International Journal of Security and Its Applications
Vol.9, No.9 (2015), pp.359-370
http://dx.doi.org/10.14257/ijsia.2015.9.9.31

Network Intrusion Detection System Model Based On Artificial Immune

Zhang Yanbin[1,2]

[1]*Dongbei University of Finance and Economics, Dalian, Liaoning, 116025, China*
[2]*Shandong Youth University of Political Science, Jinan, Shandong, 250103, China*
Zybjsj@126.com

Abstract

At present, the intelligent intrusion detection technology has become a new intrusion detection technology and its development direction. Among them, the biological immune principle is introduced into the idea of intelligent intrusion detection technology, offers a new way for the study of intelligent intrusion detection system. This paper builds a network intrusion detection system based on artificial immune principle of the new model. Through the optimization algorithm, the model could improve the ability of the immune response of the system. Finally the KDD CUP99 Data Set simulation a dynamic network environment, the simulation experiment, the new model of the new model based on optimization algorithm and the traditional intelligent intrusion detection model comparing the experimental results. On the basis of the new model to make quantitative analysis and qualitative evaluation, the next step is improving ideas were proposed.

Keywords: *artificial immune, network intrusion, detection system, model, biological immune principle*

1. Introduction

The advent of biologic immune system, a complex pattern recognition system, has brought inspiration to people on the study of computer security. The researchers have found that there are many similarities between intrusion detection system and biologic immune system, such as their function, environment, and detection method. Therefore, the corresponding algorithm models for functions of the immune system have come into being and been successfully applied into security architecture of computer information system [1-2]. In this paper, negative selection algorithm for immune theory is used in the research of network intrusion detection system. "Self" and "non-self", which are defined in this model, respectively indicate normal host behavior in the network or normal network communication mode and abnormal host behavior in the network or abnormal network communication mode[3]. Representations of both adopt binary string collection, which is of generality in computer [4-5]. The self-collection needs to be initialized in that it is intended as a learning sample and the detector collection needs to be produced in a form of binary string with negative selection algorithm in the model [7]. With lymphocyte in biologic immune system corresponding to detector in the model, the process of affinity maturation is achieved by depending on the detector's mechanism of activation threshold. On the strength of r-contiguous bits matching rule for calculation, one calculator will be put into the detector which will be activated provided the number on calculator reaches the pre-set value within a certain period of time. The detector generation process is as follows: Negatively select the randomly-generated binary string by means of negative selection algorithm and delete it in the event that it matches the Self in tolerance phase, otherwise convert it into a mature detector which will be activated as long as the number on calculator reaches threshold value [8-12]. At this time, with the capability to execute subsequent immune responses, the detector can also be transformed into a memory

ISSN: 1738-9976 IJSIA
Copyright ⓒ 2015 SERSC

detector with permanent life circle. In this case, it is easy to detect the same non-self string when it emerges, which improves the detection efficiency. The use of memory detector is inspired by secondary response mechanism [13]. Nowadays, with the growing complex network environment, the system has set high demand on intrusion detection in fast-developing technology domain, so to keep the monitoring system continuously updated mainly upon overlaying detector is an important approach in dealing with incessant network changes [14]. Through continued perfection of the detector generation, the intrusion detection system is able to detect a wider range of intrusion behavior with the original detectors being replaced by those can detect new intrusions, in which case there will be no redundant detector and the intruders will fail to escape from the detection system [15].

2. Design of Intrusion Detection System Model based on Artificial Immunity

2.1. Basic Structure of the Model

Network intrusion detection model based on artificial immunity presented in this paper primarily consists of three important modules: Self set generation module, mature detector set generation module and memory detector generation module. The concept of life cycle has been introduced into these three modules, under which circumstances the modules will be of dynamic nature. Basic structure frame of the modules is shown in Figure 1.

