TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.2, February 2014, pp. 1502 ~ 1508
DOI: http://dx.doi.org/10.11591/telkomnika.v12i2.3826

Network Intrusion Detection Based on PSO-SVM

Changsheng Xiang*[1], Yong Xiao[2], Peixin Qu[3], Xilong Qu[1]
[1]Department of Computer and Communication, Hunan Institute of Engineering, Xiangtan 411104, China,
[2]Orient Science & Technology College of Hunan Agricultural University, Changsha, 411008, China,
[3]School of Information and Engineering, Henan Institute of Science and Technology, Xinxiang, 453003, China,
Corresponding author, e-mail: cx5243879@sohu.com, Qupeixin@163.com, quxilong@126.com

Abstract

In order to improve network intrusion detection precision, this paper proposed a network intrusion detection model based on simultaneous selecting features and parameters of support vector machine (SVM) by particle swarm optimization (PSO) algorithm. Firstly, the features and parameters of SVM are coded to particle, and then the PSO is used to find the optimal features and SVM parameters by collaboration among particles, lastly, the performance of the model was tested by KDD Cup 99 data. Compared with other network models, the proposed model has reduced input features for SVM and has significantly improved the detection precision of network intrusion.

Keywords: *network intrusion detection, features selection, model parameters, PSO*

1. Introduction

With the tremendous growth of network-based services and users of the Internet, it is important to keep the data and transactions in the Internet more secure. Intrusion Detection System (IDS) can detect the intrusions of someone who is not authorized to the present computer system automatically, so Intrusion detection system has emerged as an essential component and an important technique for network security.

Support vector machines (SVM) is a machine learning method, it has shown growing popularity and has been successfully applied to network intrusion detection [1]. Feature selection is used to identify a powerfully classified subset of network intrusion detection features and reduces the number of features presented to the mining process. By extracting as much information as possible from a given data set while using the smallest number of features, we can save significant computation time and establish network intrusion detection model that generalizes better for data set with all of the features [2]. The dimension of network features are very high and contains redundant and useless features, only a part of features influence the intrusion results, so these redundant and useless features need to deletion [3]. When SVM is used to establish network intrusion detection model, the parameters that should be selected include penalty parameter C and the kernel function parameters [4]. When SVM is used to establish network intrusion detection model, in addition to the feature selection, proper parameters setting can improve the network intrusion detection precision. These two problems are crucial in network intrusion detection modeling, because the feature subset choice influences the appropriate kernel parameters and vice versa [5]. Therefore, obtaining the optimal network intrusion detection feature subset and SVM parameters must be selected simultaneously [6]. In the literature, the Grid algorithm is an alternative to select the best parameters of SVM, however, this method is time consuming and does not perform well [7]. Moreover, the Grid algorithm can not perform the feature selection task. And a few algorithms have been proposed for network feature selection such as GA (genetic algorithm), PSO (particle swarm selection) algorithm, and other algorithms [8]. However, these feature selection algorithms focused on feature selection and did not deal with parameters selection for the SVM classifier.

In order to improve intrusion detection precision, this paper proposed a network intrusion detection model (PSO-SVM) based on simultaneous selection features and SVM

Received July 4, 2013; Revised August 29, 2013; Accepted September 30, 2013

parameters by PSO algorithm, and the performance of the model was tested by KDD Cup 99 data.

2. Research Method

2.1. Principle of SVM

Let the given training data sets be represented by(x_i, y_i), i=1,2…,n, where $x_i \in R^d$ is an input vector, $y_i \in R$ is its corresponding desired output, and n is the number of training data. In SVM a linear function is constructed:

$$f(x) = \omega^t g(x) + b \tag{1}$$

Where, ω is a coefficient vector and b is a threshold.

