http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (3): 514-517, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.514.517

Optimized Multilevel Immune Learning Algorithm in Abnormal Detection

Shan Chen
School of Media Communications, LinYi University, LinYi 276002, China

Abstract: Artificial immune algorithm is a kind of intelligent learning algorithm which simulates the biology immunity systems and is widely applied in anomaly detection. There are many techniques for anomaly detection. Among these approaches, Multilevel Immune Learning Algorithm (MILA) is a prominent one due to better discrimination ability and a higher detection rate. However, the limitation of MILA is the T suppressor (Ts) detector generation mechanism which fails to match the coverage of the high-dimensional self space well. In comparison with MILA, an optimized multilevel immune learning algorithm is presented. The optimized algorithm takes a novel variable-recessive threshold model into the process of detector generation and achieves a better coving effect. The experimental results indicate that the optimized algorithm is useful to the anomaly detection with very good detection rate and lower false alarm rate.

Key words: Negative selection, multilevel immune learning, artificial immune system, anomaly detection

INTRODUCTION

Artificial Immune System (AIS) has been applied to many applications, such as network security and data fusing, etc. Most existing research about AIS has focused on the algorithms and their applications (Gonzalez *et al*., 2003; Luo *et al*., 2005, 2006; Zhang *et al*., 2005). To resolve the conflicts of different anomaly detectors, the selective Conserved Self Pattern Recognition Algorithm (CSPRA) applies both domain knowledge and randomization technique and achieves better performance (Yu and Dasgupta, 2009, 2011). This study focuses the important anomaly detection algorithm originated from the Negative Selection Algorithm (NSA) that is the first AIS algorithm (Forest *et al*., 1994). However, NSA is a single level approach that always fails to distinguish self space from the nonself space in the problem space due to the variation of self space and the changing problem space. The Multilevel Immune Learning Algorithm (MILA) utilizes immunological mechanisms including T helper (Th) cell, T suppressor (Ts) cell, B cells and APCs (Antigen Presentation Cells). MILA consists of four phases: Initialization phase, Recognition phase, Evolutionary phase and Response phase. Initialization phase is to make Th and B detectors cover the nonself area and make Ts detectors cover the boundaries of the self space. A multi-level recognition is performed by these three kinds of detectors in the Recognition phase. Evolutionary phase is an evolutionary process that makes the activated detectors more efficient for detecting nonself. The response phase prevents the recognized anomalies. Since the Ts detector is significantly to cover the self space and is an important factor to the false alarm rate. However, the generation mechanism of Ts detector in MILA is not suitable to the high-dimensional detection vector and fails to generate enough detectors. In this paper we provided an optimized Multilevel Immune Learning Algorithm to solve the limitations in MILA. The experiments can prove that the optimized version of MILA will be success to apply into various anomaly detection applications.

ANALYSIS OF MILA

In general, the self/non-self space used by artificial immune algorithms corresponds to an abstraction of a specific problem space. MILA uses real-valued vectors or strings with Euclidean distance rule. It has three different detectors and is better than the NSA in the covering the self space aspect.

Each element in the problem space of MILA is a real-valued string that its length is L. Additionally, Euclidean distance is used in matching processes. The problem space can be represented by one dimensional space. Let r0 be the distance threshold used by Euclidean distance rule and PM(r0) the probability that a given string matches with a random string in the one dimensional space. The probability PM(r0) is the ratio of the volume in one dimensional sphere with radius r0 and the whole space of U. The latter is 1 because the space corresponds to the set [0.0, 1.0]. The condition denoted as Conl in one dimensional space is as follows:

$$\mathrm{Con}_l : \sum_{i=1}^{l} x_i^{\,2} \le r_0^{\,2} (r_o > 0), \mathrm{PM}(r_o) = \int \underset{\mathrm{con}_l}{\int} \int dx_1 dx_2 \ldots dx_l \tag{1}$$

This can be deduced to:

$$\mathrm{PM}(r_o) = 2\pi^{\frac{l}{2}} \Big/ \Gamma\left(\frac{l}{2}\right) \times r_0{}^{l} / l \qquad (2)$$

where, Γ(1/2) is the gamma-function. The probability PM(r0) in MILA is also the death rate of immature detectors while meeting the first training vector. By means of this probability, some problems are found in MILA: (1) The calculated probability excludes the possibility of applying MILA with so small thresholds in high-dimensional space or using real-value representation with high length. (2) The Th detectors seem to work well in lower-dimensional space, but the number of Ts detectors is not large enough for detection. It is due to the fact that a small amount of Ts detectors is not able to cover the boundary of the self.

