ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0071

Research on Intrusion Detection System based on Improved Negative Selection Algorithm

Zhihao She, Mengqian Li

China University of Mining and Technology, Xuzhou 221116, China.

Abstract

Through the learning and research of the immune negative selection algorithm, after learning the traditional partial matching rules, namely Hamming matching and r-continuous bit matching, summarize its advantages and improve its disadvantages to create an improved matching scheme to deal with immune negatives Select the generation of the detection set in the algorithm and the process of intrusion detection. At the same time, this paper also established an intrusion detection model based on the improved scheme and a traditional intrusion detection model, and compared the matching rate and false alarm rate (the probability of black holes) between the two models through simulation experiments. It is concluded that the matching rate and false alarm rate of the improved scheme are better than those of the traditional scheme.

Keywords

Intrusion Detection; Artificial Immune; Negative Selection Algorithm; Custom Coding.

1. Introduction

With the rapid development of science and technology, computer and Internet technology have brought great convenience to our lives, but at the same time, there are also many problems that threaten Internet security. For example, in 2017, Globelmposter ransomware spread through RDP protocol loopholes [1] and social engineering in domestic hospitals. It used RSA2048 encrypted files, which caused the files to be undecrypted, which caused a huge blow to the hospital system. Similar and more representative of the famous ransomware GandCrab, also commonly known as "poisonous crabs"[2], the virus spread wildly around the world from 2018 to 2019 and became the most popular ransomware virus in 2018. It has infected more than 500000 computers have caused a total of more than 300 million US dollars in economic losses. There are many such cases, and we can also see the importance of network security and computer security to society and the country.

Although there are many defensive measures, such as firewall technology and anti-virus software, most of them are passive defensive, which can be used to cope with relatively simple attacks, but the continuous improvement of attack methods is obviously inadequate [3]. Obviously, this approach can no longer meet the needs of contemporary network security. In this context, a more proactive technology-intrusion detection system (IDS) came into being [4]. The intrusion detection system is mainly used as the second line of defense behind the firewall technology. It mainly performs security detection and analysis on the data passing through the firewall, and discovers whether the data is offensive in real time. Once the attack data is found, it will respond and adopt corresponding strategies. Process the attack data. Intrusion Detection System (IDS) has become one of the important technologies and research directions in the security field in the past 20 years of development [5].

With the deepening of the cognition and understanding of the immune system of organisms, the characteristics of the immune system, such as specific recognition ability, memory function, self-regulation ability, etc., all coincide with the expected characteristics in engineering practice. Drawing lessons from the immune system of organisms and using the adaptive characteristics of the immune

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0071

system of organisms as the basis to solve many problems encountered in practical engineering and applications, the artificial immune system (AIS) was born. A new subject with vitality. Therefore, the introduction of the relevant principles and properties of the artificial immune system into the intrusion detection system is also of great help to improve the detection efficiency of the traditional intrusion detection system. The intrusion detection system based on artificial immune system has broad research prospects.

Negative selection algorithm (NSA) was proposed by Forrest in 1994 based on the principle of immune tolerance process in the immune system and applied in the research of intrusion detection system [6]. Intrusion detection based on immune negative selection algorithm can be divided into two main parts: one is the process of generating mature detection set, and the other is the process of intrusion detection. Simply put, the generation process of the mature test set is to simulate the process of immune tolerance in nature to generate mature T lymphocytes. This is one of the core contents of the immune negative selection algorithm. First, define a self-set, and then randomly generate some immature test sets to be tested. Then let these immature detection sets and self-sets perform corresponding partial matching. If the matching is successful, it means that this detection set is an immature detection set and needs to be eliminated. Conversely, if it does not match, it means that the immune tolerance is successful and can become a mature test set.

