Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Design of Cybersecurity Threat Prediction Model Based on Support Vector Machines and Data Mining Algorithms

Guo Li[1,✉], Minghua Wang[2]

[1] *College of Intelligent Manufacturing and electrical engineering, Nanyang Normal University, Nanyang, Henan, 473000, China*

[2] *Shandong Gete Aviation Technology Co., Ltd, Jinan, Shandong, 250000, China*

ABSTRACT

In order to strengthen the construction of network security defense system and effectively respond to new types of threat attacks appearing in the network environment, this paper constructs a network security threat prediction model using data mining algorithms. The network security threat posture needs to be assessed before the security threat prediction. Accordingly, this paper assesses the four security threat postures of services, vulnerabilities, weaknesses, and hosts on the basis of the quantitative assessment method of hierarchical security threat posture. After that, a network security threat prediction model is constructed based on the support vector mechanism, and a genetic algorithm is used to optimize the parameters of the model. The three evaluation index values of MAE, RMSE and MAPE for the GA-SVM-based cybersecurity posture prediction method proposed in this paper are 0.0106, 0.0133 and 0.0222, respectively, which are better than those of the ABC-SVM-based and PSO-SVM-based prediction methods. It indicates that the method in this paper has smaller error and higher accuracy in cyber security posture prediction. This shows that the method in this paper usually achieves better accuracy in cyber security threat posture prediction.

Keywords: Data Mining, Support Vector Machine, Genetic Algorithm, Cyber Security Threat Prediction

1. Introduction

While the rapid development of the Internet has brought convenience to life, work and study, cyber attacks also exist. These cyber-attacks target the vulnerabilities existing in the system and software, submit some processed special data to realize the exploitation of the vulnerabilities, thus threatening individual or collective assets. There are four main forms of cyber-attacks: viruses and malware, fake websites and phishing, data leakage, and distributed denial-of-service attackers [1-4]. Whereas, cyber security has several types of threats such as property damage, personal privacy leakage, counterfeit fraud, information destruction and loss [5-8]. For this reason, the prediction of network security

✉ Corresponding author.
E-mail address: airforce1980@126.com (G. Li).
Received 20 March 2024; Revised 05 June 2024; Accepted 20 November 2024; Published Online 15 April 2025.
DOI: 10.61091/jcmcc127a-233

threats can discover potential network threats in time, grasp the development trend of threatening events, predict possible future threats, and provide a strong basis for network security decision-making.

Network security threat prediction refers to the analysis of past network attacks, through the analysis of the network environment, the attack mode and the behavior of the attacker, etc., to predict the possible future network attacks [9-11]. Network security threat prediction can provide reference in preventing and solving network security problems, and effectively improve the level of network security. There are 3 main aspects of network security threat prediction. First, data analysis. The past cyber-attack events are collected and data analysis is performed to find the attacker's patterns and behavioral patterns, and by learning the attacker's patterns, the possible future cyber-attack events can be better prevented [12-13]. The second is event prediction. It is based on machine learning, support vector machine, data mining and other technologies to predict possible future cyberattacks, predictive auditing can not only effectively improve the efficiency of responding to cybersecurity events, but also provide accurate guidelines and methods for cybersecurity defense [14-17]. Third, risk modeling. Establish a set of perfect network risk model, the network environment, risk assessment, attack patterns and other aspects of modeling, in order to better improve the accuracy of network security prediction [18-19]. Network security threat prediction is not a one-time work, but a cyclical and progressive analysis and response process. Only through long-term data collection, analysis and processing can we better improve the level of cybersecurity and reduce cybersecurity risks.

In this context, security researchers use techniques such as support vector machines and data mining algorithms to analyze and predict complex network situations [20-21]. Through cybersecurity threat prediction, we can better understand the evolution of cyber threat patterns and provide help and guidance for cybersecurity defense and attack practices.

The study of predicting cybersecurity threats requires first analyzing information related to cybersecurity such as services, vulnerabilities, weaknesses, hosts, etc., and reasonably calculating the value of cybersecurity threat posture. On this basis, a support vector machine-based prediction model is constructed to predict the change trend of the resulting sequence. At the same time, the genetic algorithm is used to optimize the parameters of the support vector machine to accelerate the speed of network security threat posture prediction. Finally, simulation experiments are conducted on the prediction model to compare the single model and other combined prediction models, and the feasibility and accuracy of the GA-SVM-based prediction model are proved according to the results of evaluation indexes.

2. Network security threat posture assessment

The assessment and quantitative calculation of network security threat posture is the premise and foundation of security threat prediction. In this chapter, on the basis of the hierarchical security threat posture quantitative assessment [22] method, the definition of the four security threat postures and the quantitative calculation formula are given by comprehensively considering the information in the host related to the alarm mechanism of the intrusion detection system.

2.1. Service Threat Landscape

Definition 1: Service threat posture (Rs) is the extent to which the security policy of a service of a certain level of importance on a host is violated after a certain number and intensity of attacks.