OPTIMIZED MULTI-LEVEL IMMUNE LEARNING ALGORITHM (OMILA)

Aiming at improving the Ts number in high-dimensional space and covering the boundaries of self space better, an optimized Multi-Level Immune Learning Algorithm is proposed.

In MILA, R2 represents the Ts detector set. The Ts detectors can cover the self in the low dimensional space well at initialization phase. However, they fail to match the coverage of the high-dimensional self space due to the fact that it is difficult to generate enough detectors.

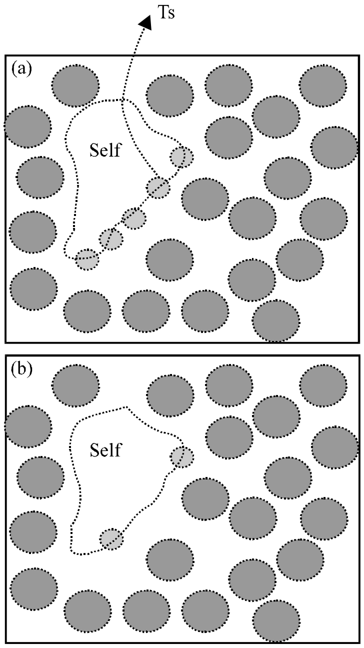

Fig. 1(a-b): Coverage of Ts detectors, (a) Ideal state and (b) Unideal state

As shown in Fig. 1, the more detectors in set R2, the more precise description of the self space. It is obviously that 5 detectors in Fig. 1a make more qualified for covering the self space than 2 detectors in Fig. 1b. The number of Ts detectors plays a critical role on false alarm rate.

The threshold T2 used in initialization phase of MILA is an important parameter to generate the Ts detector. In MILA, a stable threshold value T2 is used in initialization phase. While a novel variable-recessive threshold model is taken into the process of T2 generate in OMILA. Since the threshold T2 can be constantly recessed through the increment of training sample, the number of Ts for detection is insured. The threshold calculation of Ts is defined as following:

$$T_s(t_i) = e^{-\alpha t_i} \qquad (3)$$

EXPERIMENTS AND RESULTS

Our current experiments are to test MILA and OMILA and find out other features of anomaly detection using real-valued representation. The experiments include two parts: the stimulation experiment and the real data set to apply experiment.

The simulation experiment are studied by using the time series data sets.

As shown in Fig. 2, x-axis represents the scale of training sample in initialized phase. The generation trend of Ts in MILA vs. OMILA is studied with the growth of iteration number. It is shown that when the training sample accumulates to 10000 in 6-dimensional space, the number of Ts detectors seems to be saturated. Since, the threshold is adjusted to variable-recessive model in OMILA, the number of Ts almost increases linearly.

