I.J. Engineering and Manufacturing **2011, 5, 52-58**
Published Online October 2011 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijem.2011.05.07

Available online at http://www.mecs-press.net/ijem

Study on the Application of Artificial Immunity in Virus Detection System

DENG Daping[a,*1], DENG Xiaohong[a,b,*2]

[a] *College of Applied Science Jiangxi University of Science & Technology,Ganzhou 341000, China*
[b] *College of Information Science and Engineering,Central South University, Changsha, 410083, China*

Abstract

Artificial immunity, as a new technology, has been applied widely in virus detection system for its advantages. This paper emphasizes on these works as follows. Firstly, we analyze the disadvantages of traditional virus detection methods and new functions of artificial immune technology, and then review some typical algorithms of the existing virus detection system based on artificial immunity. Finally, a universal evaluating scheme is proposed. The purpose for this paper is to study the existent artificial immune methods and promote the new schemes emergence for virus detection system effectively.

Index Terms: Artificial Immunity; Virus Detection System; Artificial Intelligence; Antivirus

1. Introduction

With the rapid development of computer and network technology, it provides a very convenient way to share a large mount of information among different people; however it also gives chances to malware activities, such as propagating malicious programs, including computer viruses. Computer virus is a program, which invades your computer, as same as a biological virus injects an organism. It executes malicious functions to damage computer system.

In order to detect viruses and protect computer system's security, many different virus detection systems and methods have been proposed. However, these traditional schemes become more and more inefficient because of the multiversity and variation of viruses. Artificial immunity [1-5], as an important and new branch of artificial intelligence, is inspired by the natural immune system, which is a dynamic, adaptive and distributed learning system. It protects organisms against antigen invasion by distinguishing foreign antigens (pathogens and tumor cells) from organisms' own healthy cells and tissues and eliminating foreign antigens. Similarly, the functionality of computer security systems is to recognize and eliminate virus, so that the natural immune system has provided with an inspiration to develop such kind of antiviral systems. Actually, many researchers

* Corresponding author:
E-mail address: *1 270701689@qq.com; *2 dxh_lizi@sohu.com

have proposed various heuristic detection methods, including artificial immune system, to improve the effectiveness of virus detection system.

The rest of this paper is organized as follows. In Section II, the defects of traditional virus detection methods are discussed. In Section III, some typical algorithms of the present virus detection scheme based on artificial immunity are studied. Section IV is devoted to the performance evaluation of virus detection system. Conclusion is given in Section V.

2. Traditional Methods for Virus Detection System

The traditional detection methods utilized an anomaly detection system to recognize malicious programs could be divided into three major categories: Behavior-based, Data-based and the Virtual Memory Machine techniques.

Behavior-based methods use the operating system's application programming interface (API) sequences, system calls or other kinds of behavior characteristics to identify the purpose of a program [2]. These methods at first construct profiles during the legitimate operations of the monitored programs. During the detection process, any system call sequence or argument that does not comply with the previously generated "normal" profiles is regarded as a sign that the system is compromised. The corresponding program will be stopped and then classified as a malicious. The malicious programs are identified when the computer are already damaged, so many methods use a virtual environment (called sandbox) to simulate real system environment where the unclassified programs are running. The drawback of these methods is cost too much time and effort to build a less-error sandbox, but the sandbox can not simulate the same environment as real operating system. Although these approaches have produced promising results, they can produce high rates of false positive errors, an issue which has yet to be resolved.

Data-based methods can detect virus before they are executed, they utilize the binary data extracted from the program files. The traditional methods extract signatures from virus samples [3], scanners compare these signatures with unclassified files to determine whether they are virus of not. These methods are effective in the past. As novel advanced technologies are widely used in manufacturing new virus, polymorphic virus can change their signatures while spreading. So it is getting harder both for experts to extract signatures and to detect them.

With the emergence of unknown and polymorphic viruses , the Virtual Memory Machine techniques [4], as the common technique that used to detect Polymorphic viruses, which depends on emulation and sequential stages solution to decide the infection of a file or a program. This technique is not practical for high level language written viruses. Also, the emulation-based virus detection is considered as quite slow when it analyzed a file.

