# A New Method of Improving the Traditional Traffic Identification and Accuracy

Wang Zhongsheng
School of Computer Science and Engineering
Xi'an Technological University
Xi'an, 710021, Shaanxi, China
e-mail: wzhsh1681@163.com

Gao Jiaqiong
Department of ComputerScience
SichuanVocational andTechnicalCollege
Suining, 629000, Sichuan, China
e-mail: 516719510@qq.com

***Abstract*—As the traffic generated by the increasing number of applications on the Internet is becoming more and more complex, how to improve the quality of service and security of the network is also increasingly important. This paper studies the application of Support Vector Machine (SVM) in traffic identification to classify network traffic. Through data collection and feature generation methods and network traffic feature screening methods, SVM is used as a classifier by using the generalization capability, and the parameters and the kernel functions of SVM are adjusted and selected based on cross comparison ideas and methods. Using the cross-validation method to make the most reasonable statistics for the classification and recognition accuracy of the adjusted support vector machine avoids the situation that the classification accuracy of the support vector machine is unstable or the statistics are inaccurate. Finally, a traffic classification and identification system based on SVM is implemented. The final recognition rate of encrypted traffic is up to 99.31%, which overcomes the disadvantages of traditional traffic identification and achieves a fairly reliable accuracy.**

***Keywords-Support Vector Machine (SVM); Traffic Classification; Feature Extraction; Kernel Function***

## I. INTRODUCTION

Due to the rapid development of the Internet, Internet business has greatly facilitated and enriched people's lives, learning and work, and has attracted more and more users. With the new application patterns (such as P2P) and application demand emerging in the Internet[1], the pressure of huge data transmission is becoming more and more heavy, and the occurrence of network failures is becoming more and more frequent, which leads to a series of network failures, such as packet loss, network congestion, and time delay in the process of data transmission. The maneuverability of the network is greatly reduced, the normal operation of the network is affected, and huge economic losses are incurred. Therefore, how to identify and classify the network traffic in real time helps the Internet service provider to understand the network operation status and optimize the network operation and management. It is of great significance.

The current popular network traffic identification technologies include traffic identification algorithms based on known ports [2]; traffic identification based on Deep Packet (DPI) [3-5]; traffic identification algorithms based on data flow behavior pattern [6]; traffic identification algorithms based on machine learning and so on.

The traffic classification method [7] has been widely proposed in the past few years. Initially, the type of data transmitted over the Internet is relatively small. The traffic identification technology is mainly based on port identification. That is, the general network protocol port number [8] is used to roughly classify traffic. For example, the protocol uses a fixed port. However, with the development of the Internet, merely relying on port identification technology has been insufficient to distinguish between more and more network applications and protocols. In 2004, an application layer load signature recognition method, the DPI technology [9], was proposed to extract the data message samples and determine whether the traffic belongs to the application by matching the signature of the unknown traffic. In recent years, the proportion of network traffic transmitted by encrypted text is increasing. DPI

DOI: 10.21307/ijanmc-2019-006

technology has been powerless for this part of the traffic. At present, the method of network traffic recognition based on machine learning [9] shows a higher accuracy.

Machine learning is an important tool for the study of network traffic identification. Dong S and others described the current popular machine learning method [10]. After comparing and evaluating the clustering algorithm [11], it was found that the feature selection algorithm [12] was better for supervised machine learning[13,14]. DBSCAN algorithm [15] of unsupervised clustering algorithm has higher precision.

Since the development of a complete classification architecture [16] for real-time work on high-capacity links is limited, Este A[18] and others after demonstrating the computational time and the optimization steps required to handle different traffic traces, used machine learning techniques(SVM model[17]) to improve system performance and enable real-time traffic identification for high-speed networks. Zhao X proposed a P2P network traffic classification method based on support vector machine [19], using a statistical principle to divide the network traffic of four different types of P2P traffic applications (file sharing BitTorrent, media streaming PPLive, Internet phone Skype, instant messaging MSN), and studied network traffic statistics and SVM methods. The overall framework of P2P traffic classification based on SVM was introduced, how to obtain traffic samples and processing methods were described, and the traffic classifier was constructed, with an average accuracy of 92.38%.