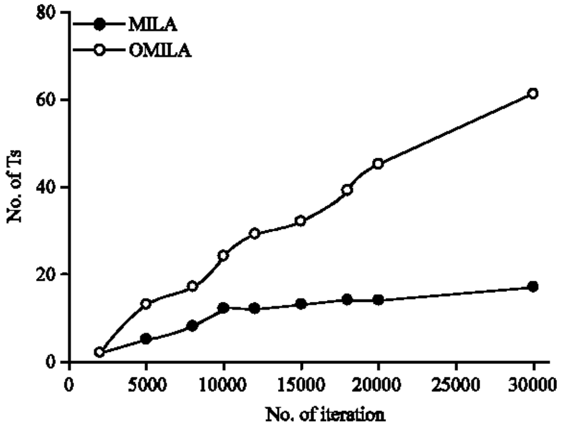

Fig. 2: Comparison of Ts number between MILA and OMILA

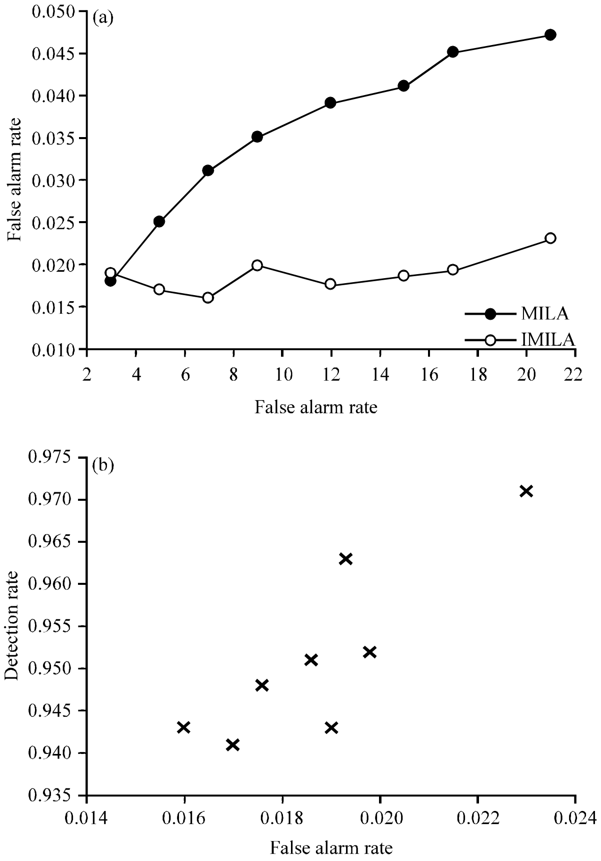

Fig. 3(a-b): Comparing MILA and OMILA in DARPA 99, (a) False alarm rate trend and (b) Detection rate trend

The practice experiment uses the DARPA 99 network intrusion detection system test data set. There are large numbers of records from five weeks of network traffic, where each one is labeled as either normal or as an attack. The 1st week and the 3rd week contained with normal network traffic while the other three weeks filled with several attack behaviors. The 1st week data set is adopted for training phase and the 2nd week data set for testing phase. Three fields are extracted from the data set: source IP address, destination IP address and destination port. Since the IP address can be split into 4 independent bytes, a 5- dimensional space is structured. For instance, the connection from IP equal to 201,197,91,234 to port 80 can be described as a five-tuple. Being preprocessed, the total number of training samples is 16000 and the number of testing sample reaches to 24297, among which 11 attack behaviors were related.

As shown in Fig. 3a, the false alarm rate of MILA gradually increases with the increment of the Detector Vector number. Figure 3b shows the detection rate of OMILA under the different false alarm rate conditions. The best detection rate of OMILA in our experiments is about 0.97 with the worst false alarm rate. Furthermore, when the false alarm rate is lower than 0.02, the detection rate of OMILA is about 0.945. This experimental result demonstrates that OMILA has better detection performance in practice.

CONCLUSIONS

According to the results of our experiment, detection rate and false alarm rate are affected not only by the parameters of detectors but also by the clustering situations of self/nonself. The main advantage of IMILA is to define the profile of self subspace well. That led very good detection rate and false alarm rate. Also the new algorithm has its limitations. Finding new methods to generate detectors, studying relationship between the clustering of self set and the parameters of detectors and applying OMILA to other applications are the main directions of our future work.

ACKNOWLEDGMENTS

We would like to thank the support of National Nature Science Foundation of China (No. 60970012), the Innovation Program of Shanghai Science and Technology Commission (No. 09511501000) and Shanghai leading academic discipline project (S30502).

REFERENCES

Forest, S., S. Hofmeyr, A. Somayaji and T. Longstaff, 1994. Self-nonself discrimination in a computer. Proceedings of the Symposium on Computer Security and Privacy, May 16-18, 1994, IEEE Computer Society, USA., pp: 202-202.

Gonzalez, F., D. Dasgupta and L.F. Nino, 2003. A randomized real-valued negative selection algorithm. Proceedings of the 2nd International Conference on Artificial Immune Systems, September 1-3, 2003, Edinburgh, UK., pp: 261-272.

Luo, W., J. Wang and X. Wang, 2005. Evolutionary negative selection algorithms for anomaly detection. Proceedings of the 8th Joint Conference on Information Sciences, July 21-26, 2005, Salt Lake City, UT., USA., pp: 440-445.

Luo, W., X. Wang, Y. Tan and X. Wang, 2006. A novel negative selection algorithm with an array of partial matching lengths for each detector. Proceedings of the 9th International Conference on Parallel Problem Solving from Nature, September 9-13, 2006, Reykjavik, Iceland, pp: 112-121.

Yu, S. and D. Dasgupta, 2009. An empirical study of conserved self pattern recognition algorithm: Comparing to other one-class classifiers and evaluating with random number generators. Proceedings of the World Congress on Nature and Biologically Inspired Computing, December 9-11, 2009, Coimbatore, Tamil Nadu, India, pp: 403-408.

Yu, S. and D. Dasgupta, 2011. An effective network-based intrusion detection using conserved self pattern recognition algorithm augmented with near-deterministic detector generation. Proceedings of the IEEE Symposium on Computational Intelligence in Cyber Security, April 11-15, 2011, Paris, France, pp: 17-24.

Zhang, H., L. Wu, Y. Zhang and Q. Zeng, 2005. An algorithm of r-adjustable negative selection algorithm and its simulation analysis. Chin. J. Comput., 28: 1614-1619.