Journal of Computers Vol. 32 No. 1, 2021, pp. 144-153
doi:10.3966/199115992021023201012

Network Security Situation Prediction Method Based on Support Vector Machine Optimized by Artificial Bee Colony Algorithms

Gang Ke[1], Ruey-Shun Chen[1*], Yeh-Cheng Chen[2], Jyh-haw Yeh[3]

[1] Department of Computer Engineering, Dongguan Polytechnic, Dongguan, Guangdong, China
kegang95@126.com, chen1868@gmail.com

[2] Department of Computer science, University of California Davis, CA, USA
ycch@ucdavis.edu

[3] Department of Computer science, Boise State University, Idaho, USA
jhyeh@boisestate.edu

Received 1 June 2019; Revised 1 November 2019; Accepted 6 January 2020

Abstract. The validity and accuracy of the network security situation prediction algorithm is of great significance to network security. Aiming at the shortcomings of the basic artificial bee colony algorithm, such as easy to fall into the local optimal solution and slow convergence in the late stage of the algorithm, this paper proposes a network security situation prediction model based on support vector machine(SVM) optimized by improved artificial bee colony algorithm(I-ABC), using I-ABC algorithm for SVM. The penalty factor a and the kernel parameter b are optimized. Finally, the simulation test is performed using real network security situation data. The simulation results show that the proposed algorithm can accurately track the change of situation value and effectively improve the prediction accuracy of network security situation.

Keywords: artificial bee colony algorithm, support vector machine, network security, situation prediction

1 Introduction

With the development of Internet technology, network attack means are increasingly diversified, and various network security incidents emerge in endlessly. At present, intrusion detection, firewall, anti-virus and other technologies belong to passive defense means, so it is difficult to grasp the security state of network system and take corresponding preventive measures in time. Network security situation prediction is to extract the security-related elements in the network for analysis and understanding to perceive the network security state, and predict the future network security situation. Therefore, it has become a research hotspot in the field of network security [1-4].

Since Time Bass first proposed the concept of network security situation in 1999, scholars at home and abroad have done a lot of research on the prediction of network security situation. Yu, et al. [5] used grey relational model to predict the network security situation, which can better reflect the development trend of network security situation; Li and Cao [6] used ARIMA model to predict the network security situation, which is easy to implement and has high prediction accuracy. However, the traditional linear forecasting methods have great limitations because of the characteristics of non-linear and abrupt changes in network situation. You, Ling and Hao [7] uses Elman neural network with dynamic memory function and sensitivity to historical data to predict network security situation, and achieves high prediction accuracy. Li, et al. [8] proposed a situation prediction method based on the modified Radical

* Corresponding Author

Basis Function (RBF) neural network, the performance advantage of the radial basis function is used to improve the convergence speed and accuracy of the prediction model. However, the neural network has some shortcomings, such as complex parameter setting, easy to fall into local minimum, slow convergence speed and so on. Support Vector Machine (SVM) method based on statistical theory can solve the problems of small samples and non-linearity better, which has been proved in network security situation prediction. It is superior to ANN and other methods, but this method has the problem of parameter optimization.

Artificial bee colony algorithm (ABC) is a swarm intelligence algorithm with fast operation speed and strong local search ability [9-10]. In this paper, ABC is used to optimize the parameters of SVM, which improves the shortcomings of traditional SVM and speeds up the convergence speed of the network. Finally, simulation experiments verify the feasibility and effectiveness of the algorithm.

This paper is organized as follows. In Section 2, the network security situation theory is described. In Section 3, the principle of SVM algorithm are illustrated in detail. In Section 4, ABC optimizes SVM parameters was presented. In Section 5, experimental results and discussions are demonstrated. In Section 6, the conclusions are given.

2 Network Security Situation Theory

2.1 Principle of Network Security Situation Prediction

Network security situation is a kind of data collected in time sequence, so it can be processed as a time series [11]. The former time series situation value is the input variable of the prediction model, and the output is the situation value of the next period. Suppose there is a time series of network security situation value:

$$x = \{x_i | x_i \in R, i = 1, 2, ..., L\}$$

Network security situation prediction is to predict the following M situation values through the situation value of the first N moments in the sequence. It includes three steps: situation element extraction, situation value calculation and situation prediction [12]. The prediction principle is shown in Fig. 1.

