Research on adaptive network theft Trojan detection model

Ting Wu

Guangdong Teachers College of Foreign Language and Arts, Guang Zhou, 510640, China

Keywords: Trojan, Communication behavior, SVM, Incremental learning, Active probing.

Abstract. This paper intends to establish theft Trojans detection system capable of adaptive dynamic feedback learning. To achieve this goal, this paper first studies the characteristics of the network data stream and theft Trojans communication data stream, then introduces support vector machine algorithm based on incremental learning, proposes the construction method of incremental learning samples, designs and implements the theft Trojan detection system with self-adaptive capability. Finally, the experimental results in a real network environment verify the effectiveness and efficiency of the system.

Introduction

With the rapid development of information technology, Internet technology has been applied widely, the Trojan horse technology as the main means of the attacker to steal secret information [1], which has been a serious threat to the network security.

At present, according to the two kinds of detection technique are: communications of Trojan horse detection based on feature matching and detection based on communication behavior analysis. Detection methods based on feature matching, as a part of IDS, has the serious lag, cannot detect the unknown trojan, can not adapt to the rapid changes in the network environment.

Communication behavior analysis of Trojan Detection Based on statistical features of communication data stream distinguish Trojan horse communication and network communication based on normal, can detect unknown trojan, has good versatility, has begun to attract attention. This kind of thought initially proposed by Yin Zhang and Vern Paxson [2], they use the data packet size and inter arrival time feature to detect theft Trojan behavior, but they are for different protocols SSH, FTP algorithm used to detect different, not with the common good. 09 years M. Dusi[3] et al. Proposed a method of statistical classification, statistic features extraction layer communication data of the IP detection of HTTP and SSH tunnel in the network gateway, this method is not only effective and it can detect encrypted tunnel. But with the increasing of network applications, to detect HTTP tunnel will produce a lot of false positives only through the packet size and time interval, and the system does not have the ability of adaptive learning, can not adapt to the complex network environment. Colin Pattinson[4] et al. flow characteristics are obtained with the IDS/IPS intrusion detection system to detect Trojans, combined with the host port information, their method saves a lot of resources, make full use of the existing equipment, but the method must establish a trusted port base, construct the detection result still rely on trusted library, has lagged behind.

Based on the research of Trojan horse communication behavior and the existing detection technology, proposed a classification model theft Trojan using incremental SVM algorithm, and the detection results of heuristic analysis processing, based on feedback learning models of alarm information online incremental learning and establish adaptive.

Based on support vector machine incremental learning algorithm

Establishment of detection system with adaptive ability, must get the alarm information and new samples to learn, and learn the results of feedback to the detection model. Therefore in the adaptive detection system, machine learning is an important part of the autonomous. Network data flow, fast transmission speed, with quick arrival and huge scale characteristics, traditional data mining algorithms can not play the role of an ideal. Incremental learning algorithm of [9] according to the

new arrival data stream samples for incremental learning, dynamic adjustment of the original classification model, has the following two significant superiority:

One is the incremental learning without reservation to all the historical data, reduce the data storage space; one is the incremental learning based on the historical training, without the need to re learn all the historical data, significantly reduces the training time. In view of the network data stream change, rapid arrival, huge scale characteristics, incremental learning is more suitable for learning and classification of network data flow.

Support vector machine algorithm: Support vector machine (Support-Vector Machines, SVM) [3-4] statistical learning theory and structural risk minimization principle based on, both the training error and the generalization ability, so that SVM has good generalization can. The basic idea of support vector machine is first to divide the input space to a high dimensional feature space by kernel function to generate the nonlinear mapping, and then in the feature space to obtain the optimal classification sample linear separate surface. The classification algorithm using interval control linear learning machine capacity, thus makes the structure risk minimum, also make it has stronger generalization ability in finite samples. The main mechanism is that KKT conditions, and KKT conditions is also performed incremental learning basis.

Given data samples $(x_i, y_i)\ i = 1, 2, \cdots n$ kernel function, and adjusting the parameter C, Kuhn-Tucker (KKT) condition is the most advantages should be sufficient and necessary condition of full, that all the training data samples shall meet:

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

$$0 < a_i < c \Leftrightarrow y_i f(x_i) = 1 \tag{2}$$

$$a_i = c \Leftrightarrow y_i f(x_i) \leq 1 \tag{3}$$

Among them, with respect to output the data samples for the SVM, KKT condition diagram as follows:

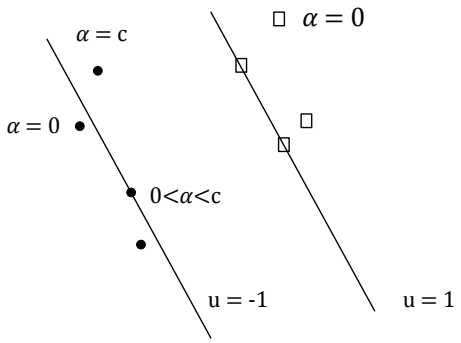

Fig. 1 Schematic diagram of the KKT condition

SMO algorithm is the decomposition of the problem to the minimum size may achieve, each optimization problem dealt with only two data samples. Its greatest advantage is that by the analytical method for solving optimization problems of minimum size of each one, avoid the iterative algorithm, the computation speed improves the total, and the algorithm does not need to handle large matrices, no additional requirements on the storage space, so this algorithm is suitable for processing network data flow. Based on this algorithm, in order to realize the adaptive dynamic feedback mechanism, this paper adopts the adaptive learning algorithm based on SMO algorithm, to join the forgetting mechanism in the algorithm, avoiding the infinite increase of sample storage space the number of samples tends to infinity.

Based on SMO incremental learning algorithm: The classification error is a specific sample data in violation of the KKT condition, so the SMO incremental learning algorithm based on KKT [2], will determine whether the conditions as new data samples to update existing SVM classification function is based on the new data samples, including changing the SVM classification function and new data like this does not change the two processes SVM classification function. When

determining the type of kernel function and its parameters, the characteristics of support vector classification can completely describe the whole data set, and the set of support vectors is only part of the data set. Therefore, if the classification information of new data samples containing raw sample data set does not contain, incremental learning support vector will change after. The whole process of the adaptive learning method only test whether a violation of the KKT condition for all data samples, the need for new data sample to study alone, get the new SVM classification function. If the SVM classification function $f(x_i)$, $y_i \in \{+1,-1\}$ is category, then $y_i f(x_i)$ are contrary to the KKT condition. The basic idea of [3] algorithm is as follows:

Input: the initial classification sample set x°, the initial SVM classification function $f(x)$, incremental samples x^{i+1}, support vector set SV;

Output: the set of support vectors

The algorithm process are:

(1)Inspect the sample of x^{i+1} whether accord to the KKT condition of SVM classification function $f(x)$.

If (accord) Stop; $f(x)$ Are the results, the set of support vectors does not change?

(2) by violating KKT conditions of data training sample set x_o^{i+1} to get a new classification function $f'(x)$ and support vector set;

(3) inspect the sample of x° whether accord to the KKT condition of the new SVM classification function $f(x)$.

If (accord) Stop; $f'(x)$ Are learning results, Support vector set sv'.

Else. Set x° into against the classification function new KKT conditions of the sample set x_e^{o} and meet the new classification function condition of KKT samples x_o^{o};

(4) Support vector set is obtained as the result of incremental learning.

(5) The algorithm terminates. Support vector machine classification algorithm complexity by the number of training set size L, a description of the sample vector dimension D and the number n_{sv} to support vector machines influence, its computational complexity is $O(n_{sv}^3 + Ln_{sv}^3 + DLn_{sv})$. Set the size of the sample set L' for incremental learning, incremental learning to get the number of support vectors n_{sv}, $n_{sv'} / n_{sv} \approx 1$, so the complexity of incremental learning are $O(n_{sv}^3 + DLn_{sv}^2)$, complexity is far less than the new learning, support vector machine classification algorithm for incremental learning and is more suitable to process network data real-time arrival flow.

The overall structure and design:Theft Trojan detection system in real time from the switch mirror port collecting network data flow, through credible whitelist filters reduce test data volume. Construction of 4 layer list according to the identification information source IP, destination IP {, source port, destination port, protocol} flow data, establish a session information, according to the session information pack update data acquisition.

Adaptive feedback detection system theft Trojan has three main modules: data acquisition and processing module, detection module, alarm information processing and incremental learning module. The data acquisition and processing module through the establishment of trusted IP white list, using the Winpcap packet filtering mechanism to capture the underlying technology, filter the data flow, reduce the burden of the rear detection. Trojan horse detection in real time according to the characteristics of the collected data to construct a data list from each session, five tuple information and alarm information will be the session session character appears as a new sample.

