PSVM: Quantitative Analysis Method of Intelligent System Risk in Independent Host Environment

Shanming Wei[1,2], Haiyuan Shen[1], Qianmu Li[2(✉)], Mahardhika Pratama[3], Meng Shunmei[2], Huaqiu Long[4], and Yi Xia[5]

[1] Jiangsu Zhongtian Technology Co., Ltd., Nantong 226463, China
[2] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
qianmu@njust.edu.cn
[3] Nanyang Technological University, Singapore 639798, Singapore
[4] Intelligent Manufacturing Department, Wuyi University, Jiangmen 529020, China
[5] PT. Sinoma Engineering Indonesia, Jakarta Utara 14440, Indonesia

Abstract. Quantitative risk analysis of security incidents is a typical non-linear classification problem under limited samples. Having advantages of strong generalization ability and fast learning speed, the Support Vector Machine (SVM) is able to solve classification problems in limited samples. To solve the problem of multi-classification, Decision Tree Support Vector Machine (DT-SVM) algorithm is used to construct multi-classifier to reduce the number of classifiers and eliminate non-partitionable regions. Particle Swarm Optimization (PSO) algorithm is introduced to cluster training samples to improve the classification accuracy of the constructed multi-classifier. In the ubiquitous network, the cost of information extraction and processing is significantly lower than that of traditional networks. This paper presents a quantitative analysis method of security risk based on Particle Swarm Optimization Support Vector Machine (PSO-SVM), and classifies the flow data by combining the way of obtaining the flow data in ubiquitous networks, so as to realize the quantitative analysis of the security risk in ubiquitous networks.

In the experiment, KDD99 data set is selected to verify the effectiveness of the algorithm. The experimental results show that the proposed PSO-SVM classification method is more accurate than the traditional one. In the ubiquitous network, this paper builds an experimental environment to illustrate the implementation process of security risk analysis method based on PSO-SVM. The risk analysis results show that the analysis value of risk in ubiquitous network fits well with the change trend of actual value. It means quantitative analysis of risk can be achieved.

Keywords: DT-SVM · PSO-SVM · Ubiquitous network

1 Introduction

Risk analysis of security incidents is a typical nonlinear classification problem with finite samples. Support Vector Machine (SVM), widely used in situational awareness of

X. Zhang et al. (Eds.): CloudComp 2019/SmartGift 2019, LNICST 322, pp. 544–552, 2020.
https://doi.org/10.1007/978-3-030-48513-9_43

traditional network, provides a good classification model for this kind of problem. Compared with traditional network node, which can only get the relevant information of the transmitted packets, the application layer in ubiquitous network architecture can obtain all flow data in the control domain through the control layer. The cost of information extraction and processing can be reduced significantly, making the application prospect of security analysis based on stream data much wider.

As a machine learning algorithm, SVM is based on statistics. SVM has the advantages of strong generalization ability and fast learning speed. It can solve the problem of smaller training set error and larger test set error in traditional machine learning algorithm. SVM can use limited sample information by finding a balance between learning ability and learning accuracy to solve the classification problems of limited samples. At the same time, DT-SVM has fewer sub-classifiers and eliminates unclassifiable region. It is a good method of multi-classification SVM construction. However, it may cause accumulation of misclassification, which is mainly due to the error of the upper classifier. To improve the performance of the classifier, we propose in this paper to cluster the training samples, construct an optimal binary tree classification, and construct DT-SVM using the classified samples and tree structure. The traditional clustering algorithm has good applicability to the clustering problem of low-dimensional data, but its ability to process high-dimensional data and massive data is not enough. The clustering method based on the Particle Swarm Optimization (PSO) can deal with clustering problem in high-dimensional data with quick convergence, thus obtain the global optimal solution. This paper presents a quantitative analysis method of security risk based on PSO-SVM, and classifies the flow data by combining the way of obtaining the flow data in ubiquitous networks, so as to realize the quantitative analysis of the security risk in ubiquitous networks.

The paper is organized as follows: Sect. 1 is the introduction of the paper, Sect. 2 introduced multi-classification problems; Sect. 3 is about the Improvement of DT-SVM by PSO algorithm; Sect. 4 tells the process and result of experiment; Sect. 5 is the conclusion of this paper.