The normal access to the service varies under different time periods, and given a time analysis window Δt, the threat posture of service S_j at moment t is Rs_j^t, which is computed as:

$$Rs_j^t = N_{s_j}^t \alpha_{s_{ji}=1}^{\text{mum}} (\text{count}_{ji}^t \times 10^{b_i - \text{signs}(i)}) \tag{1}$$

Where: $N_{s_j}^t$ is the normal number of visits to service S_j at the moment of t. α_{s_j} is the weight of service S_j. num is the total number of different attacks. $count_{ji}^{'}$ is the number of attacks i on the service S_j from t to $t+\Delta t$ time. b_i is the severity of the attack i. The schematic function $signs(i)$ is defined as:

$$signs(i)=\begin{cases}1 & i \text{ indicates invalid attack}\\ 0 & \text{other}\end{cases} \tag{2}$$

Definition 2: An attack event A is said to be an invalid attack if the attack event A relies on a specific vulnerability T and the vulnerability T is not in the system S being attacked.

The discriminant of an invalid attack is:

$$A_{\mathrm{DIP}}\in K\wedge A_{DP}\in P\wedge T\notin S\Rightarrow A \tag{3}$$

Where: K is the set of legitimate addresses of the system. A_{pwp} is the destination port of the attack. A_{DP} is the source port of the attack. P is the set of open ports of the system. S is the set of vulnerabilities that the system already has.

If there are m services in host k and the set number of time analysis windows is h, its service threat posture Rs_k is:

$$Rs_k=\sum_{t=1}^{h}\sum_{j=1}^{m}\sum_{i=1}^{m}R_{s_j}^{'} \tag{4}$$

The greater the value of Rs_k, the higher the threat level of the host.

2.2. Vulnerability Threat Landscape

Definition 3: Vulnerability Threat Posture (Rl) refers to the extent to which a vulnerability of a certain level of importance on a host has its security policy violated after a certain number and intensity of attacks.

Consider the threat of a vulnerability to a host in terms of confidentiality, integrity, and availability. Denoting the vulnerability threat posture of host k by Rl_k, its quantitative computational formula is:

$$Rl_k=\begin{bmatrix}C^k & I^k & A^k\end{bmatrix}\begin{bmatrix}C_1 & C_2 & \cdots & C_n\\ I_1 & I_2 & \cdots & I_n\\ A_1 & A_2 & \cdots & A_n\end{bmatrix}\begin{bmatrix}e_1\\ e_2\\ \vdots\\ e_n\end{bmatrix} \tag{5}$$

Where: n is the number of vulnerabilities. $C_i,I_i,A_i(i=1,2,\cdots,n)$ is the weight of vulnerability i in terms of confidentiality, integrity and availability respectively. C^k,I^k,A^k is the weight of the host k in terms of confidentiality, integrity, and availability. $e_i(i=1,2,\cdots,n)$ is the complexity of attacking the vulnerability. The higher the value of Rl_k, the higher the level of threat to the host.

2.3. Vulnerability threat posture

Definition 4: Vulnerability Threat Posture (Rr) refers to the extent to which a vulnerability with a certain degree of importance exists in the host software or system configuration that has had its security policy violated after a certain number and intensity of attacks.

If there are m vulnerabilities on host k and the weight of vulnerability r is β_r, the number of times vulnerability k has been attacked by attack i and the severity of attack i on vulnerability k are $Rfep_i^{'}$ and $\chi_i^{'}$, respectively, and assuming that the total number of different classes of attacks is u, the vulnerability threat posture Rr_k of host k is:

$$Rr_k = \sum_{r=1}^{m} \beta_r \sum_{i=1}^{n} (Rfreq_i^r \times 10^{\chi_i^r - signr(i)}) \tag{6}$$

For an attacker, the purpose of attacking by utilizing weaknesses in the system is to gain more privileges, and from the property of weakness attacks, weaknesses are only meaningful if they give more privileges to the attacker. Therefore, the severity of a weakness attack that has not gained privileges is downgraded by means of a schematic function when calculating its threat posture value. The definitions for determining whether attack i is a weakness attack and schematic function $signr(i)$ are respectively:

$$(P(gst) \wedge F(gst, i, r) \rightarrow P(\neg gst)) \Rightarrow i \text{ Get more permissions} \tag{7}$$

$$signr(i) = \begin{cases} 1 & i \text{ did not get more permissions} \\ 0 & \text{other} \end{cases} \tag{8}$$

Where: $P(gst)$ represents the set of permissions initially set by the client gst by the administrator. $F(gst, i, r)$ is the client gst to the weakness r using i type of attack. The higher the value of Rr_k, the higher the threat level to which the host is exposed.

2.4. Host threat posture

Definition 5: The host threat posture Rh is the extent to which the security policy of a host has been violated due to multiple services, vulnerabilities, and weaknesses of varying degrees of importance on the host after a certain number and intensity of attacks.

