An Improved SVM Method for Internet Traffic Classification

Siyuan Wu[a], Dengyin Zhang[b,*], Fei Ding and Min Zhang

Institute of Internet of Things, Nanjing University of Posts and Telecommunications University, Xinmofan Road, Nanjing, China

[a] 9507542@qq.com, [b] zhangdy@njupt.edu.cn

*Corresponding author

Keywords: support vector machine (SVM); features select； network traffic classification; classification accuracy

Abstract: With the development of network technology, network traffic classification technology plays an increasingly important role in network service and management. In terms of classify traffic flows, Support vector machine (SVM) has shown that more effective than traditional methods. However, classification accuracy will have different performance due to different features. Therefore, we proposed a new method to choose the best combination of features. Moreover, a novel SVM is proposed, which can eliminate the influence of noise on classification accuracy. Experimental results show that the method proposed in this paper has a better and more stable performance.

1. Introduction

The accuracy of the network traffic classification is the premise and foundation of analyzing network user behavior, detecting network abnormal behavior and improving network QoS. Therefore, network traffic classification has attracted great attention from academia. Many traditional businesses use fixed ports assigned by IANA, and port analysis can be used to accurately recognize port numbers directly [1]. However, with the widespread use of dynamic ports [2], it is inefficient for some unreliable ports. Moore et al. [3] proposed a signature matching method. Although it has a higher classification accuracy, but it can't handle continuously updated signature patterns and its inability of handing encrypted packets limit the application. Therefore, using machine learning method to classify network traffic becomes an important research direction of network measurement.

Recent years, many researchers have applied the method of machine learning to network classification. The machine learning method classifies traffic flow based on the statistical characteristics of the traffic. Due to its independence in network address, load encryption and dynamic port, network traffic classification method based on machine learning has widely used. This method is mainly based on the classification of the flow statistical characteristics of the transport layer, according to the experience and experiment, the corresponding feature attributes are extracted, and various algorithms of machine learning are used to realize the traffic classification. At present, it is divided into two main method including network traffic classification based on supervised learning and network traffic clustering method based on unsupervised learning. Yang et al. [4] proposed based on C4.5 and AdaBoost + C4.5 network traffic classification method, and using association-based filtering method to abstract features. Wang et al. [5] proposed an SVM-based P2P traffic identification method, but the proposed method relies on the connection mode of the application, and the stability of the classification result is greatly affected by the network environment. McGregor et al. [6] proposed an unsupervised learning network traffic classification method based on EM clustering. Such methods do not need to use the type of training samples in the clustering process, so they can identify part of new types network traffic that have not been defined.

At present, many feature selection methods have been proposed. In order to improve the accuracy of classification by machine learning. We adopted a Fuzzy SVM-based recursive feature

  DOI: 10.25236/systca.18.090

elimination (FSVM-RFE) selection algorithm to obtain the best combination of features for support vector machine. This option set of discriminations not only yields high accuracy, but also overcome the problem of reduced recognition accuracy caused by the complexity of network traffic statistics and the avoidance and identification of various application protocols.

2. A Fuzzy SVM-based Recursive Feature Algorithm

Aiming to reduce the large amount of noise in network traffic and the excessive redundant features in sample. A fuzzy support vector machine based on recursive feature elimination (FSVM-RFE) method is proposed in this paper. By introducing the fuzzy factor, the method can effectively eliminate the influence of noise and outlier samples on classification accuracy. Moreover, recursive feature elimination method uses classifiers to evaluate the importance of a feature subset and, generally, can achieve better performance than traditional feature selection methods, which can effectively eliminate the influence of weak correlation features and redundant features on classification accuracy.

A set S of l training samples, each represented are given as (x_i, y_i, μ_i) where x_i is the feature vector, y_i is the class label, and μ_i is the fuzzy membership. Each training sample can be divided into two classes. And we can use a label $y_i = \{-1, +1\}$ to describe it. A fuzzy membership $0 \le \mu_i \le 1$ with $i = 1, ...l$, corresponding to x_i.