2 Multi-classification Problems

Multi-classification problem of SVM is a hots pot in the field of machine learning. Basic SVM only aims at binary-classification problem. Faced with multi-classification problems, there are many solutions:

One-Versus-Rest SVM: 1-v-r SVM trains k binary-classification SVM for k classes. When constructing the SVM of the first class, the training samples belonging to the first class and the rest belong to another class. When test sample x input into the K binary-classification SVM, and x belongs to the SVM class with the maximum value.

One-Versus-one SVM: 1-v-1 SVM constructs $k(k-1)/2$ binary-classification SVM for k classes. The test samples are inputted into the binary-classification SVM, Then the value of each class is accumulated, and X belongs to the class with the highest accumulated value. These two kinds of SVM may exist unclassifiable regions, so they only used in the case of fewer classes.

Multi-class Support Vector Machine: M-SVM solves k SVMs simultaneously for k classes. It belongs to an optimization problem, but because the objective function is too

complex and the computational complexity is too high, so the practicability is not high when the accuracy is guaranteed.

Decision directed acyclic graph SVM: The construction method of DAG-SVM is the same as that of 1-v-1 SVM. But in the test phase, each binary classification SVM is used as a node to generate a binary directed acyclic graph. $k(k-1)/2$ internal nodes correspond to $k(k-1)/2$ SVMs and k leaf nodes correspond to k classes. Thus, for a problem with k classes, to estimate the class of a test point, we need to evaluate the output of $k-1$ classifiers. DAG-SVM has faster decision-making speed and fewer classifiers, but root-node classifiers have a greater impact on the classification results. Different root-node classifiers have different results which lead to great uncertainty in the final classification results.

Multi-class Support Vector Machine Based on Decision Binary Tree: DT-SVM constructs $k-1$ binary-classification SVMs for k classes. When constructing the i-th class of SVM, the training samples belonging to the i-th class are divided into the first class. The other class includes $i+1, i+2, \ldots, k$ class.

A decision binary tree is constructed from the root node. DT-SVM constructs fewer binary-classification SVMs and eliminates unclassifiable regions. The nearer the leaf node has the fewer the total training samples of SVM. It is a better method to construct multi-classification SVM. There are also some problems in DT-SVM. The error classification of a root node will lead the error to the next leaf node, which will lead to the accumulation of errors.

3 Improvement of DT-SVM by PSO Algorithm

The error accumulation problem of DT-SVM is mainly caused by the error of upper classifier. In order to improve the performance of classifiers, we should follow the construction principle of "easy before difficult". Firstly, the training samples are clustered and an optimal decision binary tree classification is constructed. On this basis, DT-SVM is constructed. The traditional clustering algorithm has good applicability to the clustering problem of low-dimensional data, but it is not effective when facing high-dimensional data and massive data. PSO: Particle Swarm Optimization can deal with clustering problem in high-dimensional data better. Its converge speed is faster and it can get optimal solution more easily.

3.1 PSO Algorithm

PSO algorithm is a kind of swarm intelligence evolutionary algorithm. Its basic idea is to find the optimal solution through cooperation and information sharing among individuals in a group. The parameters of PSO algorithm is less and it can be realized easily. It is widely used in various optimization problems, such as the optimal classification of samples. In this paper, PSO clustering algorithm is used to find the optimal two-class partition in the sample set. The algorithm is as follows:

3.2 Improved DT-SVM Construction Algorithms

Algorithm: PSO: Particle Swarm Optimization

Input： Training sample

Output： Two Classes of Clustering Results of Samples

1. Initialize particle swarm optimization, randomly classify particles in two classes, and calculate cluster centers of different classes as position coding of particles.
2. Calculate particle fitness and initialize particle velocity to 0.
3. Calculate the individual and global optimal positions of each particle based on fitness
4. Update the velocities and positions of all particles
5. Cluster analysis of particles according to the nearest neighbor rule and cluster center coding
6. Calculate the clustering center and update the particle fitness based on the new clustering division,
7. Update individual and global optimal positions of particles
8. If the maximum number of iterations is reached, the run ends with the output of the classification result, otherwise jump 4

