International Journal of Hybrid Information Technology
Vol.8, No.11 (2015), pp.101-112
http://dx.doi.org/10.14257/ijhit.2015.8.11.08

A Novel Negative Selection Algorithm for Recognition Problems

Yuan Tao[1], Min Hu[2] and Yanlin Yu[3]
[1]*Computation Center, Shanghai University, Shanghai 200444, China*
SHU-UTS SILC Business School, Shanghai University, Shanghai University,
[2,3]*Shanghai 201800, China*
[1]*taoyuan103@shu.edu.cn,* [2]*minahu@163.com*

Abstract

In this paper, a novel negative selection algorithm for recognition problems was given. Compared with the traditional negative selection algorithm, a co-stimulation signal was added to start the detectors, which a key factor in immune response. Co-stimulation signal was calculated by the techniques of the statistics and the sliding window, which not only reduced time complexity of algorithm but also improved accuracy of the algorithm. Entropy was adopted to evaluate the density of detectors for optimizing the coverage of nonself area. Experiment results proved high accuracy and efficiency of the proposed algorithm.

Keywords: *Artificia1 Immune System, Negative Selection Algorithm, Entropy*

1. Introduction

In the field of the immunology, people have been continuously trying to establish appropriate theories that could best explain the behavior of the biological immune system, such as the self-nonself (SNS) model [1], the infectious-nonself (INS) model [2] and the danger model [3]. Meanwhile biological immune system provides a source of inspiration for Artificial Immune Systems (AIS) to effectively solve complex problems. AIS are defined as a new computational paradigm based on metaphors of the biological immune system [4]. In AIS, immune inspired systems and technologies are applied to diverse real world applications [5-7].

The field of AIS has obtained a significant degree of success as a branch of Computational Intelligence since it emerged in the 1990s by proposing and developing the Negative Selection Algorithm (NSA) [8]. Since then, the immune techniques have been showing their excellent characteristics for solving recognition problems. Negative selection algorithm (NSA), based on the self-nonself (SNS) model, is one of the earliest and the most successful AIS models and attracts widespread interest in solving real world problems [9-11]. Its successful applications include anomaly detection, fault detection, and network intrusion detection and so on.

Since NSA has been attracting the attention of many researchers, a diverse family of negative selection algorithms has been developed. Though different variations of negative selection algorithms have been frequently proposed, the main characteristics of this method described in [8] still remain, including negative representation of information, distributed generation of the detector set, and one-class classification [12].

Anomaly detection aims at building an appropriate profile of the system that reflects the normal behavior, and at classifying a given state of the system into the normal or abnormal state. The states of the system can be represented by a set of features [13]. NSA is based on the principles of self/nonself discrimination in the natural immune system, which supposes that the immune system distinguish between self and nonself, tolerating self and attacking nonself [14].

ISSN: 1738-9968 IJHIT

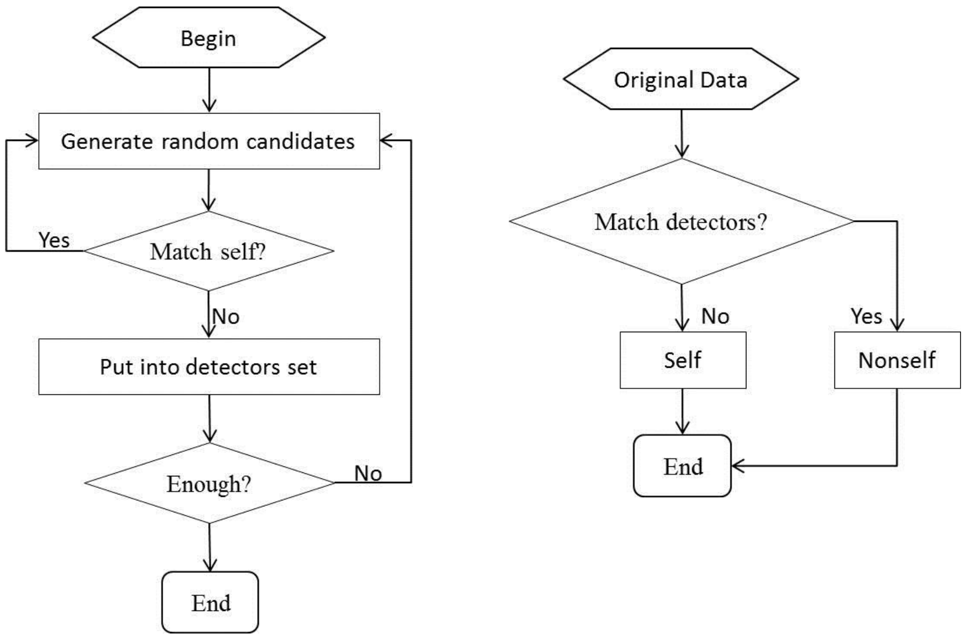
NSA for anomaly detection has been adopted widely because of its biological analogy with body resistance in the immune system provided against agents which causes diseases. NSA opened the door as an important addition within the confines of anomaly detection [15].

