Computer Network Flow Recognition Method Based on Improved Support Vector Machine

Shoubai Xiao1,*

1NanChang Institute of Science and Technology, Nanchang, China, 330108

*Corresponding author e-mail: 64349855@qq.com

Abstract. Computer network traffic recognition based on improved support vector machine is a defect of current mainstream network traffic algorithm, designed a network traffic prediction algorithm based on improved support vector machine. This paper mainly introduces the computer network traffic identification method based on improved support vector machine. This article mainly analyzes the related content of network traffic prediction, including the linear and nonlinear characteristics of network traffic, the theoretical basis of network traffic prediction and the method of obtaining traffic data. This paper studies support vector machine theory and least square support vector machine, and proposes an improved algorithm for least square support vector machine. The purpose of this article is to design a network traffic identification and analysis system. On the one hand, by monitoring the network traffic, we will be able to grasp the operation of the entire network in real time; on the other hand, the system statistically analyzes the results of different stages, We have a more comprehensive understanding of the operational efficiency of network resources, network performance and the rationality of network configuration. The experimental results in this paper show that the recognition efficiency of traffic based on the improved support vector machine method has been significantly improved. Under this method, the security problem of traffic has been increased by 14%, and the efficiency of traffic has been increased by 24%. The improved support vector machine will be the future computer network The development trend of traffic identification direction.

Keywords: Support Vector Machine, Computer Network, Flow Identification, Flow Analysis, Prediction Algorithm

1. Introduction

1.1. Background and Significance
Because of the rapid development of my country's computer network technology, the scale of the network is getting larger and larger, and the network structure is becoming more and more complicated. In a complex network structure, it is not only the local bandwidth that restricts network traffic, but the narrowest bandwidth in the entire network result. Then, it is necessary to have a reasonable network load prediction and planning to improve the overall network performance\cite{1}. By analyzing the defects of the existing network traffic prediction algorithms, the support vector machine algorithm is used to predict the network traffic. At present, the performance of support vector machines is sensitive to the selection of parameters, and the current research on the optimization of support vector machines is not perfect. In practical applications, most of the parameters are determined by experience\cite{2}.

At present, there is a lack of technically effective methods for the detection and control of computer network traffic, especially for network applications that do not have fixed traffic characteristics and encrypted data streams. Therefore, the research on network traffic detection technology has important practical significance. The identification and analysis of network traffic can quickly and effectively help enterprises to discover the source and characteristics of network traffic, thus providing a basis for enterprises to maintain and adjust the network\cite{3}.

1.2. Research Status at Home and Abroad

With the increasing prosperity of the Internet, operators and Internet companies are paying more and more attention to the network traffic of their servers\cite{4}. On the one hand, by observing the changes in network traffic, you can understand the operation of the device, and then you can make more reasonable company network Planning; on the other hand, operators also need to have a detailed understanding of the company’s traffic sources and user behavior characteristics in order to reformulate marketing strategies\cite{5}.

Foreign research institutions are also actively conducting similar research. Haffner. P and others use machine learning to identify application layer protocols. The process is to distinguish TCP data streams according to the upstream and downstream directions\cite{6}, and reassemble these data as training. Vector, use the classic AdaBoost, Naive Bayes and other algorithms in machine learning to train samples to generate a detection model. Subhabrata S and Paul Barford and others have proposed a method based on the identification of payload features. This method identifies the application layer protocol by extracting the feature code of the protocol\cite{7}.

1.3. Related Work Content

The kernel parameters and penalty parameters of the support vector machine have a great influence on the complexity and accuracy of the prediction model. Tharwat A proposed a support vector machine parameter optimization method based on the Bat algorithm to reduce the classification error. In order to evaluate the proposed model, the experiment used 9 standard data sets obtained from the UCI machine learning data warehouse. However, due to the complexity of the evaluation, the results obtained are not very accurate\cite{8}. This paper mainly studies the theory of network traffic prediction algorithms. Analyze the various existing prediction algorithm models, study their respective advantages and disadvantages, and derive the theoretical feasibility of a new network traffic prediction algorithm based on the support vector machine model\cite{9}. It also introduces the specific implementation
of the system, which mainly includes the realization of unit time flow statistics and flow distribution statistics, the realization of protocol feature extraction, the realization of business flow distribution statistics, and the realization of network delay distribution. Finally, the content of this chapter is simply Summary[10].