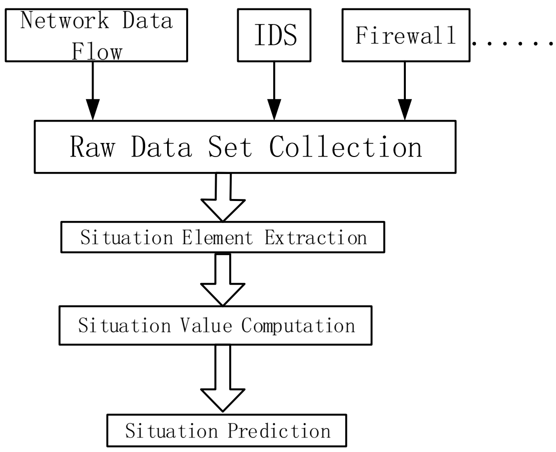

Fig. 1. Principle of Network Security Situation Prediction

Network security situation prediction firstly collects the original data of network security events from network data stream, network equipment log, host log and so on, extracts the situation elements of network security, and then calculates the situation elements weighted to get a situation value which can reflect the network operation status at a certain time. After that, on the basis of the current situation value of M moments, the situation value of N moments can be calculated by using the situation prediction model, and the situation value can be analyzed to predict the changing trend of network security situation, so as to provide decision support for network managers.

2.2 Calculation of Network Security Situation Prediction Value

This paper refers to the quantitative calculation method of network security situation value provided by Chen, et al. [13], which calculates the situation value from the service layer, host layer and system layer.

Definition 1: F_{S_j} represents the security situation index of service S_j at time t, denoted as:

$$F_{S_j}(t)=\sum_{i=1}^{k}10^{P_{ji}}C_i(t) \tag{1}$$

where, k is the number of attack types of service S_j in a certain period of time, $C_i(t)$ is the number of times that service S_j is attacked by A_{ji} in a certain period of time, P_{ji} is the severity of the attack service S_j.

Definition 2: F_{H_k} represents the security situation index of host H_k at time t, denoted as:

$$F_{H_k}(t)=\sum_{i=1}^{m}v_iF_{S_j}(t) \tag{2}$$

where, F_{S_j} is the security risk index of the service S_i s of the host H_k according to the Eq. (1); m is the number of services opened by host H_k; v_i is the weight of service S_i in various services of host H_k.

Definition 3: F_L represents the network security situation index of the current time period, denoted as:

$$F_L(t)=\sum_{l=1}^{n}w_lF_{H_l}(t) \tag{3}$$

where, $F_{H_l}(t)$ is the risk index of the host H_l calculated according to Eq. (2); n is the number of hosts in the network system; w_l is the proportion of importance of host in LAN, and the sum of importance weights of all hosts in LAN is 1.

3 Support Vector Machine Algorithm

Suppose a set of sample sets $\{(\mathrm{x}_i,y_i),i=1,2,\cdots,l\}$, $x_i\in R^n$, x_i is the input vector for the i-th sample, $y_i\in R$, represents the corresponding output vector. By using the non-linear mapping function $\varphi(\cdot)$, the samples in the original input space are mapped into the high-dimensional feature space, and then the optimal decision function is constructed in the high-dimensional feature space [14-16].

$$f(x)=w^T\varphi(x)+b \tag{4}$$

where, w is the weight vector of the feature space, and b is the offset.

According to the principle of structural risk minimization, the regression fitting problem is transformed into the following optimization problem:

$$\begin{cases}\min\dfrac{1}{2}\|w\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)\\ s.t.\begin{cases}y_i-w\varphi(x_i)-\mathrm{b}\le\varepsilon+\xi_i,i=1,2,...,l\\ w\varphi(x_i)-y_i+\mathrm{b}\le\varepsilon+\xi_i^*\\ \xi_i\ge 0,\xi_i^*\ge 0\end{cases}\end{cases} \tag{5}$$

where, C is the penalty factor, ξ_i,ξ_i^* is the relaxation variable, and ε is the linear insensitive loss factor.

