Communications and Network, 2012, 4, 41-47
http://dx.doi.org/10.4236/cn.2012.41006 Published Online February 2012 (http://www.SciRP.org/journal/cn)

An Improved Artificial Immune System-Based Network Intrusion Detection by Using Rough Set

Junyuan Shen[1], Jidong Wang[1], Hao Ai[2]
[1]RMIT University, Melbourne, Australia
[2]Nanjing University of Posts and Telecommunications, Nanjing, China
Email: junyuan.shen@student.rmit.edu.au, jidong.wang@rmit.edu.au, aihao_beibei@163.com

Received October 26, 2011; revised November 27, 2011; accepted December 10, 2011

ABSTRACT

With the increasing worldwide network attacks, intrusion detection (ID) has become a popular research topic in last decade. Several artificial intelligence techniques such as neural networks and fuzzy logic have been applied in ID. The results are varied. The intrusion detection accuracy is the main focus for intrusion detection systems (IDS). Most research activities in the area aiming to improve the ID accuracy. In this paper, an artificial immune system (AIS) based network intrusion detection scheme is proposed. An optimized feature selection using Rough Set (RS) theory is defined. The complexity issue is addressed in the design of the algorithms. The scheme is tested on the widely used KDD CUP 99 dataset. The result shows that the proposed scheme outperforms other schemes in detection accuracy.

Keywords: Intrusion Detection; Negative Selection; Artificial Immune System; KDD CUP 99

1. Introduction

Driven by the rapid growth of the computer network technologies, the security of the computer and network information is becoming increasingly important. The appearance of the new access technologies and the advanced devices has increased the possibilities of malicious attacks or service abuse by various hackers. Also, with the appearance of multimedia services (video, audio, image, text, etc.), a faster, short-delay anti-virus system is required. However, the traditional passive defence mechanisms like encryptions and firewalls cannot fully meet current security requirements. Therefore, a special attack and misuse detection system is needed. The intrusion detection system (IDS) is such a system, which is composed by a series of devices and software applications to monitor network activities in order to protect the system from malicious activities.

The general IDS detect unauthorized users or processes by comparing a user's behaviour with the user's profile. Two approaches, misuse detection and anomaly detection, are usually used in the intrusion detection process. The misuse detection is used to detect the intrusion when the behavior of the system matches with any of the intrusion signatures in the user profile. And the anomaly detection, which is also called as outlier detection [1], is used to detect the intrusion when the given data set does not match with the established normal behavior.

Various techniques have been used for building IDS, like Support Vector Machines (SVM) [2], Multivariate Adaptive Regression Splines (MARS) [3], and Linear Genetic Programming (LGP) [4], etc. Some of them give good performance in specific attack areas, while they might not detect other attacks well. In recent years, bio-inspired algorithms have been studied and applied in intrusion detection [5] aiming for better performance. Algorithms such as Genetic Algorithm (GA), Artificial Neural Networks (ANN) and Artificial Immune Systems are widely studied. AIS is a relatively new comer among them. The concept of AIS was proposed in mid 1980s. Farmer, Packard and Perelson [6], Bersini and Varela's [7] work have started the area. AIS has not become a subject of its own untill mid 90s. It has been defined as: "Adaptive systems inspired by theoretical immunology and observed immune functions, principles and models, which are applied to problem solving." by Castro and Timm is [8]. Early works of AIS based IDS can be found in [5]. A Multilayer IDS using AIS was proposed by Dasgupta [9] in order to provide systematic defense. These AIS based IDS have achieved good detection results. However, their computing complexity is quite high due to the complicated feature comparing. While, for IDS, responding time is also an important issue. The more complexity the system, the more computing time and the longer responding time will be. Large parameter set in IDS can increase the detection accuracy. However, the more parameters using, the more complex the system

will be. So, the trade off between the complexity and the accuracy is a challenge. Ourstudy on AIS based IDS is to further improve its detection accuracy while keeping a low algorithm complexity.