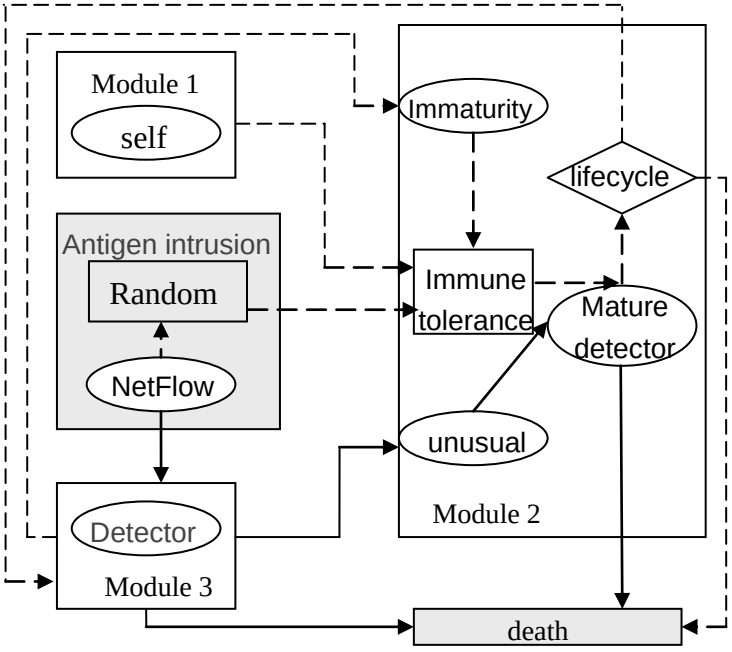

Figure 1. Basic Structure Frame of Network Intrusion Detection Model based on Artificial Immunity

(Solid line - process of antigen detection; dotted line - process of antibody detector)

The system chiefly includes two processes: 1st antigen processing (indicated by solid arrow in Figure 1); 2rd antibody evolution ((indicated by dotted arrow in Figure 1). Specific functions of each module are as follows:

 Copyright ⓒ 2015 SERSC

1). Design of Self set generation module

The initial "self-set" is finally formed by abstracting out the normal modes and defining the pattern sequence set after the data flow in the network is analyzed for normal data flow. However, with the complex changes of network, it is necessary to require a continuous change in "self-set".

2). Design of mature detector set generation module

This module mainly works to examine the random character set from internet and induce the string's immune tolerance with immune algorithm which refers to negative selection algorithm in this paper. In the process of immune tolerance, a tolerance phase within which the random string is checked whether it matches with the string is set. If they do not match, it means that the random string is an intrusion attack; otherwise, it will die out automatically. Feature of the string without a match will be extracted and the string will be turned into a mature detector. In the above-mentioned process, the concept of life cycle of a string can be introduced. Within this cycle, if cumulative frequency of attacks of an intrusive string is found to reach the predetermined threshold, it is considered that this intrusion behavior keeps attacking the system. To enhance the detection efficiency, this intrusive string will be transformed into memory detector, and this is how the mature detector is converted into memory detector. In the case that the threshold is not reached and the above deadline is not missed, the mature detector at this moment will die out automatically.

3). Design of memory detector generation module

Based on secondary response principle of biologic immune system, in this module the repeating intrusion behavior can get quick response to improve the detection efficiency. Memory detector is the detector with permanent survival date in the monitoring system and its life cycle is longer than that of detector. However, it will be deleted from the system upon matching with self. Design process of memory detector module is as follows:

Step one: Immunize the randomly generated string and memory detector. If they do not match, apply the mature detector for detection; otherwise proceed to the next step.

Step two: Detect whether the string belongs to self-set. If not, it is an abnormal behavior; otherwise, proceed to the next step.

Step three: Co-stimulation is needed. Judging from the co-stimulatory signal, if the system administrator supposes that it cannot be self, delete the identical self in self set; otherwise, deleting the memory detector t.

2.2. Design of Immune Algorithm

Negative selection process in the model can be divided into three steps:

1) Define a self-set. Self-set of negative selection algorithm is abstracted out through analysis on previous normal network behavior. The analysis process is to extract communication information, such as, service type, source address, destination address, source port, destination port, sc-bytes, cs-bytes status, time delay etc. from the normal network packets.

2) Produce a mature detector collection in which each element is unable to match that in self set. Similar to negative selection principle in immune system, produce a set of random binary sequences as candidate detectors and match them with the elements in self set. Delete the candidate detector once they match, otherwise enter it into detector collection as a mature detector.

Copyright ⓒ 2015 SERSC

3) Take advantage of the generated mature detectors to detect data on the network. It is believed that non-self is detected when a match occurs and then an alarm should be raised.