By introducing Lagrange multiplier a, the solution of e Eq. (5) is transformed into an optimization problem of unconstrained dual space, as follows:

$$\begin{cases} \max_{\boldsymbol{\alpha},\boldsymbol{\alpha}^*}[-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}(\alpha_i-\alpha_i^*)(\alpha_j-\alpha_j^*)K(\mathbf{x_i},\mathbf{x_j})-\sum_{i=1}^{l}(\alpha_i+\alpha_i^*)\varepsilon+\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)y_i] \\ s.t.\begin{cases} \sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0 \\ 0\le\alpha_i\le C, 0\le\alpha_i^*\le C \end{cases} \end{cases} \tag{6}$$

The kernel function K is defined as follows:

$$K(x_i,x_j)=\varphi(x_i)^T\varphi(x_j) \tag{7}$$

In this paper, radial basis function (RBF) was selected as the kernel function, and the SVM regression model was finally obtained, as follows:

$$f(x)=\sum_{i=1}^{l}(\alpha_i\text{-}\alpha_i^*)\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{2\sigma^2}\right)+\mathrm{b}^* \tag{8}$$

where, σ is the width of the kernel.

In the algorithm of the support vector machine, the penalty parameter c and the kernel function parameter σ have a great influence on the learning and generalization ability of the SVM. In this paper, the artificial bee colony algorithm is used to iteratively optimize the two parameters of SVM, and the appropriate c value and σ value are obtained to improve the prediction accuracy [17-19].

4 ABC Optimizes SVM Parameters

4.1 Artificial Bee Colony Algorithm

Artificial bee colony algorithm (ABC) is an intelligent algorithm to simulate the process of honey harvesting. It is simple to implement and has good robustness. The bee colony algorithm consists of three categories: employed bees, onlook bees and scout bees. The number of employed bees and onlook bees accounts for half of the total number of bees. Each food source corresponds to a employed bee. The location of the food source represents the potential solution of the optimization problem. The quality of the food source corresponds to the fitness of the solution. The process of ABC algorithm solving the function optimization problem is as follows:

(1) Bee colony initialization

In D - dimensional search space, ABC algorithm randomly generates initial solution $X_i(i=1,2,\dots,N)$, N is the number of food source, and each initial solution is a vector of D dimension. X_i represents the position of the i-th food source, i.e $X_i=\{x_{i1},x_{i2},\cdots,x_{iD}\}$

(2) Employed bee search

Employed bee is searched for a new food source near the food source X_i according to Eq. (9). According to the greedy choice to update the honey source, if the new food source is better than the original food source, the new food source replaces the original food source; otherwise, the original food source is retained.

$$v_{id}=x_{id}+\phi_{id}(x_{id}-x_{jd}) \tag{9}$$

where, v_{id} is the new food source generated by the bee searching near the food source X_i, ϕ_{id} is a random number between [-1, 1], which determines the magnitude of the perturbation.

(3) Onlook bee search

After employed bee search, it will communicate with onlook bee to exchange the food source, Onlook bee select the food source according to the probability calculated by Eq. (10) and generate new food source according to Eq. (9).

$$P_i = f_i / \sum_{j=1}^{N} f_j \quad \textbf{(10)}$$

where, P_i is the probability of selecting food source for onlook bees, f_i is the fitness of the ith solution.

$$f_i = \frac{1}{T_j} \sum_{i=1}^{T_j} (y_i^* - y_i)^2 \quad \textbf{(11)}$$

where, f_i is the fitness function, y_i^* is the predicted value, y_i is the actual value, and T_j is the number of training set samples.

