Acta Technica *62* (2017), No. 4A, 611–616

Application and implementation of svm algorithm in network traffic identification

Wu Fei1[1], Luo Fuchai[1], Su Jiangwen[2],Wang Qiuling[2],Lin Wei[2]

Abstract. With the rapid development of computer network technology and the advent of the information age, the increasing use of the network has caused the Internet data traffic's burst growth; The emergence of new applications of the network has resulted in a more flexible and mixed use of network communication protocols; Network viruses, eavesdropping and malicious attacks and other acts continue to increase, this has make the network security become a hot spots that concerned by social and government. All of these problems can be solved by network traffic identification. Therefore, people have paid more and more attention to the network traffic identification. This paper proposes a home network traffic identification method based on SVM, from the experimental results we can see that, SVM is very suitable for solving the nonlinear traffic classification problem, and has the advantages of less training samples, low computational complexity and real-time identification.

Key words. Computer Network,data Traffic,SVM, Network traffic identification;.

1. Introduction

With the rapid development of computer network technology, people entered into the information age, they can browse the web, watch the video, brush microblogging and send WeChat and so on every day, all aspects of life are filled with a lot of information, and this is an era of information big bang. As the number of Internet users grows year by year, resulting in a lack of network bandwidth, network congestion increased [1-3]. Meanwhile, communication protocols become more complex and more confusing; this makes it more difficult to analyze the type of business application through a network protocol. Now people like to shop online, using online banking to pay for business, which makes life more quickly and easily. However, the increase

[1]Workshop 1 - Fujian Yirong Information Technology Co. Ltd, Fujian 350003, China; e-mail: wu_fei@fj.sgcc.com.cn

[2]Workshop 2 - State Grid Fujian Electric Power Company, Fujian 350003, China

in network viruses, eavesdropping on the network, fraud and malicious attacks are also more and more, these are threatening the users' property and information security. In view of this, in the practical application, the role played by network traffic identification can't be ignored, and network traffic identification will become more important with the popularity of the network [4]. This paper researches the principle of SVM to solve linear and nonlinear classification problems, on this basis, a method of traffic identification based on SVM is proposed, applying SVM to Network Traffic Identification. Select the radial basis function as the kernel function of SVM, and achieve non-linear mapping from low-dimensional network flow feature space to higher dimension space [5-6]. And constructs the SVM multivalued classifier through the One-Against-One method, so that SVM can identify a variety of network application types [7].

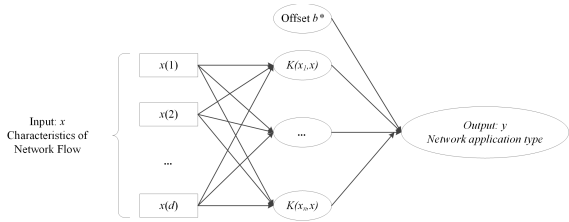

Fig. 1. Traffic Recognition Scheme Based on SVM

2. Traffic Recognition Algorithm Based on SVM

SVM is a machine learning method for small sample set; realize the nonlinear mapping by inner kernel function, and the classification method is simple and effective, and it is able to adapt to the large data and diversity of network environment. So the SVM is applied to network traffic identification [8]. In Figure 1, the SVM input the network traffic features information, and each support vector take the inner product to get the median, and then combine the middle value linearly to get the output. The whole information processing process of SVM is similar to the neural network, so the SVM is also called the support vector network.

2.1. Selection of SVM Kernel Function

The most important aspect of the SVM algorithm is how to select a kernel function. Select different kernel function$K(x_i, x)$ to process the data of different properties can form different types of nonlinear decision functions, thus forming a different SVM algorithm. In practical problems, usually choose a kernel function from the common used as a nonlinear transformation.

