Multimed Tools Appl (2017) 76:19397–19410
DOI 10.1007/s11042-015-3117-0

Towards accurate intrusion detection based on improved clonal selection algorithm

Chunyong Yin[1] · **Luyu Ma**[1] · **Lu Feng**[1]

Received: 8 November 2015 / Accepted: 23 November 2015 /
Published online: 28 November 2015

Abstract Artificial immune system constructs a dynamic and adaptive information defense system through a function similar to the biological immune system. In order to resist the external invasion of useless and harmful information and ensure the effectiveness and the harmlessness of received information. Due to the low accuracy and the high false positive rate of the existing clonal selection algorithms applied to intrusion detection, in this paper, we propose an improved clonal selection algorithm. The improved method detects the intrusion behavior by selecting the best individual overall and cloning them. Experimental results show that the improved algorithm achieves very good performance when applied to intrusion detection. And it is shown that the algorithm is better than BP neural network with its 99.5 % accuracy and 0.1 % false positive rate.

Keywords Artificial immune · Clonal selection · Intrusion detection · Biological immune · Adaptive

1 Introduction

A safe system should ensure the availability, security and integrity for the user system. However, the opening and sharing of network brings hidden danger for the security as well as it enriches the network resources. Network security events like virus intrusion, malicious attack, information leakage and illegal steal are occurring nowadays. These network security behaviors brought out a serious threat to the object system and network. However, these behaviors often occur under the normal network activities, so they are hidden and could happen without the limitation of time and area. And with the development of network, the

✉ Chunyong Yin
ycycam@163.com

[1] School of Computer and Software, Jiangsu Key Laboratory of Meteorological Observation and Information Processing, Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science & Technology, Nanjing, Jiangsu 210044, China

intrusion methods are also becoming more concealed and complex. Thus the security of network has been a subject of concern.

Intrusion detection [6, 11, 18] is a kind of security technology, which can collect information from some key points of the network or computer system and try to analysis it, in order to find whether there is a violation of the security policy or a sign of the attack in the network or computer system. In the research of intrusion detection, researchers always aimed to the high accuracy, the low false positives rate and the fast inspection speed. There are many different kinds of detection methods. According to time, intrusion detection methods can be divided into on-line detection and offline detection. The on-line detection can be realized in the process of network transmission, which has high real-time performance but lower detection accuracy and high false positives rate. The offline detection method is a kind of non-real-time detection system which analyses the audit data flow after the events and checks the intrusion activities, it has high accuracy, low rate of false positives, but it cannot find the intrusion behaviors in time. In the terms of technology, the detection methods can be divided into misuse detection and anomaly detection. Misuse detection technology assumes all of the intrusion behaviors and means can be expressed as a trait or pattern, the aim of the technology is to detect that whether the main activities conform to these pattern, if they meet the pattern, then, they will be regarded as suspicious behaviors, the technology has the advantage of low detection rate of false positives, it only can find the known attacks which have exist in library but cannot find the unknown aggressive behaviors. However, the anomaly detection technology supposes that all of the intrusion actions are different with normal behaviors, it sets up the profile of normal activities, when the main activities once violate the rule and they will be seen as suspicious behaviors. This technology can find the unknown and new intrusion behaviors, but, at the same time, it has the high rate of the false positives. In terms of the detection object, the detection methods can be divided into host-based detection and network-based detection. Host-based detection according to the monitoring and analysis of audit data of the host to determine whether abuse and intrusion events occur. This method can accurately judge the intrusion behavior, and make timely response, but this method will occupy resources of the host. The network-based detection monitors traffic of the original data flow in the key points of the network, and analyses the data to get useful information in order to further judge. Although this method consumes less resources of the host, the precision and the range of the detection is poor and smaller.

In recent years, the intrusion detection technology [20]:1) Soft computing method: mainly include neural network, genetic algorithm and fuzzy technology. 2) Mobile agent. 3) Computer immunology. 4) Data mining. 5) Protocol analysis and command parsing technology.