In addition, the traditional methods of anti-virus including character code analysis method, check-sum analysis method, behavior inspect analysis method and so on have played an important role in the history of virus detection [5]. But the traditional methods of anti-virus can not detect successful the new unknown virus with the popularity of the Internet and advanced technique of virus increasing. Faced with this situation, it is important to seek a new method of virus detection can be fast, accurate, and effective for detection.

Now, Artificial Intelligence as a new anti-virus technology has been applied in anti-virus engines. The main research field of Artificial Intelligence, such as Heuristic Technology, Artificial Neural Networks, Data Mining Technology etc, have been applied in the new generation of anti-virus detection system and play a crucial role in improving the anti-virus's performance and veracity [6-11]. First of all, artificial immune system is developed around defense. Immune cells are simply classified by antigen and antibody. Foreign cells that generate pathogen are called antigen while the cells that can recognize and kill antigen are called antibody. When externally attacked, human body will rapidly give immune response to resist the attack. These artificial immune systems draw inspirations from the biological immune system and have some similar features in functions that regard the normal programs as *Self*, suspicious programs as *Non-self*, malware detectors as antibodies or immune cells, and all procedures in the network system as antigen. Moreover, these artificial

immune systems generate qualified malware detectors to detect malware by detectors/antigens matching.

3. Study on Virus Detection System Based on Artificial Immunity

At present, many promising virus detection systems or methods based on artificial immunity have been proposed. Fig.1 gives a general model of them. This leads to some heuristic data-based methods. Among those methods, there are three basic models based on AIS principles: negative selection algorithm, clonal selection algorithm and the immune network model. It has a great similarity to computer security system. The concepts, self, non-self and so on, can be applied to the computer security system as well. Therefore, virus detection using AIS has paved a new way for anti-virus research. Based on ideas from immunology, many researchers have done a lot of works focused on two categories, including provide systematic model and specific algorithms.

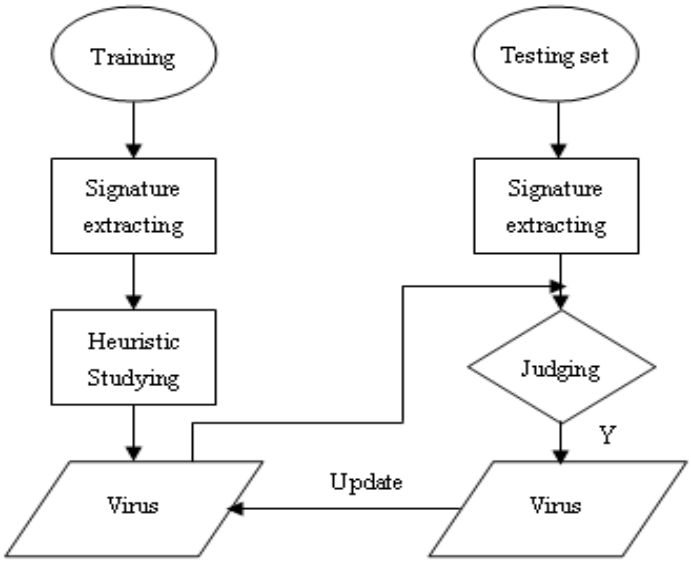

Fig. 1. A general model of virus detection systems or methods based on artificial immunity

3.1. systematic Models

From the view of systematic model, Wang et al.[3] proposed a hierarchical artificial immune system (AIS) model, which is based on matching in three layers, to detect a variety of forms of viruses. It can be divided into three layers. In the bottom layer, a non-stochastic but guided candidate virus gene library is generated by statistical information of viral key codes. Then a detecting virus gene library is upgraded from the candidate virus gene library using negative selection. In the middle layer, a novel storage method is used to keep a potential relevance between different signatures on the individual level, by which the mutual cooperative information of each instruction in a virus program can be collected. In the top layer, an overall matching process can reduce the information loss considerably. Similarly, Zhang et al.[12] also proposed a novel immunity-inspired model for malware detection (IMD). The IMD model extracts the I/O Request Packets (IRPs) sequence produced by the process running in kernel mode as antigen, defines the normal benign programs as self programs, and defines the malwares as non-self programs. By the process behavior monitoring

and the family gene analysis, the model can monitor the evolution of malware. The model generates the immature antibodies by vaccination, produces mature antibodies by clonal selection and gene evolution, and then learns and evolutionary identifies the unknown malware by the mature antibodies.

