Research on Phishing Website Detection Based on R-SVM Algorithm

Shuting Hu, Jiawei Li, Mingmeng Liu, Bin Pan*
Hunan University of Finance and Economics, Changsha 410205, China

*1808864949@qq.com

Abstract. The Internet has brought us convenience, at the same time important personal information has been exposed. Once the information is leaked, it may cause huge economic losses to individuals. Therefore, it is urgent to realize network security. The traditional method of detecting phishing websites is aimed at big data. Due to the relatively large amount of data, it indirectly causes low efficiency and often fails to solve user needs in a short time. In the detection, the operation is actually performed on the small data stream. If the small data is fundamentally classified, the efficiency will be greatly improved. In summary, based on the realization that we can identify phishing websites, we mainly solve the problems of speed and accuracy. Therefore, this paper uses the R-SVM algorithm to study the data to reduce the time spent on detecting samples and improve efficiency.

Keywords: Phishing websites detection, SVM, R-SVM Algorithm, Network Security

1. Introduction

Today, with the rapid development of information, it is extremely important to maintain network information security and prevent personal information leakage. According to incomplete statistics, about 3.5 billion people in the world use the Internet for communication and entertainment every day, and they are troubled by phishing websites 350 million times a day, and the average probability of new viruses appearing is 6 per second. While we enjoy the convenience of the Internet every day, we are also suffering from data leaks, virus attacks, privacy amplification, and cyber violence.

Currently, there are two main ways to detect phishing websites: one is to detect by URL, which is mainly operated manually, and the accuracy and efficiency are not high, and the other method is to identify through web pages. The amount is also large, which is not ideal. Now adopt the realization idea of the SVM algorithm, map the low-dimensional data sample to the high-dimensional space...
through the kernel function, find the hyperplane that can reasonably classify the data sample in the high-dimensional feature space, realize the classification of the data, and improve the detection rate and accuracy.

2. Related works
For the research of phishing webpage detection, scholars have proposed many detection algorithms in recent years. Shengli Liu[1] filters and extracts different malicious code characteristics by setting a certain threshold. Siyu Cheng[2] proposed an intelligent detection technology based on the client and the SVM algorithm. Jian Feng, Ying Zhang[3] and others proposed to detect web pages based on the document object model (DOM) structure clustering.

Foreign scholars have proposed a special method for small data classification—Support Vector Machine (SVM)[4]. The traditional SVM algorithm needs to train all the data to be learned as training samples. However, The exponential growth of data in practical applications will generate new classification requirements. The traditional SVM classification algorithm cannot meet these new classifications demand.

3. Detection based on R-SVM algorithm
The data information obtained from the KDD99 data set is shown in the following figure:

| Number | Value | Number | Value | Number | Value | Number | Value |
|--------|-------|--------|-------|--------|-------|--------|-------|
| 1      | 11    | 0      | 21    | 0      | 31    | 0      | 41    |
| 2      | tsp   | 12     | 1     | 22     | 0     | 32     | 49    |
| 3      | http  | 13     | 0     | 25     | 6     | 55     | 99    |
| 4      |      | 14     | 0     | 24     | 6     | 54     | 1,000 |
| 5      |      | 15     | 0     | 25     | 0,00  | 55     | 0,00  |
| 6      |      | 20,52  | 0     | 26     | 0,00  | 56     | 0,02  |
| 7      |      | 17     | 0     | 27     | 0,00  | 57     | 0,00  |
| 8      |      | 18     | 0     | 28     | 0,00  | 58     | 0,00  |
| 9      |      | 19     | 0     | 29     | 1,00  | 59     | 0,00  |
| 10     |      | 20     | 0     | 30     | 0,00  | 40     | 0,00  |

**Fig.1 Data set information**

In order to reduce the storage and calculation pressure, the normalization interval is set at [0, 1], and the advantages of the sparse matrix are fully utilized. The normalized transformation formula is[5]:

$$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} \cdot (U - L) + L$$

(3-1)

The R-SVM algorithm[6] is based on the pros and cons of the feature classifier performance and selects from best to worst according to the preset number of required features. Based on the SVM algorithm, it is suitable for data with high dimensions and small samples.

The basic steps of the algorithm are shown in the following figure:
The extracted features are included in 4 categories, indicating that this extraction is comprehensive data extraction. In order to ensure the improvement of accuracy in the follow-up, it is necessary to balance the relationship between reducing the number of features and adding the number of support vectors.

Fig.3 reflects the experimental results of four feature extraction algorithms based on the SVM algorithm:

![Fig. 3 Comparison of feature algorithms](image-url)
As can be seen from the figure:
In order to achieve the best detection effect, the number of samples needs to be increased to increase the detection rate and reduce the false alarm rate. However, as the number of samples increases, the prediction time will also increase significantly. From this, it is inferred that appropriate algorithms should be used to reduce the data dimension. It can improve the efficiency of detection.

4. Experiment and result analysis

4.1. Experimental data
The data set used in this experiment is KDD99. After the data is standardized, 10% of it is used as the training set, and the test set is ‘corrected’.