Among them, we will put forward intrusion detection method in computer immunology. Biological is the most effective weapon to defense pathogen intrusion, and the intrusion detection system is the most effective weapon to defense network intrusion for a computer or a network system, the former has character of distribution, robustness and diversity, it can be used in intrusion detection system as well. Artificial immune system is the simulation of biological immune system, so, it's feasible to apply artificial immune mechanism and artificial immune algorithm to intrusion detection system. We will improve the clonal selection algorithm of AIS which is used in intrusion detection. The improved algorithm is superior to the original algorithm in the detection efficiency and accuracy.

2 Artificial immune system

AIS is a variety of intelligent systems which learn and utilize the mechanism, characteristics and principle of the biological immune system, and it could be applied to all kinds of information processing technology, computing technology and engineering and science.

The University of Mexico America Forrest-Hofmeyr group [1, 4, 13] is the most representative team who let the immune theory be used in intrusion detection. In 1994, Forrest firstly proposed the negative algorithm and applied it in intrusion detection; in 1996, Forrest monitored the specific process which was required for system to realize intrusion detection; in 2000, Hofmeyr and Forrest combined with a variety of immune mechanisms to monitor network data flow in the TCP SYN package and put forward a kind of artificial immune system structure which can speed up the pace of development in the fields. Kim and Bentley [8] proposed a intrusion detection model based on artificial immune, then they proposed dynamic clonal selection algorithm [9].

At present, there are three main artificial immune algorithm applied in intrusion detection [15, 19] clonal selection algorithm, negative selection algorithm and immune genetic algorithm. Now, application of artificial immune system are mainly in network security, fault diagnosis, pattern recognition etc..., Artificial network security model recommends the working principle of the natural immune system into the research field of network security and makes network security system to be similar with the natural immune system of immune recognition and immune memory function, and makes it has the characteristics of self-learning and adaptive.

Due to the human body can effectively and automatically distinguish self and non-self, human can protect themselves from external virus, this process [14] is similar to the process of intrusion detection. But the traditional intrusion detection technology does not have these advantages. So, it's necessary and feasible to apply the artificial immune system into intrusion detection. The relationship between the network intrusion detection and artificial immune system is shown in Fig. 1.

We can see from the diagram above, when the computer attacked by virus, that is equivalent to the human body infected by external virus antigen. The human body's own antibodies, B cells and T cells is string of description of detection of computer. When the body's internal antibody matches the external antigen, that is the process of body to recognize non self, and similarly computer also matches the external intrusion and recognizes them. When the body is attacked by external antigens, it will produce self-tolerance, if the value exceeds the maximum, then coordinated stimulus signal would be produced. Also, in computer, when the intrusion behavior by matching exceeds preset activation threshold, the computer would begin to enter the negative selection. Once body's coordinated stimulus signal is sent, the body will produce large amounts of memory cells and begins to clone cells containing effective information, ultimately for foreign antigen detection and response. The computer starts to enter the memory detector by negative selection algorithm, and begin to copy contains a large amount of information detector, finally complete the virus and intrusion detection and response.

Artificial immune system based on the recognition of "self" and "not self" and thus to detect and remove invasion behavior effectively, realize the immune function of self-defense.

The immune system has characteristics of diversity, adaptability, dynamic, distribution, robustness and can self-study, recognition, memory, etc. Most existing intrusion detection systems do not have more of these features, so the artificial immune system is able to in the detection of intrusion behavior has a better effect.

