Research on Active Defence Technology with Host Intrusion Based on K-Nearest Neighbor Algorithm of Kernel

Xuedou Yu
Department of Computer Science and Technology. Dezhou University
Dezhou Shandong, China
E-mail: yuxuedou@163.com

Abstract—**In this paper, the intrusion detection system is introduced, and point out two important prerequisite that the IDS work normally must depend on, and in view of the prerequisite, the paper proposed a solution which is based on k-nearest neighbor algorithm of kernel and the active defence technology anti-host intrusion. The solution can distinguish normal event from the unknown event efficiently, and can detect unknown event more accurately.**

Key words- Intrusion Detection System, Active Defence, Kernel Function, K-Nearest Neighbor Algorithm

1. Introduction

Operating system, application software and hardware equipment, inevitably exist some security vulnerabilities, which the computer viruses and attackers intrude a computer or system by non-normal means.

The network security technology developed unceasingly in the resistance to the network attack's and experienced the development process from static to dynamic, from passive defense to active defense , from the local defense to the overall defense.

Intrusion detection provide a method of finding attack and abuse of authority legitimate users by monitoring and analyzing the status and activity of protected system, which is based on an important premise: The intrusion behavior and the legal act are may differentiate, that is to say, may judge the nature of this behavior through the extraction behavior pattern characteristic. An intrusion detection system needs to solve two problems [1]: At first, how to extract the data of behavioral characteristics entirely and reliably; the second is how to determine the nature of the act by the characteristic data efficiently and accurately.

The purpose of intrusion detection is to identify internal and external conduct which is exceed its authority, misuse and abuse, and at the same time to protect the legitimate users to make use of the resources of system efficiently.

Host-based intrusion detection system is installed in the host need to focus on detection, monitors and analyzes the host audit records. If it is found that the activities of objects are very suspicious, intrusion detection system will take corresponding measures. Host Intrusion Detection System commonly used in the analysis of "possible attacks", can provide more detailed information on the evidence that the intruders tried to implement a "dangerous order", to distinguish the specific acts of the intruders.

In this paper, a new way was present to distinguish between normal events and intrusion events more accurate. This method combines the positive features of both active defence technology and the k-nearest neighbor algorithm of kernel.

2. Overall system structure

DARPA proposed the Common Intrusion Detection Framework, which is showed in Figure 1[1].

In order to facilitate the design, the framework of active defence system against host intrusion is showed in Figure 2. The system consists of three sets, which are “normal event set” N, ”intrusion event set” I and “suspicious event set” S. The normal event signatures are stored in N, the known intrusion event signatures are stored in I, the suspicious event signatures are stored in S, and U represents full set, and the relationship between the various sets is:

$$U = N \bigcup S \bigcup I$$

$$N \bigcap I = \emptyset$$

$$\mathrm{N} \bigcap \mathrm{S} = \emptyset$$

$$\mathrm{S} \bigcap \mathrm{I} = \emptyset$$

The information of the three sets can be obtained through the following method. The flow chart of active defence anti-host intrusion system is showed in Figure 3[10].

2.1 The establishment of “normal event set”

After finishing setting up a small local area network in the no-virus and no-malicious procedure environment, we pick-up the signatures of information which generated when the computers exchange information, and encrypt through MD5 then save it to “normal event set”.

2.2 The establishment of “intrusion event set”

We pick-up the signatures of the known virus and malicious procedure by signature extraction software, and encrypt through MD5 then save it to “intrusion event set”.

2.3 The establishment of “suspicious event set”

The monitor process intercepts and captures all events from network and suspends it for a while. Then pick-up the signatures of the event by signature extraction software and encrypt it through MD5, then compare it with the background “normal event set”, if the running event is a legal event, then let it runs, if not so compare its signature

978-0-7695-3744-3/09 $25.00 © 2009 IEEE
DOI 10.1109/IAS.2009.255

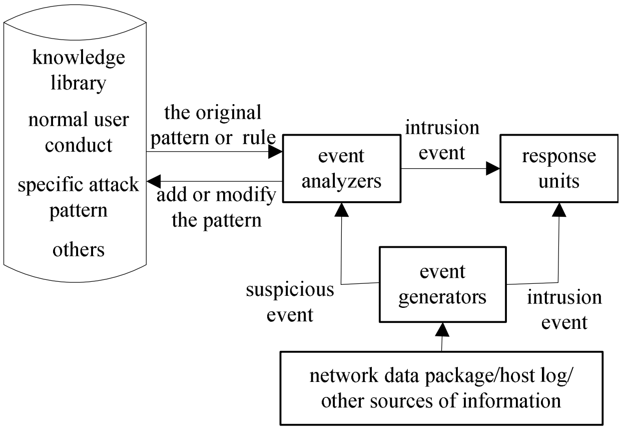

Figure 1 the common intrusion detection framework

with the signatures in "intrusion event set", if it is harmful, then kill it, otherwise suspend the process and save it to "suspicious event set".