Bernaille L and others divided the traffic classification mechanism into two phases [20]: offline learning and online classification. The offline learning stage uses the kMeans method [21,22,23] to divide the original traffic and give a description of each cluster and its application type; the online learning stage determines the application type of the new traffic according to the learning knowledge.

Ye M proposed a new method of identifying P2P traffic through data transmission behavior of P2P applications [24]. The data downloaded from the P2P host finds the shared data of the download stream and the online upload stream, and proposes a content-based partitioning scheme to divide the stream into data blocks.

Based on the above viewpoints and taking into account the excellent performance of machine learning and SVM in solving P2P traffic classification problems [25-29], this paper proposes a network traffic two classification method based on SVM. which is used to complete the network flow parameters obtained from the packet header after network traffic collection to classify Internet traffic into a wide range of application categories. In the selection process of feature vectors, it should be suitable for SVM algorithm and try to calculate independently of the protocol and port. Therefore, in this paper, we choose the number of packets, size characteristics, data flow time characteristics, flag bits and other information as a preliminary feature vector, through a plurality of classifier selection methods to obtain the optimized feature set. It is used to implement the initial identification of normal traffic in the network, reducing the workload of the feature value matching module, improving the efficiency of the network traffic identification system, and comparing with the method of identifying network traffic that only adopts the feature value matching. The experimental results using the traffic from the campus backbone network show that 99.31% accuracy can be achieved through regular biased training and test samples. When using bias-free training and test samples of the same feature set, an accuracy of 96.12% can be achieved. In addition, since all feature parameters can be calculated from the packet header, the proposed method is also applicable to encrypted network traffic.

## II. PROPOSED METHOD

### A. Support Vector Machine (SVM) model

SVM is a machine learning method that is based on one of the statistical algorithms with good generalization ability. It is mainly used to solve small samples. The feature vectors of the data stream in the network are more or less, and too many features will affect the efficiency and accuracy of the SVM algorithm. Therefore, to reduce redundant features, feature combinations with high discrimination are selected as feature vectors. After completing the support vector machine network traffic classification identification code, statistics and evaluation of the operating efficiency and accuracy of the results are also required.

The identification of network traffic is essentially a pattern classification process and is mainly divided into the following three points:

*1)* Converting the actual problem into the high-dimensional feature space through the kernel function, so that in the high-level space, the hyperplane can be used to classify the data, and the classification decision function is constructed so that the nonlinear problem of the original dimension is converted into linear separable problem. The classification decision function is a linear combination of non-linear functions with support vectors as parameters. The classification function itself is only related to the number of support vectors, so the method of this kind of kernel function is very effective in dealing with the classification problem of high dimensional feature space.

*2)* Under the condition that the number of known training samples is small, the network traffic classification is converted into secondary optimization and improve the accuracy of classification. The initial threshold is determined by iterating feature subsets using the inter-class distances and intra-class distances of the features.

*3)* The optimization problem is coded by simulating the natural evolutionary process. The key point of coding is that the code must be able to represent all possible subsets of the feature set. The optimal hyperplane is used to optimize the learning ability of the classifier. This method does not need to rely on the prior probability of the network traffic samples and has better generalization.