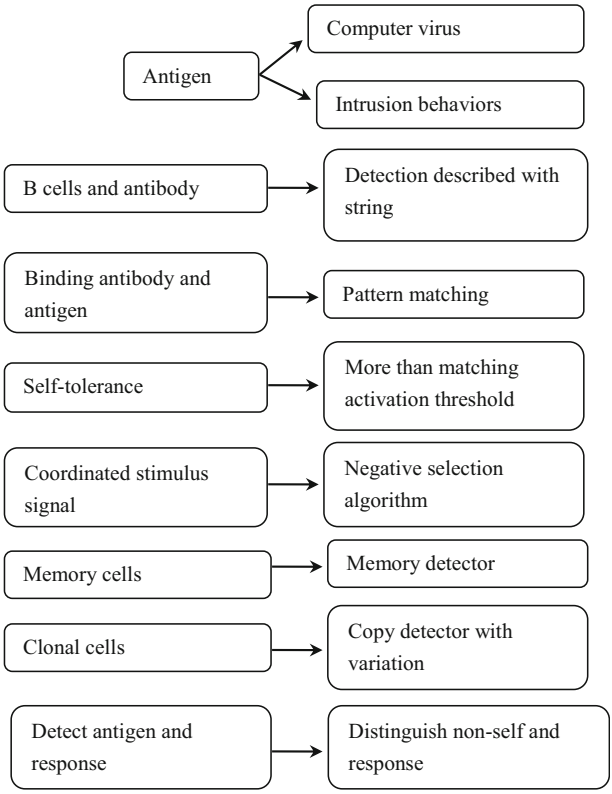


Fig. 1 The relationship between the network intrusion detection and artificial immune system

3 BP neural network

Artificial neural network [21] is a constructed neural network which can implement some functions. It's based on the understanding of human brain neural network. It's a theoretical mathematical model of neural network. It is an information processing system which is built by imitating the brain neural network structure and function.

BP neural network proposed by Rumelhart, McCelland [12] and other scientists. And it is also one of the most widely used neural network model. BP neural network is composed of positive communication of information and back propagation of error. The topology model includes input layer, hidden layer and output layer, as the following Fig. 2.

In this mode [5, 7], the neurons receive external information in input layer, and pass them to the hidden layer. Hidden layer that is internal information processing layer, the hidden layer can be divided into single hidden layer and multilayer hidden layer in the light of different needs. The hidden layer is responsible to change the received information and transfer the information to the output layer, this completes a learning process of positive spread. When the information arrives in output layer, the output layer analyzes the information and output results, when the actual output does not accord with the desired value, then, it will begin to enter the error back propagation, according to the gradient descent method, step by step a reverse adjustment, until the final result to achieve acceptable error or the maximum number of learning.

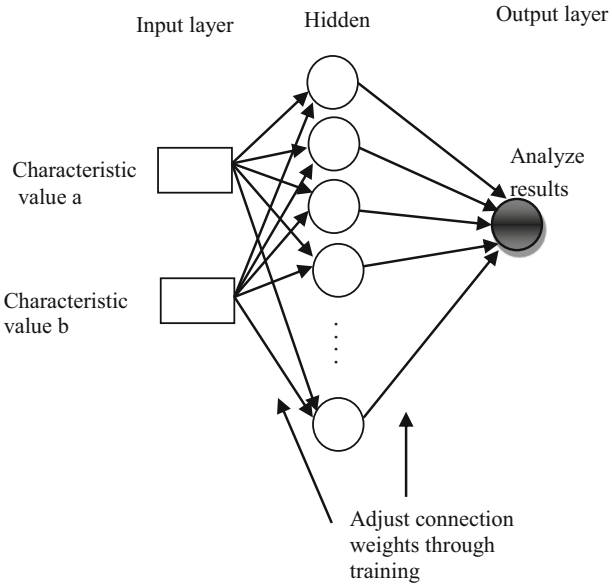


Fig. 2 BP neural network topology

The BP neural network [16] is mainly used in the function approximation, classification, pattern recognition and data compression. Although the BP neural network has been widely used, but it still exists some disadvantages, mainly includes the following aspects:

1) The learning rate is fixed, this leads to the slower convergence speed of the network.
2) The minimum values of its convergence is a local minimum, cannot guarantee for the global minimum.
3) The layer number of hidden layer of the choice has no theoretical basis for guidance.
4) The learning and memory of the network have instability, when adding new learning sample, the training of the network will have to start from scratch again.

4 Improved clonal selection algorithm

Clonal selection algorithm (CSA) [17] is a kind of intelligent method which is based on the clonal selection theory of biological immune system and simulates immune system learning process. This algorithm mainly after initialization, selection, clone and mutation, replacement of five stages.

The operation of the standard clonal selection algorithm steps are as follows:

1) In the domain of the problem randomly generated candidate set of antibody S, in the collection have N antibodies, N is the number of antibodies in the collection.
2) Calculation of each individual into the fitness function of the set S affinity, select n best collection of antibodies of P, where n < = n.
3) Clone all the antibodies in the P set and copy each antibody c, constitute a temporary clone set c.
4) Mutate C set in a certain probability m, and form a new mutated set D.