The mean value of the positive and negative samples is O^+、O^-. At this point, the distance between the sample points in the positive and negative classes to the hyperplane in each class is:

$$\begin{cases} d_{i+} = \dfrac{\varphi^T (x - O^+)}{\|\varphi\|} \\ d_{i-} = \dfrac{\varphi^T (x - O^-)}{\|\varphi\|} \end{cases} \tag{1}$$

Where φ is the normal vector of the $O^+ - O^-$, and φ^T is the transpose of the φ.

Let $D_+ = \max\{d_{i+}\}$、$D_- = \max\{d_{i-}\}$ denote the maximum distance of the sample points in the positive and negative classes from the hyperplane in their class respectively. Thus, the fuzzy factor can be defined below:

$$\mu_i = \begin{cases} -1 + 2e^{-\ln 2/(D_+ + \delta l_{i+})} \\ -1 + 2e^{-\ln 2/(D_- + \delta l_{i-})} \end{cases} \tag{2}$$

Where δ is the adjustment factor to ensure $0 < \mu_i \le 1$.

SVM-RFE with a minimum redundancy-maximum relevance method selects the features with the highest relevance to the target class and is also minimally redundant. The mutual information between two random variables X and Y is defined as [7]

$$I(X,Y) = \iint_{\Omega_Y \Omega_X} p(x,y) \log_2 \left(\frac{p(x,y)}{p(x)\,p(y)}\right) dxdy \tag{3}$$

Where Ω_X and Ω_Y are the sample spaces of X and Y, respectively, $p(x,y)$ is the joint probability density function, and $p(x)$ $p(y)$ are the marginal probability density function.

Let $F = \{f_1, f_2, ..., f_D\}$ denote the vector that is composed of all the features, and let $c \in C = \{+1, -1\}$ denote the class variable (malignant or benign). The relevance R_S of the features in subset $S \subset F$ is given below:

$$R_S = \frac{1}{|S|} \sum_c \sum_{f_i \in S} I(f_i, c) \tag{4}$$

The redundancy of feature f_i with the other features in subset S is given below:

$$Q_{S,f_i} = \frac{1}{|S|^2} \sum_{f_j \subset S, f_i \ne f_j} I(f_i, f_j) \tag{5}$$

The features f_i can be selected through this method:

$$f_i^* = \arg\max_{f_i \subset S} \frac{R_S}{Q_{S,f_i}} \tag{6}$$

The fuzzy support vector machine based on recursive feature elimination (FSVM-RFE) proposed in this paper removes the influence of noise or outlier samples on classification accuracy by introducing fuzzy factors. On the other hand, the recursive feature elimination method can eliminate the effect of weak correlation and redundant features on classification accuracy. The construction steps are as follows:

Step 1: Calculate the sample mean points O^+、O^- in the sample set to be tested, and obtain the positive and negative inner hyperplane $\varphi^T(x-O^+)=0$、$\varphi^T(x-O^-)=0$;

Step 2: Calculate the distance d_{i+}、d_{i-} between positive and negative sample points to the hyperplane in each class;

Step 3: Calculate the distance from the sample point to the hyperplane in the class to obtain the fuzzy factor μ_i;

Step 4: Construct a fuzzy support vector machine (FSVM) using the fuzzy factor μ_i;

Step 5: Using recursive feature elimination method to obtain the features f_i^*;

Step 6: Using the features f_i^* to replace the features in Step 4, and repeat the step 4-6 until obtain FSVM-RFE.

3. Experiment and Result Analysis

3.1. Data for Experiment

In order to verify the effectiveness of the proposed method, the network traffic experimental data were obtained from a backbone router of the campus network of our university. We collected a set of 8 hours traffic data on a Gbps Ethernet link within a week period. The experimental data set contains a total of 153 638 network flow samples, which are divided into 7 categories. The application names and the number of each type of network flow are shown in Table 1.

