tag{1}$$

In the formula, a_i is Lagrange multiplier, b is a given threshold according to the training samples, S_V is support vector set and $k(x, x_i)$ is kernel function. Among them, only the kernel function satisfies Mercer conditions, then it corresponds to inner product of a certain space, and the nonlinear transformation can be changed by the inner product to linear transformation of high-dimensional space.

The process of solving Equation (1) is the training process of SVM. Therefore the results through the training and evaluation are as same as the classification function used in the actual detection.

3.2 DS EVIDENCE THEORY

The evidence theory established by Dempster and Shafer, which is called the D-S theory, is an important method of uncertainty reasoning. DS theory combines the trust function from two or more evidence bodies into a new trust function as a basis for decision making. The principle is as follows:

Let Θ be recognition framework. We define function $m: 2^\Theta \rightarrow [0,1]$, if there is $m: 2^\Theta \rightarrow [0,1]$ such that: $M(\phi) = 0$ (ϕ is empty set), $\sum m(A) = 1$ ($A \in 2^\Theta$), then we call $m(A)$ as that basic probability assigning (BPA) on the framework, which means the precise degree of confidence in the proposition, which is also known as the mass function. If $m(A) > 0$, A is focal elements.

Let m_1, m_2, \dots, m_n be BPA of different evidences on recognition framework, $m(A)$ can be expressed as:

$$m(A) = \sum_{A_1 \cap A_2 \cap \dots \cap A_i = A} (\prod_{1 \leq i \leq n} m_i(A_i)) / k. \tag{2}$$

K is the normalization constant, and

$$k = 1 - \sum_{A_1 \cap A_2 \cap \dots \cap A_i = \phi} (\prod_{1 \leq i \leq n} m_i(A_i)).$$

4 The decision level fusion Trojan detection algorithm

By analysing the characteristic information of Trojans in the detection, we can get that Trojan session feature, session upload data volume / download data volume feature and distribution feature of packet size are independent of each other, so you can take advantage of DS theory, which has the ability of composite the independent evidences to combine the identification information from the SVM of different characteristics. Finally, the target type is given by decision module (Trojans or normal HTTP C/S server). The Trojan detection algorithm model is shown in Figure 2.

The basic idea of detection algorithm is as follows:

1) Trojan Detection of Single Feature SVM. According to Trojan feature extraction method mentioned before, we extract three types of single-feature, Trojan session feature $T_1 T_2 T_3$, session upload data volume / download data volume feature D and distribution feature of packet size $S_1 S_2$. Then we detect three types of single-feature above with SVM classifier.

2) The Choice of SVM Training Algorithm Using SMO (Sequential Minimal Optimization) algorithm [13] to train the training set, it is because of that the influence from the choice of kernel function parameters is much bigger than from the kernel function itself on the classification results. So this paper selects Radial Basis function (RBF) as the kernel function.

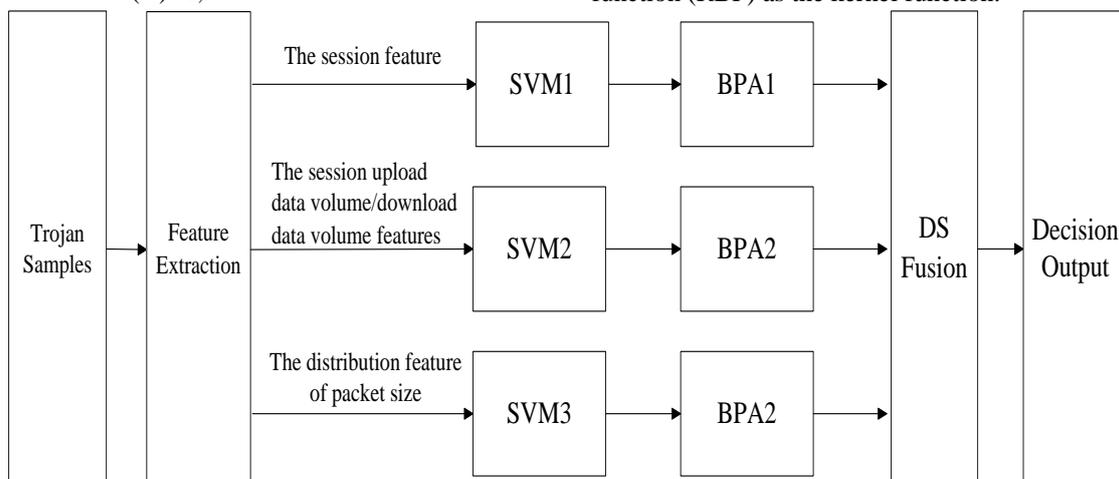


FIGURE 2 The Trojan detection algorithm model

In this method, given data sample (x_i, y_i) , $i = 1, 2, \dots, n$, kernel function $K(x_i, y_j)$ and adjustable parameters C , the necessary and sufficient conditions are Kuhn-Tucker (KKT) conditions that all training data samples should meet the following conditions:

$$a_i = 0 \Leftrightarrow y_i f(x_i) \geq 1, \tag{3}$$

$$0 < a_i < c \Leftrightarrow y_i f(x_i) = 1, \tag{4}$$

$$a_i = c \Leftrightarrow y_i f(x_i) \leq 1, \tag{5}$$

$f(x_i)$ is the output of data sample with number i . The diagram of KKT conditions is as follows

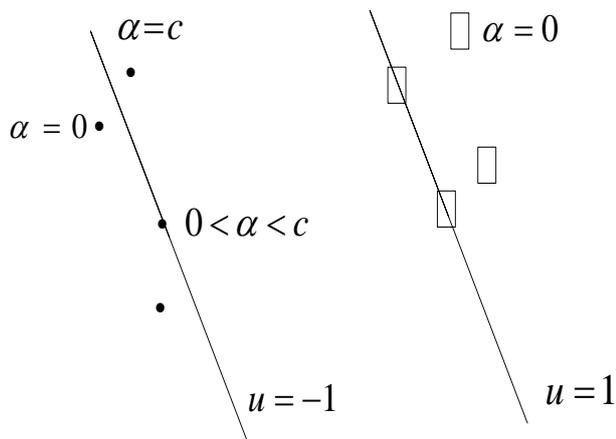


FIGURE 3 The schematic diagram of KKT conditions

The SMO algorithm breaks the problem down to the smallest size which is able to achieve, and each optimization only processes the optimization problem of two data samples. Its greatest advantage is that with analytical method for solving smallest optimization problems, avoiding the iterative algorithm and improving the overall computing speed. The algorithm does not require processing large matrix and no additional storage space is required. So this algorithm is suitable for processing the network data flow.

3) BPA Function Building. The standard output of SVM is $\{1, -1\}$, which is not a probabilistic output but a hard decision output, so we can't use it as BPA of evidence body. In order to construct the BPA of evidence theory. We build probabilistic modelling by combining many classification probabilistic results of two categories. First, assuming the given data of type 1, and for any x , we estimate the probability of matching class by sigmoid function [14], namely:

$$h_{ij} \approx p(y = i | y = i \text{ or } j, x). \tag{6}$$

For posterior probability p_i :

$$\min_p \left(\frac{1}{2} \sum_{i=1}^l \sum_{j:j \neq i} (h_{ij} p_i - h_{ij} p_j)^2 \right). \tag{7}$$

According to Equation (7), we can get the posterior probability after the training of the sample set. Then

testing the learning sample set, and you can get the accuracy of classification. So BPA function is defined as following:

$$m_j(A) = p_i q_i \tag{8}$$

4) The Decision Fusion and Judgment Rule. We can calculate the confidence of evidences by Equation (8), and the confidence of overall evidences by Equation (2).

Let $A_j (j = 0, 1)$ be sample type (data flow of Trojan or normal data flow), A_w is the target category (data flow of Trojan or normal data flow). The classification decision obeys the following rules:

1) $m(A_w) = \max\{m(A_j)\}$, the target class is the class with the greatest confidence.

2) $m(A_w) - m(A_j) > \varepsilon_1$ ($\varepsilon_1 > 0$), the difference value of target class and other class should be greater than a certain threshold.

3) $m(A_w) - m(\theta) > \varepsilon_2$ ($\varepsilon_2 > 0$), the confidence of target class should be greater than the assign value of uncertain reliability.

4) $m(\theta) < \varepsilon_3$ ($\varepsilon_3 > 0$), the uncertain reliability of target class's evidence shouldn't be too large, and the assign value of uncertain reliability should be less than a certain threshold.