When using SVM, classifiers with better generalization effects can be achieved by defining different kernel functions and relaxation factors. The optimization model is as follows:

Let the training sample set be: {（xi,yi）}，i=1，2，3，…n; map this sample set to the high-dimensional feature space and achieve regression, the following are obtained:

$$fx\omega T\emptyset x+b \qquad (1)$$

(ω is the weight vector; b is the offset vector)

Convert equation (1) to the minimization problem. The objective function of SVM regression is:

$$\min\omega2 +12Ci=1n\delta i2$$

s.t.

$$yi-\omega Tx+b=eii=1,2,3,\ldots n \qquad (2)$$

In this formula, C is the penalty parameter; ei is the regression error. Through the Lagrangian operator, the corresponding dual problem is obtained as follows:

$$L\omega,b,\delta,\alpha=\min|\omega|2+12\ \gamma i=1n\delta i2+i=1n\alpha i(\omega Tx-b+ei-yi) \qquad (3)$$

Set the kernel function Kxi,xj=(xi)T(xj), then use the nonlinear SVM regression model established by the RBF function. There are:

$$fx=i=1N\alpha iexp-xi-xj2\sigma22+b \qquad (4)$$

(σ is the width of the core)

## B. *Finding support vectors in training samples*

Introduce the following rules to distinguish. Set the threshold of the support vector decision function λ=1 or λ=-1, Assume that the decision function in the detection process isfx=sgn{i=1n[aiKxi,x-sidi§]}, f (x)≠1 or f(x)≠－1, The x vector does not belong to the support vector or the x vector belongs to the support vector.

An initial support vector library trained from known flows. After the known flow rate is trained by the data acquisition module, the feature extraction module, the data preprocessing module, and the training module, a support vector is generated to perform feature analysis, and its characteristic word information is added to the support vector library. Various known P2P traffic passes through the above process eventually forms a multidimensional support vector group, and a known support vector library is also formed. Finally, the MSVM threshold is determined. If the threshold is equal to 1 (or -1), the detected network traffic is P2P traffic; otherwise, the detected network traffic is non-P2P traffic.

When selecting P2P traffic characteristics, the feature extraction should be able to reflect the difference of P2P traffic as much as possible. Different nodes in the network have different functions: Some nodes function as servers and provide resource transmission services to other nodes in the network. Some nodes function as clients and receive various services provided by the server. The nodes in the P2P network can serve as servers to other peer nodes, and can also serve as clients to receive services provided by other peer nodes. Therefore, node traffic with different functions and providing different services presents different behavior characteristics.

### C. *Support vector machine network traffic identification process*

The network traffic identification based on vector machine is essentially making full use of the powerful capability of SVM to deal with non-linear multi-factor system to mine the internal rules and establish the complex non-linear relationship of network traffic change, so as to achieve accurate network traffic prediction. In the learning and classification process of the SVM model, the selection of kernel functions plays a decisive role in the training and classification performance. At present, several frequently studied kernel functions are: linear kernel, RBF(radical basis function) kernel and Gaussian kernel and so on. In this paper, RBF kernel is selected as the kernel function.

The overall strategy when selecting the kernel function and adjusting parameters is approximately the following steps: preparing a batch of classified data; splitting the data into two groups: a training group and a test group; using a training group to give a support vector machine for training and learning; The support vector machine predicts the classification of test group data and compares it with the actual classification of the number of test groups, calculates the classification accuracy, replaces the parameters, and then iterates again. If we do not use the cross comparison idea, it is very easy to cause the prediction result to be very good only in the case of a specific input. In other cases, the prediction of the parameter is not stable.

### D. *P2P traffic classification model based on SVM*

Figure 1 shows the classification framework based on SVM in this paper. This paper firstly extracts and analyzes the traffic to extract several main characteristics of network traffic that are suitable for recognition in the support vector machine. Then, the data is preprocessed, and the known data set for the target problem is set as a training data set, and use an iterative process to train a classification model. The parameters of the model are continuously adjusted by a method of random optimization or analysis, so that it is closest to the actual situation of the training data set. After the model is trained, it can be used to identify unknown samples and dynamically adjust the training sample data by continuously searching for useful training samples to realize the entire network traffic identification based on SVM.