For host k, its host threat posture is calculated as:

$$Rh_k = Rs_k + Rr_k + Rl_k \tag{9}$$

If ρ_i is the host weight determined by the host's environment, asset value, etc. for a network segment with $total(total > 0)$ hosts, the threat posture R is:

$$R = \sum_{k=1}^{total} \rho_k Rh_k \tag{10}$$

2.5. Application of Data Mining in Cybersecurity Threat Posture Assessment

In the process of conducting cybersecurity threat posture assessment, the simplified processing of large-scale cybersecurity event database can be effectively accomplished through data mining technology. Data is the basis of research, and the selection of data types plays an important role in the final research results. Data acquisition includes three parts, which are device log acquisition, sensor acquisition and data preprocessing in order. Device logs are mainly the collection of log data originating from different security devices generated by the network. Sensors provide more complete data and improve the reliability of the analyzed results. Preprocessing mainly preprocesses the collected data and removes some invalid data to get more accurate analysis results.

3. Cyber security threat prediction model based on support vector machine

3.1. Support vector machine algorithm

Support vector machine (SVM)-based prediction models outperform general prediction algorithms, especially for highly complex nonlinear problems [23].

A cybersecurity threat posture training sample is provided: $\{(x_1, y_1), \cdots, (x_n, y_n)\}$, where x_i and y_i denote the cybersecurity threat posture input vectors and output values, and n denotes the number of training samples. Support vector prediction is performed by a nonlinear mapping φ

function that maps nonlinear data x_i to a high-dimensional feature space H and performs linear prediction in that high-dimensional feature space, i.e:

$$f(x)=\omega^r\varphi(x)+b \tag{11}$$

where ω denotes the weight vector of the hyperplane of the support vector machine and b is the bias.

Thus, support vector prediction is to solve an optimization problem as follows, i.e:

$$\min_{\omega,b,\xi_i,\xi_i^*}=\frac{1}{2}\omega^T\omega+c\sum_{i=1}^{n}(\xi_i+\xi_i^*) \tag{12}$$

where the constraints are:

$$\begin{cases} y_i-w\cdot x_i-b\le\varepsilon+\xi_i \\ w\cdot x_i+b-y_i\le\varepsilon+\xi_i^* \quad i=1,\cdots,n \\ \xi_i\ge 0,\xi_i^*\ge 0 \end{cases} \tag{13}$$

where c denotes the penalty parameter, ξ_i,ξ_i^* denotes the slack variable, ε the insensitive loss function, and ε is defined as follows:

$$L_x(f(x)_i,y_i)=\begin{cases} |f(x_i)-y_i|-\varepsilon & ,|f(x_i)-y_i|\ge\varepsilon \\ 0 & ,otherwise \end{cases} \tag{14}$$

Passing into the introduction of Lagrange multipliers transforms the nonlinear prediction problem into the following:

$$\begin{aligned} L=&\frac{1}{2}\omega^T\omega \\ &+c\sum_{i=1}^{k}(\xi_i+\xi_i^*) \\ &-\sum_{i=1}^{k}a_i(\xi_i+\varepsilon-y_i+f(x_i)) \\ &-\sum_{i=1}^{k}a_i^*(\xi_i^*+\varepsilon-y_i+f(x_i)) \\ &-\sum_{i=1}^{k}(\eta_i\xi_i+\eta_i^*\xi_i^*) \end{aligned} \tag{15}$$

where α_i,α_i^* denotes the Lagrange multiplier.

According to the KKT condition, the support vector machine prediction problem can be solved by solving the dyadic problem of Eq. (12), i.e:

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)k(x,x_i)+b \tag{16}$$

The constraints are:

$$\begin{cases} \partial_b L=0\Rightarrow\sum_{i=1}^{i}(a_i-a_i^*)=0 \\ 0\le\alpha_i,\alpha_i^*<c,i=1,2,\cdots,l \\ \partial_\omega L=0\Rightarrow\omega=\sum_{i=1}^{i}(a_i-a_i^*)\phi(x_i) \end{cases} \tag{17}$$

where $k(x,x_i)$ denotes the support vector machine kernel function, which describes the inner product of the high-dimensional feature space.

Since the Gaussian kernel function is better than other kernel functions, this paper chooses the Gaussian kernel function as the support vector machine kernel function, and the Gaussian kernel function is defined as follows:

$$k(x,x_i)=\exp\left(\frac{\|x-x_i\|^2}{\sigma^2}\right) \tag{18}$$

Then, the final expression of the support vector machine prediction model is:

$$f(x)=\sum_{i=1}^{n}(\alpha_i-\alpha_i^*)\exp\left(\frac{\|x-x_i\|^2}{\sigma^2}\right)+b \tag{19}$$

where σ denotes the Gaussian kernel function width.

3.2. Support vector machine prediction model parameter optimization

When the kernel function of the support vector machine is a Gaussian kernel function, the parameters to be optimized for the prediction model are: ε, c and σ. The parameters of the traditional support vector machine are generally determined by empirical method, exhaustive method and network search method. The parameters selected by the empirical determination method are generally not optimal, and the prediction accuracy of the model is low, while the exhaustive method and the network search method are quite time-consuming, and it is difficult to find the optimal parameters, so in order to improve the prediction accuracy of the cybersecurity threat posture, the first step is to solve the optimization problem of the parameters of the support vector machine.