Sample labeling of incremental learning:Results the initial alarm detection model is not 100% sure, will generate false positives of this kind of noise samples. Methods using artificial markers to label all new samples is significantly enhanced the labeled accuracy, but the marking speed slow, long time, not easy to realize intelligent operation system, so this paper according to the characteristics of theft Trojan control terminal monitor common service ports, puts forward the method of active detection alarm further recognition to determine sample. Most of the existing Trojan using HTTP tunnel technology of communication, this paper studies to detect Trojan horse based HTTP tunnel. The control end of Trojan horse program opens the way to a 80 port monitor, does not establish a real Web server to respond to a HTTP request, it only receives information sent by the controlled terminal. So the normal web server sends an HTTP request detection, will receive

about server information; to Trojan control terminal HTTP request detection, will not receive the server information. Construction of HTTP HEAD or GET request packet, detection control terminal is a normal Web server.

According to the detection result, read the results returned by server configuration information {the server type, content type, content length, modification time} can be inferred information such as information control end of the alarm information. In addition, in determining the incremental learning sample set, through the connection information manual verification controlled host, rules of alarm information of comparison of the number of days can be further verified for tagged data.

Test and analysis of experiment

The theft Trojan detection system deployed in LAN switches mirroring port, to detect data flow collection land. The experimental environment of network bandwidth of 100Mb/s, testing machine 32 Taiwan, of which two host for the implantation of Trojan horse. The initial data of samples from 412 normal network session and 147 session consisting of a Trojan horse detection were carried out according to the following steps:

(1) 90% initial selected sample data as training set, the data of 10% samples as a test set;

(2) the use of the training set using the original SMO support vector machine algorithm to obtain the classifier is trained using the test set for testing to obtain the initial classifier;

(3) the classifier as the detection module to the real network data flow were detected in two sets of Trojan horse is different from the training set samples implanted in the Trojan and Trojan variant;

(4) When the alarm information to achieve the initial sample set 20%, use the method of active detection of incremental learning samples for verification, use the validation of samples for incremental learning to update classifier.

The basic research in the above, this paper implements the SMO incremental learning algorithm and detect the real network data flow(The operating environment for the Matlab7.0/VS 2008/Windows 2003 server, CPU: i5, frequency: 3.2GHz, memory 4GB). Through different sets of incremental data structure, this paper designed two experiments: one is the accuracy of detection results by the new samples are added to a different number of comparison between adaptive and non-adaptive detection system, is a more incremental learning and re learning efficiency. The experimental results such as the following two tables.

In the experiment, this paper constructs the incremental data set for the original alarm information, including false positives and noise must, on 432 alarm information using the method of active probing further marker incremental learning samples, the elimination of 76 false information, false information in accounting for the total 95%, verified the effectiveness of the active detection alarm information for verification.

Table 1 compare with adaptive and non-adaptive detection accuracy

Data set	quantity	Adaptive		non adaptive	
		rate of false alarm	rate of missing report	rate of false alarm	rate of missing report
Initial Data set	569	12.7%	3.6%	12.7%	3.6%
Incremental set 1	235	8.3%	1.2%	14.1%	4.7%
Incremental set 2	128	5.7%	1%	14.3%	4.7%
Incremental set 3	69	5.2%	0.9%	14.8%	4.9

Table 2 efficiency incremental learning and repetitive learning algorithm

Data set	quantity	Incremental learning		repetitive learning	
		Learning time (s)	detection time (s)	Learning time (s)	detection time (s)
Initial Data set	569	103.2	96	103.2	96
Incremental set 1	235	69.8	95	203.6	97.2
Incremental set 2	128	78.4	96.7	289.7	95.4
Incremental set 3	69	56.8	94.3	367.2	96.8

Summary and Prospect

The experimental study shows that, in this paper we introduce the incremental learning algorithm was used to detect the theft Trojan communication behavior, to overcome the traditional static model cannot adaptive shortcomings, and repetitive learning cost too big shortcoming, is to adapt to the method of fast, large amount of network data flow testing. Based on the incremental SMO learning algorithm, this paper designed and implemented a theft Trojan detection system adaptive, able to adapt to changes in network data flow, dynamic adjustment of the detection model, real-time detection of the high theft Trojan.

References

[1] ZHANG YIN, VERN P. Detecting backdoors: Proc. 9th USENIX Security Symposium, 2000[C].Las Vegas: Use nix. 2000: p. 543-554.

[2] COLIN P., KEMAL H... Trojan detection using MIB-based IDS/ IPS system: Information, Communication and Automation Technologies, 2009 [C]. Singapore: IEEE conference publication. 2009: p. 1-5.

[3] MOFFIE M., KAELI D... Hunting Trojan horses [R]. NUCAR Technical Report TR-01, 2006.

[4] Spigot-Search and destroy [EB\OL]. (2011-09-15) [2012-02-05]. http: //www.safer-networking.org/.