2. Support Vector Machine Prediction Algorithm

2.1. Linear Separable Support Vector Machine Algorithm

If the input of the training sample is \( (x_i, y_i) \), the linear classification decision function can be set as \( D(x) = w \cdot x + b \) in the dimensional space. Set the optimal classification line as \( w^*x + b = 0 \), normalize the classification decision function so that all samples of the two categories satisfy \( D(x) > 0 \), then the classification decision function satisfies the following constraints:

\[
y_i(w \cdot x + b) - 1 \geq 0
\]

Solving the optimal classification function is to maximize the conditions while satisfying the conditions. The Lagrange function is introduced to solve the constraint maximization problem, which is transformed into the objective function formula:

\[
\min L(w, b, a_i) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^{n} a_i [y_i(w \cdot x_i + b) - 1]
\]

(2)

It can be seen from the above that when \( a_i = 0 \) is the training sample, it has no effect on the classification problem. When \( a_i \neq 0 \), the training sample that meets the constraints is the support vector, the optimal solution and the weight vector are obtained, and the value of the optimal solution can be obtained using the original constraints. Where S is the set of support vectors. At this time, the best classification decision function is:

\[
D(x) = \sum_{i \in S} a_i^* y_i x_i^T x + b^*
\]

(3)

If \( D(x) = 0 \), it means that x is on the boundary and is inseparable. If the training result is acceptable, the area is \( \{1 > |D(x)| > -1 \} \) is the generalized area.

2.2. Linear inseparable support vector machine method

In the problem of nonlinear separability, according to the theory of universal functions, as long as a function satisfies the condition, it corresponds to the inner product in a certain transformation space, so that there is a nonlinear mapping. The sample mapping space of the input space has a high Dimensional feature), that is, the kernel function, so that using the appropriate kernel function in the optimal classification surface can linearly classify the problem after nonlinear transformation. When
high-dimensional space is involved, as long as the inner product operation is performed, the original function can be used. The function in the space is implemented, and the details of the transformation can be omitted. Although the dimension of the feature space has increased a lot, the computational complexity has not increased much. So the best classification decision function can be obtained as:

$$D(x) = \sum_{i=0}^{S} a_i^* y_i x_i^T x + b^*$$

(4)

The linear separable support vector machine method is the same. If $D(x)=0$, it means that $x$ is on the boundary and is inseparable. If the training result is acceptable, the area is $\{1|x|D(x)|>1\}$ is the generalized area.

3. Support Vector Machine Flow Detection Experiment

The data collected in this experiment is divided into 5 groups, from set A to set E, as shown in Table 1. set A and set B are pure P2P traffic and non-P2P traffic respectively, set C is the mixed traffic of P2P and non-P2P, and set D and set E are the mixed data collected in one day and one week respectively. Each data set is aggregated every 15 minutes and aggregated four times per hour. The size of all data sets and the proportion of P2P in mixed traffic are listed in the table.

|       | time interval(min) | P2P Flow ratio | P2P Flow type  | Flows |
|-------|--------------------|----------------|----------------|-------|
| Set A | 15                 | 28.50%         | Bit Torrent    | 2342  |
| set B | 15                 | 43.50%         | E Donkey       | 4321  |
| Set C | 15                 | 72.30%         | baizhao        | 1324  |
| Set D | 15                 | 64.30%         | pplive         | 2354  |
| Set E | 15                 | 34.60%         | Bit Torrent,   | 2834  |

4. Support Vector Machine Network Traffic Identification Result Analysis

4.1. Analysis of Network Traffic Distribution Results

Counting the amount of data transmission in the network at each moment is very important for the company's equipment maintenance personnel, because the network data transmission amount in each second can be used to discover problems in the network equipment in time and make maintenance adjustments.

Figure 1 Analysis results of the unit time flow in the network. The unit time flow statistics are mainly displayed using a line chart, in which the abscissa represents time in seconds. The ordinate indicates the data flow, in bytes. From the interface, you can see the amount of data sent and received
by the network at a specific time. Within 1s, the peak of traffic can reach 690k, or it can suddenly drop from the peak to tens of bytes.

![Flow rate change per unit time](image)

**Figure 1.** Flow rate change per unit time

### 4.2. Improved Support Vector Machine Analysis of Traffic Recognition Results

The improved support vector machine and it statistics on the predicted and measured values of the test samples are shown in Table 2. The comparison results of the improved support vector machine's predicted and measured values are shown in Figure 2 respectively.