5) Calculate the affinity of the antibody of D set, choose n the best antibodies, and from memory cell set M, use the set M instead of P.
6) Replace d low affinity antibodies with new generation of antibodies in set S, in order to keep the diversity of antibody group and prevent "premature" phenomenon of the algorithm Fig. 3.

We will first to select data sets to feature selection [10], remove the useless and redundant features, so as to achieve the purpose of dimension reduction, in determining the characteristics of effective standards, to find the most effective feature subset, feature selection flow chart is shown in Fig. 4.

The steps of the feature selection for the data set are as follows:

(1) Scan the data into storage, with a characteristic for the unit, each column features as a data.
(2) Pre-processing of data sets, including the character attribute value processing and numerical attribute normalization.
(3) Calculate the Euclidean distance of characteristics of each column with other characteristics, calculation formula is: $D = \text{sum}\left(\sqrt{\sum_{i=0}^{n}(x_i - y_i)}\right)$, where x and y respectively represent different characteristics of columns, n is the number of data.
(4) Identify the characteristics whose distance is the biggest, and according to the number of identified by many, at least, in turn, delete occurrences more features, leave a suitable number of characteristics, complete the extraction of feature selection.

In this paper, we improved the classical clonal selection algorithm, the aim of this algorithm is to produce a large number of memory cells from the training data, and use them to classify the test data.

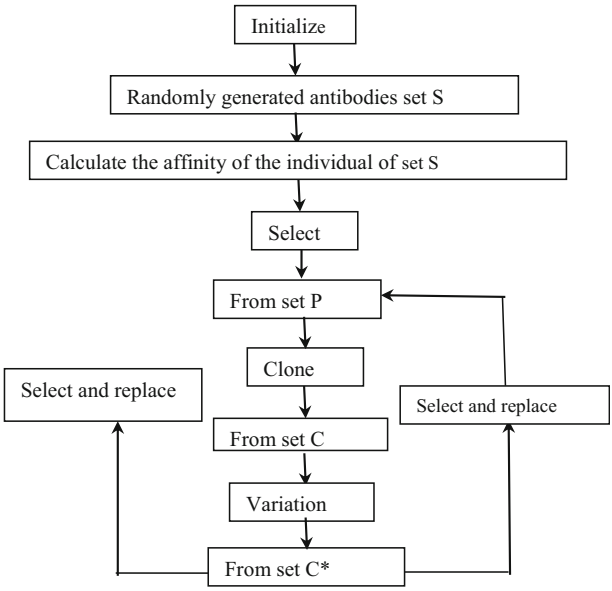


Fig. 3 The flow chart of clonal selection algorithm

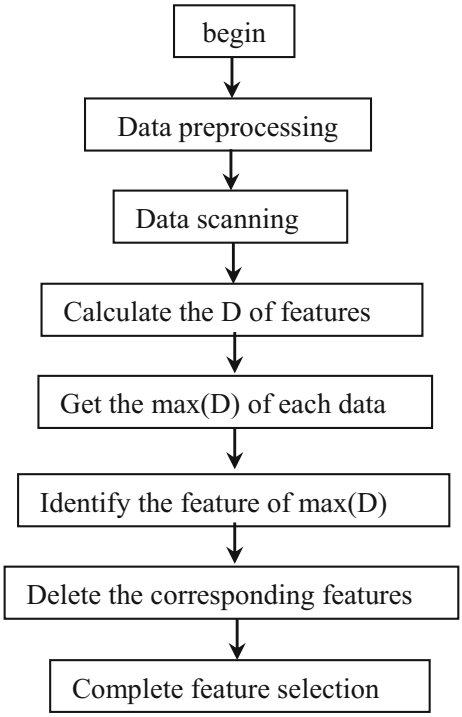


Fig. 4 The flow chart of feature selection

The improved algorithm is designed by simulating many mechanism of the immune system, including: resource competition, clonal selection, affinity maturation, memory cells retained and artificial immune system with limited resources. In this algorithm, the feature value of the training and test data is antigen, and the system units is B cell. The similar B cells in the ARBs will compete with each other for the fixed resources. Once the antigens of the training set is proposed to algorithm, the algorithm will create a unit for it. A memory cells set will be formed after the presentation of all training antigens.