Initial establishment of the model takes place in two phases: training phase and test phase. In training phase, the generated detectors should experience the negative selection process similar to self-tolerance process, that is to say, the generated detectors (candidate detector) and the training set (of self-collection) will make matching tests according to a certain matching rule. Consequently, remove those candidate detectors that matched with self-set while keep those unmatched as mature detectors. Mature detectors, actually the pattern string of non-self, form a set by the name of detector of the detection system which is stored in detected set. In test phase, the detector is utilized to detect all patterns which are abstracted out from data packet in the network. It is a certain pattern string for non-self once some detector is found out to be matched with the undetected pattern and a false alarm will be communicated to the system administrator [14].

2.3 Generation Algorithm for Detector

The concepts involved will be introduced below for the following pattern design:

N_s refers to the number of self in S set; N_{R_0} refers to the number of detectors in the immature detector collection; N_R refers to the number of detectors in the mature detector collection R.

p_m refers to probability of the intrusion detection system making the matching test between the random binary string and the mature detector.

P_f refers to the probability of detector in mature detector collection un-matching with a random non-self, i.e. $P_f = (1-p_m)^{N_A}$, f refers to the probability of string in self set un-matching with a random self-string, i.e. $f = (l-p_m)^{N_s}$.

Besides, in universal set L refers to fixed length of the binary string data. In the matching rule, the lowercase letter r is used to define matching length. R_0 and R represent immature detector collection and mature collection respectively. During the subsequent algorithm design, some other concepts may be encountered, and we will introduce them over the realization of algorithm.

Detector generation algorithm based on linear time complexity will take place in two steps. In the first step, the number of immature detectors which are unmatched with "self-set" should be calculated. In the second step, the mature detectors should be found out from the immature detector collection. Below is respective implementation process of the two steps [2].

The first step will not start with counting the number of immature detectors un-matching with "self-set" but with the simplest random string. What "simplest" means here is that the r-contiguous bits of random string equal the fixed length of string minus matching start bit plus 1, i.e. r=L-i+1. At this moment, the number of unmatched detectors is represented by C_{l-r+1}. The calculation formula is:

$$C_{i-r+1}[s] = \begin{cases} 0, & t_{l-r+1} \quad Nomatch\ in\ S \\ 1, & t_{l-r+1} \quad Match\ existing\ in\ S \end{cases} \tag{1}$$

When $1 \le i < (L-r+1)$, the calculation formula is:

$$C_i[s] = \begin{cases} 0, & t_{i,s} Match\ existing\ in\ S \\ \hat{C}_{i+1}[s \bullet 0] + \hat{C}_{i+1}[s \bullet 1], & t_{i,s} Nomatch\ in\ S \end{cases} \tag{2}$$

 Copyright ⓒ 2015 SERSC

In the second step, total number of the mature detectors which are later found out in the immature detector collection should be reckoned first. The total number is represented by T. The calculation formula is:

$$T = \sum_{s} C_1[s] \tag{3}$$

By this time, the detector space should be partitioned with division method until each detector can be numbered. In this case, the required mature detectors will be gained with ease. For instance, if the K^{th} detector is needed, it should satisfy:

$$P_1 = \sum_{s<s1} C_1 < k < Q_1 = \sum_{s<s1} C_1[s] \tag{4}$$

Determine the division range (P_1,Q_1] with formulas below：

$$P_i = \begin{cases} P_{i-1}, b_{i-1} = 0 \\ P_{i-1} + C_i[\hat{S}_{i-1}, 0], b_{i-1} = 1 \end{cases} \tag{5}$$

$$Q_i = \begin{cases} P_{i-1} + C_i[\hat{s}_{i-1}, 0], b_{i-1} = 0 \\ Q_{i-1}, b_{i-1} = 1 \end{cases} \tag{6}$$

Time complexity and space complexity of detector generation algorithm based on linear time complexity are:

Time complexity:

$$O((l-r)*N_\varepsilon) + O((l-r)*2^r) + O(l*N_\rho) \tag{7}$$

Space complexity:

$$O((l-r)^2 * 2^r) \tag{8}$$

Detector generation algorithm based on linear time complexity considerably raises the detector generation efficiency and lowers the time complexity and space complexity. The reason why the algorithm is given the name is that its time complexity is on a linear scale. However, the redundancy still cannot be avoided even though the detector generation efficiency is promoted.