A novel negative selection algorithm (NNSA) is proposed in this paper. Compared with the traditional negative selection algorithm, NNSA adopts the techniques of the statistics and the sliding window to calculate the co-stimulation signal which is a key factor for improving the performance of algorithm. The rest of the paper is organized as follows: Section 2 briefly gives related work; Section 3 describes NNSA; Section 4 shows the experimental results; finally, the conclusion is given in Section 5.

2. Paper Preparation

Discrimination between self and nonself is considered one of the major mechanisms in the complex immune system. Artificial negative selection is a computational imitation of self/nonself discrimination [9]. Similarly, recognition problem aims to identify the normal and abnormal states of a system. NSA has been developed and applied in anomaly detection.

In a negative selection algorithm, a collection of detectors, usually called detector set, is generated so not to match any self-samples (training data). The detectors are subsequently used to check whether incoming new data items are normal (self) or not normal (nonself). It is typical regarded as anomaly detection or a one-class classification method because the training data are from normal cases only [16]. Figure 1 gives the flowchart of the traditional negative selection algorithm.



(a) Generation of the Detectors (b) Outline of Recognition

Figure 1. Flowchart of the Traditional Negative Selection Algorithm

NSA has been developed and applied in numerous real world applications for it only requires normal samples for training. However, this character also brings some limitations. As a two-class algorithm, if the self-data is used as a reference, the detectors are located outside of self-region. Therefore, the main limitations are that the self and non-self is too clearly and the detector coverage is a crucial content of the classification performance for NSA. It is still difficult to satisfy the detector coverage [17].

Some efforts have been made to tackle the above limit of NSA. Many groups have been developing alternative detector generation schemes to improve the algorithm performance.

Inspired by the multi-pattern matching algorithm proposed by Aho and Corasick in 1975 [18, 19] presented a fast negative selection algorithm, which constructs a self state graph and a detector state graph according to the self set and the detector set respectively, and processes input strings using partial matching algorithm based on the state graph. The algorithm improved the time and space costs.

[20] successfully produced a good estimate of the optimal number of detectors needed to cover the non-self space, and the maximization of the non-self coverage was done through an optimization algorithm with proven convergence properties.

[21] used a novel technique to adjust the self radius and evolved the nonself-covering detectors. The approach could build an appropriate profile of the system only by using a subset of normal elements, and could adapt the varieties of self/nonself space.

[22] introduced hyper-ellipsoid detectors as an improvement to hyperspheres detectors in a negative selection algorithm. Then an evolutionary algorithm (EA) was used to optimize the set of ellipsoids. The fewer hyper-ellipsoids than hyperspheres were needed to achieve similar coverage of nonself space in a negative selection problem.

From the above we can see that the number of detectors and the coverage of nonself area are the most main concerned. They are also the weaknesses and difficulties of NSA. In accordance with the limitations of NSA, a novel artificial immune recognition algorithm (NNSA) is given in the next section.

3. A Novel Negative Selection Algorithm

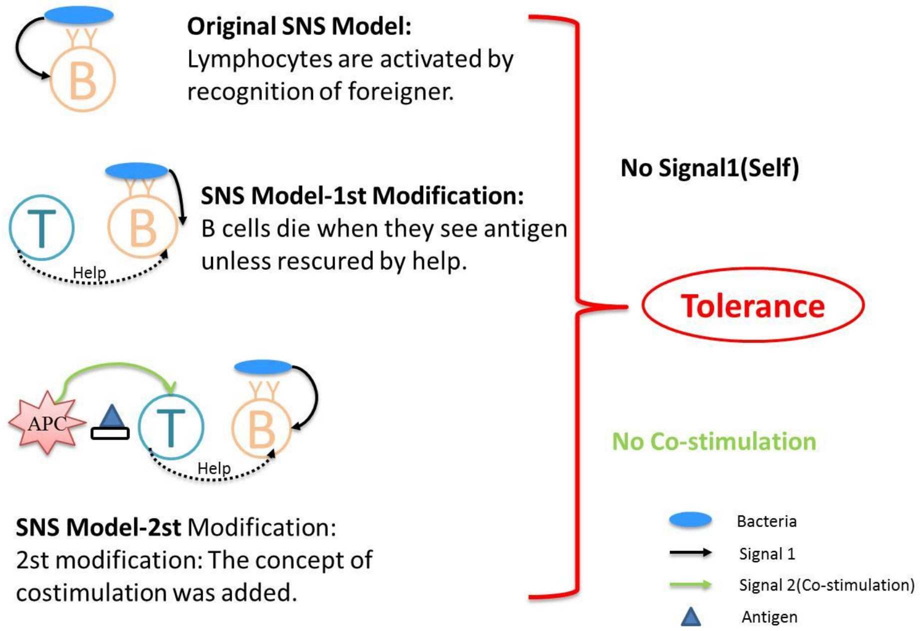


Figure 2. The Self-nonself (SNS) Model

As shown in Figure 2 [23], SNS model was modified twice in 1969 and in 1975. In 1975, Lafferty and Cunningham proposed that T cells also need a second signal (named “co-stimulation”), which they receive from “stimulator” cells (APCs) [24]. Co-stimulation principle suggested that the immune response is initiated by APCs. However, APCs are not antigen specific but they capture all sorts of self and foreign substances, then the immunity cannot be directed only against non-self. Co-stimulation signal is a key factor for immune response. However, the concept of co-stimulation was always essentially ignored in AIS.