The improved algorithm includes four stages: initialization and normalization; the generation of ARB; Resource competition and the choice of memory cells. The memory cell selection mechanism is as follows:

1) First of all, all the training antigens which are presented to memory cells belongs to the same category. We will clone the memory cell whose simulation is the highest, put it and its clones into the ARB pool. The number of clones is decided by the affinity between the memory cell and training antigen, and at the same time, the affinity is depended on the distance between the feature value of a memory cell and training antigen. The smaller the distance, the higher the affinity, the more the number of copy. On the contrary, the bigger the distance, the lower the affinity, the less the number of copy.
2) Next, in ARB pool, we calculate the corresponding values of the ARBs according to their affinity. So, the sum of all system resources is produced. At this time, we compare the sum of system resources with the maximum value of resources allowed by the system. If the

result is that the sum of system resources is bigger than maximum value of resources, then, we begin to delete the ARB whose affinity is lower until the sum of system resources is not bigger than the maximum value of resources. We calculate the affinity of remain ARBs and training antigens, if all instances of the average normalized level cannot meet the needs of the stimulation thresholds defined by system, the ARBs begin to mutate and clone in ARB pool. Repeat step 2 until the affinity meet the stimulation threshold.

3) Choose the ARB whose stimulation value is the highest as the memory cell. If its affinity for the training antigen is bigger than that of the original memory cells which are selected in step 1, then put it into memory cell set. And in addition, if the differences in the affinity of the two memory cells is less than the user defined threshold, then the initial memory cells will be removed from the pool. For each training antigen, repeat the above steps to establish the actual classification set of memory cells Fig. 5.

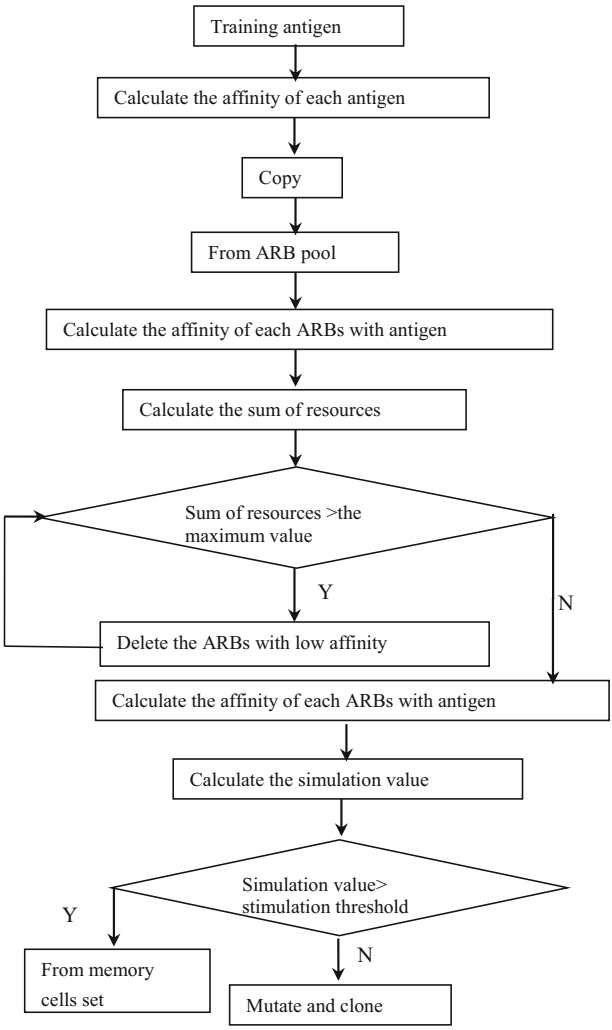


Fig. 5 The formation of memory cells

In the paper, we will choose the optimal memory cells and cloning generated can better classify a set of memory cells, make the best individual choice is not limited to local. The improved algorithm is described as follows:

1) Prepare training data set A.
2) Calculate the affinity of the antibodies in set S, set the affinity threshold for the size of the F, and copy high affinity data, copy number size to C, then form set D.
3) Compare the sum of affinity M with the total cost of the resource R, where $M = sum(S)$, if larger than R, looking for the smallest individual accessibility and delete, repeat 3) operation until $M \leq R$.
4) Generate the test set H.
5) Detect test set by using KNN method Fig. 6.