5 The prototype of detection system

In this paper, in order to verify the efficiency of Trojan communication detection method, a rapid Trojan detection system has been designed. The system consists of data acquisition module, feature extraction module, Trojan detection module and alarm response module. The structure is shown in Figure 4:

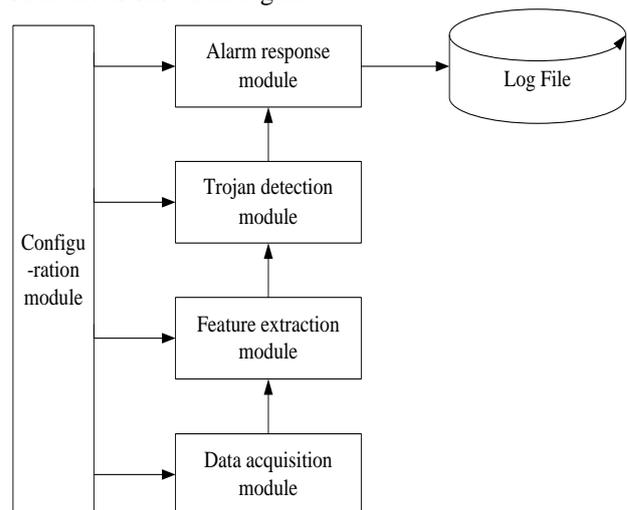


FIGURE 4 The structure of detection system

In this system, we use Winpcap to get data in data acquisition module. Feature extraction module is responsible for extracting the Trojans communication session, the session upload data volume / download data volume and packet size distribution. Trojan detection

module detects Trojan detection in real time based on Fusion Trojan detection algorithm of decision-making level, and alarm response module extract the appropriate information recorded in the log file. Configuration module is used to configure other modules.

6 Trojan detection experiments

6.1 EXPERIMENT DESCRIPTION

The experiment is conducted in a small laboratory environment. Shown in Figure 5, there are 20 hosts in the LAN access the internet and the host network bandwidth is 3Mbps. Among them, we implanted Trojans in host PC1, and build DNS in host PC5. The program containing the above-mentioned Trojan detection algorithm has been installed in a separate server host.

The operating environment:

- CPU: Xeon E5 2.3G
- Memory: 16G

In this paper, based on grey pigeons Trojan (Win32.Hack.Huigezi), we designed three types of Trojans with high concealment by increasing redundancy, dividing sending and discontinuous transmission to conceal their communication data flow features, which

are called Wake1.0, Wake2.0 and Wake3.0 respectively. Wake1.0 masks itself by increasing redundancy, Wake2.0 masks itself by dividing sending, and Wake3.0 masks itself by discontinuous transmission. The initial data sample consists of 410 normal network sessions and 145 Trojan sessions. Five groups of experiments were performed. Each of them randomly selected 90% of the initial data sample as the SVM training sample, and selected alternatively 10% of the data sample as the test sample. Then the test samples were divided into 10 test sets according to the type of data flow, and the data type of each test set was Trojan data flow or normal data flow.

First, we extracted six features from three types by using the method mentioned, which are session feature, upload data volume / download data volume feature and distribution feature of packet size, and normalized to them. Then we classified the data flow by multi-feature fusion method. In the method, we selected RBF as the kernel function of SVM model and used cross-validation method to select the error penalty parameter C and kernel parameter σ : $C = 40$, $\sigma = 2.43$. We obtained judgment threshold from a lot of experiments: $\varepsilon_1 = 0.6$, $\varepsilon_2 = 0.7$, $\theta = 0.1$.