Based on the inspiration of SNS model, a novel negative selection algorithm (NNSA) is proposed. As shown in Figure 3, compared with the traditional negative selection algorithm, co-stimulation signal receptor is added to calculate co-stimulation signal. With help with the co-stimulation, detectors can be activated to recognize antigen. The problem can be formulated as follows: define antigen set $\{AG \in R^d\}$ and antibody set $\{AB \in R^d\}$; take input data as antigen$\{Ag \in AG\}$ and reference samples as antibody$\{Ab \in AG\}$. Ag should be recognized by Ab; the proposed method should be able to calculate the recognition results.

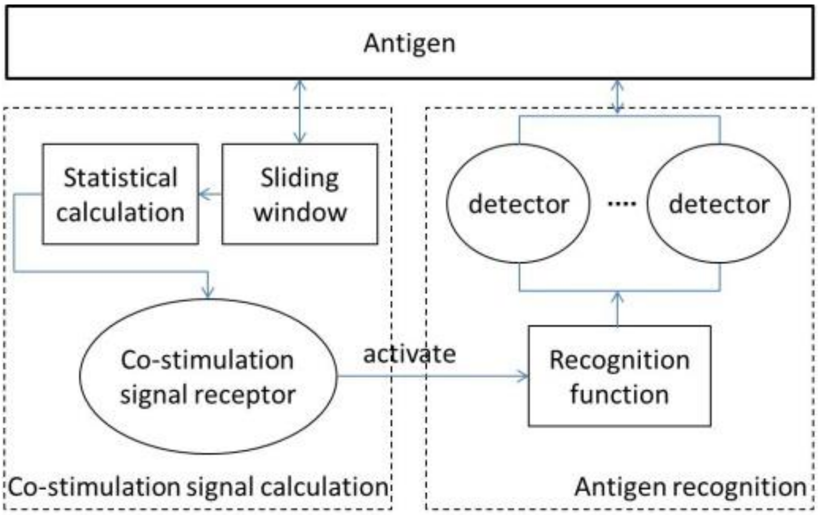


Figure 3. A Novel Negative Selection Algorithm (NNSA)

3.1. Co-stimulation Signal

Since the concept of co-stimulation signal was added to SNS model, it can explain more and more findings well. Co-stimulation signal is a very important factor in immune response. Similarly, co-stimulation signal is a key factor for the algorithm to improve the performance.

In many systems, it is difficult to model the signal which usually is not implicit in the database, and it is hard to get the abnormal feature. To overcome these problems, the sliding window and statistical techniques are adopted to calculate co-stimulation signal, the role of which is to activate detectors for recognition. One of advantages is that it is able to ignore the differences of different systems.

Relative to the abnormal feature, the normal feature is easier to get. Another advantage of the proposed method is that co-stimulation signal is calculated with normal samples. Variance measure is adopted in NNSA.

Define self set$\{SELF \in R^d\}$, the sliding window Tw is built as shown in in Figure 4. The window size is N. The big difference from [23] is that the training set is taken as the part of the sliding window.

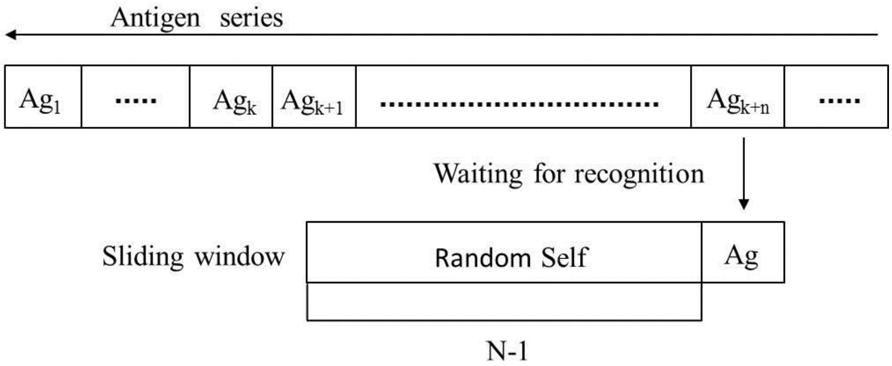


Figure 4. Sliding Window

Variance measures how far each number in the set is from the mean [25]. Variance is calculated by taking the differences between each number in the set and the mean, squaring the differences (to make them positive) and dividing the sum of the squares by the number of values in the set. Therefore, variance is a very good choice for NNSA.

Co-stimulation signal of the antigen Ag $\{Ag \in AG\}$ is defined as:

$$S_{CS}(Ag_i) = \frac{\sum_{j=1}^{N}(Tw_i(j)-\overline{Tw_i})^2}{N} \tag{1}$$

Where i means the ith dimension of Ag, $\overline{Tw_i}$ is the average of Tw.

To improve the running efficiency of the proposed algorithm, co-stimulation signal can be written as:

$$S_{CS}(Ag_i) = \frac{N\sum_{j=1}^{N} Tw_i(j)^2-(\sum_{j=1}^{N} Tw_i(j))^2}{N^2} \tag{2}$$

The proof process can see in [23] for reference. The flowchart and pseudo code of Co-stimulation signal calculation algorithm is given in Figure 5.