5 Experimental analysis

In order to verify the performance of the improved algorithm, using KDD CUP99 dataset to experiment. In 1988, DARPA intrusion detection evaluation project shall be borne by the MIT Lincoln laboratory [13], The project is designed to measure and evaluate the intrusion detection system, One of the outcome of the project is to set up a simulation of various military network intrusion security audit data set MIT LL intrusion detection data sets, But when using the data set into intrusion detection analysis, there are the following disadvantages:

1) The data set is too big, not only contains the network traffic, also contains a variety of host audit data and log information, including the host audit data and log information relate to the configuration and system running environment, is not conducive to a variety of different algorithms of intrusion detection and evaluation.
2) The tcp dump binary format data processing is not convenient.
3) Each of the data contains the characteristics of large amount of information, different protocols will produce different data format, so it's not conducive to feature selection.

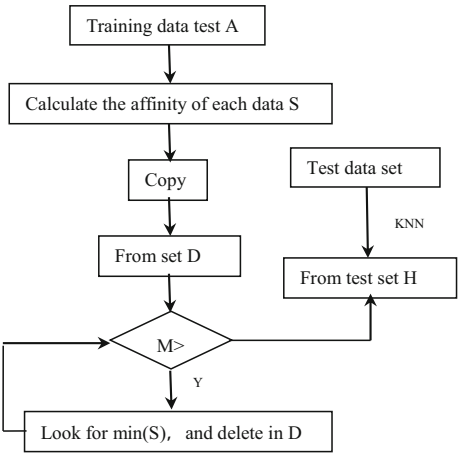


Fig. 6 The flow chart of the improved clonal selection algorithm

Based on these, we select the KDD CUP99 data set is sorted from 1988 MIT LL organize the intrusion detection data set. The data set only includes network traffic data. At present, it's a security audit data set which is generally accepted and practical. The data set includes: all data set, 10 % data set and correct.gz test data set that containing the tag type. We sample 10 % of KDD CUP99 data set randomly, extract with 1 % and 2 % of the data set again. So the experiment data with 49,402 and 98,804, each data including 41 characteristic attributes and a marker for normal or attack, impersonation attacks mainly divided into the following four categories:

1) Denial-of-service attacks (DOS), the attacker let the memory of the computer to be too busy and cannot handle legitimate requests or refuse to legitimate user's access to the machine
2) User to Root(U2R):through a user who has no legal power or low permissions utilizes loophole in the website to get root permissions directly, then log in and do illegal operation.
3) Remote to Local(R2L): Attacker via remotely login computer, use the computer account and password to access to the computer for illegal operation.
4) Probe: collect the information of computer network to bypass the security control.

First of all, we operated feature selection on data set, we achieved the result of dimension reduction by feature selection and compared processed data with unprocessed data. Our experiment through the feature selection, delete the redundant features, eventually left 21 suitable characteristics, the serial number of these characteristics as follows: 1, 3, 4, 5, 6, 8, 11, 12, 13, 23, 25, 26, 27, 28, 29, 30, 32, 33, 34, 36, 39. Under the same kind of clonal algorithm, we detected and analyzed two kinds of data by using the 10-fold cross-validation classification method. The experimental results in the following Table 1.

The experimental results show that the time of dealing with the data set is shorter after the feature selection. And at the same time, it is not hard to find in Table 1, under the condition of the same kind of algorithm of detection, the overall detection accuracy rate increased by 0.1 %, and the rate of false positives was reduced by 0.7 %, so the overall detection rate has been further improvement.

We adopt the CLONALG algorithm and the improved clonal selection algorithm to detect the nearly 50,000 and 100,000 data respectively in this paper. In the experiment, we set affinity threshold F was 0.2, the rate of replication of C was 10, the total resource was 150, the value of KNN was 3. We classified the two kinds of data by using the 10-fold cross-validation classification method and compared with the results of the clonal selection algorithm. The results obtained are shown in Table 2.