Before constructing DT-SVM, PSO clustering algorithm is used to classify training samples, and the optimal binary decision tree structure is generated to maximize the separability of samples. The classified samples and decision tree structure are used to construct the model. The construction algorithm of DT-SVM based on PSO clustering is as follows:

Algorithm: Construction algorithm of DT-SVM based on PSO clustering

Input：Training samples
Output： DT-SVM

1. Sample the initial training set
2. Use PSO algorithm to divide the sample into two sub-nodes
3. Judge whether sub-node 1 is separable, if it can be separable jump 2
4. Judge whether sub-node 2 is separable, if it can be separable jump 2
5. All nodes can't be separable, then generate decision binary tree
6. Constructing DT-SVM with the Samples and Binary Tree Structure

4 Quantitative Analysis Experiment

To verify the effectiveness of the method, this paper uses KD99 data set to test. The KDD99 data set is based on the DARPA data. And Professor Wenke Lee removes a large number of identical connection records and some duplicate data when DoS attacks occur. At the same time, the data set is cleaned and preprocessed. This data set is widely accepted as the standard of attack test data set. The data set contains normal flow data and 39 kinds of attack flow data, which belong to four attack modes: Probe, DoS, R2L and U2R.

4.1 Data Preprocessing

Each connection in the KDD99 dataset contains 41 feature dimensions and one label dimension, the format is as follows:

```
0,icmp,ecr_i,SF,520,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,511,511,0.00,0.00,0.00,0.00,1
.00,0.00,0.00,255,255,1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,0.00,smurf.
```

In the 41 feature dimensions, there are four main characteristics: 9 basic characteristics of TCP connection, 13 content characteristics of TCP connection, 9 network traffic time characteristics and 1 host network traffic characteristics. Among these features, there are 9 discrete features and 32 continuous features, and the metrics of continuous features are not exactly the same.

Firstly, is Preprocessing the data. The discrete features should be continuous and coded before we process them. For example, for the second feature, protocol type, it can be mapped to integers so that TCP = 1, UDP = 2, ICMP = 3, and other types = 4. Other discrete features are also mapped in this way. Secondly, data are standardized and normalized to facilitate analysis and processing. Finally, each connection has 41 features. In this paper, PRFAR attribute reduction algorithm proposed in document [74] is used to select attributes, remove irrelevant and redundant features, reduce data dimension and amount, and improve the training speed of classifiers.

4.2 Analysis of Experimental Results

KDD99 dataset contains many attack instances, which can be considered as a breadth test to test the coverage of the system for attack detection. The dataset used in this paper is kddcup. data_10_percent training set and test set.

The experiment implements PSO-SVM algorithm based on Python's Libsvm software package. The constructed multi-classifier is shown in Fig. 1.

The multi-classifier constructed in this paper regards all the traffic that cannot be classified as attack traffic. Although this method may misjudge normal access as attack, it has much less consequences than misjudge attack as normal traffic. The classification results are shown in Table 1.

The experimental results prove that the proposed method is feasible and PSO-SVM method has more advantages in classification accuracy. Although the training time of the proposed method is longer than that of several SVM algorithms, the construction process

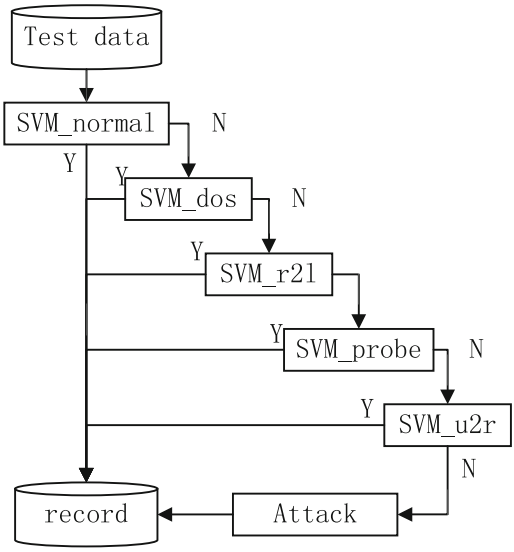

Fig. 1. Multi-classifier

Table 1. Experiment results

Accuracy	Attack categories			
	DoS	R2L	PROBE	U2R
SVM	95.3017	85.8407	98.2906	82.2951
Document[74]	100	91.558	90.3846	88.2353
PSO-SVM	99.2108	97.1722	98.7991	93.274

of DT-SVM is usually prior and will not affect the real-time classification. The controller of Ubiquitous Network Control Layer (Ubiquitous Network Control Layer) can obtain the flow data of all network nodes, and the risk analysis efficiency of Ubiquitous Network is much higher than that of traditional network. This paper uses a simple network structure to illustrate the analysis process of the analysis method proposed in the ubiquitous network environment.