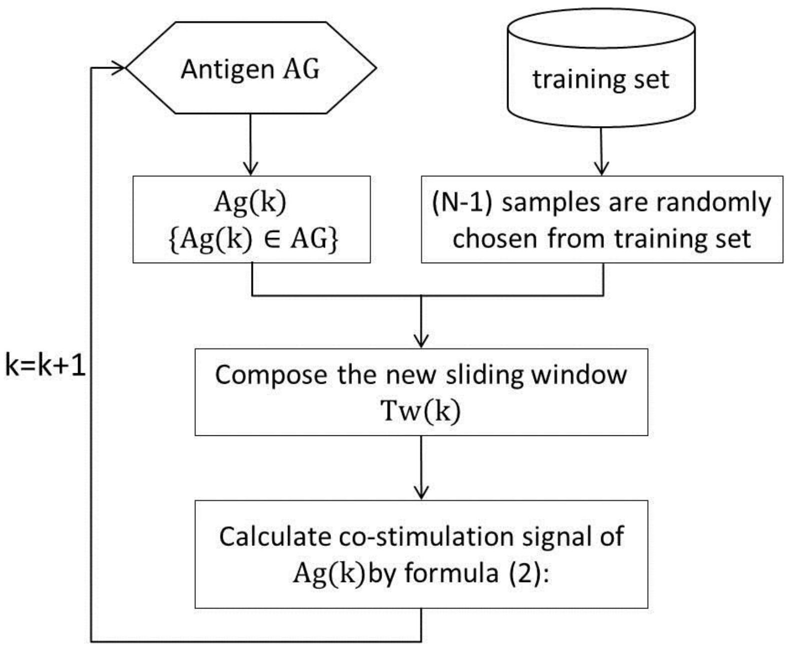


Figure 5. Flowchart and Pseudo Code of Co-stimulation Signal Calculation Algorithm

3.2. Estimation of the Detector Coverage

The detector coverage is a key content of the classification performance, which is the main difficulty for NSA. NSA requires only normal samples for training. Therefore, the exact abnormal feature is hard to get. Optimization of detection distribution and its number is difficult to satisfy the detector coverage.

In order to overcome the limitation of detector coverage, NNSA adopts the information entropy to calculate the density of detectors to estimate the necessary number and distribution of detectors for the nonself coverage. Maximizing the coverage space of nonself by keeping the detectors apart.

Consider d_i ($d_i \in$ detectors set) is one of detectors. Entropy is defined in terms of probability density. Entropy of detector d_i $E(\overrightarrow{d_i})$ is given as follows:

$$E(\overrightarrow{d_i}) = -P(\overrightarrow{d_i}) \log P(\overrightarrow{d_i}) \tag{3}$$

Where $P(\overrightarrow{d_i})$ is the probability of d_i.

$$\hat{p}(\overline{X}) = \frac{1}{|\overline{X}|} \sum_{i=1}^{|\overline{X}|} \frac{1}{\overline{\sigma}^{m}} K\left(\frac{\overline{X} - \overline{x}_i}{\sigma}\right) \tag{4}$$

The normal function is chosen as the kernel function in this paper. The advantages of this method are as follows:

1) The normal function makes estimation function change smoothly;
2) If symmetrical normal function is chosen, there will be only one parameter variation in the estimating function.

$$K(\overline{X}) = \frac{1}{(2\pi)^{m/2} |\overline{\sigma}|^{m}} \exp\left(-\frac{\overline{X}^{T}\overline{X}}{2}\right) \tag{5}$$

The calculation of the entropy of detectors in the area of nonself. Intuitively, a detector with a higher *E* will have a better contribution to the coverage of nonself-region.

3.3. Antigen Recognition

For the negative selection algorithm, if the self-data is used as a reference, the detectors are located outside of self-region. Matching rule is one of the most important comments in a negative selection algorithm. And, the distance (or similarity) measure is always preferred for antigen recognition. Therefore, one of the drawbacks of the traditional negative selection algorithm is that the self and non-self is too clearly. Antigens in the boundary of the nonself-region and self-region are more difficult to be recognized accurately. To overcome this limitation, the function of antigen recognition is defined as:

$$\begin{aligned} AR(Ab, Ag) &= \{f_m(Ab, Ag) | Ab \in AB, Ag \in AG\} \\ &= \begin{cases} 1, & f_m < \lambda \\ 0, & f_m > \lambda + \theta \\ \dfrac{\sum_{i=1}^{n-1} T_{n-i} S_{CS_{n-i}}}{n-1}, & \text{else} \end{cases} \end{aligned} \tag{6}$$

Where T_{n-i} is the state of the previous neighbor, $S_{CS_{n-i}}$ is the corresponding co-stimulation signal, f_m could be Euclidean Distance, Manhattan Distance, Hamming Distance, *etc*.

According to formula (6), in the boundary of the nonself-region and self-region, NNSA need consider the nearest neighbors and the co-stimulation signal which has triggered recognition function.