Table 1 Data set for network flow experiment

Traffic class	Application	Number of flows
Bulk	ftp, BaiduNetdisk	13,468
Mail	smtp, pop3, imap4	2368
Service	dns, Whois	17,463
WWW	http, https	47,142
P2P	bittorrent,edokey,webex,	10,478
Others	tftp, ntp, snmp, smb,etc.	61,965

3.2. Result Analysis

In order to verify the accuracy classification ability of FSVM-RFE in the network traffic with noise or outlier sample environment, this section compared the SVM-RFE [8] and FSVM-RFE. The experimental results are shown in Table 2.

Table 2 Comparison of classification accuracy between SVM-RFE and FSVM-RFE

Type	SVM-RFE(%)		FSVM-RFE(%)	
	recall	precision	recall	precision
Bulk	81.74	80.91	84.26	85.67
Mail	80.17	79.52	85.12	86.45
Service	80.73	81.87	84.32	83.56
WWW	81.25	80.67	85.28	85.67
P2P	74.83	73.15	80.43	81.49
Others	78.34	77.32	81.15	82.78
Average	79.35	78.87	83.26	84.23

From Table 2, the accuracy of classification using FSVM-RFE was improved by 9.069% for P2P service. From the average accuracy of all services, the FSVM-RFE method has an accuracy of 6.248% higher than the SVM-RFE. The reason is that a small number of noise or outlier samples will affect the performance of the support vector machine. By introducing the fuzzy factor μ_i, the FSVM-RFE assigned less weight to noise or outlier samples according to the contribution of the classification, so that the noise or the outlier samples played only a small or no effect when constructing the optimal classification surface. Effectively weakened the impact of noise or outlier samples on classification accuracy. Therefore, the FSVM-RFE proposed in this paper has a better performance in noise immunity.

In order to avoid the influence of weak features and redundant features. In this section, the FSVM [9] was compared with FSVM-RFE. The result was shown in Table 3.

Table 3 Comparison of classification accuracy between FSVM and FSVM-RFE

Type	FSVM(%)		FSVM-RFE(%)	
	recall	precision	recall	precision
Bulk	79.72	80.91	84.26	85.67
Mail	82.47	82.35	85.12	86.45
Service	78.72	81.17	84.32	83.56
WWW	83.25	83.67	85.28	85.67
P2P	76.83	76.15	80.43	81.49
Others	78.54	77.82	81.15	82.78
Average	79.56	79.57	83.26	84.23

Table 3 shows that the FSVM-RFE accuracy increased over 13% than FSVM. From the average of all businesses, using FSVM-RFE is more precise than FSVM, and the accuracy improved about 5%. The reason is that the FSVM-RFE method applies the recursive feature elimination. This method can obtain the best combination of features for classification and make the strong feature attribute play a greater weight in the classification, so a better classification hyperplane in the feature space will be founded. Therefore, the FSVM-RFE proposed in this paper has a better performance in resisting weak correlation and the accuracy of classification.

Figure 1 shows that the receiver operating characteristic (ROC) [10] analysis obtained by the FSVM-RFE method. It reveals the balance between the True Positive and the False Positive, which can reflect the performance of the FSVM-RFE classifier. It can be concluded from Figure 1 that the FSVM-RFE works in the upper left corner of the ROC curve and has a steep rising curve. For businesses such as WWW, BULK, MEDIA and P2P, the TP is mostly maintained at 0.75~0.85, while the FP at 0.03~0.05, which is in an ideal range. It is indicated that the FSVM-RFE has a good stability and can accurately identify various services in the network.