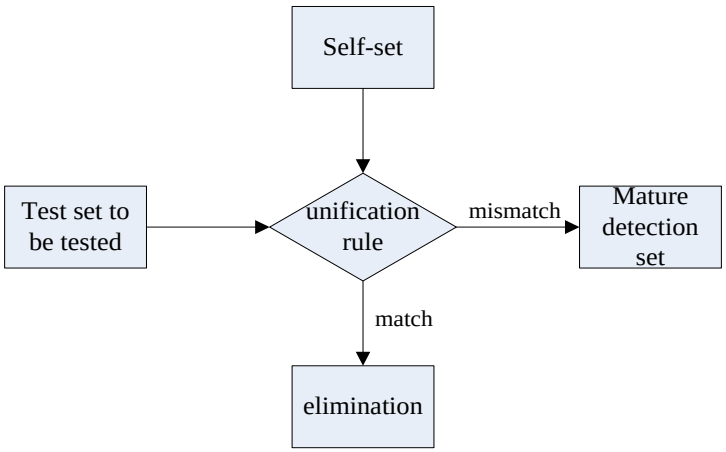

Fig. 1 Generation process of mature detection set

The other process is basically the same as the detection systems of other intrusion detection systems: after feature extraction and coding of data from the network, and the mature detection set that has undergone the immune tolerance process, it is based on a certain The partial matching scheme is used to match bit by bit. If it matches, it means intrusion data. If it does not match, it is normal data.

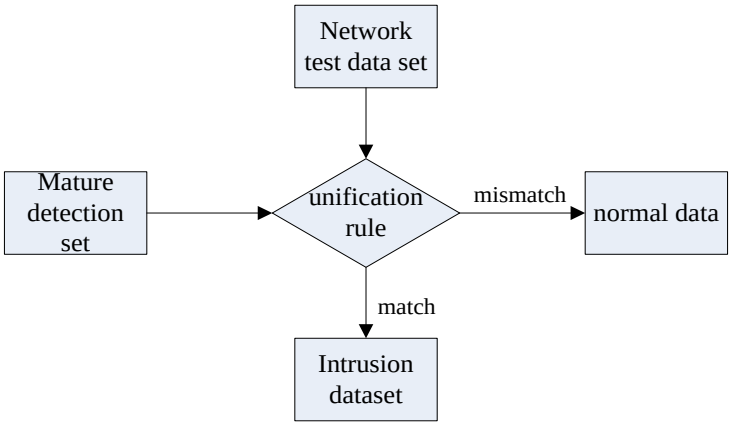

Fig. 2 Detection process of immune negative selection algorithm

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0071

2. Matching Scheme in Negative Selection Algorithm

In the negative selection algorithm, if the immature test set wants to become a mature test set, it depends on the partial matching scheme with the self-set data. Therefore, the matching scheme directly determines the efficiency and success rate of the mature test set generation, and The difference in the focus of each matching scheme will also lead to the effect after the detection set is generated.

2.1 Hamming Matching Rule

Suppose there are two strings X and Y, and they are both binary strings of length L, set the threshold σ, and its range is (0,L). The Hamming rule requires that if X and Y match under the threshold σ, the sum of the same number of bits in the corresponding positions of X and Y is required to be greater than or equal to the threshold σ. E.g:

X: 00110101 Y: 01101001

Set two strings X and Y with a length of 8, it can be seen that there are 4 bits in the corresponding positions that are the same, that is to say if the threshold setting range is (0,4], it means that the string X and the string Y are the same Matching, if the threshold is set to be greater than 4, it means that the two strings do not match. Hamming matching mainly considers the matching of the overall discrete bits, that is, it depends on the sum of the matching of the two characters at each corresponding position [7] The Hamming matching rule is a relatively simple and easy to implement matching algorithm, but it also has some problems, because the Hamming matching only pays attention to the whole but ignores some important tasks, which may lead to some relatively high-quality Mature detection sets are eliminated, and some poor detection sets become mature detection sets, which will inevitably lead to a large number of false positives and other junk data during intrusion detection.