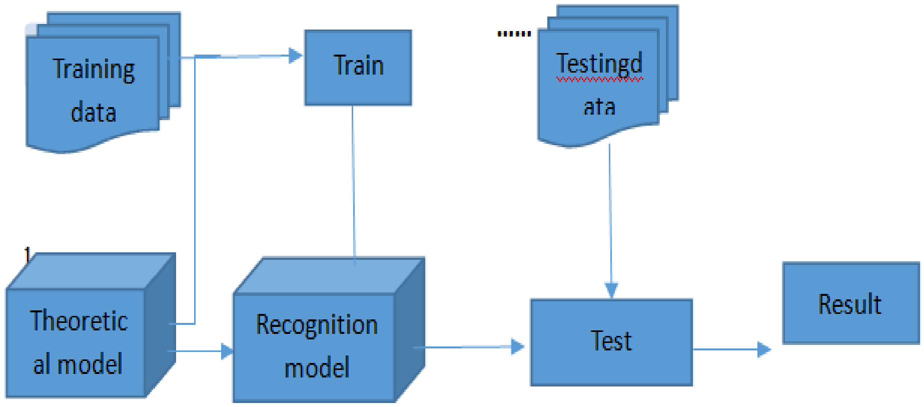

Figure 1. Classification framework based on SVM

Theoretical model

*1)* Collector: Using port mirroring method to collect data from routers and collect data as raw data and preprocess them. Multiple harvesters can be connected in parallel or in series.

*2)* Analyzer: The raw data preprocessed by the collector is subjected to a data feature extraction module to extract the characteristic function parameters. Stored in data warehouse. An analyzer can analyze the data of multiple collectors. After the data is preprocessed, the grid search method can be used to verify the optimal parameters of the RBF kernel function for the training data set. So that the analyzer can accurately predict unknown data.

*3)* After the optimal parameters are determined, the training data set can be trained to obtain the support vector machine model. The extracted parameter data is taken as the feature value of the original data, and the continuous features and discrete features existing in the data are converted, and these heterogeneous data sets are translated into machine-readable values by the data preprocessing module.

*4)* Multidimensional support vectors are generated by the data after SVM training. At the same time, the multidimensional support vectors are formed through the process of different P2P traffic data, and one support vector library is formed.

*5)* Known P2P traffic can get specific P2P type through SVM library. Unknown P2P traffic will be subjected to data preprocessing and SVM training by the data acquisition device and analyzer extraction feature extraction module, and the extracted feature information will be added to the SVM support vector library. After obtaining the specific name of the traffic, it is put into the SVM support vector library and finally identifies the specific P2P traffic.

The initial SVM support vector library is a vector library that is trained by known traffic. When the known traffic is subjected to initial data acquisition and feature extraction, data preprocessing, and SVM training, multidimensional support vectors are generated, multidimensional support vectors are characterized, and their characteristic information is added to the SVM support vector library. Known traffic can also form a multidimensional support vector group through the above process.

## III. EXPERIMENTS

### A. *Traffic data collection*

Select a network server outlet network traffic to carry on the simulation experiment, take 10ms as the sampling time, select the total number of data packets, uplink traffic ratio, average length, TCP traffic ratio and the ratio of the number of connections and different IP number five traffic characteristics as input data feature information, set up the data set as a training sample set and separate and collate, and preprocess the collected data and normalize it. The collected data samples are shown in Table 1.

TABLE I. COLLECTED DATA SAMPLES

| DataSet | Time (ms) | Total flow |
|---|---|---|
| DataSetA | 1hour | 2300 |
| DataSetB | 1 hour | 3020 |
| DataSetC | 1 hour | 1831 |
| DataSetD | 1 hour | 2290 |
| DataSetE | 1 day | 9538 |

Among them, the first four sets of data are used as input data for the training module. DataSetE is used as the data set to be tested. Three support vector machines are constructed here, namely SVM1, SVM2, and SVM3. After training the classifiers SVM1, SVM2, and SVM3, DataSetE was used as the test sample data set, and experimental results were obtained through the SVM classifier.