Saman et al.[2] presented a significative model from the human immune system's view in 2009; the model contains three basic steps as follow. First of all, the detection process is started by checking the behavior of the suspected file to know if the polymorphic behavior existing or not. Then the suspected file could be scanned to find the predefined signature or to know to what percentage there is a similarity between the new strain and the old others. Then the process of hyper mutation will be stared. After that the convergence to the goal will be checked. Finally, the model is not forgotten to memorize the generated signature that has a good affinity with the goal signature. Since there is a feed-back procedure in the system that used to update the database of signatures and new behaviors, the known polymorphic viruses and their new strains could be detected successfully. Ou et al.[13] described an agent-based computer virus detection system (CVDS) based on danger theory of human immune system. This paper embeds multiple agents into AIS-based virus detection system, where each agent coordinates one another to calculate mature context antigen value.

3.2. Immune Algorithms

Negative Selection Algorithm (NSA)

In general, NSA is used to form the virus detector set. Saman et al [2] proposed the initial Negative Selection Algorithm (NSA). The algorithm can recognize self and non-self without reference to any particular information of the non-self set, especially suitable for computer fault diagnosis in an unknown time-varying environment like virus detection. But the model has a high computational cost. The number of detectors is exponentially related to the size of self set. Wang improved the NSA [3], make the number of detectors linearly related to the self set. But it still needs too many detectors, and they are also not guided generated detectors. Due to above reasons, it is a key problem how to generate effective detectors in a quick way. Rui et al.[4] also used NSA to generate initial virus database in detector. The process of negative selection is shown in Fig.2.

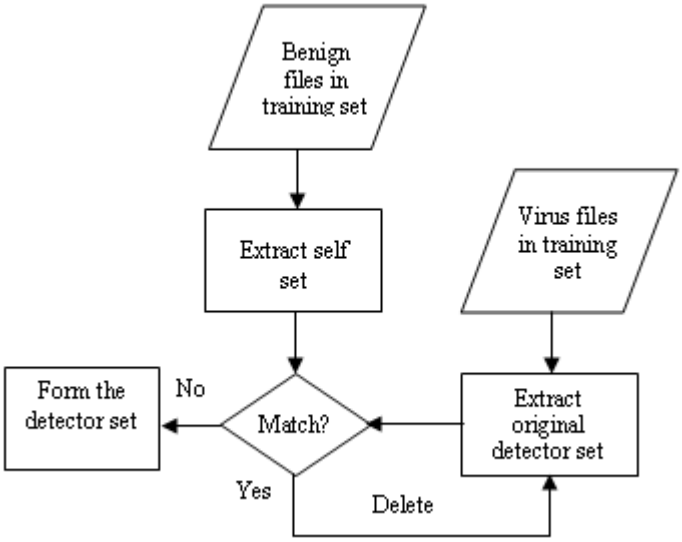

Fig. 2. Negative selection process

Clonal Selection

The clonal selection algorithm is used by AIS to define the basic features of an immune response to an antigenic stimulus [14,15]. It establishes the idea that only those cells that recognize the antigens are selected to proliferate. The selected cells are subject to an affinity maturation process, which improves their affinity to the selective antigens. Clonal selection is used to increase the diversity of detector set in the non-self space; new detectors are generated from data fragments in detector set. Rui et al.[4] gives the computing method for the number of clones generated as shown in (1).

$$|C(x_i^t)|=\frac{a}{Fx_i^t} \tag{1}$$

Where |*| denotes the number of elements, a is the coefficient of clone selection, usually a=10. x_i^t represents a single detector in detector set and Fx_i^t is the occurrence frequency of it, $C(x_i^t)$ is number of clones generated by x_i^t .