In this paper, an improved AIS based intrusion detection system with Rough Set feature selection algorithm is presented. The anomaly detection in the system is set up based on AIS negative selection algorithm. And the feature selection algorithm is used to reduce the complexity of the system. The artificial immune system and the negative selection algorithm are introduced in Section 2. The AIS based IDS is presented in Section 3. Our experiment and results are illustrated in Section 4. Section Vdraws a conclusionand somefuture works are discussed.

2. Artificial Immune System

Artificial Immune System (AIS) applies to various areas of researches that attempt to build a bridge between immunology and engineering by using the techniques of mathematical and computational modeling of immunology.

The origin of AIS is rooted in the early theoretical work of J. D. Farmer, N. H. Packard, A. S. Perelson [10, 11], F. Varela, A. Coutinho, B. Dupire, and N. Vaz [12]. It was first proposed in mid 1980s and became a subject of its own in mid 90s. Originally, AIS aimed to find efficient abstractions of processes in the immune system [13]. By carefully reviewing the efficient natural mechanism, a number of computer scientists proposed artificial immune based computer models to solve various problems ranging from virus detection, fault analyzing to clustering. Two researchers, HuguesBersini and Stephanie Forrest, played an important role in crossing the divide between computing and immunology. Bersini and Forrest did a lot of basic works rooted from immunology and their works formed a solid foundation of the area of AIS. With regards to Bersini, he was focusing on the basic theory of immune network and examining how the immune system maintained its memory and how to build a model to mimic that progress. And for Forrest, she was focusing on the application area of the AIS. She proposed the idea of introducing the immune system into the computer security area by using its ability to distinguish between self and non-self.

Negative selection, which is proposed by Forrest *et al*. [14], is inspired from the negative selection process of the adaptive immune system [8]. The important characteristic of the human immune system is that it can maintain its diversity and generality, and it can detect a large number of antigens by using a small number of antibodies. In order to make it possible, several functions will be processed [15]. One of those functions is to develop the antibodies through the gene library. The gene library will be used in creating thymus cell (T cell) and bone marrow cell (B cell). While creating a new antibody, the gene segments in the gene library will be randomly selected and assembled. As shown in **Figure 1**, large number of antibodies can be generated from combining different genes segments in the gene library.

However, there is a problem due to the full immune response above. Not only responding to harmful antigens, those new generated antibody may also react to self-cells coming from the host. In order to protect the body from self-reactive, the human immune system produces the negative selection.

In the case of an anomaly detection domain, the algorithm prepares a set of exemplar pattern detectors trained on normal (non-anomalous) patterns that model and detectunseen or anomalous patterns [17]. The principle of the negative selection is shown in **Figure 2**.

As shown in **Figure 2**, the basic idea of the negative selection is to generate a selected detector set *D* and use the detector set to distinguish the new data. In the process, a set of detectors *R* will be randomly generated, and all the randomly generated detectors will be compared with each elements of the self-string set *S*. Under certain matching algorithms, if the detector in set *R* fails to match any element in *S*, it will be saved in the detector set *D*, otherwise, it will be rejected.

In the matching process, several algorithms have been proposed to determine the difference between self and

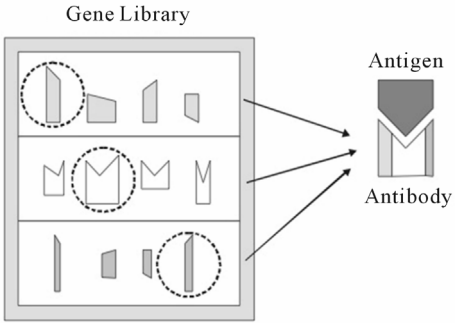

Figure 1. Gene expression process [16].