2.2 R-continuous Bit Matching Rule

Suppose there are two strings X and Y, and they are both binary strings of length L. The r-continuous bit matching stipulation is that if these two strings match with consecutive r bits from any position, then these two The string satisfies r-continuous bit matching [8]. E.g:

X:11000110 Y:10100101

Given two strings X and Y of length 8, it can be seen that they match in the three consecutive digits starting from the 4th digit to the 6th digit. Therefore, it is said that in the case of r=3, it satisfies r-continuous Bit match. But if r takes a value greater than 3, it does not satisfy r-continuous bit matching. For any two strings (x, y) of length L, the probability of satisfying r-continuous bit matching is [9]:

$$P(x, y)=2^{-r}((L-r)/2 +1) \quad (1)$$

The choice of r plays a crucial role in the matching rate. If you make some changes to r, the matching rate will be greatly fluctuated. For example, if r is replaced with r-1, the probability will probably increase by a factor of 1, so The choice of r has become a very tricky thing, too large and too small values will have a greater impact on the matching rate. Therefore, finding a suitable r value is the key factor in r-continuous bit matching [10].

R-continuous bit matching can make up for some deficiencies in Hamming's matching rules to a certain extent. r-continuous bit matching can better reflect the similarity between two strings, because it mainly compares continuous bits instead of overall bits, so r-continuous bit matching can better pay attention to the similarity of some important bits degree. But r-continuous bit matching also has some problems. r-Continuous bit matching mainly focuses on continuous bit matching. In other words, it cares about the part rather than the whole. If there are multiple intervals of continuous matching and multi-continuous matching in the matching, although the actual effect of multiple continuous matching may be better than multi-continuous matching, However, the use of r-continuous bit matching will use multi-continuous matching instead of matching with better actual effect, which will also cause some good detectors to be ignored, so r-continuous bit matching also has its limitations.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0071

However, for some matching schemes, the existence of a "black hole" is certain. The "black hole" is in the data that passes the detection, and there are some cases where the detection set cannot be detected because of the similarity of some data to the self set. Generally speaking, as the detection threshold increases, the number of "black holes" will decrease [11].

2.3 Improved Matching Rules based on Specific Code Sets

In the previous two subsections, I mainly introduced the common matching rules in the immune negative algorithm, but these rules are tried to be used in different situations. That is to say, these two solutions are proposed to solve specific problems, so how to deal with these two Proper combination and use of solutions to achieve better detection results and reduce the number of "black holes" are the main problems to be solved.

In this regard, this article proposes the following solutions. The first is to try to determine the known abnormal data set. For the data set that is already normal, let it pass the matching directly, because it must be normal data (the generated self set should also Most of them are normal data, and the purpose is to find abnormal data, so normal data can be passed directly, but it cannot be used as a detection set, that is, normal data sets must be eliminated in the detection set generation plan), and For Probing data, special treatment is also required, because it is very likely that the computer system's anti-virus software and other programs also need to monitor the computer, this kind of set a threshold r0, if its type is "Probing" type, directly compare whether it matches r0 , If it matches with the self-set, it will be eliminated, if it does not match, it can be regarded as a mature detection set. For the remaining two cases, double detection rules are required. First set a threshold interval [r1,r2], if the self-set and the detection set match according to Hamming when the threshold is r2, they will be eliminated directly, but if the self-set does not match at r1, these data will be collected , And then this batch of data will be matched by the r-continuous bit rule. The rules are still in accordance with the above rules. If at r1, the set that does not match the self set directly becomes the mature detection set, and there is another case that matches when the threshold is r1 If there is no match at r2, directly turn this batch of data into a mature detection set, because the lower threshold is the minimum found value we set, that is, it can meet our requirements in these. Just use this double rule to increase the detection degree of the scheme. The specific flow chart is as follows:

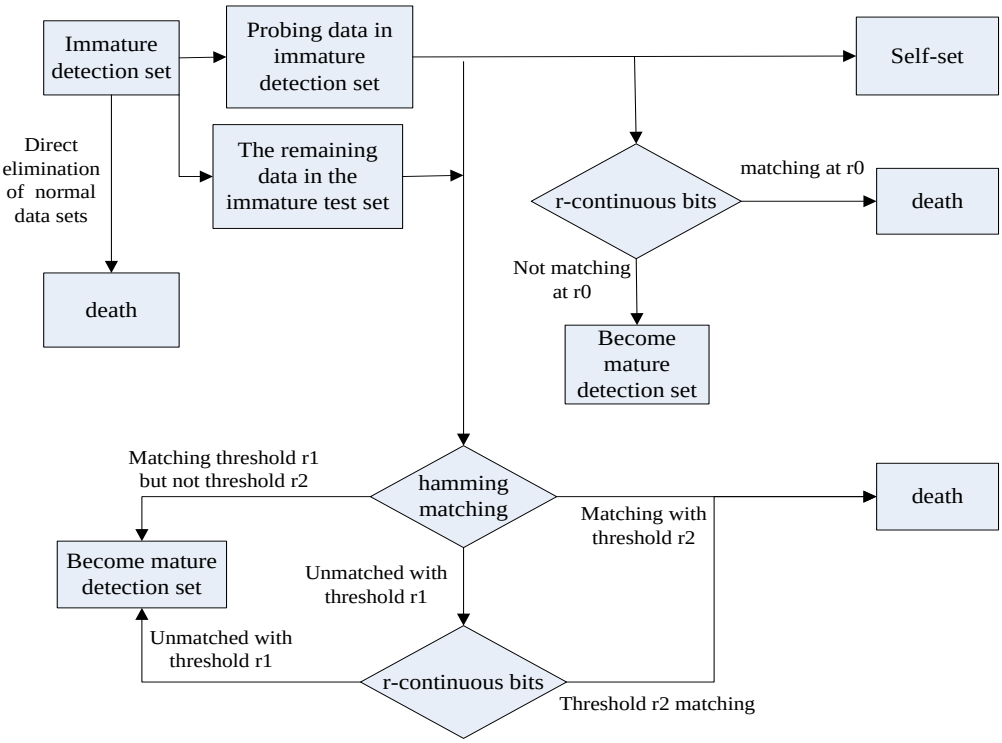

Fig. 3 Flow chart of improved algorithm

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0071

3. Intrusion Detection System Model based on Improved Negative Selection

The previously proposed a negative selection algorithm based on the two-way rule of a specific coding set, this section will establish an intrusion detection model on the previous basis, this model is designed for the detector generation process and how to detect, and perform experiments to compare Intrusion detection system and false report rate of intrusion detection system and traditional algorithm model based on this model.

Based on the learning and research of artificial immunization and negative selection algorithm above, an intrusion detection model based on an immunostatic selection algorithm is established. The model generally includes: the coding process of the data set, the generation scheme of the maturation detection set, the detection scheme of intrusion data, and the final result comparison. The basic structure is shown in Fig. 4:

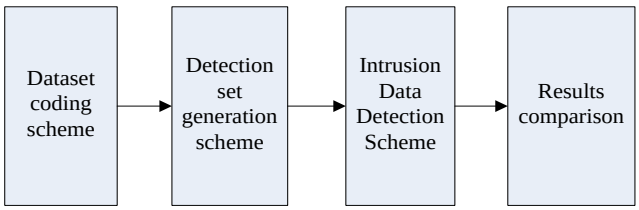

Fig. 4 Basic structure of the model

The intrusion detection system based on immune-negative selection algorithm designed herein mainly includes two detection sets: unproved detection sets and maturation detection sets. The first step is to create an auto set first, and the self-integration is randomly generated 13 digits, but considering that some threats of threatened data may occur in the autologous concentration, it is not limited to the form of autologous sets. However, this may result in differences in data results, therefore require multiple tests to find a suitable autologous set. A large amount of unproved detection sets in unproved detection concentration is randomly generated to increase the quality of the detection. The mature detection set is to be able to change the data set of multiple matching tests by immunoassay. The maturation detection set is mainly used to test whether the intrusion data set is matched with itself.