4. Experiment Results

In this section, we perform experiments on the datasets of UCI to verify the effectiveness of the proposed approach for recognition.

Breast cancer dataset of UCI has been used as the test set for many different anomaly detection approaches, which describes the test results of the potential breast cancer patients, and the last dimension describes diagnosis (benign or malignant) of them.

In order to accurately test the performance of the proposed algorithm, instances in the test set are randomly ranged. The experiment results are shown in Figure 6 and Table 1.

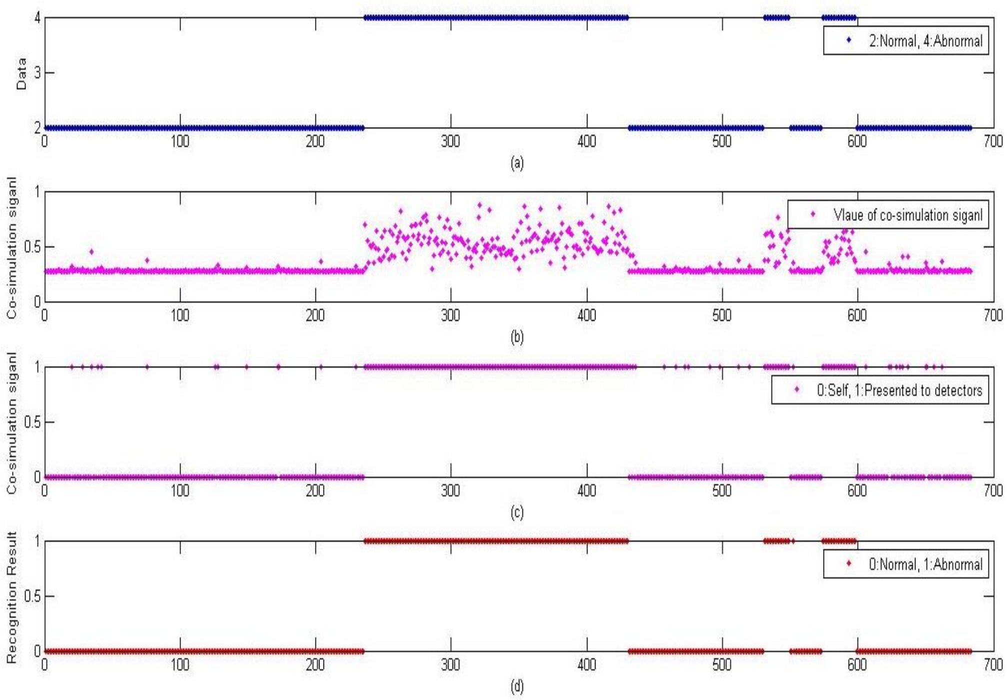


Figure 6. Breast Cancer Dataset

Figure 6(b) is the calculation value of co-simulation signal, and Figure 6(c) shows the results given by co-simulation signal. A small number of self are presented to detectors, and all of nonself are presented to detectors by co-simulation signal. Table 1 gives the exact number of the normal data presented by co-simulation signal.

Table 1. Number of Normal Data Presented by Co-simulation Signal

Number of Normal data	Presented by Co-simulation signal	Percentage Rate
444	36	8.1%

This method reduces complexity of the algorithm greatly and improves efficiency of the algorithm. Accuracy is 99.85%. Compared with [23], accuracy of the proposed algorithm has been improved. It’s worth noting that instances in the test set are randomly ranged in this paper.

Iris dataset of UCI is three classifications dataset including iris-setosa, iris-versicolor and iris-virginica. Data of iris-setosa and iris-versicolor are randomly selected in the following experiment. Take iris-setosa as normal, iris-versicolor as abnormal. In order to accurately test the performance of the proposed algorithm, recorders in the test set are randomly ranged. The experiment results are shown in Figure 7 and Table 2.