Genetic algorithm is a kind of heuristic algorithm that simulates the mechanism of natural selection and population evolution in the biological world, which has the ability of global and parallel search, and the speed of finding the optimal is quite fast [24], so this paper adopts genetic algorithm to optimize the parameters ε, c and σ of the support vector machine.

3.3. Cybersecurity Threat Prediction Process with Support Vector Machines

1) The collection of cybersecurity threat posture data and the processing of the number anomalies and undesirable data in the collection.

2) Pre-processing of cyber security threat posture data. As the network security threat posture is affected by a variety of factors, the data sometimes differ greatly from each other, and the support vector machine prediction model is most sensitive to between (0, 1), so it is necessary to normalize the original data of the network security threat posture to the range of (0, 1). The specific processing is as follows:

$$x_i'=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}} \tag{20}$$

Where x_i is the original value of network security threat posture, x_i' is the normalized value of network security threat posture, $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values of network security threat posture, respectively.

3) one-dimensional network security threat posture data by determining the embedding and time delay into multi-dimensional network security threat posture data, this paper sets the data network security threat posture time delay for 1, embedding dimension by 2, 3, ... etc. gradually try to put together, and finally determine the number of embedded dimensions of m. Set one-dimensional network security threat posture data for $\{x_1,x_2,\cdots,x_n\}$, then converted into multi-dimensional network security threat posture data is shown in Table 1.

Table 1. Multidimensional network security threat situation data

Sample input	Expected output
$x_1, x_2, \cdots, x_{m-1}$	x_m
$x_2, x_3, \cdots, x_m$	x_{m+1}
$x_3, x_4, \cdots, x_{m+1}$	x_{m+2}
...	...
$x_{n-m}, x_{n-m+1}, \cdots, x_{n-1}$	x_n

4) Divide the cybersecurity threat posture data into two parts: the training set and the test set, input and output data of the training set to the support vector machine, use genetic algorithm to optimize the parameters of the support vector machine, and use the optimal parameters to establish the optimal prediction model of the cybersecurity threat posture.

5) The optimal network security threat posture prediction model is used to predict the test set, and the prediction result is converted into the actual network security threat posture value by the inverse normalization formula of equation (20), and the current network security threat state is judged according to the network security threat posture value:

$$x_i = x_i' \times (x_{\max} - x_{\min}) - x_{\min} \tag{21}$$

3.4. Network Security Threat Prediction Model Implementation

In this paper, the specific steps to optimize the parameters of the hybrid kernel function of the support vector machine constructed above using the genetic (GA) algorithm are as follows:

1) Firstly, the mapping relationship between the phenotype and genotype is established, when the individual characteristics are complex, a large amount of coding is required to accurately describe them, so in order to improve the efficiency and accuracy of the operation, the potential solution of the problem is encoded by utilizing floating-point coding. In the genetic algorithm, each decision vector X is represented by a chromosome V :

$$X = (x_1, ..., x_n)^T \Rightarrow V = [v_1, v_2, ...v_m]^T \tag{22}$$

Where X is the phenotype, V is the genotype, a v_i represents a gene, the length of the chromosome m depends on the coding method of the chromosome, this paper adopts the floating-point coding method, x_i using the floating-point number within the $[U_{\min}^i, U_{\max}^i]$ to represent the genes v_i that make up the chromosome, and the value of the m is equal to the number of decision variables n.

2) Randomly generate a set of feasible solutions, i.e., the first generation of individuals, as the initial population, denoted as $P(0)$, set the evolutionary generation counter $t = 0$, and set the maximum evolutionary generation $T = 100$.

3) Select the fitness function. In this paper, the mean absolute percentage error (MAPE) is chosen as the fitness function, and this can measure the prediction results:

$$M(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i} \tag{23}$$

Where y_i is the true value and $\hat{y}_i$ is the predicted value. Set the individual fitness is $F(y)$, because the fitness function to meet the non-negative, single-valued, continuous, maximization requirements, and $M(y)$ is a minimization problem, so the above $M(y)$ to be processed as the fitness function, that is:

$$F(y)=\frac{1}{1+M(y)} \tag{24}$$

4) Selection strategy. Also known as the selection operator, it is the process of randomly selecting some individuals from the parent generation to survive and the rest to be eliminated according to a selected selection strategy, in accordance with Darwin's theory of evolution, with reference to the fitness function.

In this paper, we use tournament selection method as a selection strategy, repeating to select the best individual into the offspring population based on fitness, this method is generalized to maximization and minimization problems, and does not need to convert its fitness value when dealing with minimization problems, which is more convenient and easy to operate than roulette selection method.