The detection scheme of the intrusion data set is as follows:

1. First, the feature encoded is a binary string for the selected intrusion data set.

2. After the encoding is completed, the matching of the modified scheme is matched, if match, the system will record this data set. If it is a normal data set, it does not match, the system jumps to the next data to be detected to continue to match the mature detection set.

In the contrasting scheme, the detector generation scheme is made in accordance with the sea matching, and the intrusion detection scheme is generated by the intrusion data set and the mature data set according to the R-continuous bit matching.

4. Experimental Results and Analysis

The operating system used by this article Windows 10 operating system, the programming language used is Python, the IDE used is Microsoft's VSCode and VS2019. This article data processing uses Pandas and Numpy modules in Python. This article random number is generated using the Random module in Python, the CSV file in the Python used by the process module of the CSV file. This article uses the KDD CUP99 data set as the main data source of this experiment. The KDD CUP99 dataset is from a Nine multi-week network connection data acquired from a US Air Force LAN, including more than 7 million network connection data [12]. These seven million data are divided into approximately 5 million identified training data sets and two million unconssed test data sets. This test data set contains some attack types that do not appear in training data sets, which also make

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0071

intrusion data sets more authentic. The KDD CUP99 data set is composed of 42 attribute feature values. It contains 41 fixed feature properties and 1 class identity. In the KDD CUP99 training data, it includes a normal identification type NORMAL and twenty-two attack data identification types, which can be generally divided into five types.

Table 1. Data identification type

Type	Meaning	Specific classification identification
Normal	Normal record	normal
Dos	Denial of Service	back, land, neptune, pod, smurft, teardrop
Probing	Monitoring and other detection activities	Ipsweep, nmap, portsweep, satan
R2L	Illegal access from remote equipment	ftp_write, guess_passwd, imap, multhop, phf, spy, warezclient, warezmaster
U2R	General users access to super users	buffer_overflow, loadmodule, perl, rootkit

The experimental steps in this article are mainly divided into four parts:

1) First convert the data set to the CSV file and make basic processing in Python and the PANDAS module, and encode it according to the coding scheme provided in Chapter 3, and the data after encoding is re-use CSV. The module is output to a CSV file with the PANDAS module to support subsequent operations.

2) Writing a partial matching algorithm is the sea matching code, R-continuous bit matching code, and some matching schemes improved according to the encoding scheme.

3) Read the dataset after the encoding is completed, then use various schemes to achieve the production plan of the mature detection set, and the detection scheme of the intrusion data set. In the experiment, the number of autologous sets is different (because the number of individual sets can also show the difference in detection), the intrusion detection data selected by the experiment is also divided into different segments (500, 11000, 2000, 3000, 5000) 10000) The matching rate of each segment also exhibits differences, and the R value in the experiment is between 6-9.

4) The final job is to compare the visualization of the data. The comparison is mainly compared to control the variable, which is: the difference between the different segments of matching rate, respectively, and the difference in matching rate of different autologous sets between the same paragraph, And the difference between the two options.

First give a comparison of the segment detection sets of different autologous sets. (The result of a total of autologous episodes is divided into 20, 50, 100, 1000. Here, only some of the result map is compared in the paper)