2.4 How to deal with the suspicious process

Whether to allow the event whose signature was saved to "suspicious event set" to run depends on the detection by K-nearest neighbor algorithm of kernel.

2.4.1 Extract the data of behavioral characteristics

After the analysis of the virus and Trojans horse, we can sum up several common characteristics as follow.

(1) The automatic access to mobile storage media. It is an important way that Trojan horse steals information by automatically accessing to mobile storage media, so the process which can access to mobile storage media automatically is commonly a Trojan horse.

(2) The leaking secrets operation. In order to steal information from users, Trojan horse will collect users' information automatically. This is an important feature of Trojan horse.

(3) The suspicious jump structure. If the jump address is the address of core DLL, commonly we consider it as a virus or Trojan horse.

(4) The destructive operation. The virus has destructive operation generally, such as erasing or modifying certain files or even formatting the disk, so if we detect a process performing destructive operations, we can determine it is a harmful.

(5) Abnormal file access. When the harmful process infects other executable files, it is generally to traverse all the executable files, which is not the characteristic of common files, and if finding a process with the kind of characteristic, we can recognize it is a suspicious process.

(6) Self-replication operation. If a process has the self-replication code sequence, we should suspect it is virus.

(7) Abnormal stack operation. If there is a character array in local variable, and if we assign a long enough character string to the array, the address returned before will be covered. In this way, the buffer overflow happened. So if there is someone submits overlong data, we can list it as a suspected target.

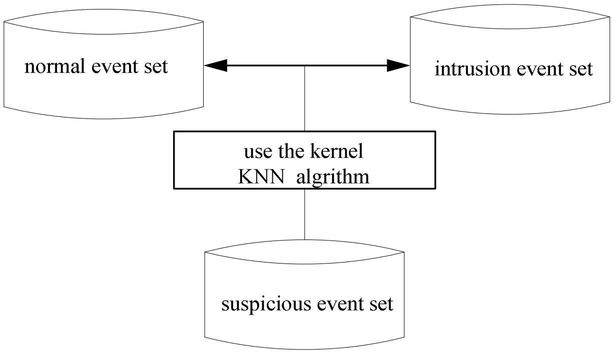

Figure 2 the framework of active defence system

(8) Abnormal memory operation. When the virus infects the common files and runs, it will erase or move or replace the memory, and the operation is the unique to the virus generally.

(9) The call of private API function. In order to reduce its volume, virus or Trojan horse often calls a large number of API functions, including private API function. The process with this feature can be classified as suspicious ones.

(10) Presence in memory. The virus must be presented in memory for a long-term in order to achieve the purpose of infection and spreading. So a process which automatically resides in memory for a long-term can be considered as a harmful process.

2.4.2 The k-nearest neighbor algorithm

The k-nearest neighbor algorithm is an analogy-based study method. The key technique of k-nearest neighbor algorithm is searching N-dimensional space to find out k training samples which are the nearest to the unknown sample, and the unknown sample is assigned to the public class of the k training samples which are the nearest neighbors and its neighbors are defined by Euclidean distance. The nearest neighbor algorithm is based on such assumption that the classification of an example is similar to the examples in the vicinity of the Euclidean space.

Based on the definition and analysis of suspicious operation sequence above, we can obtain the corresponding code feature and eigenvector of the process by scanning suspicious code in the corresponding instruction sequences operations, the definition of the eigenvector is as follow:

$$x=(t_1(x),t_2(x),\cdots,t_{10}(x)),t_i(x)\in\{0,1\},1\le i\le 10$$

Then, the distance of two arbitrary examples is defined as follow:

$$d(x_i,x_j)=\sqrt{\sum_{k=1}^{10}[w_k(t_k(x_i)-t_k(x_j))]^2}\ ,$$

$$x=(t_1(x),t_2(x),\cdots,t_{10}(x)) \qquad (1)$$

Because of every attribute has different security level and different suspicious level, and in order to make $d(x_i,x_j)$ more realistic reflection of the distance between the two examples, we can add different weighted value w_j

to different dimension $t_i(x)$ according to the characteristics that the virus use and may have. Then the formula (1) can be modified as follow:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{10} w_k (t_k(x_i) - t_k(x_j))^2} \qquad (2)$$

2.4.3 K-NN algorithm based on kernel methods

The K-NN algorithm based on kernel methods is simple, and has a broader applicability than the original K-NN algorithm, but the classification accuracy of this algorithm will be greatly reduced for some more complex issues, in particular the distribution of training samples is very irregular. The kernel methods can make the data map from low-dimensional space which is hard to divide to high-dimensional feature space which can be divided more easily through non-linear transformation, and does not need explicit non-linear transformation. The only need is that calculating the inner product only using the kernel function --the vector of the original features of the new features of the space to ensure that the conversion of space without increasing the complexity of computing, so kernel in the field of pattern recognition applications have been paid more and more attention to.