How to reduce the computing cost, time complexity and space complexity, and also increase the efficiency is a research direction for improving the algorithm. This paper mainly focuses on perfection of the generation algorithm by how to decrease the space complexity. Enhancement of generation algorithm by how to cut down redundancy of the detectors will be highlighted below.

Next comes how to eliminate the redundancy in detector collection. On the base of r-contiguous bits matching rule, the redundancy in detector collection is lowered and the detector generation algorithm based on linear time complexity is improved, ensuring the generated detectors can cover non-self space as much as possible

With the original detector generation algorithm based on linear time complexity, total number of the mature detectors is counted first of all. Then the mature detectors are found out from the immature detector collection with division method. Finally, the obtained detectors will be put together into a set named mature detector collection. However, embracing an idea that more non-self space can be covered by fewer generated detectors, the improved algorithm prevents more than one detector from matching the non-self.

With the improved algorithm, the implementation takes place in two steps. In the first step, total number of the binary strings undetected by the confirmed mature detectors is calculated. In the second step, a new detector will finally be generated by adding 0 or 1 in left-right direction of the detector template undetected by mature detectors.

Above all, D^R array represents total number of the binary string detected by the confirmed mature detectors. D^R is calculated using the formula $D_i[s] = C_i[s] * C_i'[s]$

Copyright ⓒ 2015 SERSC

with recursive computation. Then, those mature detector templates which are undetected by the known mature detectors and can match more non-self are selected. A new detector comes into being by adding 0 or 1 in left-right direction of the templates. In this process, these templates match with the new detectors and value of D^R also changes continuously. Repeat the above operation until corresponding bits of all the obtained mature detector templates in D^R array is 0.

Moreover, the total number of non-self undetected by the confirmed mature detectors can also be calculated. It is known that total number of detectors in mature detector collection is represented by N_R. Because

$$N_R \le 2^r \tag{9}$$

Therefore, combine the probability formula P_f (Here p_m is 2^{-r}):

$$(1-2^{-r})^{N_S} \tag{10}$$

Expected value of N_R is:

$$2^r * (1-2^{-r})^{N_S} \tag{11}$$

3. Simulation Experiment and Result Analysis

3.1 Acquisition of Experimental Data

The experimental data stems from evaluation dataset of the real dataset KDDcup99 [10] which originates from intrusion detection evaluation procedure of DPA (Department of Defense Advanced Research projects Ageney) in 1999. It includes various simulated intrusions in military network environment and increases characteristic attributes for partial network connection. A 10% data subset containing 494021 network connection records in which 396744 pieces are normal while 97277 pieces are abnormal is provided by KDDcup99 in this paper. Four types of attacks are included in the dataset: (1) DoS (denial of service Attacks); (2) R2L (Remote to Local Attacks); (3) U2R (User to Root) - Unauthorized access to super user permissions Attacks; (4) PROBE (probing) - port scanning and vulnerability scanning. See Table1:

Table 1. Attack List

Type	Name of Intrusion
Dos	*Smurf,pod,teardrop,back,Nepune,land,apache2,mailbomb*
Probe	*Guess_password,named,sendmail, xsnoop,ftp_write*
U2R	*Rootkit,buffer_overflow,xterm,loadmodule,httptunnel*
R2L	*Portsweep,ipsweep,nmap,mscan,saint*

After feature extraction, the obtained KDDcup99 embodies 41-dimensional characteristics, which can reflect the inner connection among multiple data connections. See Table2:

 Copyright ⓒ 2015 SERSC

Table 2. Feature Set in Need of Detection for Different Types of Attacks

Type	Feature set for detecting attacks
Dos	Basic feature set+ network traffic feature set
Probe	Basic feature set + network traffic feature set+ host traffic feature set
U2R	Basic feature set + content feature set
R2L	Basic feature set + content feature set

3.2 Experimental Design

The setting of experimental parameters is of great significance and ideal values are often selected both by experiment and experience. Below are discussions on several key parameters.