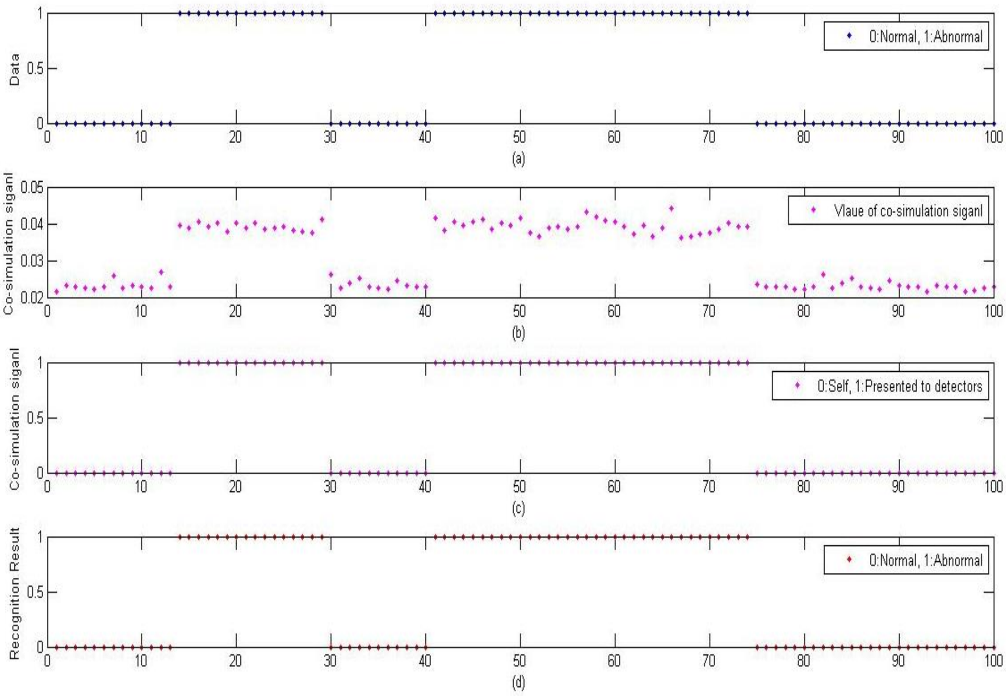


Figure 7. Iris Dataset

Figure 7(b) is the calculation value of co-simulation signal, and Figure 7(c) shows the results given by co-simulation signal. From Table 2 and Figure 7, we can see that none of self are presented to detectors, and all of nonself are presented to detectors by co-simulation signal.

Table 2. Number of Normal Data Presented by Co-simulation Signal

Number of Normal data	Presented by Co-simulation signal	Percentage Rate
50	0	0%

Accuracy of the proposed algorithm in this experiment is 100%. Experiment results show that the proposed algorithm has high accuracy for Iris data set.

To further increase test difficulty for the proposed algorithm, data of three classifications are randomly selected in the following experiment. Figure 8 and Table 3 show experiment results, and accuracy of the proposed algorithm is 100%.

Table 3. Number of Normal Data Presented by Co-simulation Signal

Number of Normal data	Presented by Co-simulation signal	Percentage Rate
22	4	18%

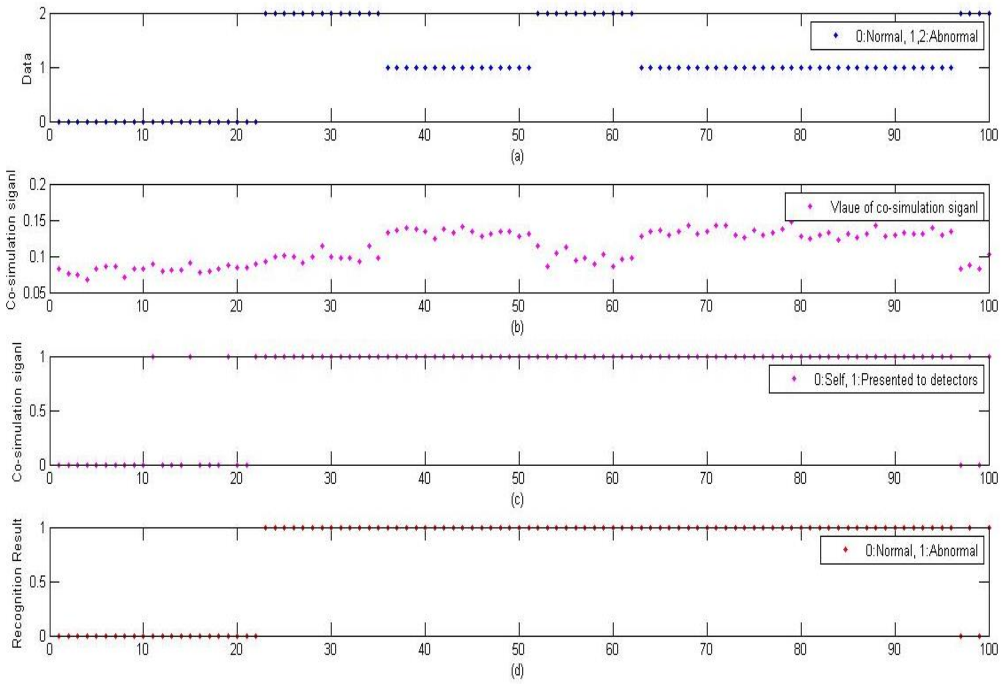


Figure 8. Iris Dataset (2)

KDD CUP 99 dataset of UCI is another popular dataset for many different anomaly detection approaches. In order to thoroughly test the performance of NNSA, the experiments are divided into two groups. One is 2 classifications, of which experiment results is shown in Table 4 and Figure 9; another is 3 classifications, of which experiment results is shown in Table 5 and Figure 10.

