Network Intrusion Detection Using Support Vector Machine Based on Particle Swarm Optimization

Li Wang[1,2], Chunhua Dong[2], Jianping Hu[2], Guodong Li[2]

1.School of Electronics and Information Engineering, Hebei University of Technology,Tianjin,300401,China

2.School of Computer & Information Engineering, Tianjin Chengjian University, Tianjin, 300384, China

Email: yeyue818@163.com

Keywords: Network Intrusion Detection; Support Vector Machines (SVM); Particle Swarm Optimization (PSO) ; Multiclass Classification

Abstract. As an important part of the study of network security, Intrusion detection has aroused special attention of scholars from home and abroad. PSO-based SVM network intrusion detection is innovatively adopted in the paper where PSO is applied to support the parameters of SVM.
Multi-classification is carried out with one versus one (OVO). The experiments on standard intrusion detection data set show that the PSO-based SVM method proposed in this paper is better than classical SVM method. Therefore, PSO -SVM test is very suitable for network intrusion detection.

Introduction

With the development of computer and internet, the network has been widely applied and its security has aroused people's attention. By detecting relevant audit data, system journals or system data for example, Intrusion Detection can decide whether there are any strategies or methods to threaten the security of network. Intrusion detection is to classify relevant data, finding out what data is normal, what is abnormal. Anomaly Detection is an Intrusion Detection technique, which is applied to test the deviation of the present situation to normal situation, according to observation result in the normal situation, and then make the intrusion detection through the analysis system or the deviation between the user behavior and normal behavior. Anomaly Detection can find out the new intrusion methods and users' misbehavior. The detecting methods include nerve system, fuzzy set theory, genetic arithmetic, and immunity theory. But methods like nerve system are low in accuracy when doing network security intrusion detection. Support Vector Machines (SVM) is proposed to avoid these disadvantages [1].

Based on statistic study theory and structure risk minimization, SVM theory is a brand-new study machine, solving the problems of partial extremum and over fitting phenomenon. Despite a lack of priori knowledge, SVM may make more accurate classification so that the whole intrusion detection system is better at detecting data. But there are some innate defects in traditional SVM, and some problems are difficult to solve, for example: the preference of kernel function, the speed of detection etc. These are main difficulties of SVM [2].

So, in this paper, various SVM parameters are optimized utilizing particle swarm algorithm. A SVM network intrusion detection model is put forward based on PSO and experiments have been conducted on subset data_10_percent of KDD CUP99 standard intrusion detection data set, and a good classification has been gained.

PSO －Based SVM Model

Support Vector Machines (SVM)

SVM is a machine-based study method, based on statistic study theory. Generalization of the study machine will be achieved according to risk minimization theory of Vapnik structure.

SVM developed from optimal separate hyper plane of detachable linearity. As for the sample set (x_i, y_i) ,i=1,2,…,n, $x_i \in \{-1,+1\}$, if it meets the condition of $y_i[(w \cdot x_i) + b] - 1 \geq 0$ (1)

, and class interval $2/\|w\|$ is the algorithm, then the problem of algorithm may become dual problem using Lagrange optimization method, at last, the optimal function goes as follows:

$$f(x) = \text{sgn}[(w \cdot x) + b] = \text{sgn}[\sum_{i=1}^{n} a_i y_i (x_i \cdot x) + b] \quad (2)$$

Under the circumstance of inseparable linearity , a smooth solution may be added to **Eq.1**, which is ξ_i , $\xi_i \geq 0$, then,

$$y_i[(w \cdot x_i) + b] - 1 + \xi_i \geq 0 \text{ i=1,…,n} \quad (3)$$

Change the target to minimize $\phi(w, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i$, among which C is penalty factor. It controls the degree of penalty of wrong classification samples. Nonlinear problems may be changed to linear problems of a certain high-dimensional space by nonlinear transformation. We may solve the optimal hyper plane in the transformation space; utilizing the nuclear function in the optimal hyper plane, we may achieve a linear classification after a certain nonlinear transformation without adding complexity of computation. The following classification function is achieved by utilizing dual theory:

$$f(x) = \text{sgn}[\sum_{i=1}^{n} a_i y_i K(x_i \cdot x) + b \ ] \quad (4)$$

among which $0 \leq \alpha_i \leq C$.