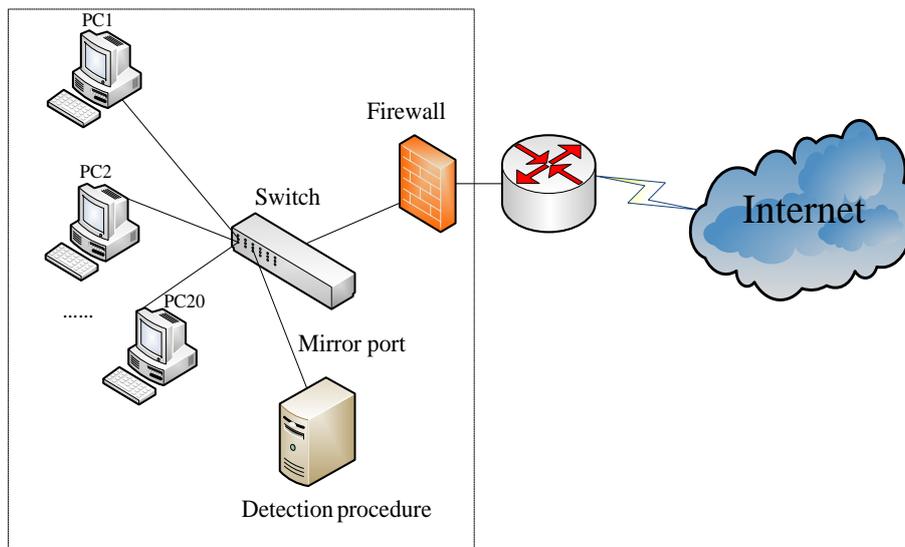


FIGURE 5 The experimental topological diagram

TABLE 2 The related records of three groups' experiments

| The types of Samples | Single feature | Confidence function value | | | Detection results |
|----------------------|----------------|---------------------------|----------|-------------|-------------------|
| | | $m(A_0)$ | $m(A_0)$ | $m(\theta)$ | |
| 2-3 Wake1.0 | m_1 | 0.5733 | 0.3125 | 0.1142 | Indefinite |
| | m_2 | 0.6673 | 0.2134 | 0.1193 | Indefinite |
| | m_3 | 0.6014 | 0.1756 | 0.2230 | Indefinite |
| | m | 0.8242 | 0.1405 | 0.0353 | Wake1.0 |
| 4-6 Normal data flow | m_1 | 0.1012 | 0.7914 | 0.1074 | Indefinite |
| | m_2 | 0.2345 | 0.5678 | 0.1977 | Indefinite |
| | m_3 | 0.1947 | 0.6235 | 0.1818 | Indefinite |
| | m | 0.0754 | 0.8997 | 0.0249 | Normal data flow |
| 5-7 Wake3.0 | m_1 | 0.5678 | 0.2565 | 0.1757 | Indefinite |
| | m_2 | 0.7912 | 0.1345 | 0.0743 | Indefinite |
| | m_3 | 0.6322 | 0.2379 | 0.1299 | Indefinite |
| | m | 0.8943 | 0.0924 | 0.0133 | Wake3.0 |

6.2 ANALYSIS AND COMPARISON

1) We recorded the value of confidence function and recognition results of single-feature and fusion multi-feature in the experiments. Table 2 shows the related records of three groups' experiments randomly selected. In which, m_1 is the session feature, m_2 is upload data volume / download data volume, m_3 is distribution feature of packet size and m is the fusion result. For example, No.2-3 means the data flow of the third test set in group 2.

Table 2 shows the experimental results:

a) Comparing the value of confidence function combined by session, upload data of session / download data of session, and distribution of data packet size with

the value of single-feature confidence function, the confidence of actual target increases a lot, which means the uncertainty of the detection is greatly reduced.

b) Three types of data samples which are unable to detect by multi-feature SVM (such as 2-3 and 5-7 in the table), can be accurately detected after the data fusion, so the multi-feature fusion method based on D-S evidence theory has good performance on accuracy and stability, which enhances the ability to detect Trojans.

2) Using the test set in the test samples as a unit, we recorded the detection rate and the multi-feature detection rate of five groups' experiments. The results are shown in Table 3. The detection rate means the percentage of correct classification in data samples, and m_1, m_2, m_3 has the same meaning of Table 2.

TABLE 3 The experimental results

| Experimental group | The detection rate of single feature | | | The fusion recognition rate of multiple features |
|--------------------|--------------------------------------|--------|--------|--|
| | m_1 | m_2 | m_3 | |
| 1 | 60.25% | 80.15% | 65.34% | 97.28% |
| 2 | 87.23% | 92.34% | 89.23% | 96.96% |
| 3 | 70.34% | 85.00% | 80.00% | 96.44% |
| 4 | 70.46% | 90.25% | 78.05% | 100% |
| 5 | 62.25% | 81.25% | 68.75% | 96.78% |
| Average | 70.11% | 85.80% | 76.27% | 97.48% |

Table 3 shows the experimental results:

1) Comparing three types of single-feature detections, because of that the great difference between Trojan data flow and normal data flow in upload data volume and download data volume, the method based on upload data of session / download data of session performs better than the method based on session and distribution of data packet size. On the other hand, the test data which is complex in fact results in the inaccurate calculation to reduce detection rate. In summary, the single-feature detection has a high error rate, poor reliability and stability.

2) The average accuracy of multi-feature fusion detection reaches 97.48% and has less volatile. Comparing with the single-feature detection, the accuracy and stability has improved significantly. Because that the DS evidence theory is based on SVM's posterior probability and credits assigned of classification accuracy structure, it combines the different feature detection information with session, upload data of session / download data of session, and distribution of data packet size, and makes full use of multi-feature information to improve the accuracy and stability of Trojan detection.

Additional, we use BotHunter detection software [15] to test the samples above. BotHunter is a kind of driver-based IDS detection software which can detect the behaviors of network scanning and file downloading during the process of bot software infection. It matches with data flow feature and gives an alert when the

infected host is highly unusual. Some of the experimental results are shown in Table 4.

TABLE 4 The detection result of BotHunter

| Experimental group | The types of Samples | The detection rate |
|--------------------|----------------------|--------------------|
| 2 | Wake1.0 | 88.3% |
| 3 | Wake2.0Wake1.0 | 90.6% |
| 5 | Wake3.0 | 85.2% |

We can get the conclusion that the detection rate of the method proposed in this paper is higher than the detection rate of BotHunter software, which has a better practical significance.