In this paper, we also compared the artificial immune system with neural network in intrusion detection system. We adopt 49,402 and 98,804 data sets after feature selection,

Table 1 The testing results of feature selection

Algorithm	Data set	TP(%)	FP(%)
Clonal	41 features with 100,000 data	71	27.5
Clonal	41 features with 500,000 data	72.1	27.3
Clonal	21 features with 100,000 data	79.5	8.3
Clonal	21 features with 500,000 data	78.5	8.8
Ocerall		+0.1	−0.7

Table 2 The results of comparison of the improved clonal algorithm with clonal algorithm

Algorithm	Data set	Detection rate TP(%)	Error detection rate FP(%)
Clonal	49,402	78.5	8.8
Improve Clonal	49,402	99.4	0.1
Clonal	98,804	72.1	27.3
Improve Clonal	98,804	99.2	0.2

selected the BP neural network as classifier, set the learning rate to 0.1, the training iteration number is 500. We compared experimental results with experimental results of the improved algorithm, the results are shown in Table 3.

We can see from the table, in the two tests, rates of detection of the improved algorithm is superior to the classic clonal selection algorithm and neural network algorithm. We also can see from the above table, in the case of data quantity increase, error detection rate of classic clonal algorithm increases sharply and leads to decrease the detection accuracy, but the improved clonal algorithm has no change in error detection rate, it means that it has less effect on the accuracy of testing large amounts of data for improved clonal algorithm. And although it's not obvious than when the data increases, the accuracy of neural network changes a little, the overall classification accuracy is lower than the improved algorithm.

6 Results analysis

In this paper, we compared the artificial immune system with artificial neural network. Artificial neural network and artificial are intelligence systems which are inspired by biological immune system and have the ability to recognize. They take advantage of learning, memory, associate, restore to solve the problem of recognition and classification tasks; But the internal mechanism of recognition and learning are completely different. Dasgupta [2, 3] analyzed the similarities and differences of artificial neural network and artificial immune system systematically. He thought that component unit and the number, interaction, recognition task, task execution, memory, learning, and the system robustness of them are similar, but they are different in composition distribution, and in the system unit communication, system control and so on. And he pointed that the natural immune system is a important source of inspiration to artificial intelligence methods. Gasper think diversity is the basic feature of the adaptive dynamic, and the maintenance of AIS is better than GA optimization method of this diversity. In this paper, the experiments have verified the artificial immune system is better than that of the artificial neural network.

Table 3 The results of the comparison of BP neural network algorithm with the improved algorithm

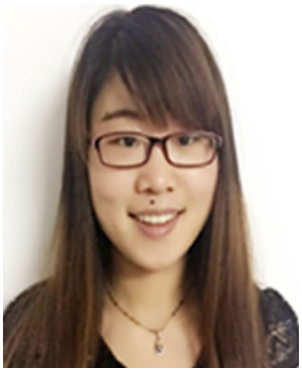
Algorithm	Data set	Detection rate TP(%)
BP	49,402	96.08
Improve Clonal	49,402	99.4
BP	98,804	96.06
Improve Clonal	98,804	99.2

Acknowledgments Foundation item: This work was funded by the National Natural Science Foundation of China (No.61373134). It was also supported by the Priority Academic Program Development of Jiangsu Higer Education Institutions(PAPD), Jiangsu Key Laboratory of Meteorological Observation and Information Processing (No.KDXS1105) and Jiangsu Collaborative Innovation Center on Atmospheric Environment and Equipment Technology(CICAEET).