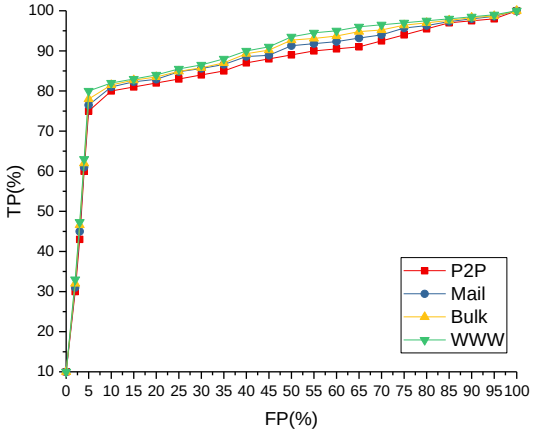

Fig.1 Receiver operating characteristic analysis

4. Conclusions

Network traffic classification technology plays an increasingly important role in network service and management. As Internet traffic becomes more complex and dynamic, more and more network applications use evasive means such as dynamic port, masquerading port and application layer encryption. It is a big challenge for identify the network traffic accuracy. Aiming at the large number of noise and redundant feature attributes in the network environment, this paper proposed a Fuzzy SVM-based recursive feature elimination (FSVM-RFE) and applied FSVM-RFE to network traffic classification. A large number of experimental results show that FSVM-RFE can improve the anti-noise ability of traditional SVM, and can eliminate the influence of redundant feature attributes on classification accuracy. In our future work, the network traffic classification method proposed in this paper needs to be further experimentally verified in the Internet environment. At the same time, how to improve the real-time performance of the network traffic classification method still needs further research.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (No. 61571241 and 61872423), the Ministry of Education-China Mobile Research Foundation, China (No. MCM20170205), Postgraduate Research & Practice Innovation Program of Jiangsu Province (KYCX18_0912), the Scientific Research Foundation of the Higher Education Institutions of Jiangsu Province, China (No. 17KJB510043 and 15KJA510002), Six talent peaks project in Jiangsu Province (DZXX-008), the Research Foundation for Advanced Talents, Nanjing University of Posts and Telecommunications (No. NY217146).

References

[1] J. Zhang, X. Chen, Y. Xiang, W. Zhou, J. Wu, “Robust network traffic classification,” IEEE/ACM Trans. Netw, vol. 23, no.4, 2015, 1257-1270.

[2] R. Raveendran, R. R Menon, “A Novel aggregated statistical feature based accurate classification for Internet traffic,” In Proc. of Data Mining and Advanced Computing (SAPIENCE), Ernakulam, India, 2016.

[3] A. W. Moore and K. Papagiannaki, “Toward the Accurate Identification of Network Applications,” Proc. PAM’05, 2015, 41-54.

[4] YANG Y X, WANG R, LIU Y, et al. Solving P2P traffic identification problems via optimized support vector machines [C]. Proc of IEEE / ACS International Conference on Computer System and Applications, Amman, Jordan, 2017, 165-171.

[5] WANG R, LIU Y, YANG Y X, et al. Solving the application classification problem of P2P traffic via optimized support vector machines [C]. Sixth International Conference on Intelligent

Systems Design and Applications, Jinan, China, 2006, 534-539.

[6] MCGREGOR A, HALL M, LORIER P, et al. Flow clustering using machine learning techniques [C]. Passive & Active Measurement Workshop, Antibes, France, 2004, 321-324.

[7] T. M. Cover and J. Thomas, Elements of Information. Hoboken, NJ, USA: Wiley. 2006.

[8] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intel., vol. 27, no. 8, 2005, 1226-1238.

[9] WANG Y Z, ZHANG H, JIANG S H, et al. Fuzzy support vector machine based on improved possibility c-means clustering algorithm in kernel space [J]. Journal of Information and Computational Science, 2010, 7(2), 585-591.

[10] WITTEN I H, FRANK E. DATA MINING: Practical machine learning tools and techniques [M]. New York: SF Morgan Kaufman, 2005, 158-171.