International Journal of Security and Its Applications
Vol.9, No.12 (2015), pp.1-10
http://dx.doi.org/10.14257/ijsia.2015.9.12.01

Technique for Intrusion Detection based on Minkowsky Distance Negative Selection Algorithm

Niu Ling, Feng Gao-feng and Peng Hai-yun

Zhou Kou Normal University, Zhoukou 466001, China
JiYuan Vocational and Technical College, JiYuan Henan 454650, China
Zhou Kou Normal University, Zhoukou, 466001, China
Niuling@zknu.edu.cn, fengjyjava@126.com,hangfan_2007@163.com

Abstract

Traditional negative selection algorithms often result in a number of black holes, which directly leads to the missing alarm drawback in the intrusion detection system. In order to settle the above problem, a novel negative selection algorithm based on Minkowsky distance is proposed. Firstly, the proposed algorithm computes the Minkowsky distance between the detectors. Then, compute the serial same numbers between the detector and self-set strings, which is helpful to improve the coverage area of the detector. Finally, the new detectors after training and renewal are put into the mature detector set to decline the number of black holes. Experimental results demonstrate that, compared with the traditional negative selection algorithms, the number of black holes and the missing alarm rate decline a lot in the proposed algorithm.

Keywords: *Artificial immune; Negative selection algorithm; Minkowsky distance; Intrusion detection*

1. Introduction

The development of computer network is rapid in recently years, which not only brought great convenience to people's production and living, but also brings many new security risks. The negative selection algorithms (NSA) come from the organism autoimmune perspective, which is proposed by Forrest. The NSA put forward a new concept and model of computer network system of self maintenance. It overcomes the defects of the traditional anomaly detection model to a certain extent. At present, NSA has been widely used in pattern recognition, intrusion detection and so on. In the whole NSA system, the key performance of the algorithm is always the generation and training mechanism of detector. Therefore, the related literature [2-10] in recent years is focused on the problem of improving generating mechanism of detector in NSA.

Aiming at the defects of real valued detector generation process in the classic NSA, the paper [2] presents an optimization algorithm. In order to solve the problem that current candidate detectors need to spend a lot of time with the whole data set of matching, the literature [3] put forward a kind of self set hierarchical clustering based on NSA, effectively improve the operating efficiency of the algorithm. As the problem that vulnerability exists widely in the real-valued NSA, The literature [4] presents a boundary detector using real-valued NSA, which effectively reduced the vulnerability. The traditional binary chaotic NSA is extended to n-dimensional chaotic NSA in the literature [5], which re-examine the negative selection problem from a multidimensional perspective, greatly improved the detector generating mechanism. The traditional state space is extended to the matrix level in literature [6], and a matrix form of NSA is put forward, and then a new detector generating algorithm is built. In view of the coverage of the mature detector and the existing detector, literature [7] proposed dual NSA, which

ISSN: 1738-9976 IJSIA

removed the cover detector samples to purify the detector set, and improved the mature detector generation efficiency and the efficiency of the algorithm.

In view of the existing literature of NSA, the Minkowsky distance model is used in the field of NSA in the study. However, there are a large number of "black hole" exist in the traditional NSA. As a result, a novel NSA based on Minkowsky distance is proposed in this paper. Firstly, the proposed algorithm computes the Minkowsky distance between the detectors. Then, compute the serial same numbers between the detector and self-set strings. And then adjust the detector coverage, so corrected detector coverage "boundless" defects in the past literature. Finally, the new detectors after training and renewal are put into the mature detector set to be integrated, effectively reduce the number of "black hole". The simulation results show that compared with the traditional NSA, the method proposed in this study can greatly reduce the number of "black hole", reduces the missing alarm rate, and greatly improving the detection efficiency.

2. Negative Selection Algorithm (NSA)

The computer immune system model of Forrest and the intrusion detection system model are very similar in the abstract. In this model, the security problem is considered as a kind of immune system of "self-set" and "non-self-set" problem. The principle of intrusion detection is to construct a boundary in the "self-set" and "non-self-set", so that, it became easy to improve the rate of the missing alarm, which has a bad state in traditional anomaly detection algorithm.