SVM learning can be obtained by the minimization of the empirical risk on the training data, and the ε-intensive loss function is used for the minimization of empirical risk. The loss function is defined as:

$$L^{\varepsilon}(x, y, f) = \left|y - f(x)\right|_{\varepsilon} = \max(0, \left|y - f(x) - \varepsilon\right| \tag{2}$$

Where, ε is a positive parameter. The empirical risk is:

$$\mathrm{R}_{emp}(\omega) = \frac{1}{n}\sum_{i=1}^{n} L^{\varepsilon}(y - f(x_i)) \tag{3}$$

Other than the ε-intensive loss, SVM tries to reduce the model complexity by minimizing $\left|\omega\right|^2$. This can be described by slack variables ξ_i and $\hat{\xi}_i$.. Subsequently, the SVM approximation is obtained as the following selection problem [9].

$$\min \frac{1}{2}\left|\omega\right|^2 + C\sum_{i=1}^{n}(\xi_i + \hat{\xi}_i) \tag{4}$$

Where, C is a positive constant to be regulated.

By using the Lagrange multiplier method, the minimization of formula (4) causes the problem of maximizing the following dual selection.

$$\max \sum_{i=1}^{n} y_i(\bar{\alpha}_i - \alpha_i) - \varepsilon\sum_{i=1}^{n} y_i(\bar{\alpha}_i - \alpha_i) - \frac{1}{2}\sum_{i,j=1}^{n}(\bar{\alpha}_i - \alpha_i)(\bar{\alpha}_j - \alpha_j)K(x_i, x_j) \tag{5}$$

Where, $\bar{\alpha}_i$ and α_i are Lagrange multipliers, and kernel $K(x_i, x_j)$ is a symmetric function, here the Gaussian function is used as kernel.

$$K(x, y) = \exp(-\frac{\left\|x - y\right\|^2}{2\sigma^2}) \tag{6}$$

Then the approximation function is represented by Lagrange multipliers, namely:

$$f(x) = \sum_{i=1}^{P}(\bar{\alpha}_i - \alpha_i)k(x_i, x_j) + b \tag{7}$$

2.2. PSO Algorithm

Inspired by the social behaviors of bird flocking, PSO algorithm was developed by Kennedy and Eberhat [10]. The particle is endowed with two factors: velocity and position which

can be regarded as the potential solution in the *D* dimension problem space. In PSO algorithm, they can be updated by following formulas:

$$v_{id}(t+1) = \omega v_{id}(t) + c_1 \times r_{1d} \times \left(p_{id}(t) - v_{id}(t) \right) + c_2 \times r_{2d} \times \left(p_{gd}(t) - x_{id}(t) \right) \quad (8)$$

$$x_{i,j}(t+1) = x_{i,j}(t) + vx_{i,j}(t+1) \quad (9)$$

Where, *w* is the inertia weight factor. r_{1d} and r_{2d} are two random numbers. $v_{id}(t)$ and $x_{id}(t)$ are the velocity and position of the current particle *i*. p_i is called "personal best", and its *d*th-dimensional part is p_{id}. The "global best" p_g is the best position found in the whole particles. c_1, c_2 are the acceleration constants.

2.3. Particle Design

When the Gaussian function is selected as kernel function, (*C*, *σ*) and features are used as input attributes. Therefore, the particle comprises four parts: *C*, *σ* and the features mask. In Figure1, C_1~C_i represents parameter *C*, σ_1~σ_j represents the parameter *σ*, f_1~f_m represents the feature mask. In the feature mask, the bit with value '1' represents the feature is selected, and '0' indicates feature is not selected.

Figure 1. Particle Design

2.4. Fitness Function

Detection precision and the number of features are used to design a fitness function. Thus, for the particle with high detection precision and a small number of features produce a high fitness value. We solve the multiple criteria problem by creating a single objective fitness function that combines the two goals into one. As defined by formula (11), the fitness has two predefined weights: (i) w_a for the detection precision, (ii) w_f for the summation of the selected feature.

$$f = w_a \times Acc + w_f \left(\sum_{i=1}^{N_f} f_i \right)^{-1} \quad (10)$$

Where, *Acc* is the network intrusion precision, f_i is defined as follow:

$$f_i = \begin{cases} 1 & feature\ is\ selected \\ 0 & feature\ is\ not\ selected \end{cases} \quad (11)$$