7 Conclusions

In order to improve the detection rate of Trojans, we propose a multi-feature fusion method based on SVM and DS evidence theory. First, we use three types of SVM classifiers respectively based on session, upload data of session / download data of session, distribution of data packet size for detecting Trojans. Then the SVM classification results of each single feature are used as evidences to construct the basic probability assigned (BPA). Finally, we make use of DS evidence theory to achieve the decision fusion and give final detection results by fusion results. The experimental results showed that the accuracy of multi-feature fusion method was 97.48% in the real network environment, which has good performance on accuracy and stability compared with the single feature method in Trojan detection.

References

[1] Wang Y Z, Jin X L, Cheng X Q 2013 Network Big Data Present and Future *Journal of Computers* **36** 1-15

[2] Liu T, Guan X, Zheng Q, Lu K, Song Y, Zhang W 2009 Prototype Demonstration Trojan Detection and Defense System *Proc of the IEEE Consumer Communications and Networking Conference* 1-2

[3] Zhang L, White G B 2007 An Approach to Detect Executable Content for Anomaly Based Network Intrusion Detection. *Proc of the IEEE Parallel and Distributed Processing Symposium* 1-8

[4] Dusi M, Croti M, Gringoli F, Salgarelli L 2009 Tunnel Hunter: Detecting application-layer tunnels with statistical Finger printing *Computer Network: The International Journal of Computer and Telecommunications Networking Archive* **53** 81-97

[5] Perdiscia R, Leea W, Feamster N 2010 Behavioral Clustering of HTTP-Based Malware and Signature Generation Using Malicious Network Traces *Proc of the 7th USENIX conference on Networked systems design and implementation* 391-404

[6] Crotti M, Dusi M, Gringoli F, Salgarelli L 2007 Detecting Http Tunnels with Statistical Mechanisms *Proc of the IEEE International Conference on Communications* 6162-8

[7] Roesch M 1999 Snort Lightweight Intrusion Detection for Networks *Proc of the LISA 99 System Administration Conference* 229-38

[8] Borders K, Prakash A 2004 Web Tap Detecting Covert Web Traffic *Proc of the 11th ACM Conf Computer and Communications Security* 110-20

[9] Pack D 2002 Detecting HTTP Tunnelling Activities *Proc of the IEEE Workshop on Information Assurance and Security* 1-8

[10] Xh X, Zl L 2005 A method for situation assessment based on D-S evidence theory *Electronics Optics & Control* **12** 36-7

[11] Li Y, Cai Y Z, Yin R P 2008 Support Vector Machine Ensemble Based on Evidence Theory for Multi-Class Classification *Journal of Computer Research and Development* **45** 571-8

[12] Bai P, Zhang X B, Zhang B 2008 Support Vector Machine and Its Application in Mixed Gas Infrared Spectrum Analysis *Xi'an Xidian university press*

[13] Platt J C 1998 Fast training of support vector machines using sequential minimal optimization *Advances in Kernel Methods Support Vector Machines Cambridge MA MIT Press*

[14] Platt J C 1999 Probabilistic output for support vector machine and comparisons to regularized likelihood methods *Advances in Large Margin Classifiers Cambridge MA MIT Press*

[15] Gu Guofei, Porras P, Yegneswaran V, Fong M 2007 Bot Hunter Detection malware infection through IDS-driven dialog correlation *Proc of the 16th USENIX Security Symposium on USENIX Security Symposium* 1-16

| Authors | |
|---|---|
|  | <p>Sheng-li Liu, born in 1973</p> <p>Current position, grades: professor of State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou. Scientific interest: applied mathematics and information security. Publications: more than 18 papers published in various journals. Experience: teaching experience of 15 years.</p> |
|  | <p>Xiang Gao, born in 1984</p> <p>Current position, grades: doctoral student of State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou. Scientific interest: information security, artificial intelligence. Publications: more than 10 papers published in various journals. Experience: researching experience of 7 years.</p> |
|  | <p>Pan Xu, born in 1988</p> <p>Current position, grades: master student of State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou. Scientific interest: information security and data mining. Publications: more than 3 papers published in various journals. Experience: researching experience of 3 years.</p> |
|  | <p>Long Liu, born in 1983</p> <p>Current position, grades: doctoral student of State Key Laboratory of Mathematical Engineering and Advanced Computing, Zhengzhou. Scientific interest: applied mathematics and artificial intelligence. Publications: more than 8 papers published in various journals. Experience: researching experience of 7 years.</p> |