The negative selection algorithm is a kind of self or non-self-recognition technology in the immune system. The mechanism is inspired by the maturing process of the T cells in thymus. If T cell recognizing self elements it will be cleared. Tlymphocytes (T cells) is Produced by the bone marrow in the biological immune system, which plays the role of the coordination each part of the immune system in defense against foreign attack. The immature T cells will develop into two categories of T cells in the thymus, One of a class of immature T cell is destroyed due to autoimmune, while another kind of mature T cells would "learned" how to fight against a specific intrusion. So "negative selection" mechanism is the process of pick up the mature T cells which can resist the invasion from all immature T cells.

The traditional NSA is descripted as follows [11]:

Step 1: The definition of "self-set". A single binary string is divided into equal length string set. Based on the "self-set" named *S*, "non-self-set" called *N* is defined which is not include in *S*. *U* is all equal length string set, $S \cup N=U$, $S \cap N=\Phi$.

Step 2: Generate detectors. Creating effective detection element set *R*. Any one one element in *R* cannot match the elements in *S* (self-set). If any one element in *R* can match the element in *S* (self-set), then the model *R* become the detector and should be removed. Otherwise, the model *R* becomes a detector and will be stored in the detection system.

Step 3: The emergence of monitoring abnormal. The element in *R* is used to monitor the detection. The detection element matching with self-mode is discarded in step 2. Therefore, when the monitoring model matches with a valid detection element, the detecting element will be activated. It suggests that something abnormal happens.

Step 4: Identify whether there is abnormity by matching rules. In the matching rules in a continuous *r*, according to the continuous matching number to determine whether the two strings are matched. When the continuous. matching digit is greater than or equal to the *r* value, the two strings are matched, otherwise they are not matched

3. Intrusion Detection Model based on Negative Selection

The core technology of this intrusion detection model is the generation of detector, which is based on negative selection. The generating algorithm of the specific detector is described as following:

Step 1: Initialization. The "self-set" M is defined the length of whose element is l .
Step 2: Randomly produce a string (named a) which length is l.
Step 3: According to the matching rule, the string a is used to compared with elements of M sequentially
Step 4: If string a matched any element of M, then string a cannot become detector and should be removed, then go to step 3.
Step 5: If string doesn't match any element of M, then string a can become a detector and will be saved in the mature detector (The new detector will be the mature detector after training and renewal).

In the intrusion detection model based on negative selection, the NSA did not match the "self" sequence which often leads to the emergence of the "black hole" phenomenon. Black holes cannot detect "non-self" mode sequence, so that the missing alarm appears. As a result, the intrusion detection system should reduce the number of black holes as much as possible. According to the "self set" M we define the non-self-set named N, "Black hole" can be defined as a sequence named b has been existed in "non-self-set" N, $b \in N$, b is matched with a sequence of "self set" M. In the detection model, the more "self set" and "non self set" are similar, the more the number of "black hole" is. Therefore, the NSA applied in network intrusion detection will produce a large number of false negatives. It will bring huge waste in time and space which resulting in low detection efficiency. Figure 1 shows the diagram of black holes.

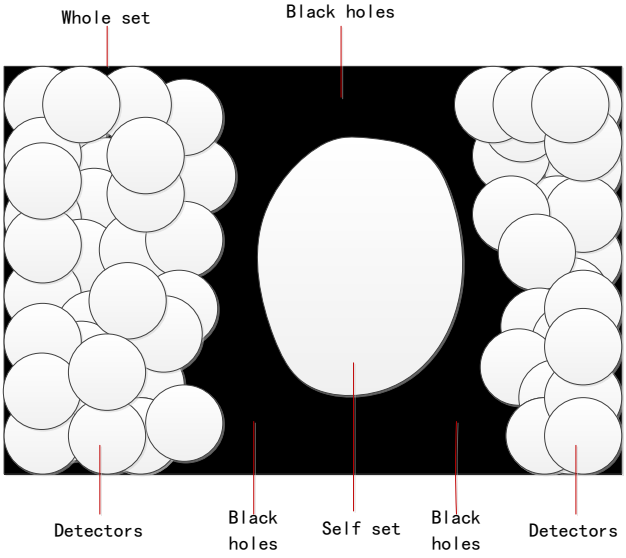

Figure 1. Schematic Diagram of the "Black Hole"

4. Negative Selection Algorithm Based on Minkowsky Distance

Minkowsky distance is wildly used in distance formula between the real vectors. It also can calculate the distance between the strings. It is Manhattan, Euclid distance formula generality expression. The core thought of NSA based on Minkowsky distance is that the parameter r is added in the matching rules based on string, which improved the detector coverage effect, and reduced the "black hole" quantity produced in the traditional NSA.