5) Crossover. Crossover is to randomly exchange certain genes of two individuals in the population through the crossover rate, so that they can produce a new combination of genes to achieve the purpose of combining good genomes together, which optimizes the combination of certain genes of the surviving parent individuals, and exchanges certain genes of the corresponding positions of the two parent individuals with each other to produce new individuals.

The crossover operators applicable to binary coding are: single-point crossover, two-point crossover (multipoint crossover), and uniform crossover, and the crossover operators applicable to floating-point or real-number coding are: discrete recombination, intermediate recombination as well as linear recombination, simulated binary crossover, and hybrid crossover.

In this paper, the floating point encoding is chosen, so the analog binary crossover is used as the crossover operator in this paper, and the crossover rate is set to 0-6.

The analog binary single point crossover method:

$$\begin{aligned}\tilde{x}_{1j}(t)&=0.5\times[(1+\gamma_j)x_{1j}(t)+(1-\gamma_j)x_{2j}(t)]\\ \tilde{x}_{2j}(t)&=0.5\times[(1-\gamma_j)x_{1j}(t)+(1+\gamma_j)x_{2j}(t)]\end{aligned} \tag{25}$$

$$\text{Among }\gamma_j=\begin{cases}(2u_j)^{\frac{1}{\eta+1}}, & u_j\le 0\\ [\dfrac{1}{2(1-u_j)}]^{\frac{1}{\eta+1}}, & other\end{cases} \tag{26}$$

$u_j\in U(0,1),\eta>0$ is the index of the distribution, and it is noted in the Agrawal paper that $n=1$ is recommended, so in this paper we use $n=1$, which converts the above equation:

$$\gamma_j=\begin{cases}(2u_j)^{\frac{1}{2}}, & u_j\le 0\\ [\dfrac{1}{2(1-u_j)}]^{\frac{1}{2}}, & other\end{cases} \tag{27}$$

6) Mutation. In this paper, the boundary variation is chosen as the variation datum because when the optimal point is located at or close to the boundary of the feasible solution, by randomly selecting one of the two corresponding boundary gene values on the genome to replace the original gene value, it makes this problem can be solved optimally.

Finally, the radius of the radial basis of the optimal parametrization kernel function δ, the order of the polynomial kernel function d, and the linear combination coefficients γ are determined to obtain the optimized support vector machine model, which is used to predict the cybersecurity threat posture and obtain the cybersecurity threat state.

4. Simulation results and data analysis

4.1. Outcome evaluation indicators

In order to better compare cybersecurity threat posture prediction methods horizontally, three types of metrics commonly used in regression prediction problems were chosen for evaluation, namely, mean absolute error, root mean square error, and mean absolute percentage error. These three types of metrics can facilitate the evaluation of the accuracy of situational prediction.

The mean absolute error (MAE) evaluates the absolute error between the test value and the actual value in the experimental dataset, and the smaller the value indicates that the mean absolute error of the model is smaller and the prediction accuracy is higher, with the following formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} | y_i^a - y_i^f | \tag{28}$$

In equation (28), y_i^a is the target actual value and y_i^f is the target predicted value.

The Root Mean Square Error (RMSE) evaluates the deviation between the test value and the actual value of the experimental dataset and is the square root of the mean square error (MSE). Its formula is as follows:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i^a - y_i^f)^2} \tag{29}$$

RMSE is extremely sensitive to anomalous data, and the larger its value, the greater the deviation and the lower the prediction accuracy.

Mean Absolute Percentage Error (MAPE) is the result of the percentage of Mean Absolute Error (MAE), and the value of MAPE represents the average percentage deviation of the predicted value from the actual value, with the following formula:

$$MAPE = (\frac{1}{n}\sum_{i=1}^{n} \frac{| y_i^a - y_i^f |}{y_i^a}) \times 100\% \tag{30}$$

4.2. Selection of cybersecurity threat posture data

The security test data of a company from March 1-April 29 and May 1-June 29, 2023 are selected and sampled four times per day, the security test data are taken as a batch of two months, 240 posture values are obtained from each batch after calculation, and experiments are carried out on both batches of data respectively with the same process, through MATLAB 7.5.

4.3. Forecasting the Cybersecurity Threat Landscape

The cybersecurity threat posture prediction model given in the paper is tested for generalizability and validity. Using the two sets of reconstructed data obtained above, the unimproved GA-SVM model and the improved model are used to make predictions respectively, and then the prediction results obtained from the two models are compared. The specific operation method is as follows:

1) The first 200 points of the two sets of reconstructed data are first used as training samples for the training of the two methods and the construction of the model. The last 40 points of the two sets of data are used as test samples for comparing the prediction results of the two models with the actual values.

2) The training samples of the two sets of reconstructed data are input into the SVM for learning respectively, and the three parameters of the SVM are optimized with the improved particle swarm optimization algorithm to obtain the first batch of data at $\sigma = 5$, $\varepsilon = 0.001$, and $c = 93$, and the second batch of data at $\sigma = 5$, $\varepsilon = 0.001$, and $c = 75.98$.