The integral operator kernel functions will form various SVM. In the characteristic space, there are various optimized decisions. Widely used kernel functions are as follows:

(1) Linear kernel function: $K(x_i, x\ ) = x_i \cdot x$

(2) Polynomial kernel function: $K(x_i, x\ ) = (\gamma x_i x\ + r)^d$

(3) RBF kernel function (RBF): $K(x_i, x\ ) = \exp(-\gamma\|x_i - x\ \|^2)$

(4)Sigmoid kernel function: $K(x_i, x\ ) = \tanh(\gamma x_i x\ + r)$

among which : γ , r和d are kernel parameters. In this paper, kernel functions are used as RBF[3], which is $K(x_i, x\ ) = (\gamma x_i x\ + r)^d$ in **Eq.4**.

Using SMO[4] training algorithm we may substitute already gained $\{\alpha_i\}_{i=1}^{n}$和b into **Eq.4** to gain classification function, then we put in the test sample to detect the data. In this paper we used PSO to optimize SVM parameter in order to improve the classification accuracy.

PSO-based SVM Algorithm

Particle Swarm Optimization (PSO) is a kind of Swarm Intelligence Algorithm, which originates from birds flock preying behavior, seeking optimal cases through individual coordination and information sharing. Each bird is defined as a particle in this algorithm; the performance of each particle is measured by a fitness set by an objective function. The particle state is determined by the following equation:

$$\begin{aligned} V_{i+1} &= w \cdot V_i + c_1 r_1 (pbest_i - X_i) \\ &\quad + c_2 r_2\ (gbest - X_i) \end{aligned} \quad (5)$$

$$X_{i+1} = X_i + V_{i+1} \quad (6)$$

among which, $pbest_i$ is the best position of one particle; *gbest* is the best positions of all particles; r_1, r_2 are the random numbers between 0 and 1; w is the inertia factor; c_1, c_2 are learning factors[5].

PSO is utilized to optimize the penalty parameter C of SVM and kernel parameter γ. PSO-based SVM parameter optimizing algorithm is as follows:

Step1: initialize particle swarm. Initialize particle swarm(C, γ), confirm swarm model, set particle swarm parameter and maximum number of inner iterations $T_{\max}$.

Step2: assess the fitness of particles. The average accuracy α_{k-cv} calculated through K-fold cross validation method is the fitness function. The process of the Cross validation (CV) goes as follows: divide the already classified sample swarms into equal-valued n parts, use one part to test the classifier trained by other parts to get classification accuracy, and then cross validate the final accuracy which is the average of n parts. So each sample is tested and classification accuracy is a ratio of correctly classified number of samples to the number of all samples. By using cross validation, we may effectively avoid over-fitting and dependence of models on the specific samples.

Step3: if the fitness of each particle precedes its best position *pbest*, then the value will be put to *pbest*; if the fitness of each particle precedes the best position of global variable *gbest*, then the fitness will be put to *gbest*.

Step4: update the speed and position of particles according to **Eq.5** and **Eq.6**.

Step5: validate the maximum iterations. If it meets the requirements, then stop iteration; or transfer to Step2.

Step6: Put out the result. When meeting the requirements to end, C and γ in accordance to maximum fitness function become the optimal combination {C, γ}.

According to the above-mentioned steps we may get optimal penalty parameter C and kernel parameter γ, so that we may make PSO-based SVM model.

Multi-class Classification Method of PSO-SVM

Originally, SVM is utilized to solve 2 types of classification problems, while we often come across various multi-class classification problems. There are various methods to utilize SVM to solve multi-class classification problems. In this paper, one versus one, (OVO) is utilized.

OVO is also defined as comparison-of- pair sorting. Find out various types of pairs from training set T(altogether k types of sets), in total $P = k(k-1)/2$; form training set T(i,j) using these two types of sample point; then get P discrimination function $f_{(i,j)}(x) = \mathrm{sgn}(g_{i,j}(x))$ utilizing SVM. Put input signal X to P discrimination function $f_{(i,j)}(x)$, if $f_{(i,j)}(x) = +1$, then X belongs to type i; type i gets one vote; or X belongs to type j, and type j gets one vote. Calculate k types of votes in P discrimination functions, the type with the most votes is the final decisive type.