2.5. Design of the Multi-classifier for Network Intrusion Detection

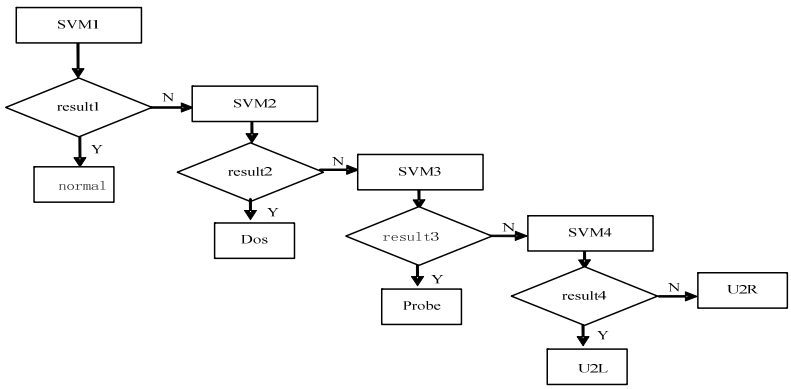

Figure2. Multi-classifier for Network Intrusion Detection

SVM is for two classifier, but the network intrusion has a variety of invasion type, thus network intrusion detection is multi-classify problem, so multi-classifier for network intrusion detection is constructed by "one" to "one" way in this paper, and is as shown in Figure 2.

2.6. The Steps of Network Intrusion Detection

Step1: The network data are collected and the initial features set are extracted.

Step2: The initial particles are produced randomly, which represented parameters of SVM and feature subset.

Step 3: The particles are decoded the parameters of SVM and feature subset, and then the network intrusion detection model is established based on to the corresponding parameters and feature subset, and calculates the network intrusion detection precision, the fitness value is obtained according to formula (10).

Step 4: each particle's fitness value is compared with the P_i, if better, and then the particle takes place the position of P_i.

Step 5: For each particle, its fitness value is compared with the P_g, if better, and then the particle takes place the position of P_g.

Step 6: The velocities and positions of the particles are updated according to formula (8) and (9).

Step 7: The iterative process doesn't stop proceeding until the number of iterations achieves the maximum number of iterations (N_{max}), and the optimal particle is decoded into the optimal parameters of SVM and features subset.

Step 8: The training samples are dealt according to the optimal features subset and are input into SVM to establish the optimal intrusion detection model according to the optimal parameters of SVM.

The work flow chart of network intrusion detection model is as following:

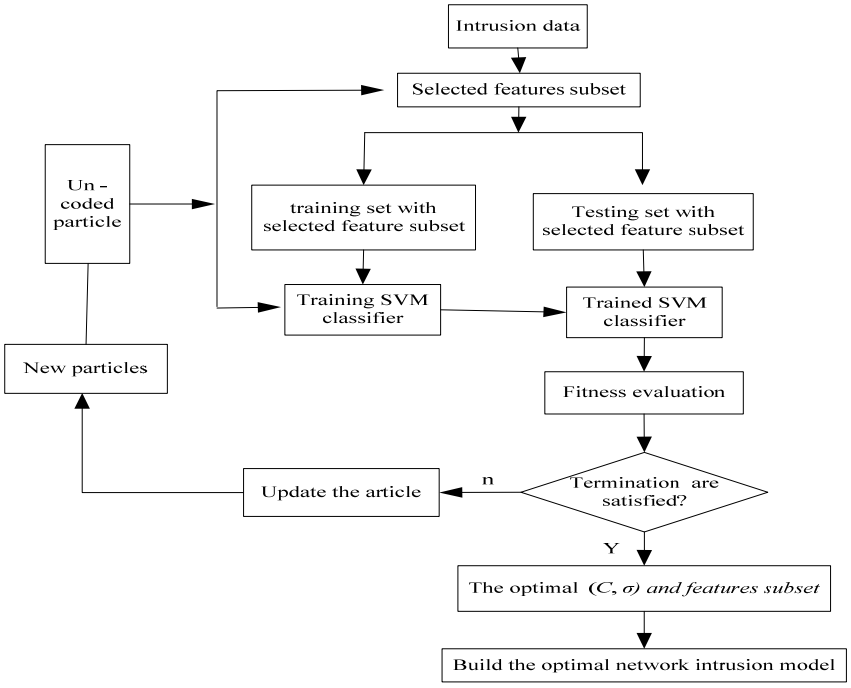

Figure 3. The Flow Chart of Intrusion Detection Model

3. Results and Discussion

3.1. Experiment Data

The experiment data are from DD Cup 99, which contain about 5,000,000 connecting records. There are four categories of attacks: DOS, R2L, U2R, Probing. The parameters of PSO algorithm are set as: the numbers of particle k=20, w=1, c_1=c_2=2, N_{max}=200.

