A Study of Detector Generation Algorithms Based on Artificial Immune in Intrusion Detection System

Jinyin Chen [1], Dongyong Yang [1], and Matsumoto Naofumi[2]
[1]Software Department, Zhejiang University of Technology, Hangzhou, China
[2]Information Department, Ashikaga Institute of Technology, Ashikaga, Japan
chenjinyin@163.com

Abstract: - Detector plays an important role in self and non-self discrimination for intrusion detection system, which makes detector generation a kernel algorithm for artificial immune system. In this paper, firstly current used binary matching rules are listed, characteristics of which are analyzed. And detector generation algorithm is divided into three main processes, including gene library, negative selection and clone selection. Evolution for gene library is explained based on the gene library theory. Several new methods are adopted to improve the performance of NSA, and finally cooperative co-evolution detector generation model is constructed which is a novel structure for intrusion detection system. This paper is aimed for researchers to focus problems on three main ideas concluded in last chapter.

Key-Words: - Detector generation algorithm, artificial immune, intrusion detection system, NSA, CSA, GA, Co-operation cooperative, maturation algorithm

1 Introduction of detector generation based on artificial immune systems

With the development of network, traditional network protection system cannot meet the demand of intrusion detection. Artificial immune system is an emergent bio-inspired research field [1-3], which has been proved efficient for network intrusion detection especially for anomaly detection and related applications [4-6].

Detectors play an important role in Immune Detection System (IDS), which makes the algorithm for detector generation and maturation especially significant. More than twenty papers have brought up novel detector generation algorithms in various ways, most of which are aimed at increasing TP rate and maintaining low FP rate. However most of them are still of large time complexity and space complexity.

In the following segments, matching rules are listed and analyzed, based on which gene library, Negative Selection Algorithm (NSA) and Clone Selection Algorithm (CSA) are summarized, including specific advantages and shortcomings of various novel techniques. Based on the current techniques, detector generation algorithm scheme based on cooperative co-evolution is come up to solve the problem of large time and space complexity, which is also suitable for various kinds of anomaly intrusions.

2 Binary matching rules for artificial immune system

In artificial immune system, affinity between

antibodies and antigens are calculated according to different models [7]. The detection capability of a detector mainly depends on the affinity between the detector and antigen, which makes the affinity model an important role in artificial immune system. Currently several models are adopted based on binary coded antibody and antigen.

(1) Euclidean distance

If the coordinates of an antibody are given by $< ab_1, ab_2, \ldots, ab_L >$ and the coordinates of an antigen are given by $< ag_1, ag_2, \ldots, ag_L >$ then distance(D) between them is presented in Equation (1).

$$D = \sqrt{\sum_{i=1}^{L} (ab_i - ag_i)^2} \quad (1)$$

Shape-spaces that use real-valued coordinates and that measure distance is the form of Equation (1) are called Euclidean shape-spaces. It is suitable for real-valued coordinates, however if coordinates length is much longer than regular situation, calculating distance costs much longer time. Besides in condition of larger real-value, Euclidean distance can be very complex to calculate. As a result Euclidean distance is only adopted in simple real-valued case.

(2) Manhattan distance [8]

Manhattan distance is calculated as Equation (2). Shape-spaces that use real-valued coordinates are called Manhattan shape-spaces.

$$D = \sqrt{\sum_{i=1}^{L} |ab_i - ag_i|} \quad (2)$$

Although no report of it has yet been found in the literature, the Manhattan distance constituted an interesting alternative to Euclidean distance, mainly for parallel implementation of algorithms based on the shape-space formalism.

(3) Hamming distance

In Hamming shape-space antigens and antibodies are represented as sequences of symbols. Such sequences can be loosely interpreted as peptides. The mapping between sequence and shape is not fully understood, but in the context of artificial immune systems, they are assumed to be equivalent. Equation (3) depicts the Hamming distance measure.

$$D = \sum_{i=1}^{L} \delta, where \delta = \begin{cases} 1 & if \quad ab_i \neq ag_i \\ 0 & otherwise \end{cases} \quad (3)$$

Hamming distance is only applied for binary-coded antibody and antigen. It has an obvious advantage is that hamming distance isn't a complex value which is in range of *[0, L]*.

(4) Rogers & Tanimoto distance [8]

Roger & Tanimoto distance matching rule, a variation of the Hamming distance, produced the best performance and defined as follows: given an antibody $< ab_1, ab_2, \ldots, ab_L >$ and an antigen $< ag_1, ag_2, \ldots, ag_L >$, the two matches with each other if and only if formula (4) holds.

$$\frac{\sum_i \overline{ab_i \oplus ag_i}}{\sum_i \overline{ab_i \oplus ag_i} + 2\sum_i ab_i \oplus ag_i} \geq r \quad (4)$$

This distance, inspired by biology, is widely employed to self and non-self discrimination in human body. It is obvious that the calculation of this distance takes more steps to complement which limits its application area.