(4) Scout bee search

When a food source has not changed after the trial cycle, the food source position will be abandoned, and the corresponding employed bee will be turned into a scout bee. The food source X_i will be replaced by a new food source randomly generated by the scout bee according to Eq. (12).

$$x_{id} = x_{\min}^d + rand(0,1)(x_{\max}^d - x_{\min}^d) \quad \textbf{(12)}$$

where, x_{id} is a new food source, $x_{id} \in (x_{\min}^d, x_{\max}^d)$, $d \in \{1, \cdots, D\}$, $x_{\min}$ and $x_{\max}$ represent the lower and upper bounds of the search space respectively. $rand(0,1)$ is a random number between [-1, 1].

4.2 Improvement of Artificial Bee Colony Algorithms

In the basic ABC algorithm, the employed bee and the onlook bee search for the better solution according to Eq. (9). This method has the following drawbacks: firstly, in the initial stage of the algorithm, the parameters of the search equation are random, the generation of new solutions has great randomness, the exploration ability is weak, and the algorithm is easy to fall into local optimum; At the end of the algorithm, the difference between individuals becomes smaller, the development ability becomes weaker, the convergence speed of the algorithm becomes slower and the convergence accuracy is low.

In view of this shortcoming of the basic ABC algorithm, this paper proposes an adaptive search strategy, which improves the search equation by introducing an adaptive factor. It solves the problem that the ABC algorithm requires better exploration ability in the initial stage of the algorithm and better development ability at the end of the algorithm, improved artificial colony algorithm called I - ABC. Therefore, Eq. (9) is modified to Eq. (13).

$$v_{id} = x_{id} + \psi(t)(x_{id} - x_{jd}) + w(1 - \psi(t))(x_{gbest,d} - x_{jd}) \quad \textbf{(13)}$$

$$\psi(t) = 1 - \log_{\lim it}^{t} \quad \textbf{(14)}$$

Where, $i, j \in \{1, 2, \cdots, N\}$ and $i \neq j$, $d \in \{1, \cdots, D\}$, $x_{gbest,d}$ is the optimal solution under the current cycle number of the population, $\psi(t)$ is the adaptive coefficient, limit is the maximum cycle number, t is the current cycle number, w is the global optimal guidance mediation value, when $w \in [0, 1.5]$, the optimization effect of the algorithm is the best.

4.3 Parameter Optimization of SVM Based on I-ABC

The penalty factor C and core parameter σ of SVM are optimized by using I-ABC algorithm. The specific steps are as follows:

Step 1: Initialize the control parameters in the ABC algorithm. Randomly initialize the food source $x_{id}(i = 1, 2, \cdots, N)$, $d \in \{1, \cdots, D\}$, set the number of food sources: N, the dimension: D, the maximum number of cycles: maxcycle, the maximum number of food source updates: limit;

Step 2: Employed bees is working, search according to Eq. (10) to generate new honey sources; and the fitness calculated according to Eq. (11), if the fitness of the new food source is lower than that of the original food source, the new food source is replaced by the original food source by employed bees.

Step 3: Onlooker bees is working, choose food source according to the probability calculated by Eq. (12). The new food source was searched near the food source according to Eq. (10), and the new fitness was calculated. If the fitness of the new food source was lower than that of the original food source, the new food source was replaced by the original food source by the onlooker bees.

Step 4: If the food source x_{id} is not optimized after limit cycles, the food source will be abandoned, and the corresponding onlooker bee will be transformed into the scout bee, and a new food source will be generated according to Eq. (9).

Step 5: If the number of calculation cycles reaches maxcycle or the optimal fitness reaches the preset accuracy, the training ends and the optimal solution is output. Otherwise, go back to step 2.

Step 6: Transform the obtained optimal solution into SVM's parameter penalty factor C and kernel parameter σ.

The pseudo code for SVM parameter optimization based on I-ABC is shown as Algorithm 1.