Practical Usage of Network Intrusion Detection

To validate the advantage of this method, experiments are conducted on subset data_10_percent[6] of KDD CUP99 standard intrusion detection dataset. The model of PSO-based SVM intrusion detection is as follows: Figure 1.

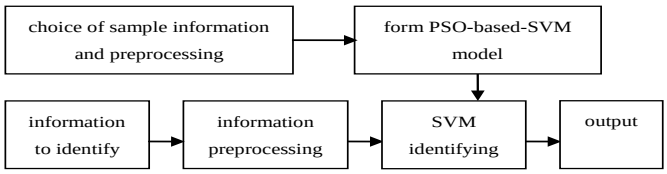

Fig.1. PSO-based SVM Intrusion Detection Model

(1) choice of sample information and preprocessing

Samples are taken from the subset data_10_percent of KDD CUP99 standard intrusion detection dataset, in which there are 42 properties in each item; the first 41 properties are features, which are as follows:

duration,protocol_type,service,flag,src_bytes,dst_bytes,land,
wrong_fragment,urgent,hot,num_failed_logins,logged_in,
num_compromised,root_shell,su_attempted,num_root,
num_file_creations,num_shells,num_access_files,
num_outbound_cmds,is_hot_login,is_guest_login,count,
srv_count,serror_rate,srv_serror_rate,rerror_rate,
srv_rerror_rate,same_srv_rate,diff_srv_rate,srv_diff_host_rate,dst_host_count,dst_host_srv_count,dst_host_same_srv_rate,
dst_host_diff_srv_rate,dst_host_same_src_port_rate,
dst_host_srv_diff_host_rate,dst_host_serror_rate,
dst_host_srv_serror_rate,dst_host_rerror_rate,
dst_host_srv_rerror_rate.The last property label is attack model; in this paper there are 5 attack models: normal, back, neptune, satan, portsweep. Among the 5 attack models, normal is normal network behavior; the other four are abnormal intrusion behavior. Take 200 data from each of property Label; 90% are training samples, while 10% are test samples, namely there are 900 training samples and 100 test samples. Define normal as the first class, back the second class, neptune the third, satan the forth, and portsweep the fifth class. Because of the character attribute data in this data set, first we convert data to a format that is suitable for SVM utilizing programming tools of C language. To avoid calculation saturation, then we normalized the sample data to make the input sample data between [0, 1].

(2) SVM Modeling

Train the training samples with SMO training algorithm， and optimize the two SVM parameters C and γ with PSO to get well-trained SVM prediction model. One-versus-one, OVO is utilized to set 10 SVM prediction models to identify 5 types of attack models in the data set. Put the test samples into 10 SVM models to get the final decision class which is determined by voting method—the one with the most votes.

(3) Comparison between Identification Results

Test accuracy and false alarm rate are utilized to describe the intrusion detection performance：

Test accuracy =number of correctly classified samples/number of all samples；

False alarm rate=number of false alarmed samples/number of all samples.

Table 1 shows the performance of PSO-SVM model proposed in this paper and classical SVM model.

Table 1. Performance Comparison

Model	***Test Accuracy***	***False Alarm Rate***
SVM	96%	4%
PSO-SVM	99%	1%

Figure 1 shows that the test accuracy and false alarm rate of PSO-based SVM proposed in this paper are apparently better than those of classical SVM models.

Figure 2 and Figure 3 are respectively the prediction classification diagram on 10% test sample set of SVM model and PSO-SVM model, in which the actual classification results are marked.