4.3 Risk Calculation

Taking the risk analysis calculation principle proposed by GB20984-2007 as an example, the risk value can be calculated in the following formula:

$$R = R(A, T, V) = R(L(T, V), F(L_a, V_a)) \tag{1}$$

In the formula, R represents the value of security risk; A represents assets; T represents threats and V represents vulnerabilities, and L_a,V_a represent the value of assets and

the severity of vulnerabilities affected by security incidents. *L* represents Indicate the possibility of security incidents and *F* represents the loss caused by security incidents.

Based on the ubiquitous network structure proposed in this paper, a simple ubiquitous network environment is built to illustrate the security risk analysis process based on PSO-SVM. The experimental environment topology is shown in Fig. 2. The risk application runs on Server. Floodlight in the control layer obtains the flow data in the network, and the attacker sends the flow to Server.

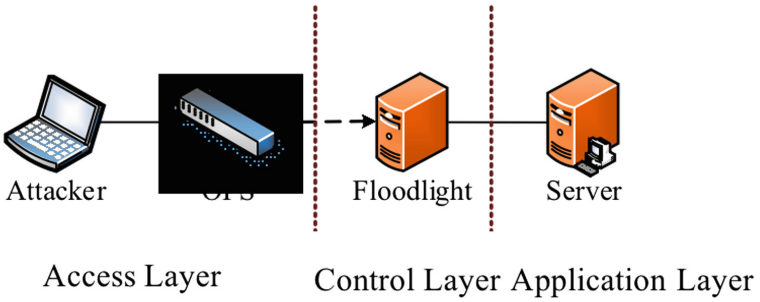

Fig. 2. Experimental Network environment topology

36,000 data are extracted from the test set to reduce the percentage of attack traffic to total traffic, adjust the proportional relationship, and randomly scramble it as simulation traffic data. The specific amount of traffic is shown in (Table 2).

Table 2. Flow data table

Label	Kind	Number
0	DoS	7000
1	R2L	4000
2	PROBE	800
3	U2R	200
4	Normal	24000

In the simulation environment, more than 1000 network traffic data are replayed at a speed of about 1000 bars per second. Record the flow data per minute and calculate the security risk in one minute, then calculate the security risk in one hour, and classify the traffic using the trained multi-classifier as the analysis result. Because the test data of DRAPA organization does not give the network topology description, the following settings are made when completing the simulation experiment: 20% of the total traffic is used as the threshold value, the ratio of an attack traffic to the threshold value is used as the probability of the network being threatened by such attack, and the probability of exceeding the threshold value is 1; Vulnerability is 1, that is, all attacks have corresponding vulnerability; asset value is 1, that is, all assets are equally important; the proportion

of an attack traffic to the total attack traffic is regarded as the weight of the attack, and because vulnerability is equal to the attack weight and the severity of vulnerability, the calculation formula can be simplified as follows:

$$R = R(T \cdot V_a) \tag{2}$$

At the same time:

$$T(Attack_i) = \begin{cases} \frac{Attack\_traffic_i}{\varepsilon} & if(Attack\_traffic_i < \varepsilon) \\ 1 & otherwise \end{cases}$$
$$V(Attack_i) = \frac{Attack\_traffic_i}{\sum_{i=1}^{l} Attack\_traffic_i} \tag{3}$$

The comparison between quantitative analysis value and actual value of specific attack risk is shown in Fig. 3, Fig. 4. It can be seen that the fitting degree between quantitative analysis value and actual value curve is very good, and the quantitative analysis value is slightly higher than the actual value. Because classifiers tend to classify unknown traffic as attack traffic.

