ATLANTIS PRESS
Advances in Computer Science Research (ACRS), volume 54
International Conference on Computer Networks and Communication Technology (CNCT2016)

Network Situation Awareness Model Prediction Method Based on Genetic Optimization Support Vector Machine

Wei-peng CHEN[1], Zhi-gang AO[2,*], Yi-qiang TU[3], Xing-dang KANG[4] and Zhen-nan ZHAO[5]

[1,2,3,4,5] College of Field Engineering, PLA University of Science and Technology, nanjing 210007, China

*Corresponding author: chenweipeng1987@163.com

Keywords: Support vector machine, Genetic algorithm optimization, Network situation awareness, Model, Prediction

Abstract. The support vector machine model is based on the network security situation has strong randomness, is affected by many factors, and the number of types of network security incidents is uncertain, the reference sample is small, the prediction model of need "intelligent", according to SVM forecast algorithm. In order to select the parameters of SVM, genetic algorithm is introduced into the parameter selection in support vector machine, genetic algorithm optimization based on support vector machine structure (GA - SVM) situation awareness prediction model and method to measure data dimensionality reduction using principal component analysis, through simulation experiments, it is proved that this model is the prediction higher precision than neural network, classification and regression tree and cluster analysis prediction model.

Introduction

Support vector machine (SVM) is a new general learning method based on the finite sample statistical learning theory, Effectively solve the small sample, nonlinear, high dimension and other learning problems.[1]However, as a new kind of learning machine, SVM also has some problems to be improved, and its parameter selection is one of the problems to be improved.[2]Due to the lack of theoretical guidance, the method of selecting parameters by repeated experiments often requires the guidance of human experience, and the need to pay a high price.[3]

This paper presents the application of genetic algorithm (Algorithm Genetic, GA) to support vector machine parameter selection, a network situation awareness model prediction model based on genetic algorithm optimized support vector machine (GA-SVM),and other prediction models are compared and analyzed, the validity and feasibility of the model are verified.

Support Vector Machines

Support vector machine is a machine learning method based on statistical learning theory, and it has become a major achievement in machine learning research.[4]Support vector machine (SVM) prediction algorithm in the prediction accuracy, real-time and so on are more suitable for the characteristics of network security situation prediction: Small samples, nonlinear, high dimension.SVM basically can be said to be the best supervised learning algorithm.[5]Its core idea: For classification problems in n-dimensional Euclidean space Rn (or regression),By looking for a real valued function on Rn g(x),In order to use decision function

f(x)=sgn[g(x)] to infer the input X corresponding to the output value of Y.

The method of determining g(x) is to construct a nonlinear programming problem which is dual to the primal problem and to solve the problem. To solve the nonlinear programming problem when the original Euclidean space variable in Rn X through the transformation Phi mapped into a high dimensional space, such as the type (1), in order to get and solve the linear programming problem.

$$Rn—>Hilbert,\ x—>x=\phi(x) \tag{1}$$

The kernel function of support vector machine in K (x, x') of the role is to achieve the phi transform through inner product transform, i.e.

$$K(x,x')=\phi(x)\cdot\phi(x') \tag{2}$$

The decision function of the original Rn space at this time is turned into formula (3):

$$f(x)=sgn[\omega^{T}\cdot\phi(x)+a] \tag{3}$$

In the formula, ω is the weight and a is the Y threshold value respectively. Therefore, SVM prediction is to solve the following an optimization problem:

$$\min_{\omega,b,\xi_i,\xi_i^*} = \frac{1}{2}\omega^{T}\omega + c\sum_{i=1}^{n}(\xi_i+\xi_i^*) \tag{4}$$

The constraint conditions are:

$$\begin{cases} y_i-\omega\cdot x_i-b\le\varepsilon+\xi_i \\ \omega\cdot x_i+b-y_i\le\varepsilon+\xi_i^* \\ \xi_i\ge 0,\xi_i^*\ge 0 \end{cases} \tag{5}$$

Among them, c indicate the penalty parameter, ξ_i, ξ^*_i indicate slack variables, ε indicate insensitive loss function.

Support vector machine is a kind of nonlinear prediction technology, It is very clear about the advantages of high complexity nonlinear problem. [6]The traditional way to support the quantization parameter is Empirical determination method, Exhaustive method and Network searching method. Empirical determination method is not the best choice of the parameters of the model, the prediction accuracy of the model is low; The enumeration method, network search method is quite time-consuming, it is difficult to find the optimal parameters. Therefore, to improve the prediction accuracy of network security situation, we must first solve the optimization problem of support vector machine parameters.