Classify Strategy

Since the virus detector set depends on the self and non-self characters, it is very important to form a vector with two elements of concentrations for characterizing the program efficiently and fast. Several classifiers including k-nearest neighbor (KNN), RBF neural network and support vector machine (SVM) with this vector as input are then employed to classify the programs. Wang et al. [14] proposed a clonal particle swarm optimization (CPSO) algorithm for optimize the selection of detector library determinant and parameters associated with a certain classifier. Yu [15] gave some related works about K-nearest Neighbor Algorithm of Kernel, the disadvantage of the virus detection system based on k-nearest neighbor algorithm is that the classified effect whose distribution of class is the non-Gaussian distribution and the non-ellipse distribution is bad, and because its time complexity is O(N2) (N is the total number of data point), the method cannot meet the demand of the project when the data point number is large.

However, the immune systems above still have some shortcomings: (1) the definitions of Self and Non-self lack flexibility, which do not reflect the dynamic evolution of network environment; (2) low generation efficiency of antibody (detector), which do not make use of the reservation of the best antibody genes; (3) insufficiency of antibody diversity; (4) antigen extraction method is simple and does not reflect the process behavior [12]. These deficiencies led to the big cost of antibody (detector) generation, long detection time, and high false-positive rate. Wang et al.[16] introduced how to use single layer neural classifier to detect boot viruses; its structure has been given in Fig.3.

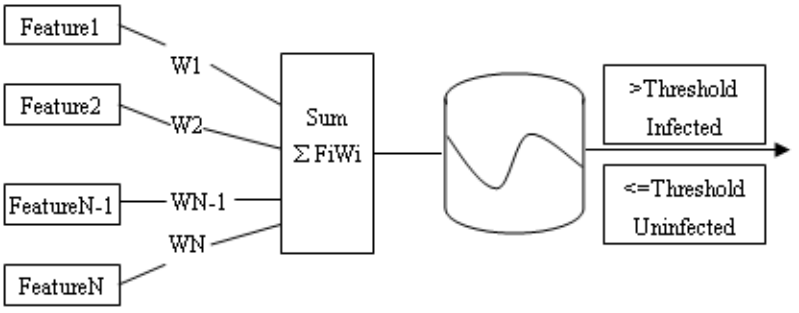

Fig. 3. Single layer neural classifier

4. Performance Evaluation

To evaluate the performances of artificial immunity-based virus detection system, four important metrics are considered: the detection rate *DR*, false positive rate *FPR*, false negative rate *FNR* and generalization. The detection ability means the virus detection system can detect how many testing files, which include benign and malicious program. False positive, in other words, the benign files has been wrongly identified as virus. In Contrast, false negative denotes the virus files has been recognized to be benign. The generalization metric means that the detector of virus detection system can be used in different environment, such as abnormal al constitution between benign and virus files, different operating system and so on. These former three indicators can be quantified by following formulae:

$$DR = \frac{|N_D|}{|N_T|} \tag{2}$$

$$FPR = \frac{||N_{D\_V}| - |N_V||}{|N_B|} \tag{3}$$

$$FNR = \frac{||N_{D\_B}| - |N_B||}{|N_V|} \tag{4}$$

Where, $|N_D|$ and $|N_T|$ denote the number of testing files to be successfully detected and the total number of the testing files, respectively. N_B and N_V represent the benign and malicious programs in testing files. Obviously, $|N_T|=|N_B|+|N_V|$. is the number of benign files that have been detected, $|N_{D\text{-}B}|$ and $|N_{D\text{-}V}|$ are the number of viruses that have been detected.

In general, in order to measure the system's generalization, the training files were randomly divided in to different parts. In addition, the minimum length L of signature was appointed a fixed number according to the instructions, for example, As the length of a single assemble code instruction varies from 1 byte to 7 bytes, L is not necessary bigger than 64-bit to contain at least one entire instruction. So the experiments choose L = 64-bit and L = 32-bit.

From experimental results in many literatures, we can get some general conclusion as follows: (1) The less the length of L, the more accuracy and generalization ability in detecting virus in the testing sets, because that the small data fragments contain enough virus characteristics information for detection, and the big one contain too much benign codes, reducing the thickness of virus information.(2) With the increase of training set, the detection rate of virus files in testing set increases.(3) In order to simulate the practical situation, in the dataset, the size of benign files is much larger than that of virus files. It is a general situation in the computer software environment.