1) Tolerance phase T of immature detector

If the tolerance phase T of immature detector is long, the generated immature detector grouping will cover a more comprehensive antigen space, which helps to reduce the system's false positive rate. However, too long tolerance phase will impact the number of mature detectors, which will influence the detection efficiency. T=25 in this experiment.

2) Activation threshold A of mature detector

Activation threshold A of mature detector has massive effect on the system's detection efficiency. A of small value will lead to a larger number of generated memory detectors of poor quality, though. On the contrary, A of big value will result in a smaller number of memory detectors, which makes it impossible to effectively detect. Therefore, an ideal value needs to be selected depending on both experiment and experience. Here A=15. Other parameters: initial self set Sn=60; use the normal data (the end property is normal) in training set to initialize the self set. Total number of the non-memory detectors (immature detector Ib and mature detector Mb) M=600, i.e. Ii+Mb=600; Ib and Mb are empty in initial phase. Threshold of the sum of memory detectors P=250; life cycle of mature detector L=25 generations; total iterative times N=6000 generations. Input each detection parameter of the detection system; click "Training" and the system will begin to initialize. After several generations, the memory detector collection is nonempty and the gene library starts to be constructed; click "Detection" and the system begins the formal detection [9].

3.3 Experimental Results and Analysis

Aimed at proving the efficiency and feasibility of intrusion detection model proposed in the paper, the experimental design steps are as follows:

1) Dataset processing

It is needed to perform simulated attack, mainly DoS Attacks here, such as Ping of Death, SYNHoo etc. on the user to verify efficiency of the model on account that most

Copyright ⓒ 2015 SERSC

data on internet are normal. Subsequently, pre-process the captured dataset and transform them into real numbers between 0 and 1 for experience convenience.

2) Analysis of experimental results with improved algorithm

Choose an ideal value by both experiment and experience. Set the initial self set Sri=60, tolerance phase of immature detector Ti=6 generations, activation threshold of mature detector A=15, life cycle of mature detector L=10 generations, threshold of the sum of memory detectors P=50, total number of non-memory detector M=250, and total iterative times N=500 generations. In order to check performance of the new model, experiments are conducted for 10 times to contrast the improved detector algorithm based on gene library and the traditional algorithm based on linear time complexity, and the results are averaged.

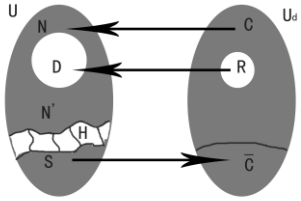

Figure 2. Comparison of the Time for Mature Detector Generation with Two Algorithms

Detector generation algorithm based on linear time complexity still remains the property that the time complexity is on a linear scale. Therefore, the improved algorithm is also with this original feature. However, unprecedented achievements have been made in decreasing redundancy. By means of the improved algorithm, the experimental results clearly show that the number of immature detectors declines, even more rapidly with the algorithm further working. It is predicable that the algorithm will have more evident effects in lowering the redundancy in detector collection and promoting the generation efficiency as the algorithm continues to work.

3) Analysis of experimental results with the new model

Experiment one: Demonstrate the efficiency and feasibility of the proposed model with FP value and TP value. Select 500 network records from the training set each time in the training phase to form the antigen collection. With the generations circulating, the gene library starts to be constructed and the inactivated mature detectors and deleted memory detectors will be added into it. The whole training phase will last for 500 generations. After the training phase, there comes the test phase that also lasting for 500 generations. At this moment, the TP value and FP value are used to evaluate the model and compared with those produced with traditional algorithm based on linear time complexity. The system continues to operate and presents the diagram of TP value and FP value with proposed algorithm and traditional algorithm based on linear time complexity respectively.