Table 4. Number of Normal Data Presented by Co-simulation Signal (2 Classifications)

Number of Normal data	Presented by Co-simulation signal	Percentage Rate
88	0	0%

Table 5. Number of Normal Data Presented by Co-simulation Signal (3 Classifications)

Number of Normal data	Presented by Co-simulation signal	Percentage Rate
51	7	13.7%

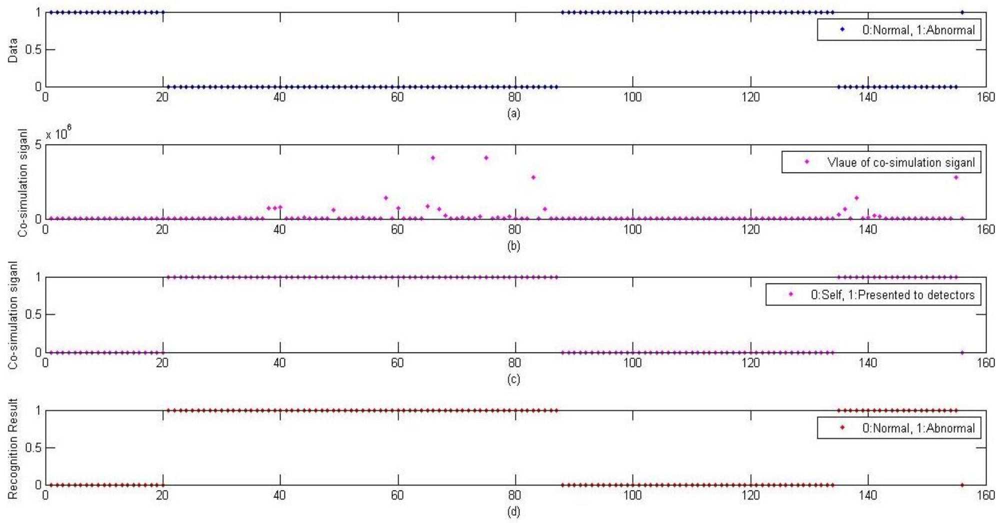


Figure 9. KDD CUP 99 Dataset

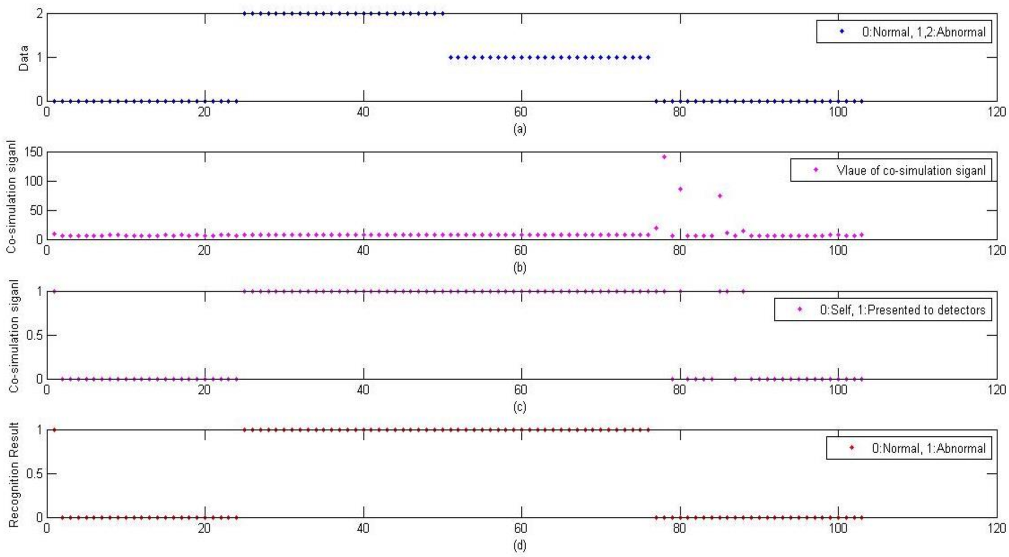


Figure 10. KDD CUP 99 Dataset (2）

Experiment results show the proposed algorithm performs well for this dataset. Accuracy of the proposed algorithm in the experiments is 100%.

5. Conclusion

This paper proposed a novel negative selection algorithm (NNSA). Compared with the traditional negative selection algorithm, the major contributions of NNSA are as follows:

Co-stimulation signal, which is a key factor in immune response, is added in the algorithm to active detectors. It improves the efficiency of the proposed algorithm.

Statistical calculation and sliding window are adopted to calculate co-stimulation signal, which is calculated with normal samples and ignores the differences of different systems.

The training set satisfies the demand of co-simulation signal calculation. The calculation method of co-stimulation signal overcomes the limits of resources required. Moreover, compared with the abnormal feature, the normal feature is more accurate.

Entropy was adopted to evaluate the density of detectors. Maximizing the coverage of nonself-region by keeping the detectors apart, and optimize the coverage and distribution of nonself-region.

Acknowledgments

This paper is supported by "Science and Technology Innovation Plan" Major Project of Shanghai Committee of Science and Technology under Grant No.13511504803; Yong University Teachers Training Plan of Shanghai Municipality under Grant No. ZZSD12060; Innovation Fund of Shanghai University under Grant No. SDCX2012017.