References

1. Aickelin U, Bentley P, Cayzer S, Kim J, McLeod J (2003) Danger theory: The link between AIS and IDS? In: Timmis J, Bentley PJ, Hart E (eds) Artificial immune systems. Second International Conference, ICARIS 2003, Edinburgh, UK, September 1–3, 2003. Lecture Notes in Computer Science, vol 2787. Springer, Berlin, Heidelberg, pp 147–155
2. Aickelin U, Dasgupta D, Gu F (2014) Artificial immune systems. In: Burke EK, Kendall G (eds) Search methodologies. Introductory Tutorials in Optimization and Decision Support Techniques. Springer, US, pp 187–211
3. DasGupta D (1999) An overview of artificial immune systems and their applications. Springer, New York
4. Forrest S, Hofmeyr S, Somayaji A, Longstaff T (1996) A sense of self for unix processes. Proceedings of the IEEE Symposium on Security & Privacy 11(30):120–128
5. Gu Y, Shi Y, Wang J (2012) Efficient intrusion detection based on multiple neural network classifiers with improved genetic algorithm. J Softw 7(7):1641–1648
6. Gu B, Sheng VS, Tay KY, Romano W, Li S (2014) Incremental support vector learning for ordinal regression. IEEE Trans Neural Netw Learn Syst 26(7):1403-1416
7. Gu B, Sheng VS, Wang Z, Ho D, Osman S, Li S (2015) Incremental learning for ν-Support Vector Regression. Neural Netw 67:140–150
8. Kim J, Bentley PJ (2001) Towards an artificial immune system for network intrusion detection: An investigation of clonal selection with a negative selection operator. IEEE 1244–1252
9. Kim J, Bentley PJ (2002) Towards an artificial immune system for network intrusion detection: an investigation of dynamic clonal selection. IEEE 1015–1020
10. Li Y, Xia J, Zhang S, Yan J, Ai X, Dai K (2012) An efficient intrusion detection system based on support vector machines and gradually feature removal method. Expert Syst Appl 39(1):424–430
11. Liao H-J, Lin C-HR, Lin Y-C, Tung K-Y (2013) Intrusion detection system: A comprehensive review. J Netw Comput Appl 36(1):16–24
12. McClelland JL, Rumelhart DE, Group PR (1986) Parallel distributed processing. Explorations in the microstructure of cognition, vol 1. MIT Press, Cambridge
13. McHugh J (2000) Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by Lincoln Laboratory. ACM Trans Inf Syst Secur 3(4):262–294
14. Ou C-M (2012) Host-based intrusion detection systems adapted from agent-based artificial immune systems. Neurocomputing 88(7):78–86
15. Qing S-h, J-c J, Ma H-t, W-p W, X-f L (2004) Research on intrusion detection techniques: a survey. Journal-China Institute of Communications 25(7):19–29
16. Schmidhuber J (2015) Deep learning in neural networks: An overview. Neural Netw 61:85–117
17. Ulutas BH, Kulturel-Konak S (2011) A review of clonal selection algorithm and its applications. Artif Intell Rev 36(2):117–138
18. Yin C (2014) Towards Accurate Node-Based Detection of P2P Botnets. Sci World J 2014:425491
19. Yin C, Zou M, Iko D, Wang J (2013) Botnet detection based on correlation of malicious behaviors. International Journal of Hybrid Information Technology 6(6):291–300
20. Zhang F, Wang D (2013) An effective feature selection approach for network intrusion detection. IEEE 307–311
21. Zhang B, Zhang S, Lu G (2013) Evaluation model research of 100 meters sprint exercise capacity based on fuzzy neural network. J Chem Pharm Res 5(9):256

Chunyong Yin Dr. Chunyong Yin is currently an associate Professor and Dean with the Nanjing University of Information Science & Technology, China. He received his Bachelor (SDUT, China, 1998), Master (GZU, China, 2005), PhD (GZU, 2008) and was Post-doctoral Research Fellow (University of New Brunswick, 2010). He has authored or coauthored more than thirty journal and conference papers. His current research interests include network computing, data mining and network security.



Luyu Ma She obtained her B.S. degree in the Computer and Software Institute from Nanjing University of Information Science and technology, China in 2013. Now, she is working toward the M.S. degree in the Computer and Software Institute. Her current research interests are in network security and intrusion detection.

Lu Feng He received his bachelor degree in 2013 from Nanjing University of Information Science & Technology. He is a graduate student at the School of Computer and Software of Nanjing University of Information Science & Technology. He is working toward the M.S. degree in the Computer and Software Institute. His current research interests are data mining, data-stream classification and feature extraction algorithm.