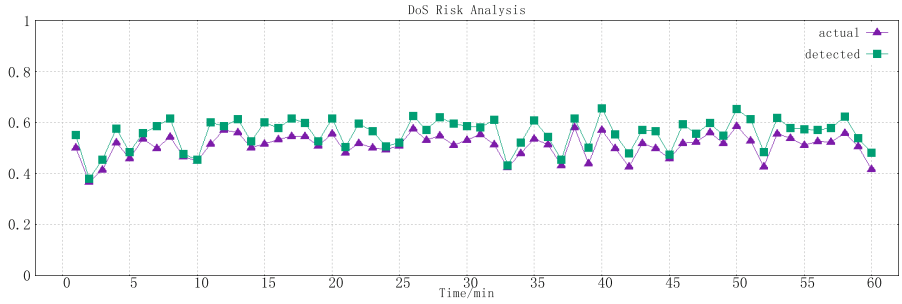

Fig. 3. DoS attack risk analysis

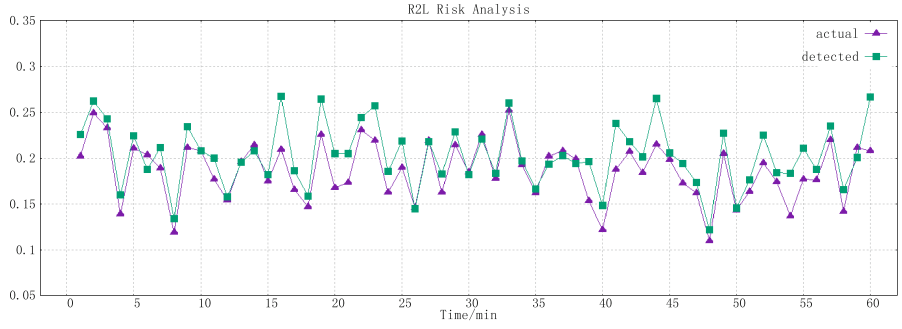

Fig. 4. R2L attack risk analysis

5 Conclusion

Quantitative risk analysis of security incidents is a typical non-linear classification problem under limited samples. Firstly, the related theory of SVM is introduced. SVM has strong generalization ability, fast learning speed, and can solve the classification problem under limited samples very well. To solve the problem of multi-classification, DT-SVM algorithm is used to construct multi-classifier to reduce the number of classifiers and eliminate unclassifiable regions. PSO algorithm is introduced to cluster training samples to improve the classification accuracy of the constructed multi-classifier. In the experiment, KDD99 data set is selected to verify the effectiveness of the algorithm. The experimental results show that the proposed PSO-SVM classification method is more accurate than the traditional one.

This work was supported in part by the Fundamental Research Funds for the Central Universities (30918012204), Military Common Information System Equipment Pre-research Special Technology Project (315075701), 2019 Industrial Internet Innovation and Development Project from Ministry of Industry and Information Technology of China, 2018 Jiangsu Province Major Technical Research Project "Information Security Simulation System", Shanghai Aerospace Science and Technology Innovation Fund (SAST2018-103).

References

1. Xu, X., Liu, Q., Zhang, X., Zhang, J., Qi, L., Dou, W.: A blockchain-powered crowdsourcing method with privacy preservation in mobile environment. IEEE Trans. Comput. Soc. Syst. 1-13 (2019)
2. Qi, L., et al.: A QoS-aware virtual machine scheduling method for energy conservation in cloud-based cyber-physical systems. World Wide Web 5(2019)
3. Qi, L., et al.: Finding all you need: web APIs recommendation in web of things through keywords search. IEEE Trans. Comput. Soc. Syst. **6**, 1–10 (2019)
4. Li, Q., Meng, S., Wang, S., Zhang, J., Hou, J.: CAD: command-level anomaly detection for vehicle-road collaborative charging network. IEEE Access **7**, 34910–34924 (2019)
5. Li, Q., Meng, S., Zhang, S., Hou, J., Qi, L.: Complex attack linkage decision-making in edge computing networks. IEEE Access **7**, 12058–12072 (2019)
6. Li, Q., Wang, Y., Pu, Z., Wang, S., Zhang, W.: A state analysis method in smart internet of electric vehicle charging network time series association attack. Transportation Research Record (2019)