Algorithm 1. Algorithmic Pseudo-Code for I-ABC Optimizing SVM

Input: number of employed bee: B, the maximum number of cycles: maxcycle, the maximum number of individual updates: limit, the number of individual updates: trial, dimension: D, cycle counts: s, the range of parameters C and σ to be optimized, the setting accuracy of optimal fitness: ε

Output: C, σ

In the search space, B food sources X_i are randomly generated according to Eq. (12), s=1, D=2;
According to Eq. (11), fit, the fitness function value of B food sources, was calculated and the global optimal value was written down.
While(globalfit<= ε **or s>maxcycle) DO**
  FOR (i = 1 ,…, B) DO   **//Employed bee search**
    A new food source V_i is generated according to Eq. (13) and the fitness value $f(V_i)$ is calculated.
    IF ($f(V_i) > f(X_i)$)
      $X_i = V_i$, $trail(i)=1$
    ELSE
      trial=trail+1
    ENDIF
  ENDFOR
Calculate the fitness value of the updated honey source, and calculate the selection probability P_i of food source X_i according to Eq. (10)
FOR (i = 1, …, B) DO   **//Onlook bee search**
  IF (rand(0, 1)< P_i)
  Generate a new food source V_i according to Eq. (13) and calculate the fitness value $f(V_i)$
    IF ($f(V_i) > f(X_i)$)
      $X_i = V_i$, $trail(i)=1$
    ELSE
      $trail(i) = trail(i)+1$
    ENDIF
  ENDIF
ENDFOR
IF (max($trail(i)$)>linit)   **// Scout bee search**
Generate a new food source X_i according to Eq. (12)
ENDIF
s=s+1
END WHILE
Output optimal solution C, σ

4.4 Comparison of Algorithm Performance Before and After Improvement

In order to verify the effectiveness of the improved ABC algorithm, four classical multimodal functions are selected for comparative experiments. The theoretical optimum values of the test function is 0. The parameters of the experiment are as follows: Population size:N=100, the maximum number of cycles: maxcycle=5000, the maximum number of individual updates: limit=200, dimension: D=40, The experimental results are shown in Table 1, Among them, Best represents the optimal value, Worst represents the Worst value, Mean represents the Mean value, and Cycle represents the average number of cycles

Table 1. Test Result

Functional equation	algorithm	Best	Worst	Mean	Cycler
Sphere Function	ABC	4.9162 E-016	3.062 7 E-016	4.53120 E-016	1076
	IABC	2.1236E-019	3.8779E-022	7.38270E-021	432
Rastrigrin Function	ABC	0.0653	4.7231 E-010	0.0078	1432
	IABC	4.9765E-014	3.745 12E-017	1.71370E-014	983
Griewank Function	ABC	0.1756	7.56211E-004	0.1182	1324
	IABC	0.1286	6.33461E-004	0.1023	796
Rosenbrock Function	ABC	3.85642E-011	8.91237E-08	7.11235E-09	3462
	IABC	2.14521E-016	7.92354E-012	5.85632E-013	1762

As can be seen from Table 1, for the test functions adopted in this paper, IABC algorithm shows higher convergence accuracy and faster search speed, and its optimization performance is higher than the basic ABC algorithm.

5 Experimental Simulation

5.1 Experimental Data Set and Parameter Settings

This paper selects a network security monitoring data set of a school campus network from March 1st to June 30th, 2018 as the situation value calculation data source. According to the method provided by section II, 120 network security situation values are calculated. The previous 96 situation values were used as training samples, and the next 24 situational values were test samples. At the same time, in order to avoid the network prediction error due to the large difference in the magnitude of the input and output data, it is necessary to normalize the calculated situation value.

The normalization formula is as follows:

$$\mathrm{x}_k = \frac{(x_k - x_{\min})}{(x_{\max} - x_{\min})} \tag{15}$$

Where, $x_{\max}$ and $x_{\min}$ are the maximum and minimum values of the network security situation, respectively, and x_k is the current network situation value.

5.2 Analysis of Prediction Results

In order to verify the effectiveness of the proposed algorithm, the ABC-SVM and PSO-SVM algorithms are compared with the proposed algorithm. The predicted results are shown in Fig. 2. From the degree of fitting of the curve, the IABC-SVM algorithm is significantly better than the other two algorithms.