 Copyright ⓒ 2015 SERSC

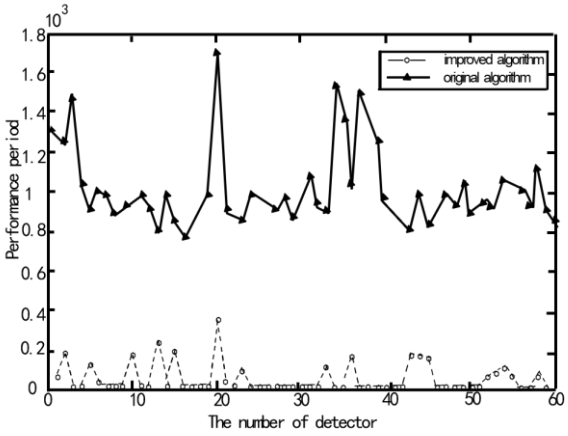

Figure 3. Comparison of Detection Rates

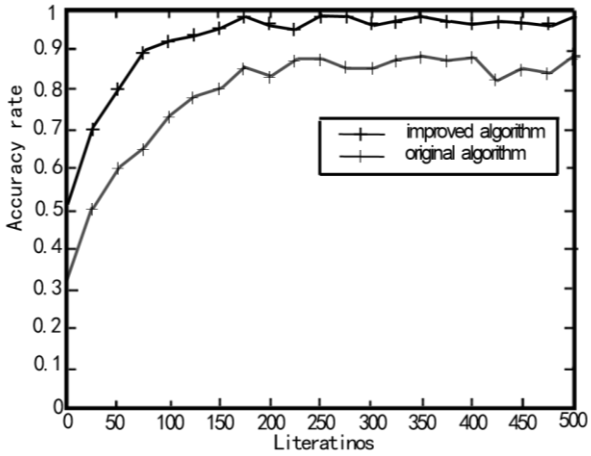

Figure 4. Comparison of False Positive Rates

Experiment two: Verify superiority of updating strategy for the proposed memory detector.

Through contrast experiments between traditional algorithm and updating strategy for the memory detector put forward in this paper, confirm the average effectiveness (X) is higher by means of the improved algorithm with the evolutional generations increasing. Experimental results as shown in the Figure below:

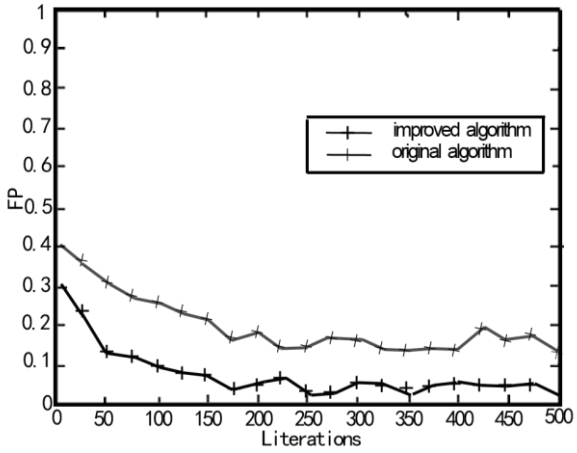

Figure 5. Diagram of Updating Strategy for Memory Detector

Copyright ⓒ 2015 SERSC

Analysis of experiment one: The Figure 3 shows that detection rate of the new model increases more markedly (steeper slope) and tends to be stable after the evolutional generation reaches approximately 100 when it is significantly higher than that in original model. At this moment, the gene library has initialized successfully and can produce evolutionary detectors at high proportion. Besides, updating strategy of the memory detectors begin to work by eliminating memory detectors of low affinity. The high-efficient memory detector grouping can rapidly detect the abnormal in the present environment, leading to a higher detection rate. The Fig. 4 shows that with the new model, the false positive rate is decreased even though its amplitude is not as high as that of detection rate, which indicates that the new model raises the detection rate at the expense of false positive rate for a faster detection speed.

Analysis of experiment two: The Figure 5 shows that the updating strategy for memory detector guarantees its completeness and multiplicity. As the evolutional generations increase, average effectiveness of the memory detector is higher than that using original algorithm.

5. Conclusions

This paper firstly introduces basic idea of intrusion detection system model in view of artificial immunity. With the established negative selection algorithm being foundation for modeling, a reasonable frame for intrusion detection system built on immune theory is set up, during which process what matters most is the detector generation. The detector generation algorithm based on linear time complexity is improved and the detector generation algorithm efficiency is promoted through removing redundancy, ensuring the generated detectors can cover non-self space as much as possible. The high efficiency of improved algorithm is verified by taking experiments.