All of this will improve the efficiency of detection. Process of the algorithm is described as follows:

(1) generate a detector sequence (named *d*) of length *L* randomly;
(2) According to the matching rule, the sequence of *d* compared with the elements of "self-set" *M* sequentially;
(3) If *d* belong to *M*, *d* cannot be detector and will be deleted, then got to step (2);
(4) If did not belong to *M*, *d* is a new detector, and will be saved in the detector set *M*, ($R \cup \{d\} \rightarrow$R);
(5) When the value of *R* changes to the threshold which is set according to the matching rule, the process will be stopped; otherwise jump back to step (1).

The matching rule is the affinity calculation is to calculate the Minkowsky distance between the strings, which descript the similarity between antibody and antigen in the immune system model. In this kind of algorithm based on string matching rules, testing the sequence matching efficiency and detection failure rate should be done at first, and then system detection performance can be evaluated. So, the definition of sequence matching efficiency (named E_m) and the definition of detection failure rate (named E_f) are described as follows:

$$E_m = \sum_{i=r}^{l} C_l^i \cdot \left(\frac{1}{2}\right)^i \cdot \left(\frac{1}{2}\right)^{l-i} = \frac{1}{2^l}\left(\sum_{i=r}^{l} C_l^i\right) \quad (1)$$

$$E_f = (1 - E_m)^{N_R} \quad (2)$$

In the Formula (1), (2), E_m and E_f are respectively for sequence matching efficiency and the detection failure rate, *C* is the sequence matching balance factor, *l* is the length of the detector which is generated randomly, *i* is the length of detection sequence, *N* is a "non-self" sequence set, *R* is the threshold of matching rules set in advance. If the Minkowsky distance of two detector sequence is less than the value of *l-r*, it means the detector sequence matches a sequence of "self set" *M*.
The detailed steps of NSA based on the Minkowsky distance are followed:

Input: self-set *M*, non-self-set *N*, sequence matching balance factor *C*, a preset threshold *R*, self set string $\phi_r (1 \leqslant r \leqslant N_\phi)$

Output: sequence matching efficiency E_m, and the detection rate of failure E_f.

Step 1: initialization. Self set M became the effective detector after training, the threshold is set as *R*, the equilibrium factor *C = l-r*, the parameter *m= r-1;*
Step 2: according to self-set string φ which is randomly selected while input, detector (named d) generatd effectively.And the uncertain sequence digit of *D* is *i = r*-1;
Step 3: calculating the Minkowsky distance *l* between detector *d* and detection sequence in Self set *M*. if $l > N_\varphi$, The detector *D* will be added to the effective detector set *M*, till the element number of set *M* reaches the threshold R;
Step 4: If the two test sequence Minkowsky distance *L* is less than or equal to *N*, $l \leq N_\varphi$. Calculating consecutive identical digits(j) at the corresponding positions between detector(d) and string φr:
if *j = r*， delete detector *d*, and turning to step2;
if *j = r*-1， The uncertain sequence bits of detector d are replaced by the corresponding anti - bit of $\varphi_{r,}$,the parameter *m* is set to 0, and turned to step3;

If $j < r\text{-}1$, and $j + m \leq r\text{-}1$, the digit of the assured sequence of detector d does not changed, and turning to step3;

If $j < r\text{-}1$, and $j + m > r\text{-}1$, The detector d and φ_r randomly generate a new detector sequence t with the equilibrium factor C = 1-(r-1-j), and $t = d$.

Step 5: According to the new test sequence t, Em and Ef can be obtained. The calculation formula updated is described as follows:

$$E_m = \sum_{i=r\text{-}t}^{l\text{-}t} C_{l\text{-}t}^{i} \cdot \left(\frac{1}{2}\right)^{i} \cdot \left(\frac{1}{2}\right)^{l-i} \tag{3}$$

$$E_f = (1 - E_m)^{N_R} \tag{4}$$

5. Intrusion Detection Test

In the above algorithm, for the same detection sequence d, select a different test sequence length of L, parameter r, sequence matching balance factor C and the preset threshold R *etc*, then get different Em and Ef. That is, all the above parameters can affect the performance of the algorithm. As for the negative selection algorithms (NSA), the test sequence length l and parameter r are the dominant effect on performance of the algorithm. So when testing the detection performance of the algorithm, you can select different values of l and r for comprehensive evaluation, and obtain the practical evaluation results.