3.2. Comparison Models and Evaluation Criterion

In order to make the detection results of PSO-SVM comparable and persuasive, three comparison models are chosen, SVM1: which features are selected by PSO algorithm while the parameters of SVM are selected randomly, SVM2: which all of the features are selected while the parameters of SVM parameters are selected by PSO algorithm, SVM3: which the features are selected by PSO algorithm firstly, and then the parameters of SVM are select by PSO algorithm. The performances of models are evaluation by precision, recall, and train times.

3.3. Results and Analysis

The PSO algorithm is a heuristic algorithm, the experiment results are random. The numbers of features are appeared in 5 times experiments, and the results are shown in Table 1.

Table 1. The Appeared Times of Each Feature

appeared times	The number of feature
1	11, 14, 17, 19, 38, 16, 39
2	5, 8, 10, 13, 15, 18, 19, 22, 21, 27, 28, 37, 41
3	2, 3, 7, 9, 12, 36, 23, 26, 32, 35
4	4, 6, 20, 25, 29, 33, 34, 40
5	1, 24, 31

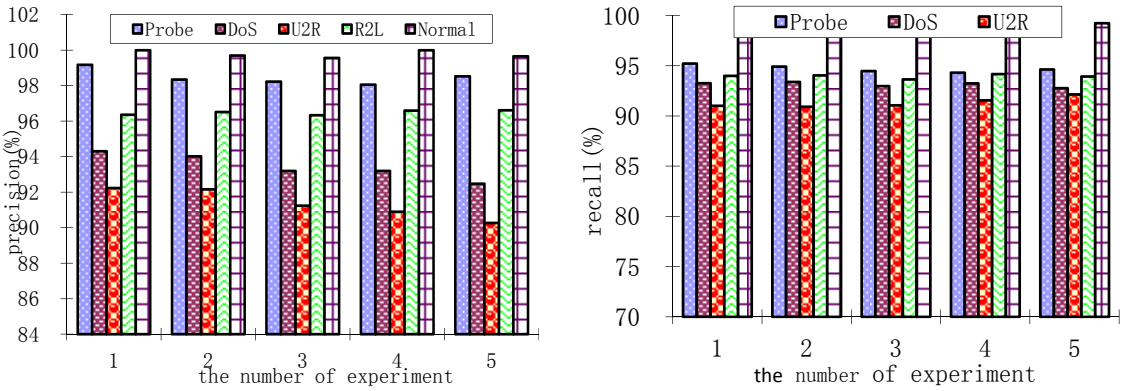

Figure 4. The Detection Precision and Recall of PSO-SVM

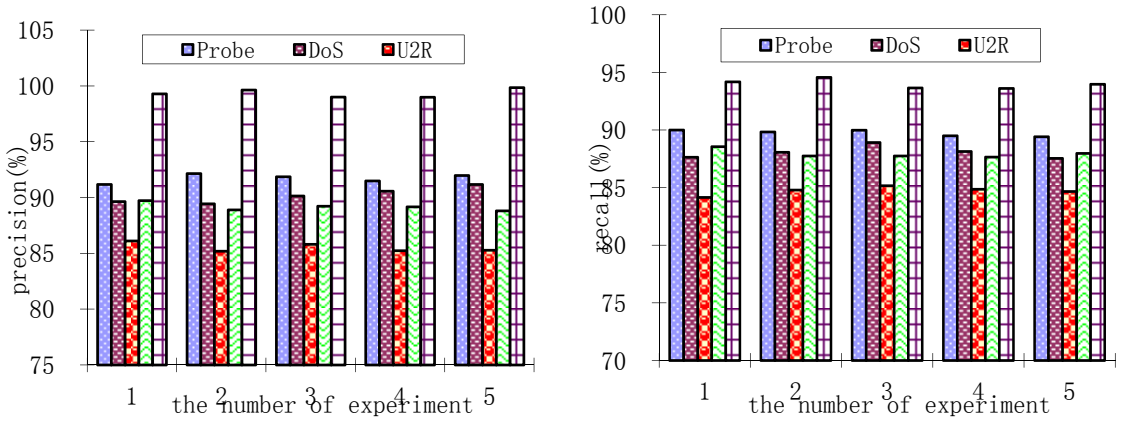