We can choose all kinds of kernel methods, but in theory, the polynomial approximation can meet any characteristics of the space transform in the appropriate parameters P. In order to enable the experimental result to have the generality and the widespread compatibility, the paper uses polynomial kernel. The kernel, such as the definition of formula (3) as shown.

$$K_p(x, y) = (1 + < x, y >)^p \qquad (3)$$

In order to solve the problems facing K-NN based on traditional Euclidean distance, the literature [8] gives a K-NN algorithm based on kernel methods. This method uses the kernel distance as the distance of samples to be classified and the samples of training set. Experiments show that the algorithm is more effective than the traditional classification of K-NN. However, a major drawback of the algorithm is that the kernel parameter P is difficult to determine, mainly depend on one's experience to determine. By optimizing the kernel parameters, we can access to the best of the space conversion, to obtain the best classification results.

The kernel distance defined as follows:

$$KD(x, y) = K_p(x, x) - 2K_p(x, y) + K_p(y, y) \qquad (4)$$

2.4.4 Experiment research

In order to verify the validity of K-NN algorithm based on kernel methods, in LAN based on Windows2000 Server environment, we selected 100 normal event of system as the

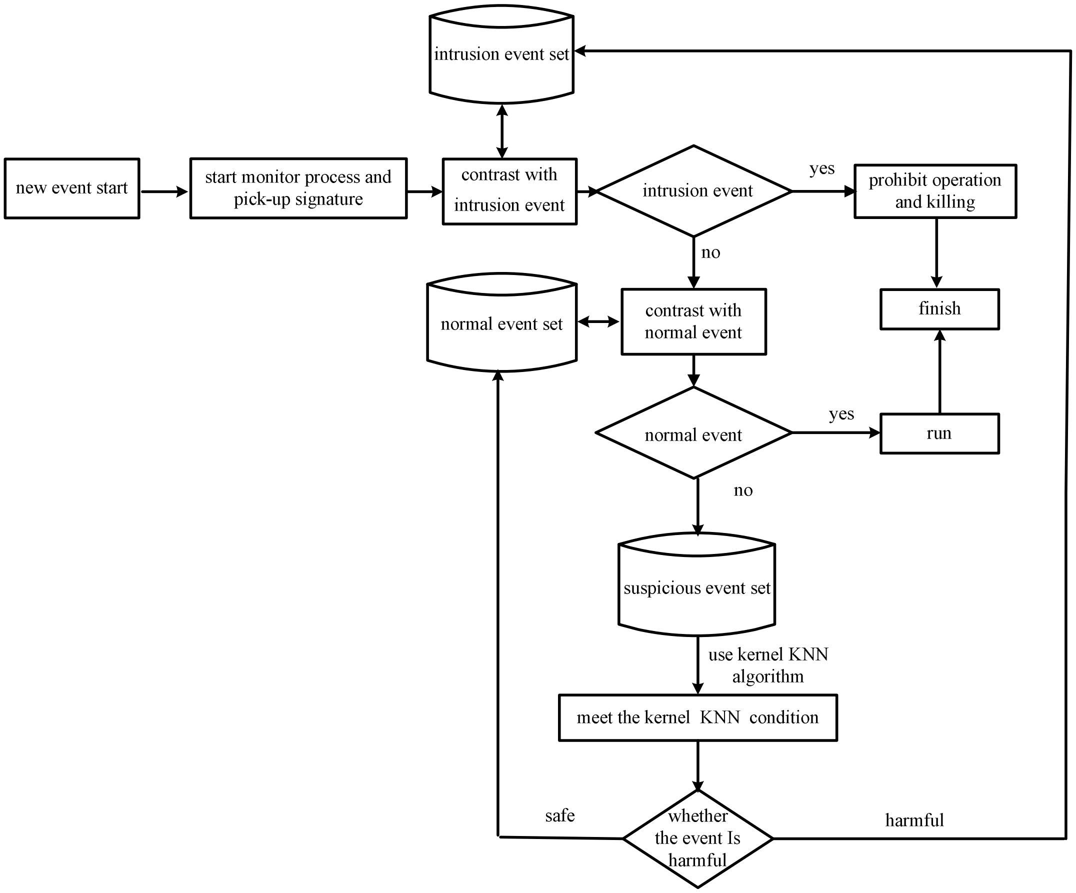
intrusion event set
new event start
start monitor process and pick-up signature
contrast with intrusion event
intrusion event
yes
prohibit operation and killing
no
normal event set
contrast with normal event
finish
normal event
yes
run
no
suspicious event set
use kernel KNN algorithm
meet the kernel KNN condition
safe
whether the event Is harmful
harmful

Figure 3 the flow chart of active defence system with host intrusion normal event set, taken another 60 normal event of system as the samples of classification. According to the event signature provided by Kingsoft, we obtained 90 intrusion files through various channels, of which 70 were randomly selected as the intrusion event set, and 20 samples as the samples of classification.