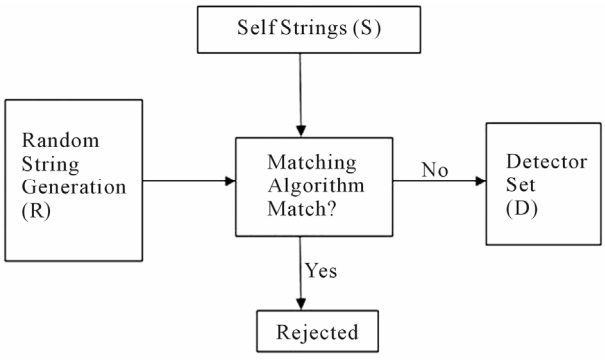

Figure 2. The principle of negative selection.

non-self like Eucliden distance, hamming distance, re-contiguous bit rule algorithm, etc. In this paper, the Manhattan Distance will be used because of its simplicity. The affinity (difference) of the set R and S are related to the distance between them. The definition of the distance is shown as follows

Let $R = <R_1, R_2 \cdots R_m>$, $S = <S_1, S_2 \cdots S_m>$,

Manhattan $D = \sum_{i=1}^{m} |R_i - S_i|$

In the intrusion detection process, for any pattern to be checked, it needs to be compared with all the patterns in the detector set. If it matches to any pattern in the detector set it will be considered as a non-self element, otherwise, it will be considered as self. AIS has been found applications in many areas such as optimization, data analysis, machine learning, pattern recognition, etc and network intrusion detection which is the focus of this paper.

3. AIS Based IDS

Generally, network intrusion detection is based on the examination of monitored network parameters. Different examine algorithms lead to different IDS. The general AIS based IDS [18-20] can be divided into two parts, *i.e.* detector set generation and the live detection. To form the detection set, negative selection algorithm is applied, as discussed in Section 2. In the live detection stage, a monitored network parameter pattern is compared with detectors in the detector set. If it is matched with any detector, then a network intrusion is detected.

A compact and effective detector set can reduce the algorithm computing complexity. For detectors that do not contribute any detection in a period of time, they should be removed or put to a sleeping state. Therefore, all the mature detectors will have a time-to-live (TTL) parameter. Whenever detection is occurred, all detectors' TTLs will be deducted by one except for the detector which detects the intrusion. Its TTL will be reset to the maximum. When a detector reaches its lifetime, *i.e.* its TTL becomes zero; this detector will be become inactive.

1) The definition of immune elements in AIS

Antigens (Ag): numerical character strings with Lelements, where L is the number of features selected from the dataset. Ag contains two subsets that are Self (normal patterns) and Nonself (abnormal patterns):

$$\text{Ag} = \{\text{Self}, \text{Nonself}\} \quad (1)$$

$$\text{Self} \cap \text{Nonself} = \phi \quad (2)$$

Detectors (Antibodies): The Antibodies Ab should have the same number of elements as the antigens Ag. *Ab is expected to be representatives of all Nonself.*

Affinity: The measurement to judge the matching between two patterns. Generally, distance is used to measure the affinity of two patterns. The shorter the distance, the closer these two patterns are in the defined Lspace.

In our project, a normalized Mahhatan distance is used for its simplicity. It is defined as following.

$$D(A,B) = \frac{1}{L}\sum_{i=1}\left|\frac{a(i)-b(i)}{r(i)}\right| \quad (3)$$

where $A = \{a(1), a(2) \cdots a(L)\}$, $B = \{b(1), b(2), \cdots b(L)\}$ are the two patterns to be measured.

$R = \{r(1), r(2), \cdots r(L)\}$, where $r(i)$ represents the range of the ith parameter in the detection feature subset.

Two thresholds are defined. Ta is the threshold, used for detector set generation in the negative selection algorithm. Let $X \in$ Self, Y is a pattern generated randomly, If $D(X,Y) <$ Ta, then A and B are considered matching and B will be rejected. Otherwise B will be added to the detector set Ab. The second threshold, Td, is for live detection. Whenever a live pattern matches any of the patterns in Ab, the alarm will be raised. In our scheme, different Td has been tested to find a trade-off between the attack detection accuracy and false alarm rate.

2) Parameter Quantization

As shown in **Table 1**, the KDD Cup 99 features are in one of the following formats, *i.e.* continuous, discrete, or symbolic. To prepare the parameters in the detection subset for AIS, they should be quantized or normalized. For symbolic features such as protocol_type (3 symbols), service (70 symbols), and flag (11 symbols), they are mapped to numerical values ranging from 0 to N-1 where N is the number of symbols.