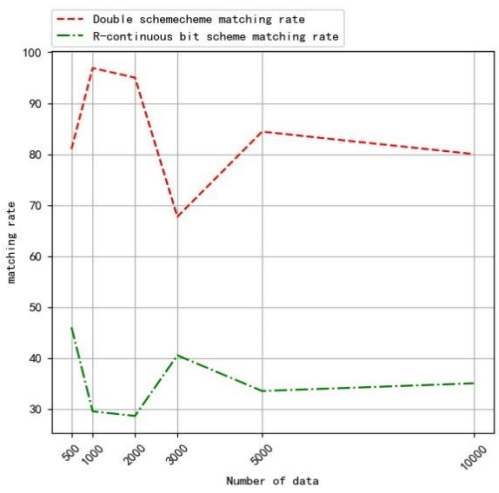

Fig. 5 Detection rate of autologous set 20

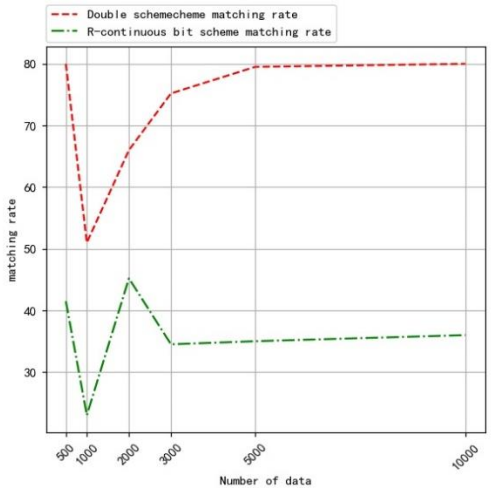

Fig. 6 Detection rate of autologous set 50

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0071

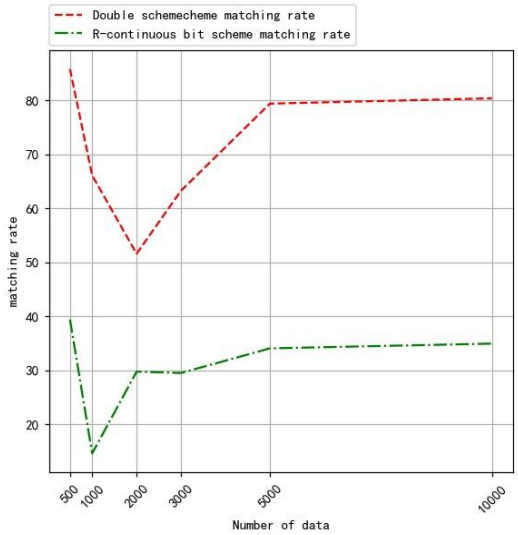

Fig. 7 Detection rate of autologous set 100

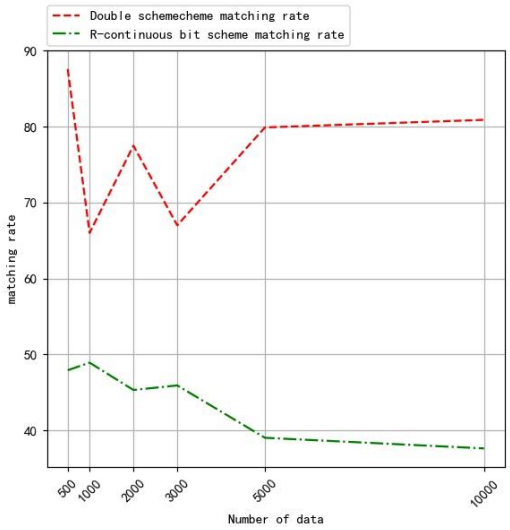

Fig. 8 Detection rate of autologous set 1000

Through these experimental results, the differences in the two programs can be apparent that the matching rate of the Double scheme is much higher than the matching ratio of the R-continuous matching scheme. When the amount of intrusion data is less, the fluctuation of the matching rate is still large. However, after the number of detection sets increases, the matching rate tends to be relatively stable. The average matching rate in the improvement scheme is approximately 80%, while the average matching ratio of the R-continuous bit matching scheme is approximately 40%. Thus, their differences and superior are huge. However, when the autologous set does not, the difference between the segments in the same case is not particularly large.