3) Using the three parameters obtained from the two data sets then input the two reconstructed data sets into the support vector machine for learning and training respectively to obtain the prediction results of the new model.
4) Comparing the original unprocessed values, the prediction results obtained by the method given in the paper and the prediction results obtained by the unimproved GA-SVM, the comparison results of the threat posture values are shown in Fig. 1. The prediction results obtained by the improved GA-SVM are basically consistent with the trend of the original values, and the average error is 0.3635.

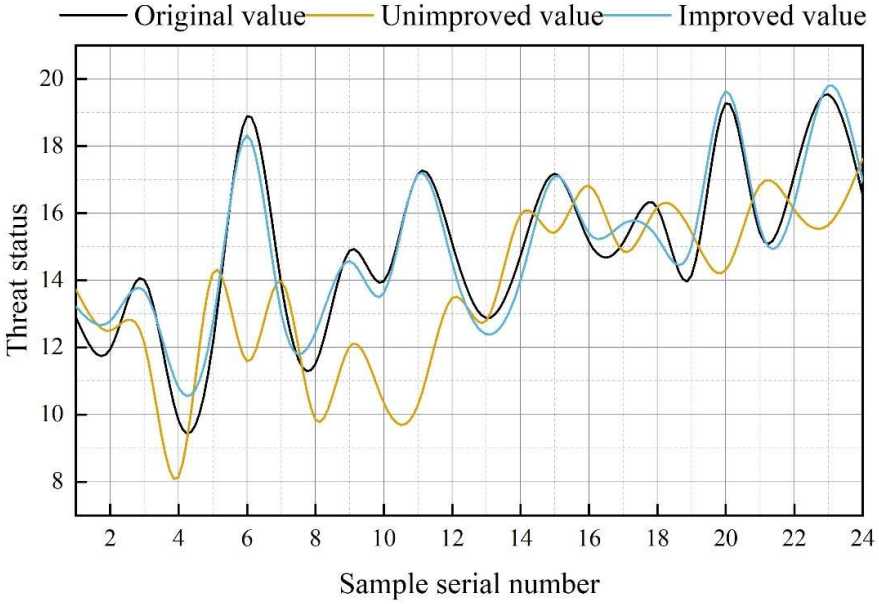

Fig. 1. The comparison of the threat situation value

The comparison of the errors of the two methods compared to the original values is shown in Table 2. The error between the threat posture values predicted based on GA-SVM and the original values is less than 1 on different samples, while the error value of SVM prediction is maximum 7.298, which is a large difference from the original values. It shows that the future cybersecurity threat state is predicted with high accuracy and low error using the cybersecurity threat posture prediction method given in the paper.

Table 2. The error comparison of the two methods

Sample number	Unimproved value	Improved value	Sample number	Unimproved value	Improved value
1	0.815	0.308	13	0.066	0.481
2	0.558	0.842	14	1.21	0.731
3	1.884	0.328	15	0.053	0.074
4	1.683	0.986	16	0.149	0.097
5	2.082	0.6	17	0.165	0.064
6	7.298	0.58	18	0.097	0.092
7	0.102	0.865	19	0.055	0.069
8	1.665	0.938	20	0.041	0.096
9	2.849	0.31	21	0.178	0.054
10	3.634	0.346	22	0.084	0.028
11	6.868	0.017	23	0.184	0.164
12	1.625	0.585	24	0.018	0.069

4.4. Comparison of prediction results with other combined models

Further comparing this paper's method with the PSO-SVM-based posture prediction model and the ABC-SVM-based posture prediction model, the experimental prediction results are shown in Fig. 2. Although the ABC-SVM-based prediction method is basically fitted to the actual posture values on the whole, the latter is better fitted to the actual posture values compared with the GA-SVM-based method proposed in this paper.