4. KDD CUP 99 with Rough Set Theory

The data set used in our experiment is the KDD Cup 99 data set, which is the most widely used data set for network-based intrusion detection. This data set is built based on the data captured in DARPA'98 IDS evaluation program [21]. The data set contains 24 training attack types and 14 additional attack types in the test data only. It has 41 parameters in each data record and the data type is shown as follows:

- 0,tcp,http,SF,181,5450,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0, 8,8,0.00,0.00,0.00,0.00,1.00,0.00,0.00,9,9,1.00,0.00, 0.11,0.00,0.00,0.00,0.00,0.00,normal.
- 0,icmp,ecr_i,SF,1032,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0, 511,511,0.00,0.00,0.00,0.00,1.00,0.00,0.00,255,255, 1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,smurf.

Each parameter in the data string has its own meaning which is shown in **Table 1**. The complexity is so high if all the 41 parameters are used and the responding time of the IDS will be slow. A feature selection is needed to minimize the data set.

Rough Set Theory (RST), first proposed by Polish

Table 1. KDD CUP 99 parameter.

No.	Feature	No.	Feature
1	Duration	2	Protocol_type
3	Service	4	flag
5	src_bytes	6	dst_bytes
7	land	8	wrong_fragment
9	Urgent	10	hot
11	num_failed_logins	12	logged_in
13	num_compromised	14	root_shell
15	su_attempted	16	num_root
17	num_file_creations	18	num_shells
19	num_access_files	20	num_outbound_cmds
21	is_hot_login	22	is_guest_login
23	count	24	srv_count
25	serror_rate	26	srv_serror_rate
27	rerror_rate	28	srv_rerror_rate
29	same_srv_rate	30	diff_srv_rate
31	srv_diff_host_rate	32	dst_host_count
33	dst_host_srv_coun	34	dst_host_same_srv_rate
35	dst_host_diff_srv_rate	36	dst_host_same_src_port_rate
37	dst_host_srv_diff_host_rate	38	dst_host_serror_rate
39	dst_host_srv_serror_rate	40	dst_host_rerror_rate
41	dst_host_srv_rerror_rate		

computer scientist Zdzislaw I. Pawlak, is an extension of conventional set theory that supports approximations in decision-making [22]. It is a mathematical tool for decision support and suits well for the classification of objects. A lot of researches have been focused in the RST-based machine learning and decision area recently. The major advantage of the RST compared with other feature selection algorithm is its simplicity. A minimal rule set can be generated by using RST. That makes Rough Set Theory suitable for real-time decision tasks.

The work of Zhang *et al.* [4] has shown that RST showed high detection accuracy and feature ranking was fast in determining the categories of the attacks in IDS. And Zainal *et al.* [23] has shown that the IDS have performed well by using RST and the six highest rank features by RST were Service, flag,src_bytes, srv_count, dst_host_count, dst_host_srv_rerror_ratein **Table 1**. But unfortunately, the false alarm rate in Zainal's research is relatively high.

In our scheme, an improved rough set theory is introduced. By using the six parameters chosen from the KDD Cup 99 data set, each parameter is associated with a "weight". The weights for the six parameters are different because of the different contributions of these parameters to the system performance. A range of the weights have been tested in our experiment in order to find a suitable one for the AIS based intrusion detection system.

5. Result and Discussion

The raw dataset that we used to generate detectors contains about five million connection records, 700 million bytes. Meanwhile, the testing data we choose contains 300,000 records, and about 45 million bytes. In our scheme, as described in Section IV, different parameters chosen by the Rough set need to give different weight. Before a weight is finalized, an "influence factor" is tested for each parameter.

Originally, an AIS based intrusion detection system is built based on the C++ platform. In the negative selection process, each parameter has a weight of "6" for all the six parameters chosenbased on rough set theory, and the total weight is equal to 36. Then, one parameter will change by the step size of 1 and the other five will change by the step size 0.2, which keep the total weight of the equation 36 unchanged. According to the test data we use in the KDD Cup 99 data set, 239,237 attacks are contained. And the attack detection quantity for different parameters is shown in the **Figure 3**.

As shown in **Figure 3**, by changing the weight of each single parameter and keep the others the same, the attack detection number will change in the meantime. The **Table 2** shows the different attack detection accuracy for each single changed parameter.