Table 1 Features of Network Traffic

Number	Time Span	Features
1	1s	downstream packets
2	1s	upstream packets
3	1s	downstream data volume
4	1s	upstream data volume
5	1s	ratio of downstream and upstream packets
6	1s	ratio of downstream and upstream data volume
7	15s	variance ratio of downstream and upstream packets
8	15s	variance ratio of downstream and upstream data volume
9	15s	number of IP for large data in the downstream
10	15s	proportion of the amount of data in the peak area
11	15s	proportion of the number of samples in the stable area

Table 2 Application Type of Network

Number	Types of Application for Network Traffic Identification	Test Case
1	P2P multimedia or download	Storm video, Thunder download
2	Non-P2P multimedia or download	Website Youku video, web page download
3	WWW (Web browsing)	Sogou browser, IE browser
4	Online Game (Client Game)	New Tian Long Ba Bu (client game)
5	Video call / Conference	QQ Video call
6	File sharing (LAN)	QQ transfer files, file group sharing

In network traffic identification, the radial basis function is chosen as the kernel function of SVM. There are two main reasons for this: on the one hand, the radial basis function can map the input network traffic Features to a certain high dimensional space nonlinearly, the nonlinear problem becomes more convenient to solve; on the other hand, when choosing a certain function parameter, the radial basis function can approximate the recognition accuracy of the linear kernel function and the polynomial kernel function, and it is more applicable to a wider range. The radial basis function is as follows:

$$K(x_i, x) = \exp\{-\frac{\|x - x_i\|}{\sigma^2}\} \tag{1}$$

2.2. Construction of SVM Multivalued Classifier

The substance of the classical SVM is only used to identification two types of network applications, and is a simple binary classifier. However, network traffic identification is a typical multi-classification problem; therefore, SVM needs to be constructed as a multi-valued classifier. This paper chooses a One-Against-One approach as a solution for multi-categorization of network traffic identification, mainly because this paper divide the applications has a large bandwidth requirements into 6 types, and the types needs to be identified is not a lot, and the One-Against-One approach has no other shortcomings. After a comprehensive consideration, choosing the One-Against-One approach is more appropriate.

For test results, voting method is often used, that is, the test sample belongs to the category of the most votes. The C_k^2 decision functions are constructed as follows

$$\begin{aligned} f_{1,2}(x) &= \omega^{1,2} \cdot x + b^{1,2} \\ &\cdots \\ f_{i,j}(x) &= \omega^{i,j} \cdot x + b^{i,j} \\ &\cdots \\ f_{k-1,k}(x) &= \omega^{k-1,k} \cdot x + b^{k-1,k} \end{aligned} \tag{2}$$

The specific procedure for identifying the i-th category by voting is as follows:

$$\begin{aligned} &D_i(x) = \sum_{j=1,j\neq i}^{n} \frac{f_{i,j}(x)+1}{2} \quad (i = 1, ..., k) \\ &x \in class = \arg \max_i \ D_i(x) \end{aligned} \tag{3}$$

3. Implementation of Traffic Identification Based on SVM

3.1. Capture Network Packets

Real-time capturing of network traffic is carried out by the home network management. The router maps the WAN port address to the LAN port address, and connects to each home device respectively. In the router, the WAN is entered from the network device eth0, LAN out from the network equipment br-lan. The network traffic identification program embedded in the gateway needs to capture packets from br-lan, through this it can captures each packet through the router.

3.2. Generate network Traffic Features

In this paper, the network application in the family is divided into six types, and by the time window detection method to produce 11 network flow features. Time window detection method intercepts the change information of the network traffic in a certain period of time, 11 features of the network traffic can be generated from the time window. According to many times of test analysis, set the time window of the network traffic to 15 seconds, take each second of the data as a sample, and it has 15 samples.