For better horizontal comparison, this paper uses two indicators, RMSE (Root Mean Squared Error) and MAPE (Mean Absolute Percent Error), which are often used to describe prediction error, to evaluate prediction accuracy. Their calculation formula is as follows:

$$RMSE = \sqrt{\frac{1}{S}\sum_{i=1}^{S}(\hat{y}(t) - y(t))^2} \tag{16}$$

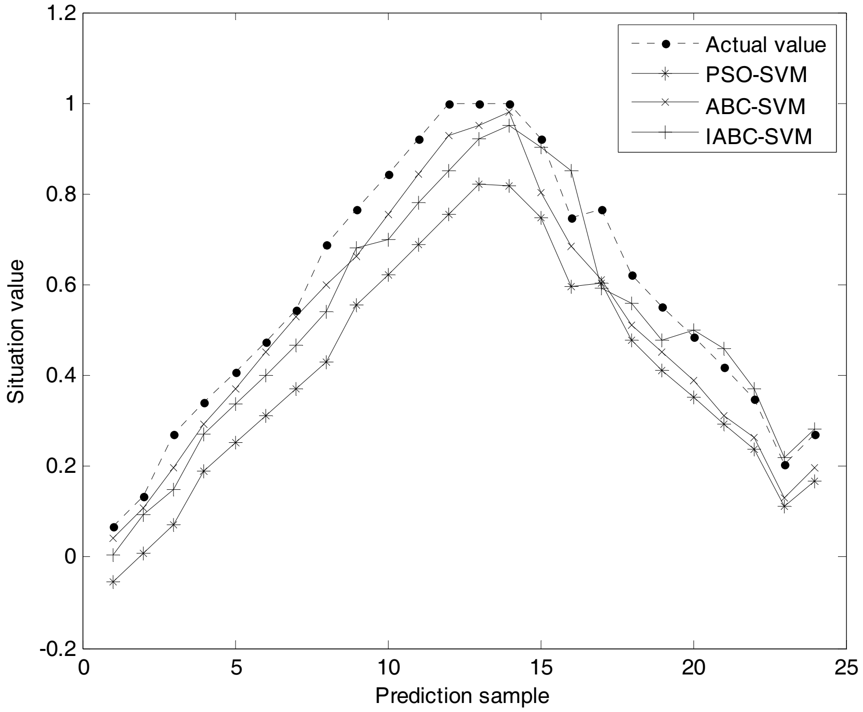

Fig. 2. Comparison of prediction curves of three algorithms

$$MAPE = \frac{1}{S}\sum_{i=1}^{S}\left|\frac{(\hat{y}(t) - y(t))}{y(t)}\right| \times 100\% \qquad \textbf{(17)}$$

Where $\hat{y}(t)$ represents the true value of the network security posture, $y(t)$ represents the predicted value, and S represents the predicted number of network security situation samples.

According to Table 2, it can be found that the error of IABC-SVM prediction is the smallest, the RMSE is 0.52%, the MAPE is 0.46%, and the error of PSO-SVM is the largest, the RMSE is 0.87%, and the MAPE is 0.81%. It can be seen that IABC-SVM has better prediction accuracy and generalization ability than ABC-SVM and PSO-SVM.

Table 2. Comparison of various methods to predict performance

Prediction algorithm	RMSE (%)	MAE (%)
PSO-SVM	0.87	0.81
ABC-SVM	0.65	0.62
IABC-SVM	0.52	0.46

6 Conclusion and Discussion

This paper proposes a network security situation prediction model based on support vector machine optimized by improved artificial bee colony algorithm. The model improves the search equation by introducing adaptive factors to accelerate the convergence speed and improve the convergence accuracy of the algorithm, aiming at the shortcomings of basic ABC algorithm, such as easy to fall into local optimal solution and slow convergence speed in later stage. The simulation results show that compared with PSO-SVM algorithm and ABC-SVM algorithm, IABC-SVM algorithm has higher prediction accuracy and is an effective network security situation prediction algorithm. However, there is still room for optimization in this algorithm. The future research focuses on improving the efficiency of the algorithm and the prediction accuracy of SVM algorithm.