Currently, many intruders carry out the attacks depending on security vulnerabilities in network protocol. Therefore, under the circumstance of high-speed broadband network, it is necessary to further decrease the packet loss, enhance the accuracy of detection and update the self collection timely. What is more, it is imperative to reinforce the study on co-operability and performance of communication protocol inside the intrusion detection model based on artificial immunity mechanism, and study on how to develop efficient detector by improving the detector generation algorithm.

References

[1] S. Forrest, A S. Perelson and L. Allen, “Self-nonself Discrimination in a Computer: In Proceedings of IEEE Society Symposium on Research in Security and Privacy and Privacy”, 1994. Massachusetts, USA, (**1994**), pp. 202-212.
[2] F. Gonzalez, D. Dasgupta and L F. Nino, “A Randomized Real-valued Negative Selection Algorithm: In Proceedings of Second International Conference on Artificial Immune Systems”, 2003. Edinburgh, UK, (**2003**), pp. 261-272.
[3] C. Gao and J. Liu, “Modeling and Restraining Mobile Virus Propagation”, IEEE Trans. Mobile Computing, Early access article, no. 99, (**2012**).
[4] P. Wang, M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi, “Understanding the Spreading Patterns of Mobile Phone Viruses”, Science, vol. 324, no. 22, (**2009**) May, pp. 1071-1076.
[5] J. M. Heffernan, R. J. Smith, and L. M. Wahl, “Perspectives on the Basic Reproductive Ratio”, J. Royal Soc. Interface, vol. 2, no. 4, (**2005**), September, pp. 281-293.
[6] A. Bose and K. G. Shin, “On Mobile Viruses Exploiting Messaging and Bluetooth Services”, Proc. Conf. Securecomm and Workshops, (**2006**), pp. 1-10.
[7] J. Hollis, “Focus on London 2010: Population and Migration”, O.f.N. Statistics, (**2010**).
[8] L. Wenke, S J Stolfo and K W Mok, “A data-mining framework for building Intrusion detection models [J]”, The IEEE Symposium on Security and Privacy, Oakland, CA,(**1999**), pp. 23-29.
[9] J. Wycross, “Integrated innate and adaptive artificial immune systems applied to process anomaly detection”, Nottingham: Notting-ham University, (**2007**).
[10] C. Stallings, J. McClary, D. DuBois and J. Ford, “NADIR: An automated system for detecting network intrusions and misuse”, Computers and Security, vol. 12, no. 3, (**1993**), pp. 253–248.

 Copyright ⓒ 2015 SERSC

[11] Y. Geng, J. Chen and K. Pahlavan, "Motion detection using RF signals for the first responder in emergency operations: A PHASER project", 2013 IEEE 24nd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), London, Britain (**2013)** September.
[12] Y. Geng, J. He and K. Pahlavan, "Modeling the Effect of Human Body on TOA Based Indoor Human Tracking [J]", International Journal of Wireless Information Networks 20(4), 306-317 Yu L, Liu H. Efficient Feature Selection via Analysis of Relevance and Redundancy. Journal of Machine Learning Research, vol. 5, (**2004**), pp. 1205-1224.
[13] W M. Hu, M. Hu and S. Maybank, "Adaboost based algorithm for network intrusion detection", IEEE Transactions on Systems, Man and Cybernetic, Parb B: Cybernetics, vol. 38, no. 2, (**2008**), pp. 577-583.
[14] L. Khan, M. Awad and B. Thuraisingham, "A new intrusion detection system using support vector machines and hierarchical clustering", The VLDB Journal, vol. 16, (**2007**), pp. 507-521.
[15] C L. Huang and C J. Wang, "A GA-based feature selection and parameters optimization for support vector machines", Expert Systems with Applications, vol. 31, no. 2, (**2009**) August, pp. 231-240.

Author

Zhang Yanbin, He was born in Shandong, China, in 1980. He received B.S and M.S. degrees in computer science and technology from Ji'nan University, Shandong, China, in 2003 and 2010, respectively. His research interests include intelligent computing theory and application, bioinformatics and systems biology, etc. Meanwhile, he is a lecturer with College of Information Engineering, Shandong Youth University of Political Science, Ji'nan, Shandong, China. Now he is studying for Ph.D. degree at Dongbei University of Finance and Economics.

Copyright ⓒ 2015 SERSC

Copyright ⓒ 2015 SERSC