As shown in **Table 2**, for the parameter service, src_bytesand dst_host_count, the detection quantity changed more obviously than the other three. To reduce the computing complexity, in our scheme, the other three parameters will keep invariant. In order to find a best parameter weight combination for the rough set theory, the exhaustive method is used. All the combination of the chosen parameters (service, src_bytes and dst_host_count) is tested and the detection accuracy is shown in **Figure 4**.

In **Figure 4**, Series 1 represents the attack detection accuracy, and Series 2 represents the normal detection accuracy. As shown in **Figure 4**, the system detection accuracy shows a significant improvement with different weighting factors of the parameter. The true positive rate (TP rate) can up to 98.25% (with TN rate 99.90%), and the true negative rate (TN rate) up to 99.97 (with TP rate 82.03%).

In general, compared with the original rough set based

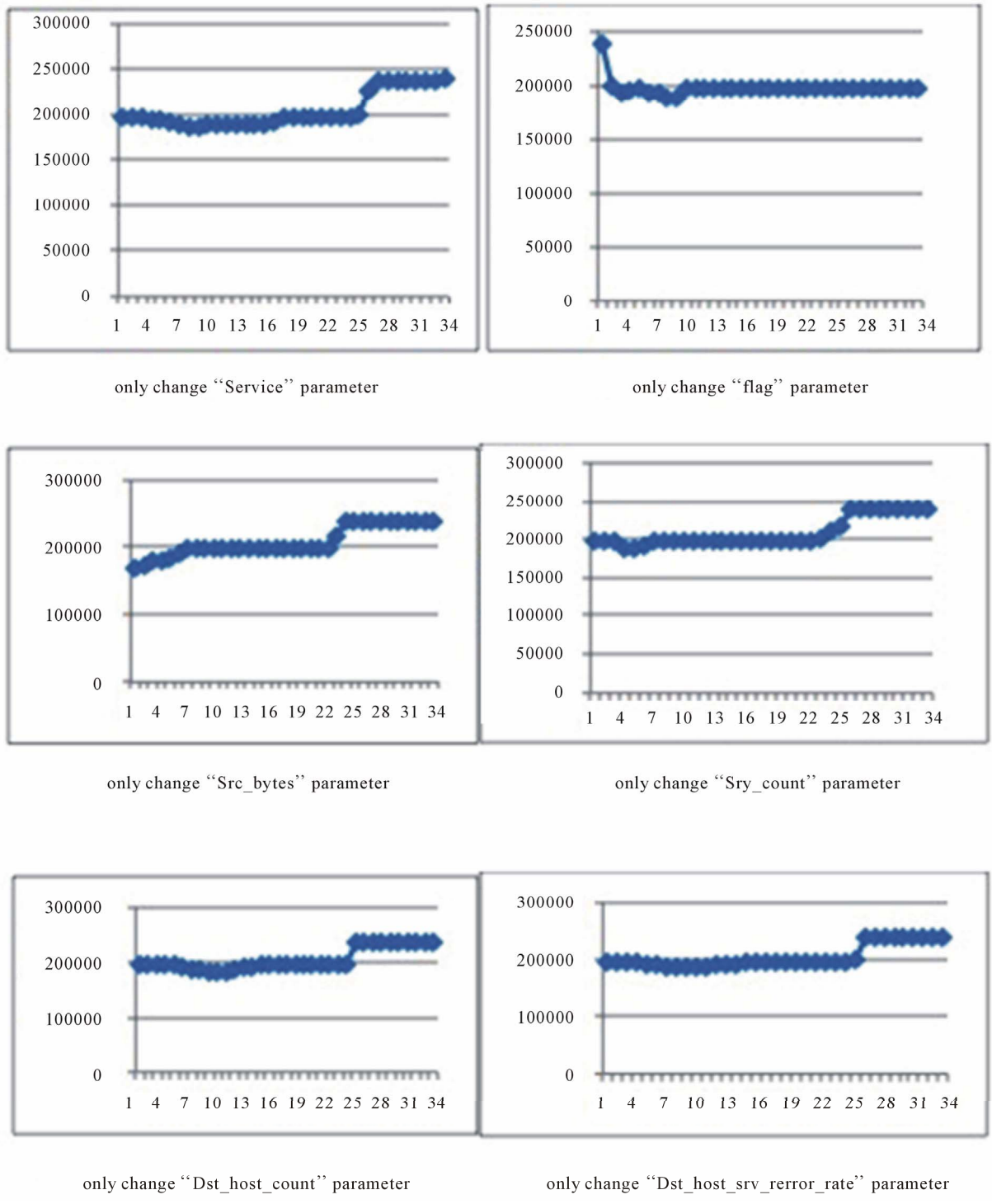

Figure 3. Attack detection quantity.

Table 2. Attack detection range for each single changed parameter.

Parameter Type	Service	Flag	src_bytes	src_count	dst_host_ count	dst_host_srv_ rerror_rate
Attack Detection Range	185,843 to 239,237	187,200 to 237,280	169,610 to 239,237	186,953 to 239,237	181,493 to 239,237	186,076 to 239,237

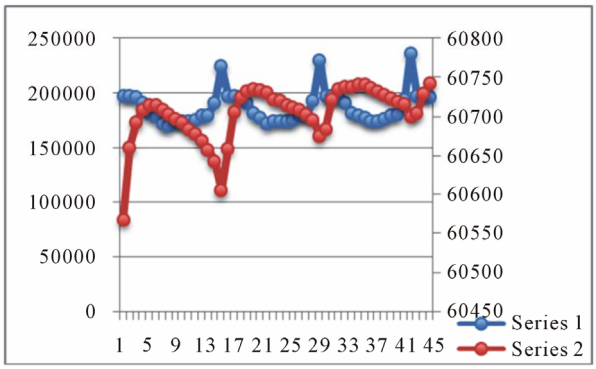

Figure 4. System detection accuracy.

IDS [26], by introducing the "weight" scheme, the proposed IDS provided a better TN rate (above 99% compared with 89.95%), and relatively high TP rate. Fine tuning the algorithm in feature selection and parameter quantization could lead to further improvement on detection rate and complexity.

6. Conclusion and Future Work