The guiding ideology is to find out k nearest neighbor of x respectively in normal event set N and intrusion event set I, then solving the center vector yn and yi respectively in N and I, calculating the distance of x , yn is $KD(x, yn)$, and x , yi is $KD(x, yi)$, if $KD(x, yn) < KD(x, yi)$, then $x \in N$, else $x \in I$. Table 1 lists the results of the experiment.

TABLE 1 THE CLASSIFY RESULT OF NORMAL EVENTS AND INTRUSION EVENTS

arithmetic	correct classification rate	
	normal events	intrusion events
K-NN	53.3	60.0
K-NN based on kernel	60.0	75.0

2.4.5 The disposal of "suspicious event set"

If an event is suspicious, we can pick-up the eigenvector of it according to the above method. We regard "suspicious event set" as test subset KT, "normal event set" and "intrusion event set" as reference subset KR. In reference subset KR, we can get k nearest neighbors relatively fast by spatial sphere algorithm. Using kernel K-NN algorithm, we can tell whether the suspicious event is harmful. If it is harmful, the monitor process adds its eigenvector to "intrusion event set", and then kills it, otherwise if it is safe, the monitor process adds its eigenvector to "normal event set", and let it run.

3. Conclusion

This paper proposed the active defence technology anti-host intrusion based on k-nearest neighbor algorithm of kernel. The system can not only judge normal event efficiently, but also can detect unknown event more accurately. The system can judge whether an event is harmful, and can store the eigenvector of suspicious event to "normal event set" or "intrusion event set" automatically.

References

[1] Teng Shaohua, a study on object-monitoring-based distributed and collaborative intrusion detection[D]. The degree of Doctor of Philosophy, Faculty of Electromechanical Engineering Guangdong University of Technology. (in Chinese)

[2] WANG Haifeng,XIA Honglei,SUN Bing.The Study of Anti -Virus Engine Base On the Analysis Of Virus Behaviors[J]. Computer Systems Applications, 2006.No. 5.29-31. (in Chinese)

[3] ZHANG Boyun,YIN Jianping,ZHANG Bingxing,HAO Jingbo.Unknown Computer Virus Detection Based on K-Nearest Neighbor Algorithm[J]. Computer Engineering and Applications, 2005. No.6. 7-10. (in Chinese)

[4] XIE Jinjing,ZHANG Yibin. Computer Virus Detection Based on Improved K-Nearest Neighbor Algorithm[J]. Modern Electronics Technique,2007.No.3.51-53. (in Chinese)

[5] WANG Yejun,NI Xizhen,WEN Weiping, JIANG Jianchun. Research on Principle and Defense of Buffer Overflow Attacks [J]. Application Research of Computers,2005.No.10.101-104. (in Chinese)

[6] M. Kwiatkowska, V. Sassone, " Science for Global Ubiquitous Computing," Grand Challenges in Computing (Research), edited by T. Hoare and R. Milner, 2004.

[7] S. W. Smith, Trusted Computing Platforms: Design and Applications,Springer, New York, 2005.

[8] WEI Wei , ZHAN G Li-yan , ZHOU Lai-shui, A Spatial Sphere Algorithm for Searching K-Nearest Neighbors of Massive Scattered Points[J]. ACTA AERONAU TICA ET ASTRONAU TICA SINICA, Sept . 2006. Vol127 No15.944-948. (in Chinese)

[9] RAO Xian , YAN G Shao-quan , WEI Qing , DON G Chun-xi. Kernel nearest neighbor rule and its appl ication in intrusion detection. Systems Engineering and Elect ronics. Mar . 2007. Vol . 29 No. 3.470-471. (in Chinese)

[10] YU Xue-dou, Research on Active Defence Technology with Virus Based on Improved K-Nearest Neighbor Algorithm, ITESS'2008.

[11] DENG Lujuan, LIU Tao, GAN Yong, XIONG Kun. Active Defence Technology with Virus Based on Differentiation and Hiding Process [J]. Computer Engineering, March 2007. Vol.33 No.5.117-119. (in Chinese)