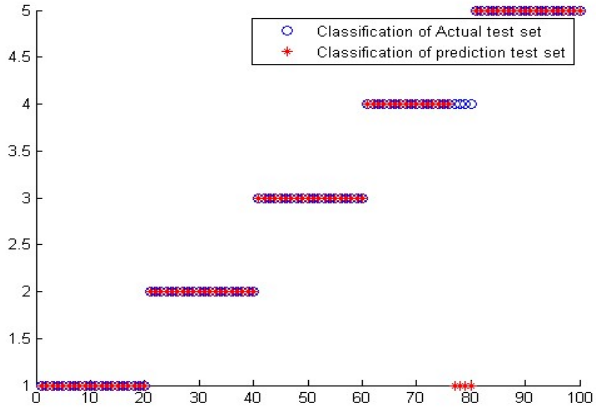

Fig.2. SVM Prediction Classification Chart

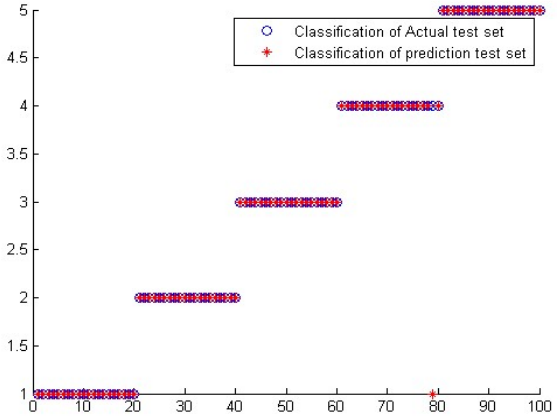

Fig.3. PSO-SVM Prediction Classification Chart

We can easily see from figure 2 and 3 that： it is easy to identify the 4 classes: normal, back, neptune, and portsweep; while it is hard to identify to the class satan. In classical SVM model, four samples of satan are misclassified as normal， while only one sample of satan in PSO-SVM model is misclassified as normal. So we can safely draw a conclusion that with a good performance in network intrusion detection, PSO-SVM is better than classical SVM.

Surely, we may get different classification due to different penalty parameter C and kernel parameter γ， shown in table 2. If the parameter is(1.2， 2.8) instead of PSO-SVM， then the SVM test accuracy is only 74%.

Table 2. Classification Result with Different C and γ

Identifying Method	**c**	γ	**Test accuracy**
SVM	1.2	2.8	74%
SVM	2	0.02	96%
PSO-SVM	39.1463	0.01	99%

Conclusion

Based on the global search character of PSO and classification character of SVM, we proposed PSO-based SVM network intrusion detection method. Compared with classical SVM method, PSO-based SVM network intrusion detection method has solved the problem of parameter optimization with a good modeling effect, and reached a high accuracy in prediction classification so that it reduces the false alarm rate of network intrusion detection.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No.2015BAK01B03), Tianjin Natural Science Foundation (No.11JCYBJC00900), Hebei Province Foundation for Returned Scholars (JFS-2012-13001) and Higher Educational Science and Technology Program of Tianjin (No. 20110814).

References

[1] Xiaoming Li, Zhihan Lv, Baoyun Zhang, Weixi Wang, Shengzhong Feng, Jinxing Hu. XEarth: A 3D GIS Platform for managing massive city information. 2015 IEEE International Conference on Computational Intelligence and Virtual Environments for Measurement Systems and Applications(CIVEMSA 2015).

[2] Liguo Zhang, Binghang He, Jianguo Sun, Mingzhu Lai, Zhihan Lv. "Double Image Multi-Encryption Algorithm based on Fractional Chaotic Time Series". Journal of Computational and Theoretical Nanoscience. 2016.

[3] Wei Luo, Zhiyong Wang, Zhihan LV. Method to Acquire a Complete Road Network in High-resolution Remote Sensing Image Based on Tensor Voting Algorithm. EXCLI JOURNAL, vol.14. Nov. 2015.

[4] Yishuang Geng, Jin Chen, Ruijun Fu, Guanqun Bao, Kaveh Pahlavan, Enlighten wearable physiological monitoring systems: On-body rf characteristics based human motion classification using a support vector machine, IEEE transactions on mobile computing, 1(1), 1-15, Apr. 2015

[5] Jie He, Yishuang Geng, Fei Liu, Cheng Xu, CC-KF: Enhanced TOA Performance in Multipath and NLOS Indoor Extreme Environment, IEEE Sensor Journal, 14(11), 3766-3774, Nov. 2014