In this paper animproved artificial immune systembased intrusion detection system by using rough set is presented. In order to find a best combination of the six parameters chosen by rough set theory, a number of tests have been conducted. The results are compared using the KDD Cup 99 dataset. The rough set theory proposed is aiming to reduce the complexity and maintain the detection performance. The system has shown excellent detection accuracy. The improved rough set theory can significantly increase the TN rate, and keep relatively high TP rate in the meantime. For future work, an adaptive mechanism will be introduced to AIS, so that the detector set will be adaptively updated so that the system can adapt to changes in the network situation. Also more feature selection algorithms can be tried for the performance improvement and verification.

REFERENCES

[1] H.-P. Kriegel, P. Kröger and A. Zimek, "Outlier Detection Techniques (Tutorial)," 13*th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Bangkok, 27-30 April 2009.

[2] N. Cristianini and S. J. Taylor, "An Introduction to Support Vector Machines," Cambridge University Press, Cambridge, 2000.

[3] J. H. Friedman, "Multivariate Adaptive Regression Splines," *Annals of Statistics*, Vol. 19, No. 1, 1991, pp. 1-141. doi:10.1214/aos/1176347963

[4] W. Banzhaf, P. Nordin, E. R. Keller, F. D. Francone, "Genetic Programming: An Introduction on the Automatic Evolution of Computer Programs and Its Applications," Morgan Kaufmann Publishers, Inc., Waltham, 1998.

[5] S. A. Hofmeyr and S. Forrest, "Immunity by Design: An Artificial Immunb System," *Proceedings of the Genetic and Evolutionary Computation Conference*, Morgan Kaufmann, San Mateo, 13-17 July 1999, pp.1289-1296.

[6] J. D. Farmer, N. Packard and A. Perelson, "The Immune System, Adaptation and Machine Learning," *Physica D*, Vol. 2, No. 1-3, 1986, pp. 187-204. doi:10.1016/0167-2789(86)90240-X

[7] H. Bersini and F. J. Varela, "Hints for Adaptive Problem Solving Gleaned from Immune Networks," *Parallel Problem Solving from Nature, First Workshop PPSW* 1, Dortmund, 1-3 October 1990.