**Table 2.** Network traffic prediction results

|   | actual value | Support vector machine | BP neural network |
|---|--------------|------------------------|------------------|
|   |              | predicted value | APE(%) | predicted value | APE(%) |
| 1 |              | 5.5               | 7.3    | 35.21          | 7.3    | 42.1 |
| 2 |              | 7.8               | 5.30   | 43.25          | 7.4    | 25.6 |
| 3 |              | 8.2               | 7.1    | 45.32          | 8.5    | 34.2 |
| 4 |              | 9.7               | 9.4    | 3.56           | 11.2   | 23.4 |
| 5 |              | 8.7               | 6.5    | 4.32           | 9      | 2.11 |

![Comparison of predicted and measured values](image)
Figure 2. Improved support vector machine traffic prediction results

It can be seen from the chart:

The prediction result of BP neural network algorithm is achieved, and the prediction result of improved support vector machine algorithm is that the prediction effect based on improved support vector machine algorithm is obviously better than that of neural network algorithm. This is because the neural network algorithm is easy to cause over-fitting of training and lead to large prediction errors. For individual points, the predicted value deviates greatly from the actual value, and the overall prediction effect is not ideal. The support vector machine is based on the principle of structural risk minimization, and the global optimal solution is obtained, and its generalization performance is higher than that of neural network. The total time required to improve the support of the machine method differs greatly from the total cost of the network algorithm. In terms of training time, the training and prediction time of the support vector machine algorithm is much faster, but the parameter optimization takes more time. However, the improved algorithm is to optimize the model parameters through genetic algorithm, without prior knowledge and insensitivity to the initial parameters, so it will not fall into the local minimum, as long as the appropriate population size is set, the search time can be greatly shortened.

5. Conclusion

The research content of this paper is based on the improved network traffic algorithm of support vector machine. The article first summarizes the research progress of network traffic, analyzes the existing linear and nonlinear prediction algorithm models, points out the characteristics and deficiencies of each algorithm, and also discusses the use of support vector machine algorithm for network traffic The advantages and feasibility of prediction.

This article mainly introduces that network traffic contains a large amount of information data between users and merchants, so it is very necessary to build a comprehensive traffic statistics analysis system. At present, most of the functions of the traffic statistics system are not perfect. For example, the business statistics cannot be collected, and the transmission delay in the network is not accurately counted. Most traffic statistics systems are desktop application software, and the system cannot fully utilize the computing power of the server.

This paper designs a new improved support vector machine algorithm for the problems of the support vector machine algorithm, that is, the introduction of genetic algorithm to achieve the optimization of the parameters of the support vector machine, so that the support vector machine can be determined more quickly and accurately Various parameters. Finally, by using the existing network traffic instance data for testing and comparing with the prediction results of the neural network algorithm, the results prove that the improved algorithm studied in this paper has much higher prediction accuracy and has certain advantages.

Acknowledgments

This work was supported by the Science and technology project of the Jiangxi Provincial Education Departmen(No. GJJ171095).
References

[1] Tharwat A , Hassanien A E , Elnaghi B E . A BA-based algorithm for parameter optimization of Support Vector Machine[J]. Pattern recognition letters, 2017, 93(jul. 1):13-22.

[2] Zhang X , Wang J , Zhang K . Short-term electric load forecasting based on singular spectrum analysis and support vector machine optimized by Cuckoo search algorithm[J]. Electric Power Systems Research, 2017, 146(2):270-285.

[3] Wang W , Xi J , Chong A , et al. Driving Style Classification Using a Semisupervised Support Vector Machine[J]. Human Machine Systems IEEE Transactions on, 2017, 47(5):650-660.

[4] Liu Zhongxiang. Research on practical teaching of computer network major in secondary vocational school based on cloud classroom% [J]. Journal of Jiamusi Institute of education, 2017, 000 (006): 425-426.

[5] Yang S . Analysis for the reliability of computer network by using intelligent cloud computing method[J]. International Journal of Computers & Applications, 2019, 41(3-4):306-311.

[6] Schlueter-Kuck K L , Dabiri J O . Identification of individual coherent sets associated with flow trajectories using coherent structure coloring[J]. Chaos An Interdisciplinary Journal of Nonlinear Science, 2017, 27(9):091101.

[7] Keshavan J , Humbert J S . An Optical Flow-Based Solution to the Problem of Range Identification in Perspective Vision Systems[J]. Journal of Intelligent & Robotic Systems, 2017, 85(3-4):651-662.

[8] Gandomi M , Hassanpour H . Behavioral analysis of traffic flow for an effective network traffic identification[J]. International Journal of Engineering Transactions B Applications, 2017, 30(11):1740-1745.

[9] Kosunalp S . An energy prediction algorithm for wind-powered wireless sensor networks with energy harvesting[J]. Energy, 2017, 139(nov. 15):1275-1280.

[10] Jirsik T , Cernak M , Tovarnak D , et al. Toward Stream-Based IP Flow Analysis[J]. IEEE Communications Magazine, 2017, 55(7):70-76.