### B. *Finding optimal parameters*

The algorithm based on the cross-validation idea is used to select an optimal parameter value C for the RBF kernel function and optimal parameters C and R for the training data set. The labels of the two categories are -1 and 1, which are iterated 51 times. The trained model is saved in the data.Model file. The following information can be obtained from this file: The svm type used for training is c_svm, the kernel function is the radial basis function RBF, the R value is 0.5, the total number of support vectors is 43, and the value of the decision function constant term B is 0.421. Each type of support vector is 22, 20, 21. After the training is completed, the model can be used for SVM type prediction.

Read the file to be predicted, the model file, and then call the function prediction and output the result to a file.

*1)* After cross test the data, the prediction accuracy is 99.31%.

*2)* When choose the best parameters (C, R), If the cross validation method of grid search is not adopted, the result of cross validation is not adopted with the default value of 1. According to the method described above, the prediction accuracy is 93.31% obtained by predicting the unknown data through the obtained model. It can be seen that the choice of optimal parameters (C, R) can improve the prediction accuracy of the results.

*3)* Repeated training and learning. In order to reflect the learning process of SVM, a total of 10 experiments were conducted, by continuously capturing data, the captured data are preprocessed, trained, and predicted. With continuous learning, the accuracy of predictions continues to increase, reaching 91.12%, 93.42%, 94.67%, 95.34%, 95.56%, 96.78%, 97.12%, 97.23%, 97.31% and 97.65% respectively. It can be seen that multiple learning is conducive to classification judgment. However, the learning process also needs to be controlled. Excessive learning will bring negative effects on classification.

## IV. Discussion

The model obtained after training can be used for SVM traffic identification. Various P2P traffic and accuracy are identified from packet capture, preprocessing, recognition, learning and training, and compared with the recognition accuracy based on the Bayesian traffic identification model, the recognition method of the SVM has obtained higher accuracy than the original traffic recognition method in practical application.Figure 2 shows the comparison of different traffic models.

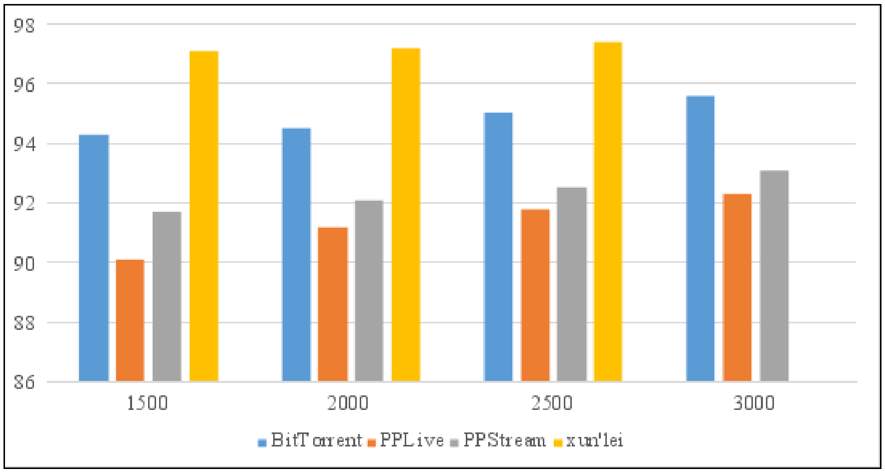

Figure 2. Comparison of different models

From Figure 2, we can see that for the four kinds of P2P traffic in this experiment, the classification and recognition rate of this classifier is all above 90%, so the effect of this MC-SVM classifier on application layer classification of P2P traffic is very good.