5.1. Test and Analysis of Algorithms

(1) When the values of l and r changed, experimental comparison of Em of two algorithms.

20 different groups (L, R) parameter value is selected, According to the traditional negative selection algorithm (NSA) and the negative selection algorithm based on Minkowsky distance (MNSA) proposed in this paper, the obtained experimental results of the Em as shown in Figure 2.

From the above results comparison chart of Em we can see that the matching efficiency of the MNSA is slightly higher than the NSA. When the L value is same, with the increase of R, the matching efficiency of the two algorithms will decline. This fully shows that the regularity control effect of the parameter r is indeed on the coverage of the detectors.

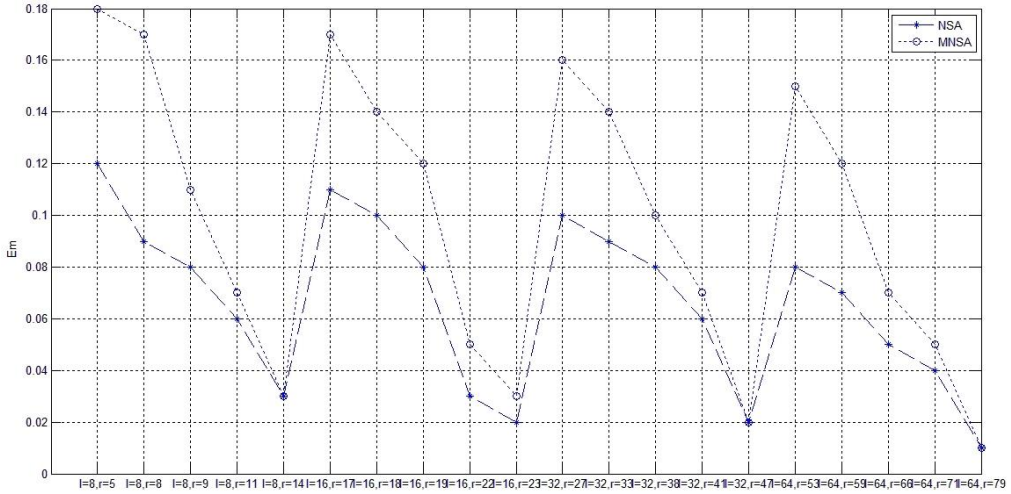

Figure 2. Comparison of Em Results of Two Algorithms

We can easily observe two conclusions in Figure 2 as follows.

(a) When the parameter l is a constant, the sequence matching efficiency of both NSA and MNSA decreases with the parameter r gradually increasing, because there is relation between the r value and the filtering number of the algorithm. The greater the value of r, the more filtering number of the algorithm with an exponential form, and hence the sequence matching efficiency E_m will decline accordingly. If the value of r is small, the E_m value of MNSA is obviously better than that of traditional NSA algorithm. However, with the r value growing namely the filtering number increasing, the E_m value of MNSA is finally approximate to that of traditional NSA algorithm, as the parameter pairs including (l=8, r=14), (l=32, r=47) and (l=64, r=79) shown in Figure 2.

(b) When the value of parameter l grows, the detecting sequence length will also increase. As a result, the detecting efficiency E_m will also fall. However, regardless of the values of the parameter pair (l, r), the sequence matching efficiency of the proposed MNSA algorithm is better than that of traditional NSA algorithm.

(2) When the values of l and r changed, experimental comparison of the "black hole" number.

Select 20 different groups (L, R) parameter value, According to the traditional negative selection algorithm (NSA) and the negative selection algorithm based on Minkowsky distance (MNSA) proposed in this paper, the "black hole" number of the experimental results as shown in Figure 3.