Genetic Optimization Support Vector Machine Prediction Model

Profit from SVM's excellent nonlinear forecasting ability, it is able to find the hidden rules of network security situation in a large number of data, But the traditional support vector machine generally adopts experience determination method and exhaustive method and network search method to determine the parameters, this is usually not the best, but the efficiency is very low. [7]The parameters mentioned here refer to the parameters of the kernel function of the SVM algorithm and the determination of the penalty coefficient. In this paper, genetic algorithm is used to guide the determination of important parameters of SVM algorithm.

Genetic Algorithm

Genetic algorithm (GA) using the characteristics of biological genetics (Darwin's theory of biological evolution), Combining with the characteristics of natural selection and random information exchange, By using the mechanism of random selection, crossover and mutation, the evolution of population is realized in the process of optimization, is a kind of search technology which can quickly search for the global optimal solution in complex search space. It has been widely used in the field of machine learning and parallel processing. [8]

The basic algorithm of genetic algorithm is as follows:

1) initialization: randomly generated initial population;

2) fitness calculation: for each individual in the population to calculate their fitness;

3) select the operation: the selection operator to the population. The purpose of selection is to direct the optimization of the individual to the next generation or through a pair of cross generation to generate new individuals and then to the next generation. The selection operation is based on the evaluation of the fitness of individuals in the species;

4) crossover operator: crossover operator to population, which is the essential part of genetic algorithm;

5) mutation operation: the mutation operator in the population. That is to change the gene value of the individual in the population. After selection, crossover and mutation operation, the population of the next generation is obtained.

6) the termination condition: if the maximum algebra is reached, the maximum fitness of the evolutionary process is obtained as the optimal solution output.

The specific flow chart is shown in figure 1.

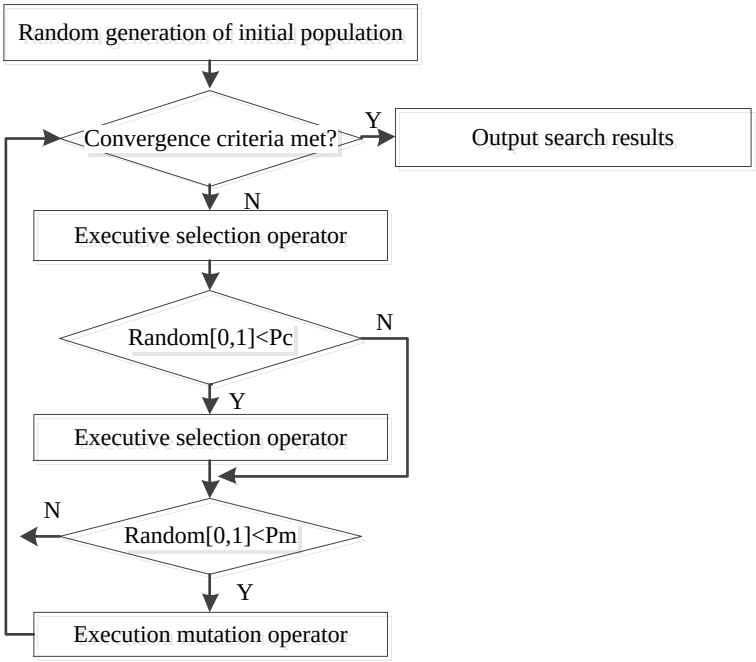

Figure 1 general process of genetic algorithm

Support Vector Machines

The key of the SVM algorithm is the selection of kernel function and the determination of the parameters of kernel function and penalty coefficient.[9]In this paper, we use the radial basis kernel function as the kernel function, such as the

formula (6).

$$k(x_i, x_j) = \exp(-\left\| x_i - x_j \right\|^2 / \sigma^2) \quad (6)$$

The SVM parameter consists of penalty factor c, kernel function parameter σ, insensitive loss function parameter ε. Which c value determines the degree of learning, σ determines the functional analysis ability of the model, ε determine the number of support vectors and the number of the base by complexity, this is also the need to optimize the SVM parameters.

Genetic Optimization Support Vector Machine Prediction Model

In this paper, genetic algorithm is added to the support vector machine parameter optimization process, and build a genetic optimization support vector machine model, the specific flow chart as shown in figure 2:

1) network security situation value collection: the historical data from the situation assessment model;

2) Data normalization processing: The penalty coefficient and kernel function parameters of SVM are binary coded.

3) Fitness calculation method: using the correct rate of classification of SVM samples as the calculation method of individual fitness.

4) Genetic manipulation: select one crossover and one mutation.

5) Termination condition: comparison with maximum genetic algebra.

6) Accuracy test: the best individual as the SVM algorithm parameters to predict the accuracy of the test, if the accuracy is not enough to continue to use the GA parameter optimization.

7) SVM prediction.