Figure 5. The Detection Precision and Recall of SVM1

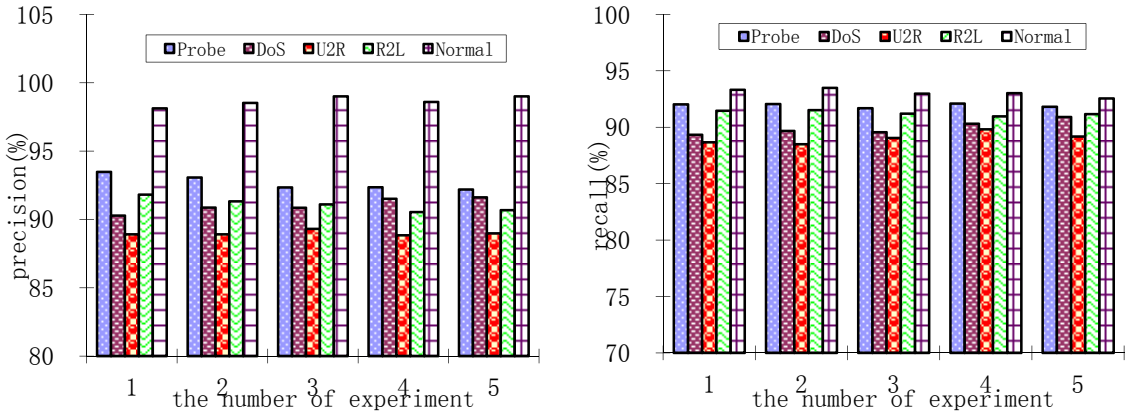

Figure 6. The Detection Precision and Recall of SVM2

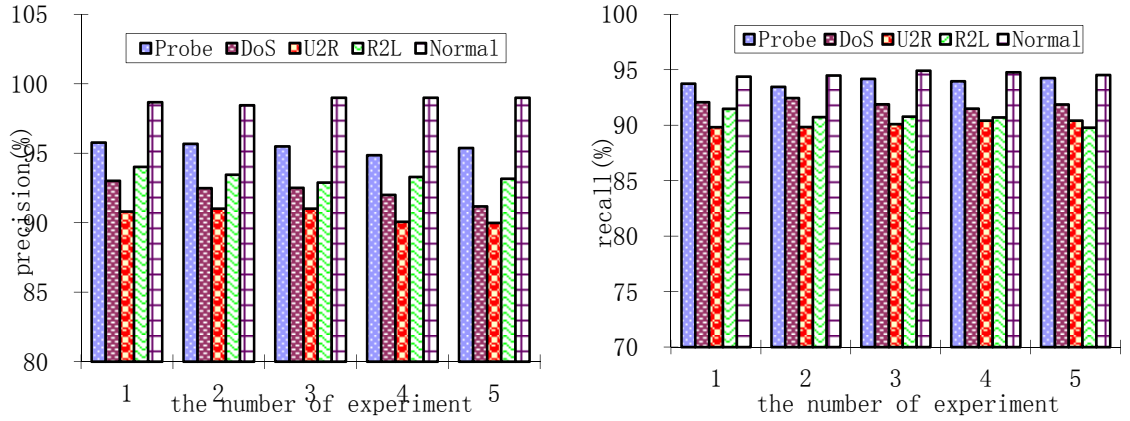

Figure 7. The Detection Precision and Recall of SVM3

In order to determine times of the features appeared have different influence on the result of classification, we carried out testing experiments on three groups of features, the first group: they appear five times plus four attributes: 1, 4, 6, 20, 21, 25, 29, 31, 33, 34, 40. The second group: features appear five times, four times and three times: 1, 2, 3, 4, 6, 7, 9, 12, 20, 23, 25, 26, 29, 31, 32, 33, 34, 35, 40. The third group: all of the 41 features. The results show that the optimal features are 2, 4, 9, 20, 21, 24, 29, 31, 33, 34, 40 and the corresponding optimal SVM parameter C =107.12, σ=1.05, in terms of classification precision and train time. The optimal features subset and the optimal parameters of SVM are used to establish to the network intrusion detection model, the detection results are shown in Figure 4. The detection results of comparison models SVM, SVM2, SVM3 are shown in Figure 5~7.