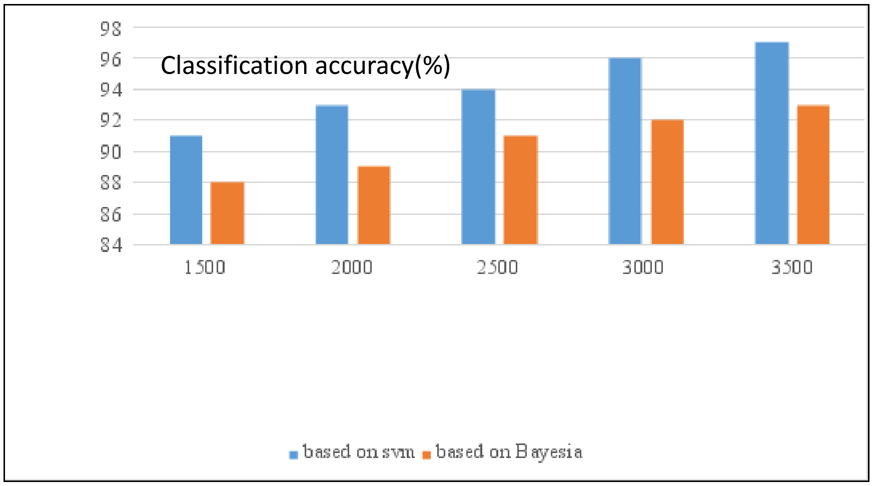

Figure 3. Comparison of stability between Bayesian and SVM

Figure 3 is by using a P2P traffic recognition model based on Bayesian and SVM. With the increase of training data sets, the average classification accuracy can still maintain a certain stability, and the accuracy of recognition reaches 97. 8%. It can be seen that the recognition method of SVM has higher accuracy than the original traffic recognition method in practical application.

## V. CONCLUSIONS

SVM algorithm is suitable for nonlinear time series modeling and prediction, so it can well identify the trend of network traffic changes. This paper conducts empirical experiments on the actual data of network traffic. The results show that, compared with the commonly used prediction methods, the recognition model based on SVM can solve the traffic identification. At the same time, it can identify the unknown and large traffic P2P types, and has good effect on the identification of encrypted P2P traffic, and has higher prediction accuracy and better adaptability.

## ACKNOWLEDGEMENTS

Fund support: National natural science foundation, ResearchonSub-NyquistSamplingofshortpulsesbasedXamplingunderGaborFrames, [61501493]; Shaanxi Education Department Special Fund, Projectnumber: Shaanxi Education Finance[2013] 23.

## REFERENCES

[1] Schulze H, Mochalski K. Internet study 2007[J]. Ipoque Gmbh, 2007.

[2] Madhukar A, Williamson C. A Longitudinal Study of P2P Traffic Classification[C]// IEEE International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems. IEEE, 2006:179-188.

[3] Ma J, Levchenko K, Kreibich C, et al. Unexpected means of protocol inference[C]// ACM SIGCOMM Conference on Internet Measurement. ACM, 2006:313-326.

[4] Moore A W, Papagiannaki K. Toward the accurate identification of network applications[C]// International Conference on Passive and Active Network Measurement. Springer-Verlag, 2005:41-54.

[5] Haffner P, Sen S, Spatscheck O, et al. ACAS: automated construction of application signatures[C]// ACM SIGCOMM Workshop on Mining Network Data. ACM, 2005:197-202.

[6] Huang K, Zhang Q, Zhou C, et al. An Efficient Intrusion Detection Approach for Visual Sensor Networks Based on Traffic Pattern Learning[J]. IEEE Transactions on Systems Man & Cybernetics Systems, 2017, PP(99):1-10.

[7] Yuan R, Li Z, Guan X, et al. An SVM-based machine learning method for accurate internet traffic classification[J]. Information Systems Frontiers, 2010, 12(2):149-156.

[8] I ANA. Internet assigned numbers authority[EB／OL]. http : //www.iana.org/assigu mens/port-numbers.