(5) r-contiguous distance

The first version of the NSA [9] used binary strings of fixed length, and the matching between antibody and antigen is determined by a rule called r-contiguous matching. The binary matching process is defined as follows: given an antibody $< ab_1, ab_2, \ldots, ab_L >$ and an antigen $< ag_1, ag_2, \ldots, ag_L >$. The antibody matches the antigen if and only if $\exists i \leq L - r + 1$ such that $ab_j = ag_j$ for $j = i, \ldots i + r - 1$ holds.

(6) r-chunk distance

This matching rule subsumes r-contiguous

matching, that is, any r-contiguous antigens and antibodies. The r-chunk matching rule is defined as follows: given an antibody $< ab_1, ab_2, \ldots, ab_L >$ and antigen $< ag_1, ag_2, \ldots, ag_L >$, with $m \le n$ and $i \le L - m + 1$, the antibody matches the antigen if and only if $ab_j = ag_j$ for $j = i, \ldots, i + m - 1$ holds.

3 Detector generation algorithms

Detector generation algorithm is extremely important for artificial immune system [9-11] and various improved detector generation algorithms have been brought up ever since negative selection algorithm put forward. According to the principle of detector generation, the process could be divided into three processes and most of the betterments for generation algorithm are a part of the process.

3.1 Gene library evolution

Gene library is firstly used for generating initial premature detectors and usually in static form, in other word static gene library is unchangeable through the whole detector mature process [13]. In static gene library algorithm, each part of detector code is derived from specific part of the gene library as shown in figure 1.

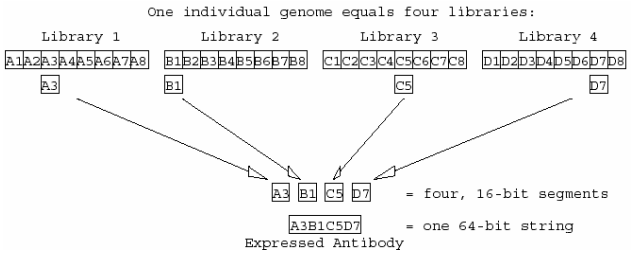

Fig. 1. Static gene library for detector generation

In current research, J' Kim found it is necessary to evolve the gene library during the detector generation process [13]. Based on the availability analysis of gene library evolution, there are two methods employed by the currently available AIS in order to evolve their gene libraries. The first approach directs gene library evolution through the Baldwin effect and the second approach allows provision of direct feedback from learning results to a gene library. So the clone selection was extended by eliminating memory detectors and evolving gene library. This extended system has higher TP and lower FP compared to traditional clone selection algorithm [13]. However according to the experiments results, it costs much more CPU time, in other word, it has larger time complexity which needs to be improved immediately.

3.2 Negative selection algorithm

Negative selection algorithm, firstly brought up by Forrest, is successful for self and non-self discrimination in detector maturation [12-13]. However the traditional NSA has large time cost complexity and space complexity, aiming at which various techniques are adopted to improve the performance. Four types of mended NSAs are listed as follows.

3.2.1 Negative selection with detector rules (NSDR)

This algorithm uses a genetic algorithm to evolve detectors with a hyper-rectangular shape that can cover the non-self space. These detectors can be interpreted as *If-Then* rules, which produce a high-level characterization of the self/non-self space. The initial version of the algorithm [14] used a sequential niching technique to evolve multiple detectors. And NSDR is an improved version of the algorithm using deterministic crowding as the niching technique. The algorithm was applied to detect attacks in network traffic data.

GA-based NSDR is come up in [7] genetic algorithm is used to evolve rules to cover the non-self space. The goodness of a rule is determined by various factors: the number of normal samples that it covers, its area, and the overlapping with other rules. This is a multi-objective, multi-modal optimization problem. A niching technique is used with GA to generate different rules. Experiments results testified that positive characterization appears

to be more precise, but it requires more time and space resources.

3.2.2 Negative selection with fuzzy detector rules (NSFDR)

NSFDR is extended NSDR algorithm with fuzzy rules. This improves the accuracy of the method and produces a measure of deviation from the normal that does not need a discrete division of the non-self space.

3.2.3 Real-valued negative selection (RNS)

This algorithm takes as input a set of hyper-spherical antibodies (detectors) randomly distributed in the self/non-self space. The algorithm applies a heuristic process that changes iteratively the position of the detectors driven by two goals: to maximize the coverage of the non-self subspace and to minimize the coverage of the self samples. This algorithm was combined with a hybrid immune learning algorithm [15] and applied it to different data sets.