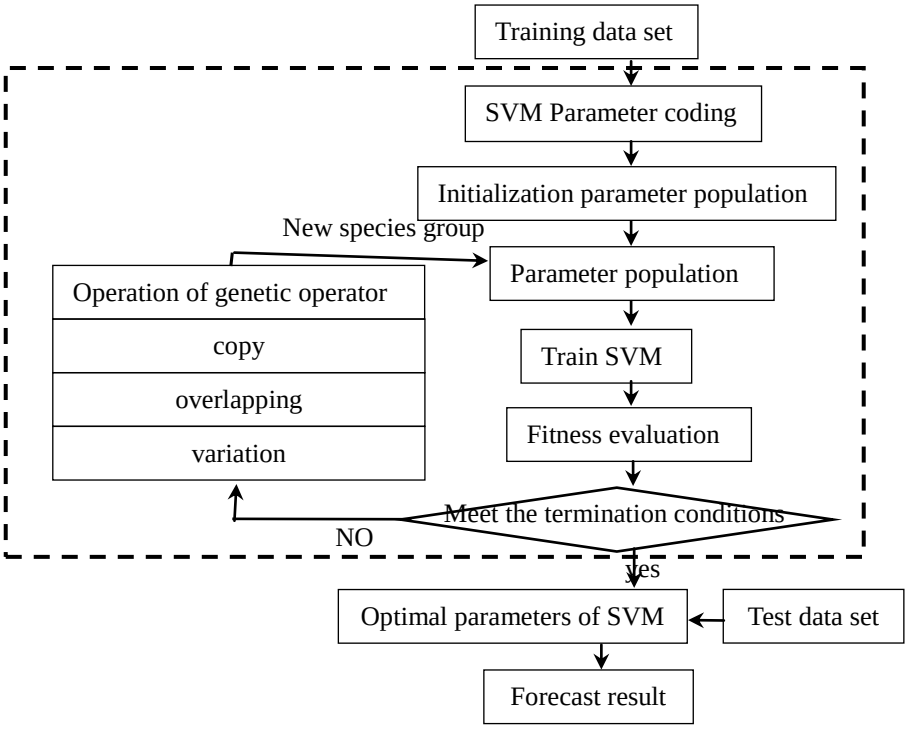

Figure 2 genetic optimization support vector machine algorithm process

Data Simulation

In order to further verify the feasibility and rationality of the network security situational awareness model, the model test and system performance test are carried out.

Model Test

The model is implemented on a PC machine configuration for Redhat Linux9.0/P4 3.0/1024M/160G.The experimental environment is the HUAWEI S5000 multilayer routing switch, IDS and Firewall each one,LAN host a total of six units, the configuration is Redhat Linux9.0/P4 3.0/512M/80G.Using the network security situation prediction model to predict the network security situation,Sample data from the network security situation assessment algorithm in the network of the University of science and technology to run 90 days of situation data(about 2000 data),Forecast the next thirty hours of network security situation value.The data used in the experiment is a random sample of 500 of these data.

In the experiment, the number of data is more than 10, and the number of error is 0~10. For the measurement data of 500 data used in the experiment, error prone data 103, less error prone data for 397.

Data Preprocessing

Principal components analysis method is used to pre process the measurement data of 500 data which are selected well, and the 3 main components are obtained as table 1.

Table 1 3 principal components

principal component	characteristic value	contribution(%)	Cumulative contribution rate(%)	Principal component weight(%)
F1	7.3332	84.72	84.72	87.23
F2	0.8098	9.58	94.30	9.85
F3	0.3135	3.71	98.01	3.97

From table 1, we can see that the cumulative contribution rate of 3 main components is 98.01%, that is, 98.01% of the original data is retained, Significant representation.F1 principal component weight is the biggest, is 84.72%, is the most important influence factor. In the experiment of this paper, the 3 main components of PCA method will be used as the input of SVM - GA, the experimental data can reduce the dimension but retains 98.01% of the original data information, to achieve a good result. Reducing the dimensionality of the experimental data will simplify the SVM structure, speed up the computing speed, and the basic information of the original data is retained, will not result in the distortion of the experimental results.

GA-SVM Prediction Model Construction

SVM classification model using LIBSVM toolbox to build, Program is configured for CPU:P4 2.9GHz, memory: 512MB running on the computer, Training and testing of samples using GA as MALAB and SVM operating platform. Genetic algorithm parameters are set as shown in table 2. Genetic algorithm parameters are set as shown in table 2.