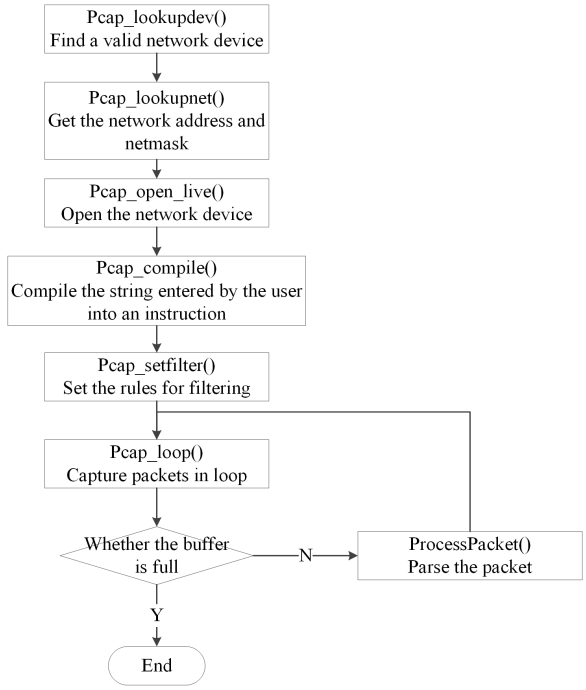

Fig. 2. Call flow of the Libpcap library

3.3. Analysis of Identification Results

Four sets of sample sets are acquired in four different time periods: Setl, Set2, Seta, and Set4. Select any one group as the test sample set, and the remaining 3 groups are used as training sample sets, so there are four combinations, and SVM algorithm can be tested 4 times.

Through the verification, SVM takes the penalty parameter of 1.5 and the kernel function parameter is 0.125. As can be seen from Figure 3, for the six types of network applications, all of the4 time's identification accuracy of SVM is 100%. High accuracy of SVM network traffic identification illustrates that: The 11 network traffic features generated by the time window detection method are very representative, and it can describe the six types of network applications very well; and the SVM algorithm is very suitable for solving the nonlinear traffic classification problem.

4. Conclusion

Firstly, this paper introduces the basic principle of SVM in binary linear classification. For non-linear classification problems, an improved method of adding slack variables and replacing the vector inner product with kernel function in the constraint condition is proposed, so as to realize the nonlinear mapping of SVM and generate the optimal hyperplane in high dimensional space. Then, the SVM-based

traffic identification method is described in detail. SVM has a strong adaptability in traffic identification, the SVM kernel function selects the radial basis function, and the SVM multivalued classifier uses a One-Against-One constructor. Finally, use SVM to train and identify the network traffic sample set, and the performance of SVM-based traffic identification method is analyzed according to the experimental results.

References

[1] Callado. A, Kelner. J, Sadok. D: *Better network traffic identification through the independent combination of techniques* . Journal of Network & Computer Applications *33* (2010), No. 4, 433–446.

[2] Su. M. Y: *Using clustering to improve the KNN-based classifiers for online anomaly network traffic identification.* Journal of Network & Computer Applications *34* (2011), No. 2, 722–730.

[3] Best. D . M,Hafen. R. P, Olsen. B. K: *Atypical behavior identification in large-scale network traffic.* Large Data Analysis and Visualization. IEEE(2011), 15–22.

[4] Ichino. M, Maeda. H,Yoshiura. H: *Score Level Fusion for Network Traffic Application Identification.*Ieice Transactions on Communications *99* (2016), No. 6,1341–1352.

[5] Flouri. K,Beferull-Lozano. B,Tsakalides. P: *Training a SVM-based classifier in distributed sensor networks.* Signal Processing Conference , 2006, European. *29* (2015),1–5.

[6] B.. Wu, H. Shen ,K. Chen: *DIAL: A Distributed Adaptive-Learning Routing Method in VDTNs.* Proc. of the IEEE International Conference on Internet-of-Things Design and Implementation (IoTDI)(2016).

[7] B. Wu,H. Shen: *A Time-Efficient Connected Densest Subgraph Discovery Algorithm for Big Data.*Proc. of the 10th IEEE International Conference on Networking, Architecture, and Storage (NAS)(2015).

[8] B. Wu ,H. Shen,K. Chen: *Exploiting Active Sub-areas for Multi-copy Routing in VDTNs.*Proc. of the 24th International Conference on Computer Communications and Networks (ICCCN)(2015).

Received November 16, 2016