3.2.4 Randomized real-valued negative selection (RRNS)

Like the RNS algorithm, the goal of this algorithm is to cover the non-self space with hyper-spherical antibodies. The main difference is that the RRNS algorithm has a good mathematical foundation that solves some of the drawbacks of the RNS algorithm. Specifically, it can produce a good estimate of the optimal number of detectors needed to cover the non-self space, and maximization of the non-self coverage is done through an optimization algorithm with proved convergence properties. The algorithm is based on a type of randomized algorithms called Monte Carlo methods. Specifically, it uses Monte Carlo integration and simulated annealing.

However there are issues that prevent NSA from being applied more extensively, as scalability, low-level detector presentation, sharp distinction exists between the normal and abnormal and other immune-inspired algorithms use higher level representation (e.g. real valued vectors).

3.3 Clone selection algorithm

Kim and Bentley adopt such a strategy as a clone selection operator with negative selection operator for network intrusion detection. They conclude that the embedded negative selection operator plays an important role. Yajing Zhang proposed a niching colon selection genetic algorithm (NCSA). The main idea is that for those valid detectors generated, if a bit or several bits are changed, their fitness score will not vary in a large extent. Thus more valid detectors will be obtained in a short time. The flow chart is shown as follows.

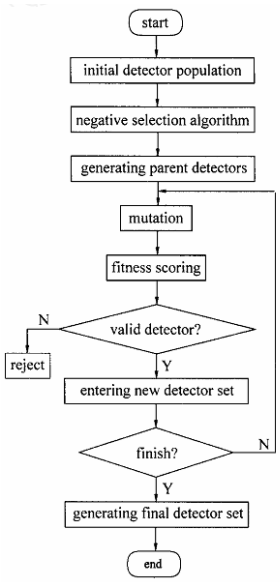

Fig. 2. Flow chart of NCSA

4 Multi-detectors cooperative co-evolution based on evolution algorithms

A different approach based on a cooperative and co-evolution model of the immune system is brought up recently [16]. Genetic algorithms are very successful for optimization problems even though in most of the cases they may not lead to the best answer. A genetic algorithm repeatedly modifies a population of individuals while seeking for the best

possible choice. An extended approach used in here, is a cooperative co-evolution genetic algorithm method. Co-evolution is the simultaneous evolution of two or more genetically distinct populations with coupled fitness landscapes.

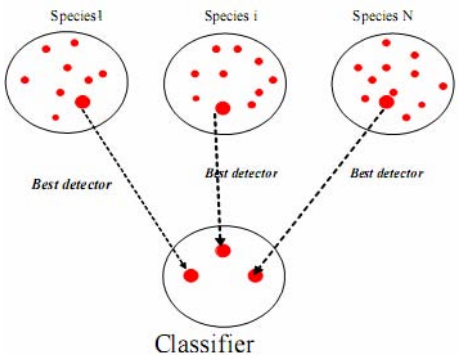

Fig. 3. The cooperative co-evolutionary based detectors generation method

As shown the cooperative co-evolution includes some subcomponents which are represented as genetically isolated species and evolves in a parallel mechanism. Individual member from each species collaborate with other members and improves its fitness according to specific objective function.

The cooperative co-evolution immune system consists of several species and each species contains several detectors, collection of a selected detector in each species forms a detector set. Each species represents only a partial solution, i.e. a collection of similar detectors. The representation of individual as follows.

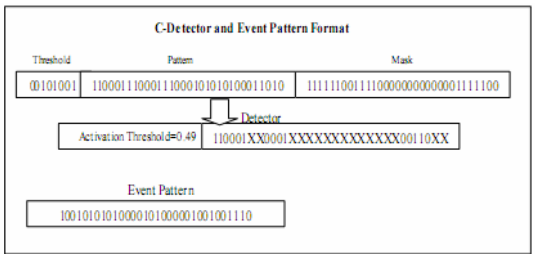

Fig. 4. Detector and event pattern representation

Three segments have different functions explained in paper [16]. The following relations present, T_a as a set of a_i bit string of attacking traffic and T_b as a set of b_i bit string of normal traffic (non-attack). The goal is to find M that is set that matches as strongly as possible to T_a and T_b. When a detector evolves in one species, it collaborates with representatives of all other species in the system. The selected representatives have the best fitness compared to other members with target set. The individuals from multiple different species collaborate to find the optimum detector for the target set. As a result, the population would converge into a collection of non-similar detector which guarantees the diversity of detectors. Applications of this method on Jini grid platform has been improved efficient [16].

5 Conclusions

The summary of detector generation methods can attract computer scientists research on immune system approaches to intrusion detection. An increasing amount of work has been published on this topic recently and here we have collated the algorithms used, the development of detector generation for immune system. The paper focused on providing an overview of detector generation in immune system for intrusion detection system for researchers to identify suitable intrusion detection research problems.