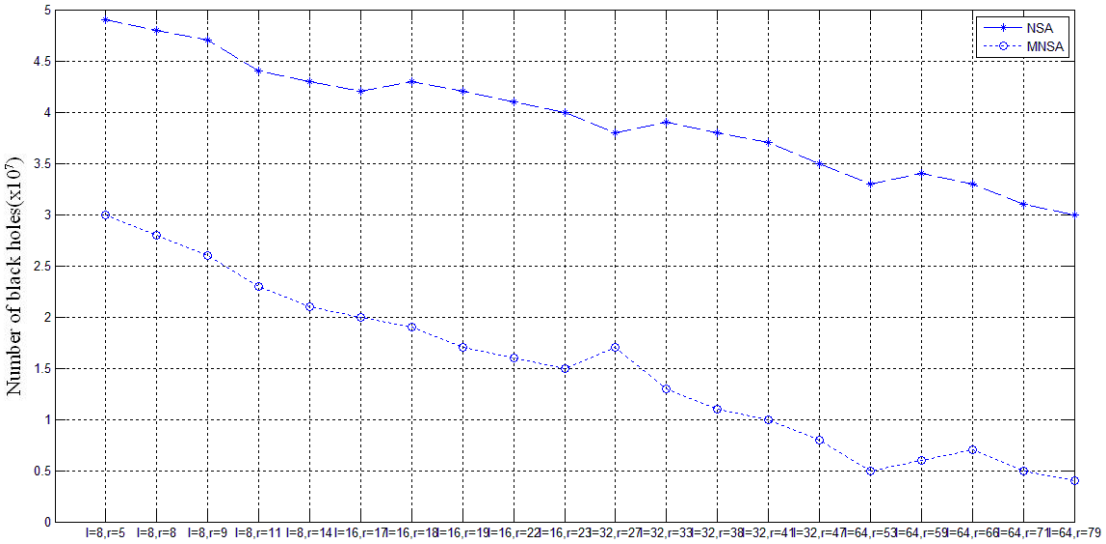

Figure 3. The Comparison Chart of two Algorithms of the "Black Hole" Number

As shown in Figure 3, five aspects can be obtained as follows.

(a) When parameter l is a constant and the value r gradually increasing, the filtering numbers of two algorithms both raise. Therefore, more and more random detectors are included in the effective detectors set denoted by M, which greatly enhances the non-self-coverage and detecting performance of M, and the number of black holes and corresponding areas gradually reduce. The

falling trends of the number of "black holes" in two algorithms are shown in Figure 3.

(b) Although the sequence matching efficiency will reduce with the value of the parameter pair (l, r) growing, the filtering number of MNSA algorithm will increase overall. Moreover, the conditions that random detectors are included in the effective detectors set will be much stricter, which will enhance the non-self-coverage and detecting performance of M, and the number of "black holes" will also necessarily narrow.

(c) Compared with the proposed MNSA algorithm, traditional NSA algorithm has only one filter. Hence, although the number of "black holes" based on the traditional algorithm reduces with the value of the parameter pair (l, r) growing, its downward trend is relatively slower than that of MNSA.

(d) The sharp falling of the "black hole" number indicates that more and more non-self-spaces are covered and detected by effective detectors. In other words, the missing alarm towards non-self during the detecting course presents an obvious downward trend, which reflects that the proposed MNSA algorithm has remarked advantages in terms of reducing the missing alarm rates.

(e) Although the number of "black holes" has an overall falling trend, the values corresponding to several parameter pairs still fluctuate accordingly, which is due to the random distribution of self-sets and non-self-sets in the experiment.

5.2. Experiment and Analysis of the Performance of Intrusion Detection Method

In this paper, the performance evaluation data of Lincoln Laboratory intrusion detection is as the experimental data, Validation and comparison of three kinds of intrusion detection model performance is done, the three models are the intrusion detection model based on traditional NSA, the one based on Hamming distance NSA and the one based on Minkowsky distance NSA. In addition, we used a total of 1000 training data and 500 test data, most intrusion attack mode has two kinds, respectively is Dos and Probe attacks. The realization environment of three algorithms are all in Windows XP (Professional) test system, The development platform is Visual Studio 2010, CPU frequency is Intel(R) Core(TM)i5-3337U, 1.8GHz, 4.00GB memory. Three performance indexes of the experiments were obtained as shown in Table 1.