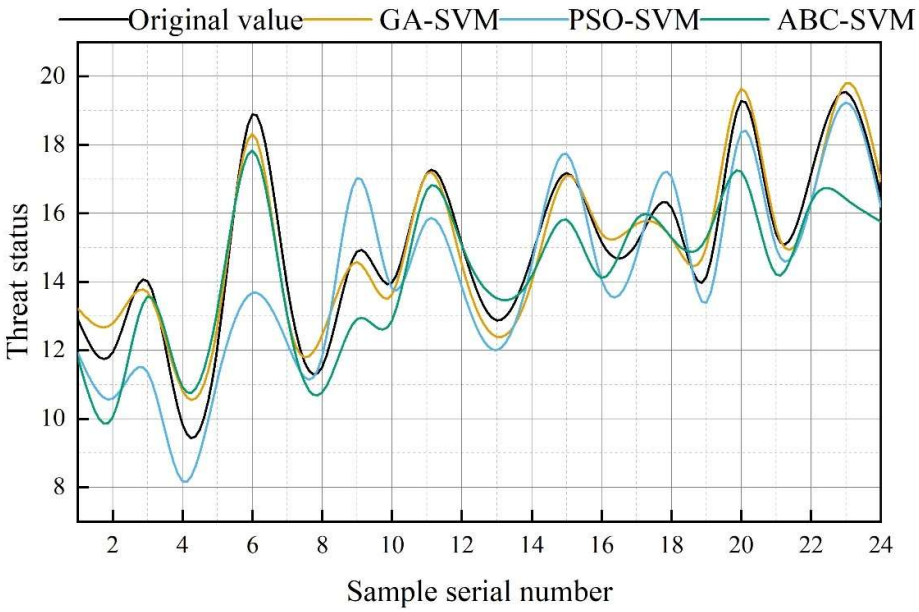

Fig. 2. Experimental prediction

In order to compare the posture prediction accuracy of the three combined models more clearly, the absolute value of the error with the original value is analyzed, and the comparison results are shown in Table 3. The average error values based on PSO-SVM and ABC-SVM are higher than GA-SVM by 1.283 and 0.727, respectively.

Table 3. Prediction accuracy of the three composite models

Sample number	GA-SVM	PSO-SVM	ABC-SVM	Sample number	GA-SVM	PSO-SVM	ABC-SVM
1	0.308	0.983	1.119	13	0.481	0.869	0.647
2	0.842	1.351	1.891	14	0.731	0.201	0.577
3	0.328	2.657	0.448	15	0.074	0.561	1.365
4	0.986	1.656	1.099	16	0.097	1.1	1.033
5	0.6	0.985	1.086	17	0.064	0.348	0.706
6	0.58	5.212	1.065	18	0.092	0.897	0.872
7	0.865	1.66	0.865	19	0.069	0.76	1.126
8	0.938	0.32	0.741	20	0.096	0.906	2.057
9	0.31	2.149	1.972	21	0.054	0.416	1.179
10	0.346	0.162	1.103	22	0.028	0.831	0.829
11	0.017	1.378	0.49	23	0.164	0.297	3.116
12	0.585	1.246	0.028	24	0.069	0.308	0.747

In order to more conveniently analyze the difference in accuracy between them and the method of this paper, the values of each indicator were calculated and retained four decimal places in combination with the calculation method of evaluation indicators given above, and the comparison of

the values of each indicator of the prediction results is shown in Figure 3. The values of MAE (0.0106), RMSE (0.0133) and MAPE (0.0222) of the GA-SVM-based method proposed in this paper are smaller than those of PSO-SVM and ABC-SVM, which indicates that the average absolute error and deviation of this paper's method are small, and the degree of deviation of the predicted value from the actual value is low, so the prediction accuracy is good.

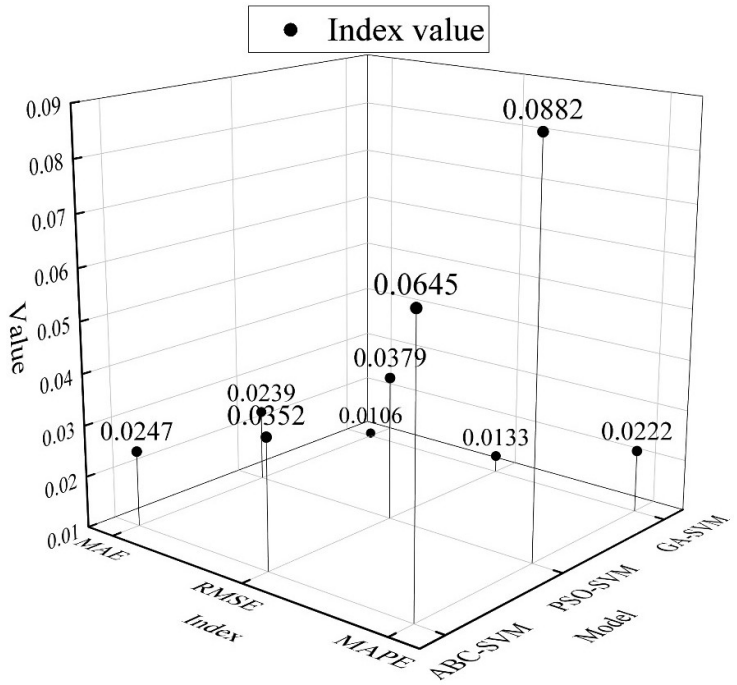

Fig. 3. Evaluate the index value contrast

5. Conclusion

The study constructed a cyber security threat prediction model based on support vector machine and improved the model using genetic algorithm. The predictions of the support vector machine based cyber security threat prediction model and GA-SVM model are compared. The average error between the threat posture values predicted by the optimized GA-SVM and the actual values is 0.3635, which is basically consistent with the trend of the original values. It shows that the optimization method designed in this paper can improve the prediction accuracy. The GA-SVM model is compared with ABC-SVM and PSO-SVM models respectively. The MAE, RMSE and MAPE values of this paper's method are 0.0106, 0.0133 and 0.0222, respectively, which are better than the prediction results of the other two combination models. It shows that the method of this paper has better prediction effect compared with other combination models.

References

[1] Park, J. H., Singh, S. K., Salim, M. M., Azzaoui, A. E., & Park, J. H. (2022). Ransomware-based cyber attacks: A comprehensive survey. Journal of Internet Technology, 23(7), 1557-1564.