Acknowledgements

This work was supported by the Technical Skills Expert Project of DongGuan Polytechnic (2019JY05); Guangdong University Student Science and Technology Innovation and Cultivation Fund (phdj2019b0900).

References

[1] G. Xu, C. Yan, Y. Ren, X. Li, Z. Feng, Network Security Situation Awareness Based on Semantic Ontology and User-Defined Rules for Internet of Things, IEEE Access 5(99)(2017) 21046-21056.

[2] C. Wang, Y. Zhang, Network Security Situation Evaluation Based on Modified D-S Evidence Theory, Journal of Wuhan University 19(5) (2014) 409-416.

[3] H. Xiao, X. Sang, Network security situation prediction based on hyper parameter optimization of relevance vector machine, Journal of Computer Applications 35(7)(2015) 1888-1891.

[4] J. Li, Z. Zhang, Network security situation prediction method based on harmony search algorithm and relevance vector machine, Journal of Computer Applications 36(1)(2016) 199-202.

[5] J. Yu, Z. Lin, B. Xie, Network Security Situation Awareness Prediction Method Based on Grey Correlation Model, Research and exploration in laboratory 38(2)(2019) 31-35.

[6] K. Li, Y. Cao, Forecasting network security threat situation based on ARIMA model, Application Research of Computers 29(8)(2012) 3042-3045.

[7] M.-Y. You, J. Ling, Y.-J. Hao, Prediction Method for Network Security Situation Based on Elman Neural Network, Computer Science 39(6)(2012) 61-63.

[8] F.-W. Li, B. Zheng, J. Zhu, H.-B. Zhang, A method of network security situation prediction based on AC-RBF neural network, Journal of Chongqing University of Posts Telecommunications 26(5)(2014) 576-580.

[9] W.F. Gao, L.L. Huang, S.Y. Liu, C. Dai, Artificial Bee Colony Algorithm Based on Information Learning, IEEE Transactions on Cybernetics 45(12)(2017) 2827-2839.

[10]D. Karaboga, B. Basturk, On the performance of artificial bee colony (ABC) algorithm, Applied Soft Computing 8(1)(2008) 687-697.

[11] Z. Fan, Z. Tan, C. Tan, X. Li, An FCM-based Hierarchical Method for Evaluating Network Security Situation, Journal of Computers 29(3)(2019) 56-72.

[12] H. Feng, The Research on Computer Network Security Defense System Based on GABP Neural Network Algorithm, Journal of Computers 30(3)(2019) 314-319.

[13] X. Chen, Q. Zheng, X. Guan, C. Lin, Quantitative Hierarchical Threat Evaluation Model for Network Security, Journal of Software 17(4)(2006) 885-897.

[14] W.H. Chen, S.H. Hsu, H.P. Shen, Application of SVM and ANN for intrusion detection, Computers Operations Research 32(10)(2005) 2617-2634.

[15] M. Pal, G.M. Foody, Feature Selection for Classification of Hyperspectral Data by SVM, IEEE Transactions on Geoscience Remote Sensing 48(5)(2010) 2297-2307.

[16] F. Friedrichs, C. Igel, Evolutionary tuning of multiple SVM parameters, Neurocomputing 64(1)(2005) 107-117.

[17] L.-F. Huang, C.-L. Ye, C. Feng, H.-B. Li, Y.-M. Zhang, A Detection System of Android Malware Based on SVM Algorithm, Journal of Computers 30(4)(2019) 151-158.

[18] C.-Y. Tsai, W.-H. Chung, Multiclass Object Classification Using Covariance Descriptors with Kernel SVM, Journal of Computers 29(5)(2019) 244-249.

[19] F.-M. Yin, H.-H. Xu, H.-H. Gao, M.-J. Bian, Research on Weibo Public Opinion Prediction Using Improved Genetic Algorithm Based BP Neural Networks, Journal of Computers 30(3)(2019) 82-101.