Table 1. The Performance of Three Kinds of Intrusion Detection Model Comparison

Intrusion detection model	The number of attack detection	The number of false alarm	False negative rate (%)	Detection rate (%)
The traditional negative selection algorithm model	66	16	21.43	60.12
Hamming distance model based on negative selection algorithm	74	13	7.65	79.89
Minkowsky distance model based on negative selection algorithm	81	11	7.71	90.21

From Table 1, we can see that the false negative rate of the intrusion detection model based on traditional NSA is highest, the false negative rate of the intrusion detection model based on Hamming distance NSA and the false negative rate of the intrusion detection model based on Minkowsky distance NSA is almost equal; however, the detection rate of the intrusion detection model based on Minkowsky distance NSA is highest (the effective detection rate under certain conditions can

reach nearly 90%), the detection rate of the intrusion detection model based on Hamming distance NSA is lower, the detection rate of the intrusion detection model based on traditional NSA is lowest. The above data analysis shows that the new intrusion detection system model is proposed in this paper is effective and feasible.

5. Conclusion

This article conducts the research to the "black hole" problem widely exists in the traditional NSA. The Minkowsky distance was successfully introduced into the field of NSA, presents a new intrusion detection method based on Minkowsky distance NSA. The method uses Minkowsky distance to calculate the detector sequence, compute the serial same numbers between the detector and self-set strings, to modify the coverage of detector, and formed a series of new detector model, finally put it in mature detector set. Theoretical analysis and experimental data show that the method proposed in this paper can effectively reduce the "black hole" number and the false negative rate, greatly improve the detection rate.

Acknowledgements

The authors thank the anonymous reviewers and editors for their invaluable suggestions. This work was supported by the Basic and Frontier Project of HeNan Science and Technology Department under grant No.142300410432.

References

[1] Forrest S, Perelson A S, All L, *et al*. Self_nonself discrimination in a computer[C]. Procedure of IEEE Symposium on Research in Security and Privacy, Oakland. CA: IEEE Press, 1994: 202-212.
[2] Chai Zheng-yi, Wang Xian-rong, WangLiang. Real-value negative selection algorithm for anomaly detection[J] Journal of Jilin University (Engineering and Technology Edition) ,2012, 42(1): 176-181.
[3] Chen Wen, Li Tao,Liu Xiao-jie. A negative selection algorithm based on hierarchical clustering of self set[J]. China Science (Information Science), 2013, 43(5): 611-625.
[4] Wang Da-wei,,Zhang Feng-bin.Real-valued Negative Selection Algorithm with Boundary Detectors[J], China Science , 2009, 36(8): 79-81.
[5] Zhang Fengbin,Wang Tianbo.Real value negative selection algorithm with the n-Dimensional chaotic map[J] Research and development of computer, 2013, 50(7): 1387-1398.
[6] Zhang Xiongmei,Yi Zhaoxiang,Song Jianshe.Research on Negative selection algorithm based on matrix representation[J]. Journal of Electronics & Information Technology, 2010, 32(11): 2701-2706.
[7] Zheng Xufei,Fang Yonghui,Li Tao.Dual negative selection algorithm[J]. China Science (Information Science), 2013, 43(4): 529-544.
[8] Jin Zhangzan,Liao Minghong,Xiao Gang.Survey of negative selection algorithms[J]. Journal of China Institute of Communications, 2013, 34(1): 159-170.
[9] Wang Hui,Yu Lijun,Bi Xiaojun.Adjustable fuzzy matching negative selection algorithm with vaccine operator[J]. Journal of Harbin Institute of Technology. 2011, 43(6): 141-144.
[10] Zhang Pengtao,Wang Wei,Tan Ying. A malware detection model based on a negative selection algorithm with penalty factor. China Science (Information Science) 2011, 41(7): 798-812.
[11] Artificial immune system [OL]. Http://www.dca.fee.unicamp.br/~Inunes/immune.html.
[12] DASGUPTA D, KRISHNA K. Negative selection algorithm for aircraft fault detection [A]. Proceedings of Third International Conference on Artificial Immune Systems (ICARIS 2004) [C]. 2004, 1-13.

Authors

Niu Ling, she received the B.Eng degree in Computer science from Henan normal university and M.Eng degree in Computer science from Chengdu University of Technology. She is currently researching on computer application technology.

Feng Gao-feng, she received the computer science degree from Henan Normal University, China, and the master degree in computer science from Beijing University of Posts and Telecommunications. He is a member of China Computer Federation and Association of Fundamental Computing Education in Chinese Universities in Beijing, China.

Peng Hai-Yun, she Yun received the B.Eng degree in Computer science from Henan University and M.Eng degree in Computer science from Huazhong University of Science and Technology. She is currently researching on computer application technology.