References

[1] P. Bretscher and M. Cohn, Science, vol. 169, no. 1042, **(1970)**.
[2] C. A. Janeway Jr., Cold Spring Harbor Symp. Quant. Biol, vol. 54, no. 1, **(1989)**.
[3] P. Matzinger, "The danger model: a renewed sense of self", Science, vol. 296, no. 5566, **(2002)**, pp. 301-305.
[4] C. A. Laurentys, R. M. Palhares and W. M. Caminhas, "A novel Artificial Immune System for fault behavior detection", Expert Systems with Applications, vol. 38, no. 6, **(2011)**, pp. 6957-6966.
[5] D. Dasgupta and F. Nino, "Immunological computation: theory and applications", CRC Press, **(2008)**.
[6] M. Vella, M. Roper and S. Terzis, "Danger theory and intrusion detection: possibilities and limitations of the analogy", In ICARIS 9th International Conference on Artificial Immune Systems. Springer: Lecture Notes in Computer Science, **(2010)**, pp. 276-289.
[7] V. Chandola, A. Banerjee and V. Kumar, "Anomaly detection: A survey", ACM Computing Surveys (CSUR), vol. 41, no. 3, **(2009)**, pp. 15.
[8] S. Forrest, A. S. Perelson, L. Allen, *et al.*, "Self-nonself discrimination in a computer", Research in Security and Privacy, Proceedings IEEE Computer Society Symposium on IEEE, **(1994)**, pp. 202-212.
[9] Z. Ji and D. Dasgupta, "Revisiting Negative Selection Algorithms", Evolutionary Computation, vol. 15, **(2007)**, pp. 223-251.
[10] X. Z. Gao, S. J. Ovaska and X. Wang, "Genetic Algorithms-based Detector Generation in Negative Selection Algorithm", presented at IEEE Mountain Workshop on Adaptive and Learning Systems, **(2006)**.
[11] X. Z. Gao, S. J. Ovaska, X. Wang and M.-Y. Chow, "Clonal Optimization of Negative Selection Algorithm with Applications in Motor Fault Detection", presented at IEEE International Conference on Systems, Man and Cybernetics, SMC '06, Taipei, **(2006)**.
[12] D. Dasgupta, S. Yu and F. Nino, "Recent advances in artificial immune systems: models and applications", Applied Soft Computing, vol. 11, no. 2, **(2011)**, pp. 1574-1587.
[13] J. Zeng, X. Liu*, T. Li, C. Liu, L. Peng and F. Sun, "A self-adaptive negative selection algorithm used for anomaly detection", Progress in Natural Science, vol. 19, Issue 2, **(2009)** February 10, pp. 261-266.
[14] J. N. Tournier and A. Q. Hellmann, "Host Pathogen Interactions", A Biological Rendez-Vous of the Infectious Nonself and Danger Models?.
[15] A. Lasisi, et al., "Comparative Performance Analysis of Negative Selection Algorithm with Immune and Classification Algorithms", Advances in Intelligent Systems & Computing, **(2014)**, pp. 441-452.
[16] Z. Ji and D. Dasgupta, "V-detector: An efficient negative selection algorithm with "probably adequate" detector coverage", Information Sciences, vol. 179, **(2009)**, pp. 1390–1406.
[17] D. Li, S. L. Liu and H. L. Zhang, "A negative selection algorithm with online adaptive learning under small samples for anomaly detection", Neuro computing, vol. 149, Part B, **(2015)** February 3, pp. 515-525.
[18] A. V. Aho and M. J. Corasick, "Efficient String Matching: An Aid to Bibliographic Search", Communications of the Acm, vol. 18, no. 6, **(1975)**, pp. 333-340.
[19] W. Luo, X. Wang and X. Wang, "A novel fast negative selection algorithm enhanced by state graphs", presented at 6th International Conference on Artificial Immune Systems (ICARIS), Santos, Brazil, **(2007)**.
[20] F. Gonzalez, D. Dasgupta and L. F. Nino, "A randomized real-value negative selection algorithm", in: Proceedings of Second International Conference on Artificial Immune System (ICARIS), **(2003)**, pp. 261–272.
[21] J. Q. Zeng, X. J. Liu, T. Li, *et al.*, "A self-adaptive negative selection algorithm used for anomaly detection", Prog. Nat. Sci., vol. 19, no. 2, **(2009)**, pp. 261–266.
[22] J. M. Shapiro, G. B. Lamont and G. L. Peterson, "An evolutionary algorithm to generate hyper-ellipsoid detectors for negative selection", presented at the conference on Genetic and evolutionary computation, Washington DC, USA, **(2005)**.
[23] Y. Tao and M. Hu, "A Fast Immune Recognition Model based on Immune Response", International Journal of Multimedia and Ubiquitous Engineering (IJMUE), vol. 9, no. 11, **(2014)**.

[24] K. J. Lafferty and A. Cunningham, Aust. J. Exp. Biol. Med. Sci., vol. 53, no. 27, **(1975)**.
[25] K. M. Abadir, W. Distaso and F. Zikes, "Design-free estimation of large variance matrices", Working paper, Imperial College London, **(2012)**.

Authors

Yuan Tao, received her PhD in Computer Application Technology from Shanghai University in 2011. She is a lecturer in Computation Center at the Shanghai University, China. Her main research interests are in the areas of intelligent information processing, artificial intelligent and pattern recognition.

Min Hu, received her PhD in Control Theory and Control Engineering from Shanghai University in 2006. She is an associate professor in SHU-UTS SILC Business School at the Shanghai University, China. Her main research interests are in the areas of data mining, risk identification method, artificial intelligent and computer application.

Yanlin Yu, he is a master student in SHU-UTS SILC Business School at the Shanghai University, China. His main research interests are in the areas of artificial intelligent and outlier detection.