4.2. SVM experiment based on R-SVM dimensionality reduction
When establishing the classification model and testing the test set, 10,000 data were selected as samples. The composition of the sample set is shown in the following figure:

| total | Normal sample | Abnormal sample |
|-------|---------------|-----------------|
| 2000  | 1780          | 220             |
| 5000  | 4420          | 580             |
| 10000 | 8960          | 1040            |

**Fig.4 Sample set composition**

The performance comparison of each algorithm is shown in the following figure:

| Number sets | Algorithm | Training Time | Accuracy | Detection Rate | False Alarm |
|-------------|-----------|---------------|----------|----------------|-------------|
| 2000        | RFA       | 2             | 0.946    | 0.9178         | 0.032       |
|             | AAC       | 2             | 0.9356   | 0.9094         | 0.023       |
|             | SVM       | 2             | 0.956    | 0.9348         | 0.019       |
| 5000        | RFA       | 16            | 0.954    | 0.9237         | 0.028       |
|             | AAC       | 12            | 0.9308   | 0.9146         | 0.033       |
|             | SVM       | 8             | 0.9631   | 0.9526         | 0.0161      |
| 10000       | RFA       | 47            | 0.956    | 0.9338         | 0.018       |
|             | AAC       | 36            | 0.9387   | 0.9221         | 0.023       |
|             | SVM       | 25            | 0.9621   | 0.9589         | 0.0137      |

**Fig.5 Algorithm performance comparison**

4.2.1. Comparison of detection accuracy
It can be seen from the above figure that when the number of samples is 2000, the R-SVM algorithm is superior to the existing single pre-selection algorithm based on Random Forest[7] and Ant
Clusterin[8] in terms of accuracy, detection rate, and false alarm rate. Only by increasing the number of samples, the other two algorithms can achieve the same detection accuracy.

4.2.2. Comparison of training time
As shown in Fig. 8, When the number of samples is relatively small, the training time of the three algorithms is basically the same, but as the number of samples increases, the training time based on the R-SVM algorithm is significantly less than the two algorithms based on Random Forest and ant Ant clustering.

![Fig.6 Training comparison](image)

4.3. Analysis of results
From the above experiments and the analysis of the data results: in the case of ensuring high detection accuracy, the R-SVM dimensionality reduction algorithm is compared with the pre-selection algorithm based on Random Forest or Ant Clustering. It is highly desirable in terms of dimensionality reduction, data reduction, and reduction of the number of samples. The R-SVM algorithm is superior to the other two algorithms in terms of accuracy, detection rate, and false alarm rate. When the number of samples increases, the training time is also significantly less than the other two algorithms.

5. Conclusion
With the rapid development of information today, the development of network security has a long way to go. This article analyzes the commonly used SVM-based intrusion detection methods and improves the related ideas of the existing methods. The recursive support vector machine (R-SVM) data preselection method is mainly used to effectively reduce the dimensionality of large-scale data and enhance the detection rate of the system. Compared with the traditional SVM-based intrusion detection method, the R-SVM data preselection method applied to the intrusion detection data set KDD99 proposed in this experiment more effectively verifies the feasibility of the R-SVM dimensionality reduction algorithm.
Acknowledgments

This work was supported in part by the National Natural Science Foundation of China, grant number 72073041. Open Foundation for the University Innovation Platform in the Hunan Province, grant number 18K103; 2011 Collaborative Innovation Center for Development and Utilization of Finance and Economics Big Data Property, Universities of Hunan Province. Open project, grant number 20181901CRP03,20181901CRP04,20181901CRP05. 2020 Hunan Provincial Higher Education Teaching Reform Research Project under Grant HNJG-2020-1124, HNJG-2020-1130.2020 General Project of Hunan Social Science Fund under Grant 20B16.

References

[1] Shengli L and Long Y 2012 Detection technology of stealing Trojans based on network communication behavior analysis J.Secrety Science and Technology 12 pp46-52
[2] Siyu C The design and implementation of the intelligent detection system for the client phishing website (2015 the academic degree's dissertation of North China Electric Power University)
[3] Jian F and Ying Z 2018 Phishing webpage detection method based on document object model structure clustering J.Science Technology and Engineering 23 pp81-89
[4] Peizhi L Support vector machine model optimization and its application research (2019 the academic degree's dissertation of Dongbei University of Finance and Economics) pp46-52
[5] Liu Y, Xiang C and Wang H 2017 Optimization of feature selection based on mutual information in intrusion detection J.Journal of Northwest University 5 666
[6] Du H and Zhang Y 2017 Intrusion Detection Based on Incremental Support Vector J.Machine Microcomputer Applications 7 15
[7] Hao F Research and design of intrusion detection system (2015 the academic degree's dissertation of Beijing University of Posts and Telecommunications)
[8] Bangxu D, Hao Z and Gang W 2019 Research on Network Intrusion Detection Method Based on MI and SVM J.Journal of West Anhui University 05 45