Through careful examination of literature presented in this paper, one can conclude that current methods for detector generation still have much room to grow and many areas to explore. And the research in this area has shown a clear focus on three major ideas:

1. Methods inspired by gene library evolution that employ gene evolve with detector maturation [13].
2. The negative selection paradigm combined with current new techniques such as niching, fuzzy, reinforcement learning, vector machines and cooperative co-evolution [9-12]. Detector set may evolve based on GA to mature optimized detectors for IDS.
3. Framework of detector generation and maturation for immune system, with younger methods based on alternative approaches still being developed [17-18].

References:

[1] L.N. de Castro, J. I. Timmis, Artificial immune systems as a novel soft computing paradigm, Soft computing, 2003, 7(8): 526-544.

[2] Tao Li, Xiaojie Liu, Hongbin Li, A new model for dynamic intrusion detection, CANS 2005, LNCS 3810, pp. 72-84, 2005.

[3] Jungwon Kim, Peter J. Bentley, Uwe Aickelin, Julie Greensmith, Gianni Tecesco, Jamie Twycross, Immune system approches to intrusion detection a review, Proceeding international conference on artificial immune systems [C]. Catania, Italy, 2004.316-329.

[4] Yanxin Wang, Smruti Ranjan Behera, Johnny Wong, Guy Helmer, Vasant Honavar, Les Miller, Robyn Lutz, Mark Slagell, Towards the automatic generation of mobile agents for distributed intrusion detection system, The journal of systems and software 79(2006) 1-14.

[5] Latifur Khan, Mamoun Awad, Bhavani Thuraisingham, A new intrusion detection system using support vector machines and hierarchical clustering, The VLDB journal, volume 16, issue 4, October 2007: 507-521.

[6] Fabio A. Gonzalez, Dipankar Dasgupta, An immunogenetic technique to detect anomalies in network traffic, Proceeding of the international conference genetic and evolutionary computation (GECCO), 2002.

[7] F Gonzalez, A study of artificial immune systems applied to anomaly detection [D]. PhD dissertation, university of Memphis, 2003.

[8] Leandro Nunes de Castro, Fernando Jose Von Zuben, Artificial immune systems: part one-basic theory and applications, Theory and Applications. Technical Report-RT DCA, 1999 (1): 89.

[9] Yajing Zhang, Chaozhen Hou, Fang Wang, Limin Su, A niching negative selection genetic algorithm for self-nonself discrimination in a computer, proceedings of the first international conference on machine learning and cybernetics, Beijin, 4-5 November 2002.

[10] Yingjie Yang, Fanyuan Ma, Antropy-based unsupervised anomaly detection pattern learning algorithm, Journal of Harbin Institute of Technology (New Series), Vol. 12, No.1, 2005.

[11] Lianhua Zhang, Guanhua Zhang, Intrusion detection using rough set classification, Journal of Zhejiang University Science, 2004 5(9):1076-1086.

[12] Boping Qin, Xianwei Zhou, Grey-theory based intrusion detection model, Journal of Systems Engineering and Electronics, Vol. 17, No.1, 2006, pp: 230-235.

[13] J.Kim, P.J.Bentley, A model of gene library evolution in the dynamic clonal selection algorithm, Proceedings of the first international conference on artificial immune systems (ICARIS) Canterbury, 2002: 57-65.

[14] D. Dagupta, F. Gonzalez, An immunity-based technique to characterize intrusions in computer networks, ieee transactions on evolutionary computation, vol. 6, no. 3, pp.281-291, June 2002.

[15] F. Gonzalez, D. Dasgupta, R. Kozma, Combining negative selection and classification techniques for anomaly detection, in proceedings of the 2002 congress on evolutionary computation CEC2002, D.B. Fogel, M. A. El-Sharkawi, X. Yao, G. Greenwood, H. Iba, P. Marrow, and M. Shackleton, Eds. USA: IEEE Press, May 2002, pp: 705-710.

[16] Mohammad Reza Ahmadi, Davood Maleki, A co-evolutionary immune system framework in a grid environment for enterprise network security, SSI'2006, 8th International symposium on systems and Information security Sao Jose dos Campos, Sao Paulo, Brazil, November 08-10, 2006.

[17] Mohamed Abou-El-Nasr, Mohamed Azab, Mohamed Rizk, FPGA-based hardware implementation for network intrusion detection system system rule matching module, WSEA

transactions on circuits and systems, Issue 1, Vol 5, Jan. 2006, pp: 195-201.

[18] Wu Yang, Yong-Tian Yang, Using inductive reasoning for network intrusion detection, WSEAS transactions on systems, Issue 11, Vol 4, Nov 2005, pp:1877-1883.