[2] Sicari, S., Rizzardi, A., Miorandi, D., & Coen-Porisini, A. (2018). REATO: REActing TO denial of service attacks in the Internet of Things. COMPUTER NETWORKS, 137, 37-48.

[3] Alkhalil, Z., Hewage, C., Nawaf, L., & Khan, I. (2021). Phishing Attacks: A Recent Comprehensive Study and a New Anatomy. Front. Comput. Sci. 3: 563060. doi: 10.3389/fcomp.

[4] Alenezi, M. N., Alabdulrazzaq, H., Alshaher, A. A., & Alkharang, M. M. (2020). Evolution of Malware Threats

and Techniques: A Review. International Journal of Communication Networks and Information Security, 12(3), 326-337.

[5] Kala, E. M. (2023). The impact of cyber security on business: how to protect your business. Open Journal of Safety Science and Technology, 13(2), 51-65.

[6] Prasad, R., Rohokale, V., Prasad, R., & Rohokale, V. (2020). Cyber threats and attack overview. Cyber Security: The Lifeline of Information and Communication Technology, 15-31.

[7] Farrand, B. (2018). Combatting physical threats posed via digital means: the European Commission's developing approach to the sale of counterfeit goods on the Internet. European Politics and Society.

[8] Ali, B., & Awad, A. I. (2018). Cyber and physical security vulnerability assessment for IoT-based smart homes. Sensors (Switzerland), 18(3), 817.

[9] Husák, M., Komárková, J., Bou-Harb, E., & Čeleda, P. (2019). Survey of Attack Projection, Prediction, and Forecasting in Cyber Security. IEEE COMMUNICATIONS SURVEYS AND TUTORIALS, (1).

[10] Duary, S., Choudhury, P., Mishra, S., Sharma, V., Rao, D. D., & Aderemi, A. P. (2024, February). Cybersecurity threats detection in intelligent networks using predictive analytics approaches. In 2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM) (pp. 1-5). IEEE.

[11] Khoukhi, L., Moudoud, H., & Cherkaoui, S. (2021). Prediction and Detection of FDIA and DDoS Attacks in 5G Enabled IoT. IEEE Network, 35(2), 194-201.

[12] Qamar, S., Anwar, Z., Rahman, M. A., Al-Shaer, E., & Chu, B. T. (2017). Data-driven analytics for cyber-threat intelligence and information sharing. Computers & Security, 67(JUN.), 35-58.

[13] Tounsi, W., & Rais, H. (2018). A survey on technical threat intelligence in the age of sophisticated cyber attacks. Computers and Security, 72(C), 212-233.

[14] Al-Rubaie, M., & Chang, J. M. (2019). Privacy-Preserving Machine Learning: Threats and Solutions. IEEE Security & Privacy, 17(02), 49-58.

[15] She, C. (2017). Application-layer ddos detection based on a one-class support vector machine. Available at SSRN 4911677.

[16] Vegesna, V. V., & Adepu, A. (2024). Leveraging Artificial Intelligence for Predictive Cyber Threat Intelligence. International Journal of Creative Research In Computer Technology and Design, 6(6), 1-19.

[17] Sun, N., Zhang, J., Rimba, P., Gao, S., Xiang, Y., & Zhang, L. Y. (2018). Data-driven cybersecurity incident prediction: a survey. IEEE Communications Surveys & Tutorials, 1744-1772.

[18] Yohanandhan, R. V., Elavarasan, R. M., Manoharan, P., & Mihet-Popa, L. (2020). Cyber-physical power system (cpps): a review on modeling, simulation, and analysis with cyber security applications. IEEE Access, 8, 151019-151064.

[19] Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-Driven Cybersecurity: An Overview, Security Intelligence Modeling and Research Directions. SN Computer Science, 2(3).

[20] Myint Oo, M., Kamolphiwong, S., Kamolphiwong, T., & Vasupongayya, S. (2019). Advanced Support Vector Machine-(ASVM-) Based Detection for Distributed Denial of Service (DDoS) Attack on Software Defined Networking (SDN). Journal of Computer Networks & Communications.

[21] Lekha, K. C., & Prakasam, S. (2017, August). Data mining techniques in detecting and predicting cyber crimes in banking sector. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 1639-1643). IEEE.

[22] Mai Rui & Wu Mingzhu. (2020). Research on the Quantitative Assessment and Security Measures of Hierarchical Network Security Threat Situation. IOP Conference Series: Materials Science and

Engineering012171-012171.

[23] Guoying Han,Bin Zhou & Yazi Zhang. (2024). Application of Genetic Algorithm-Grey Wolf Optimization-Support Vector Machine Algorithm in Network Security Services Assessment and Prediction. Journal of Cyber Security and Mobility(5),941-962.

[24] Chen Fengfeng. (2021). Optimal Design of Computer Network Security Performance Based on Genetic Algorithm. Journal of Physics: Conference Series(3).