[9] Spatscheck O, Sen S, Wang D. Method and apparatus for automatically constructing application signatures: US, US 7620807 B1[P]. 2009.

[10] Yang C S, Liao M Y, Luo M Y, et al. A Network Management System Based on DPI[C]// International Conference on Network-Based Information Systems. IEEE Computer Society, 2010:385-388.

[11] Hartigan J A, Wong M A. Algorithm AS 136: A K-Means Clustering Algorithm[J]. Journal of the Royal Statistical Society, 1979, 28(1):100-108.

[12] Liu H, Yu L. Yu, L.: Toward Integrating Feature Selection Algorithm for Classification and Clustering. IEEE Transaction on Knowledge and Data Engineering 17(4), 491-502[J]. IEEE Transactions on Knowledge & Data Engineering, 2005, 17(4):491-502.

[13] Pedro S D S. Collective intelligence as a source for machine learning self-supervision[C]// International Workshop on Web Intelligence & Communities. ACM, 2012:5.

[14] Chapelle O. Semi-supervised Learning (Adaptive Computation and Machine Learning)[J]. Mit Pr, 2006.

[15] Liu S, Dou Z T, Li F, et al. A new ant colony clustering algorithm based on DBSCAN[C]// International Conference on Machine Learning and Cybernetics. IEEE, 2004:1491-1496 vol.3.

[16] Este A, Gringoli F, Salgarelli L. On-line SVM traffic classification[C]// Wireless Communications and Mobile Computing Conference. IEEE, 2011:1778-1783.

[17] Osuna E, Freund R, Girosi F. Training svm: An application to face detection[C]// 1997.

[18] Este A, Gringoli F, Salgarelli L. On-line SVM traffic classification[C]// Wireless Communications and Mobile Computing Conference. IEEE, 2011:1778-1783.

[19] Zhou X. A P2P Traffic Classification Method Based on SVM[C]// International Symposium on Computer Science and Computational Technology. IEEE Computer Society, 2008:53-57.

[20] Bernaille L, Teixeira R, Akodkenou I, et al. Traffic classification on the fly[J]. Acm Sigcomm Computer Communication Review, 2006, 36(2):23-26.

[21] JinHuaXu, HongLiu. Web User Clustering Analysis based on KMeans Algorithm[C]// 2010 international conference on information,networking and automation. 2010:V2-6-V2-9.

[22] Poornalatha G, Raghavendra P S. Web User Session Clustering Using Modified K-Means Algorithm[M]// Advances in Computing and Communications. Springer Berlin Heidelberg, 2011:243-252.

[23] Wang T Z. The Development of Web Log Mining Based on Improve-K-Means Clustering Analysis[M]// Advances in Computer Science and Information Engineering. Springer Berlin Heidelberg, 2012:613-618.

[24] Ye M, Wu J, Xu K, et al. Identify P2P Traffic by Inspecting Data Transfer Behaviour[J]. Computer Communications, 2010, 33(10):1141-1150.

[25] Tapaswi S, Gupta A S. Flow-Based P2P Network Traffic Classification Using Machine Learning[C]// International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery. IEEE, 2013:402-406.

[26] Deng H, Yang A M. P2P traffic classification method based on SVM[J]. Computer Engineering & Applications, 2006.

[27] Yang A M, Jiang S Y, Deng H. A P2P Network Traffic Classification Method Using SVM[C]// Young Computer Scientists, 2008. Icycs 2008. the, International Conference for. IEEE, 2008:398-403.

[28] Jiang W, Wang C Z, Luo H F, et al. Research on a Method of P2P Traffic Detection Based on SVM[J]. Journal of Hubei University of Technology, 2010.

[29] Zhu A. A P2P Network Traffic Classification Method Based on C4.5 Decision Tree Algorithm[M]// Proceedings of the 9th International Symposium on Linear Drives for Industry Applications, Volume 4. Springer Berlin Heidelberg, 2014:373-379.