In Matlab 2012, the tic and toc are used to count train times(s) of the all models, the results are shown in Table 2.

Table 2. The Training Times of Different Models

intrusion type	SVM1	SVM2	SVM3	PSO-SVM
Probe	0.9	1.14	1.12	0.73
DoS	1.07	1.56	0.92	0.83
U2R	1.14	1.12	1.56	0.58
R2L	1.11	1.57	1.66	0.9
Normal	1.18	1.14	1.25	1.01

We can see from Figure 3 that the PSO-SVM can obtain high network intrusion detection precision and recall, and it is an effective network intrusion detection model. The experimental results in table 2 and figure 2~6 are analyzed and conclusion can be obtained as following:

(1) compared with the SVM1, SVM2, SVM3, the network intrusion detection times of PSO-SVM significantly are shortened, and detection efficient is improved, the experiment results show that the PSO-SVM can find the optimal features subset and the parameters of SVM, and eliminates the useless and redundancy of network intrusion detection features, reduces the input vectors of the SVM and computational complexity is deceased, it can more meet the real-time requirement with network intrusion detection.

(2) Compared with the SVM1, SVM2, SVM3, PSO-SVM has improved detection precision and recall of network intrusion, the results show that, there is relation between the network feature and SVM parameters, and they are selected simultaneously by PSO which can achieve the optimal network features and parameters of SVM simultaneously, so the performance of network intrusion detection model can be greatly improved and ensure the safe of the network.

4. Conclusion

Good parameters of SVM and features set are very crucial for intrusion detection model, and the paper proposed a network intrusion detection model based on simultaneous selection features and SVM parameters by PSO algorithm. The results showed that PSO-SVM has obtained better performance than other models, which parameters and features subset are selection separately, PSO-SVM can eliminate the useless and redundancy of features and reduces the input vectors of the SVM, and has improved detection precision of network intrusion detection and has wide application prospect in the field of network security.

Acknowledgements

This research was supported by Hunan Pro Natural Science Foundation (13JJ9022)and Hunan Science & Technology Foundation (2013GK3029).

References

[1] Denning D. An Intrusion Detection Model. *IEEE Transaction on Software Engineering*. 2010; 13(2): 222-232.
[2] Zhang XF, Zhao Y. Application of Support Vector Machine to Reliability Analysis of Engine Systems. *Telkomnika*. 2013; 11(7): 3352-3560.
[3] Khan L, Awad M, Thuraisingham B. A new intrusion detection system using support vector machines and hierarchical clustering. *The VLDB Journal*. 2007; 16(2): 507-521.
[4] Palomo EJ, Dominguez E, Luque RM, et al. A new GHSOM model applied to network security. *Lecture Notes in Computer Science Springer*. 2008; 51(18): 680-689.
[5] Durga Prasad, Nikhil Pal, Jyotirmoy Das. Genetic programming for simultaneous feature selection and classifier design. *IEEE Transactions on System*s. 2009; 36(1): 106-117.
[6] Huang ChengLung, Wang ChiehJen. A GA-based feature selection and parameters optimization for support vector machines. *Expert Systems with Applications*. 2009; 31(2): 231–240.
[7] Natesan P, Balasubramanie P, Gowrison G. Improving attack detection rate in network intrusion detection using adaboost algorithm with multiple weak classifiers. *Journal of Information and Computational Science*. 2012; 8(8): 2239-2251.
[8] Saravanan C, Shivsankar M. An optimized feature selection for intrusion detection using layered conditional random fields with MAFS. *International Journal of Mobile Network Communications & Telematics*. 2011; 12(3): 79-85.
[9] Han FQ, Li HM, et al. A new incremental support vector machine algorithm. *TELKOMNIKA Indonesian Journal of Electrical Engineering.* 2012; 10(6): 1171-1178.
[10] Jie He, Hui Guo. A Modified Particle Swarm Optimization Algorithm. *TELKOMNIKA Indonesian Journal of Electrical Engineering*. 2013; 11(11): 6209-6215.