Acknowledgements

Artificial immunity has played an important role in enhancing the efficiency and performance of virus detection system. Though computer immune system and human immune system have many similarities, how to independently simulate the biology intelligence of human by computer is difficult, which is the most topic of

artificial intelligence. So, virus detection system based on artificial immunity seems to be stalled at an early stage. So there are a lot of works to be done in the future, if the accuracy and independence of self-study and detection are achieved, artificial immunity would be accepted fully by virus detection system.

References

[1] Ruan, G.C., Tan, Y.. "A Three-layer Back-propagation Neural Network for Spam Detection using Artificial Immune Concentration". Software Computing, vol.14, pp.139–150, 2010.
[2] Saman, M. A., Omar, Z.. "Devising a Biological Model to Detect Polymorphic Computer Virus—Artificial immune system (AIM) Review". 2009 International Conference on Computer Technology and Development, Kota Kinabalu: IEEE Computer Society, November 13-15, 2009.
[3] Wang, W., Zhang P. T., Tan, Y., He, X. G.. "A Hierarchical Artificial Immune Model for Virus Detection". 2009 International Conference on Computational Intelligence and Security, Beijing: IEEE Computer Society, December 11-15, 2009.
[4] Chao, R., Tan, Y.. "A Virus Detection System Based on Artificial Immune System". 2009 International Conference on Computational Intelligence and Security, Beijing: IEEE Computer Society, December 11-15, 2009.
[5] Cui, Z. L., Lu, X., Wang, J.. "Adaptive Intrusion Tolerance Strategy of the System Based on Artificial Immune".2009 International Conference on Computational Intelligence and Software Engineering, Wuhan: IEEE Computer Society, December 11-13, 2009.
[6] Han, C.W., Shou, H.S.. "Neural Networks-based Detection of Stepping-stone Intrusion".Expert Systems with Applications, vol.37, pp.1431-1437, 2010.
[7] Horng, S.J., Fan, P.Z., Chou, Y.P., Chang, Y.C., Pan, Y.. "A Feasible Intrusion Detector for Recognizing IIS Attacks based on Neural Networks". Computers & Security, vol.27, pp.84-100, 2008.
[8] Wu, C.H.. "Behavior-based Spam Detection using a Hybrid Method of Rule-based Techniques and Neural Networks". Expert Systems with Applications, vol.36, pp.4321-4330,2009.
[9] Wang, Y., Gu, D.W., Li, W., Li, H.J., Li, J.. "Network Intrusion Detection with Workflow Feature Definition Using BP Neural Network". LNCS, vol.5551, pp.60-67, 2009.
[10] Wang, Y., Gu, D.W., Wen, M., Li, H., Xu J.P.. "Classification of Malicious Software Behaviour Detection with Hybrid Set Based Feed Forward Neural Network". LNCS, vol. 6064, pp. 556-565, 2010.
[11] Han, C.W., Shou, H.S. "Neural Networks-based Detection of Stepping-stone Intrusion". Expert Systems with Applications vol. 37, pp. 1431–1437, 2010.
[12] Zhang, Y., Wu, L.H., Xia F., Liu, X.W.. "Immunity-based Model for Malicious Code Detection". LNCS, vol. 6215, pp.399-406, 2010.
[13] Ou, C. M., Ou, C.R.. "Abstraction from Dendritic Cell Algorithm with Danger Theory". LNCS, vol. 6104, pp.670-678, 2010.
[14] Wang W., Zhang P.T., Tan Y.. "An Immune Concentration Based Virus Detection Approach Using Particle Swarm Optimization". LNCS, vol.6145, pp.347-354, 2010.
[15] Yu, X. D.. "Research on Active Defence Technology with Virus Based on K-Nearest Neighbor Algorithm of Kernel". 2009 International Conference on Environmental Science and Information Application Technology, Wuhan: IEEE Computer Society, July 4-5, 2009.
[16] Wang, X.B., Yang G.Y., Li Y.C., Liu Dan.. "Review on the application of Artificial Intelligence in Antivirus Detection System". 2008 IEEE Conference on Cybernetics and Intelligent Systems, London: IEEE Computer Society, September 21-24, 2008.