The optimal parameters of SVM obtained by the genetic algorithm as σ=0.8537, c=298.6

Table 2 Table of parameter setting for genetic algorithm

Operating parameters of genetic algorithm	parameter values
Code string length	200
Crossover probability	0.63
Mutation probability	0.02
Operator cross method	roulette wheel selection
Maximal evolutionary algebra	900

Comparison and Analysis of Forecast Results

Method using 10% off cross validation,During the experiment, the data set is randomly divided into 10 subsets, one subset is used as the test set, and the rest is set as the training set, so the experiment is done in a total of 10 times, The performance of the algorithm is evaluated by the average of the 10 experiments.
Using the Accuracy,Recall and F-Measure forecast performance metric to evaluate and compare the prediction models, these metrics are derived from the confusion matrix, as shown in table 3.

Table 3 confusion matrix

actual value	predicted value	
	Not prone to errors	Prone to error
Not prone to errors	TN=Ture	FP=False
	Negative	Positive
Prone to errors	FN=False	FP=Ture
	Negative	Positive

Accuracy: The predicted results and the actual results are in accordance with the ratio of the number of data.

$$\text{Accuracy}=\frac{TP+TN}{TP+FN+FP+TN} \tag{7}$$

Precision: The data in the actual forecast error prone for error prone data proportion.

$$\text{Precision}=\frac{TP}{TP+FP} \tag{8}$$

Recall: The proportion of the error prone correct recognition data.

$$\text{Recall}=\frac{TP}{TP+FN} \tag{9}$$

F-Measure: it is the harmonic mean of Precision and Recall.

$$\text{F-Measure}=\frac{2}{\frac{1}{Precision}+\frac{1}{\text{Recall}}} \tag{10}$$

In this paper, the experimental results of prediction model and support vector machine, neural network, classification and regression tree and cluster analysis model are compared.

Table 4 Comparison of the predicted results of each model

prediction model	Accuracy(%)	Precision(%)	Recall(%)	F-Measure(%)
SVM	75.76	73.9	57.3	62.2
neural network	69.88	70.3	56.8	60.5
Classification and regression tree	76.33	71.3	58.9	63.6
cluster analysis	62.5	67.3	48.6	59.4
GA—SVM	78.24	75.6	59.9	64.1

As can be seen from the table 4,4 performance indexes of the proposed model are presented in this paper are better than the other 4 machine learning methods in the prediction model.

Clustering analysis and neural network prediction effect is poor, because these two methods mainly rely on human experience, poor generalization ability. If the neural network, if the network process to effectively use their own experience and knowledge of knowledge, may be more ideal network structure. Therefore, the advantages and disadvantages of the neural network system is different from each individual.

Classification and regression tree has better prediction accuracy, because the support vector machine is based on statistical learning theory, has strict theory and mathematical foundation. It is based on structural risk minimization principle, ensure the machine learning has good generalization ability;It is not like the structure of neural network design need to rely on human experience and knowledge, and it needs to set the parameters is relatively small;Because the traditional manual selection parameter method is time-consuming and laborious, the SVM parameters optimization of GA is introduced, which greatly improves the efficiency of parameter selection.

Conclusions

In view of the current network security situation prediction model of single prediction inaccuracy, innovative forecasting method is proposed, SVM has better generalization ability in small sample under the premise and can obtain the global optimal solution. Taking the support vector machine as a mathematical tool, the parameter selection problem is hard to be selected, and the software reliability prediction model is established by using the genetic algorithm to optimize the support vector machine. The experimental results show that the proposed prediction model has a good effect on the overall performance.

Reference

[1] Wang Jindong, Shen Liu Qing, Wang Kun. Prediction of network security situation and its application in intelligent protection [J]. computer application, 2010,30 (6): 1480-1488.

[2] Zhu Lina, Zhang Zuochang, Feng. A hierarchical network security threat situation assessment technology research [J]. computer application research, 2011,28 (11): 4303-4306.

[3] JasonShifflet.A technique independent fusion model for Network.Intrusion Detection.Proceedings of the Midstates Conference on this research in computer science and mathematics, 2005,3 (1): 13-19.

[4] Chen Xiuzhen, Zheng Qinghua, Guan Xiaohong. Quantitative assessment method of hierarchical network security threat situation [J]. Journal of software, 2006,17 (4): 885-897.

[5] Liu Mixia, network security situation analysis and survivability evaluation research [D]. Gansu: Lanzhou University of Technology, 2008.

[6] Wang Geng,Zhang Jinghui,Wu Na.Study on the application of network security situation prediction method [J]. computer simulation, 2012,2:98-101.

[7] sun Qiang,Guo Jianghong,Wang Hui. Design and implementation of security management system based on message communication [J]. computer engineering application, 2006 (10): 140-143.

[8] history, Jane,Guo Shanqing,Xie Li. Research and implementation of unified network security management platform [J]. computer application research, 2006 (9): 92-94.