The intrusion detection model of the improvement plan is increasing the matching rate, the false positive rate is also a chance to "black hole", as shown in Fig. 9:

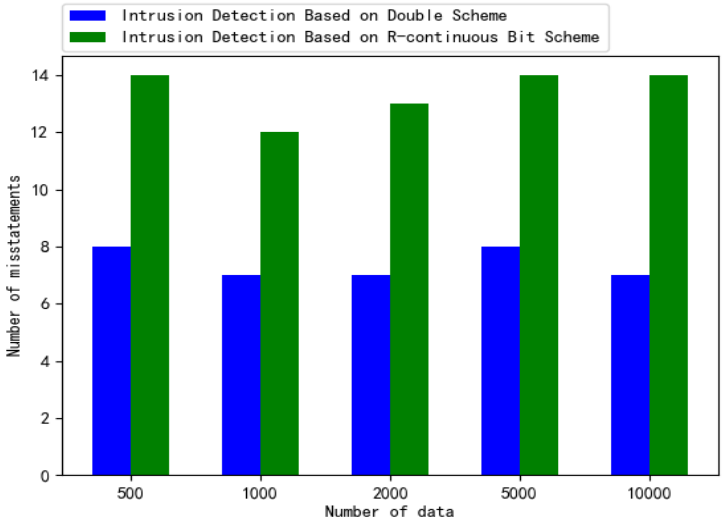

Fig. 9 Comparison of false positive rate when autologous set is 1000

This false positive rate is when the autologous set is 1000, in the case where other autologous sets are small, the generation of false packets will also have a lot of accidents, and thereby chooses a lot of autologous sets to avoid such an event. It can be seen from the histogram, and the false positive rate in the model is large, and the false report rate in the model is maintained in a higher level. This shows that the custom encoding improvement solution will have declined for "black holes".

ISSN: 2414-1895 DOI: 10.6919/ICJE.202110_7(10).0071

5. Conclusion

This paper mainly introduces the application of the immune negative selection algorithm in the intrusion detection system, explores the use of the traditional negative selection algorithm to complete the intrusion detection task, and improve the traditional negative selection algorithm, the traditional negative selection algorithm and the improved negative selection algorithm are used. In the experimental comparison of intrusion detection, the experimental results show that the intrusion detection rate and false report rate of the method model have a certain advantage.

References

[1] X.Y. Can, Y.J. Wang, Z. Xue, et al. Analysis and Verification of the Vulnerability of RDP. Cyberspace Security, 2016, 7(Z1):49-53.

[2] L.Hu. New GandCrab Leso virus protection. Computer & Network, 2019, 45(05):49.

[3] E.J.Liu. Application Analysis of Firewall Technology in Computer Network Security. Network Security Technology & Application, 2020(10):34-35.

[4] G.Z. Wang, T.G. Qu. Intrusion Detection System Research and Development Overview. Secrecy Science and Technology, 2019(02): 30-35.

[5] S.Y. Sheng, Y. Meng, S.H. Qing. A New Firewall System. Chinese Journal of Computers, 2000 (03): 231-236.

[6] Forrest S. Computer Immunology. Communications of the ACM, 1997, 40(10);88-96

[7] L. Xi. The Self Region Optimization and Detector Generation of IDS Based on Immunity. Harbin University of Science and Technology, 2009.

[8] M.Q. Zhang, J. Cheng, H.S. Kong, et al. Improved r-continual position match algorithm. Computer Engineering and Design, 2014, 35(08):2650-2654.

[9] L. Han, C.Y. Li. Research on r Contiguous Bits Matching in Immunity Network Intrusion Detection. Software Guide, 2012,11(11):60-62.

[10] X. Feng, T.L. Zhao. Research on intrusion detection system using improved artificial immune algorithm. IEEE International Conference on Computer Science and Information Technology, 2011: 636 640.

[11] Y. Zhang, X.C. Zhou, H.B. Shen. Improvement of negative selection algorithm based on hamming distance. Journal of Mechanical & Electrical Engineering, 2007(09):1-4.

[12] J.S. Wu, W.P. Zhang, Y. Ma. DATA ANALYSIS AND STUDY ON KDDCUP99 DATA SET. Computer Applications and Software, 2014, 31(11):321-325.