[8] L. N. de Castro and J. Timmis, "Artificial Immune Systems: A New Computational Intelligence Approach," Springer, Berlin, 2002.

[9] D. Dasgupta, "Immunity-Based Intrusion Detection Systems: A General Framework," 22*nd National Information Systems Security Conference*, Arlington, 27 August 2011. http://csrc.nist.gov/nissc/1999/proceedings/papers/p11.pdf

[10] J. D. Farmer, N. H. Packard and A. S. Perelson, "The Immune System, Adaptation, and Machine Learning," *Physica D*, Vol. 22, No. 1-3, 1986, pp. 187-204. doi:10.1016/0167-2789(86)90240-X

[11] A. S. Perelson, "Immune Network Theory," *Immunological Reviews*, Vol. 110, No. 1, 1989, pp. 5-36. doi:10.1111/j.1600-065X.1989.tb00025.x

[12] F. Varela, A. Coutinho, B. Dupire and N. Vaz, "Cognitive Networks: Immune, Neural and Otherwise," *Theoretical Immunology*, Vol. 2, 1988, pp. 359-375.

[13] I. R. Tizard, "Immunology: Introduction," 4th Edition, Saunders College Publishing, Philadelphia, 1995.

[14] S. A. Hofmeyr and S. Forrest, "Architecture for an Artificial Immune System," *Evolutionary Computation Journal*, Vol. 8, No. 4, 2000, pp. 443-473. doi:10.1162/106365600568257

[15] J. Kim and P. Bentley, "The Human Immune System and Network Intrusion Detection," 7*th European Conference on Intelligent Techniques and Soft Computing*, Aachen, 1999.

[16] J. Kim and P. J. Bentley, "Negative Selection and Niching by an Artificial Immune System for Network Intrusion Detection," *Proceedings of Late-Breaking Papers at the Genetic and Evolutionary Computation Conference*, Orlando, 13-17 July 1999, pp. 149-158.

[17] S. Forrest, A. S. Perelson, L. Allen and R. Cherukuri, "Self-Nonself Discrimination in a Computer," *Proceedings of the* 1994 *IEEE Symposium on Research in Security and Privacy*, IEEE Computer Society Press, Los Alamitos, 1994, pp. 202-212.

[18] S. Hofmeyr and S. Forrest, "Architecture far ao Ariificial Immune System," *Evolutionary Computation*, Vol. 7, No. 1, 1999, pp. 45-68.

[19] G. Doziert, D. Brownf, J. Hurley and K. Cainf, "Vulnerability Analysis of AIS-Based Intrusion Detection Systems via Genetic and Particle Swarm Red Teams," *Evolutionary Computation*, Vol. 1, 2004, pp. 111-116.

[20] J. W. Kim and P. J. Bentley, "Towards an Artificial Im-

mune System for Network Intrusion Detection: An Investigation of Clonal Selection with a Negative Selection Operator," *Evolutionary Computation*, Vol. 2, 2001, pp. 1244-1252.

[21] R. P. Lippmann, D. J. Fried, I. Graf, J. W. Haines, K. R. Kendall, D. McClung, D. Weber, S. E. Webster, D. Wyschogrod, R. K. Cunningham and M. A. Zissman, "Evaluating Intrusion Detection Systems: The 1998 Darpa Off-Line Intrusion Detection Evaluation," *Discex*, Vol. 2, 2000, p. 1012.

[22] A. Zainal, M. A. Maarof and S. M. Shamsuddin, "Feature Selection Using Rough Set in Intrusion Detection," *IEEE Region* 10*th Conference*, Hong Kong, 14-17 November 2006, pp. 1-4.

[23] L. Zhang, G. Zhang, L. Yu, J. Zhang and Y. Bai, "Intrusion Detection Using Rough Set Classification," *Journal of Zheijiang University Science*, Vol. 5, No. 9, 2004, pp. 1076-1086. doi:10.1631/jzus.2004.1076

