International Journal of Future Generation Communication and Networking
Vol. 9, No. 12 (2016), pp. 155-164
http://dx.doi.org/10.14257/ijfgcn.2016.9.12.14

P2P Traffic Detection Based on Particle Swarm Optimization Algorithm

Song Rong and Li Xiating

Jiangxi Vocational and Technical College of Communication, Nanchang, Jiangxi, China, 330013
39737728@qq.com, 81838225@qq.com

Abstract

Considering the shortcomings of the conventional BP neural network, such as slow learning speed, weak anti-interference ability and easy to fall into local minimum, the detection accuracy of P2P traffic detection model is low and the speed is slow,the particle swarm optimization algorithm is used to optimize it here. As the conventional algorithm's optimization ability is the initial parameters, the algorithm is easy to be early, and the convergence speed is slow. Therefore, grouping, organizing, fission and mutation operation on the conventional algorithm have been carried on in order to improve the defect of conventional algorithm. Finally, the P2P traffic detection model is built by using MATLAB software, and traffic detection experiments are carried out on Bittorrent, EMule, PPlive and PPStream 4 P2P network applications. The test data show that the average recognition rate of the recognition model is 96.14%, which is 13.3% higher than that of the conventional PSO-BP model, and9.4% higher than that of the QPSO-BP recognition model for the four P2P network applications.

Keywords*: P2P technique; flow detection; artificial neural network; improved particle swarm optimization algorithm*

1. Introduction

In recent years, the technology of P2P (peer to peer) has been developed rapidly, and it has been widely used in fields of file sharing, streaming media real-time communication and so on. The download amount of these video and files with the technology of P2P has increased notably, and this had greatly enriched the content of the Internet. According to the statistics, from 2008 to 2009, P2P traffic in Eastern Europe has more than 70% of the total network traffic. However, P2P technology brings convenience, but also brings some problem at the same time. The ratio of the rapid increase of the flow, the upload and download is roughly symmetrical to each other. This makes the congestion phenomenon of network is increasingly serious. And it also increases the burden on the network, and reduces the network's quality of service. Apart from these, a large number of malicious traffic, viruses and Trojan horses will swoop in because of the open structure of P2P, this will increase the security risks of network. So, research on the traffic detection and identification based on P2P has become a hot issue nowadays [1-2].

Presently, for P2P traffic identification and classification techniques mainly include these types: based on the port identification, based on the deep packet inspection (DPI), based on machine learning algorithm, based on network behavior and so on.

In literature [3], it uses the port identification technology to realize the detection and classification of P2P traffic. In the past, the number of the port used by P2P is relatively fixed, so we can use the port identification to detect and classify the P2P traffic. However, now, the P2P applications will use camouflage port to prevent the detection of their traffic. So this technology can't adapt to the current network environment already.

In literature [4], it uses the DPI to realize the detection and classification of P2P traffic.

ISSN: 2233-7857 IJFGCN

DPI has high recognition accuracy, and it also has a good stability and reliability of the identification. But, usually it is used to detect the non-encrypted traffic, to these encrypted traffic, the accuracy of detection is only between 30% and 70%. Meanwhile, because of the technology of DPI will threat the use's privacy, so its application and development has been restricted.

In literature [5], it realizes the P2P traffic detection and classification based on the technology of characteristics of network behavior. This technology is mainly used to identify the peers of P2P network, and not depend on the load information of the application layer. However, because it is only capable for the identification of degree and it needs a long time for network behavior extracting, so this technology is not suitable for the detection of current P2P traffic detection with high-speed network technology.

Machine Learning Based P2P traffic identification and classification technology is developed based on machine learning algorithms. Its detection performance is generally not dependent on the load on the application layer, but on the performance of the machine learning algorithms. It uses the traffic statistics feature to establish the recognition model.

In literature [6], it uses the K-means clustering algorithm to detect the protocol traffic of eDonkey, Kazaa, *etc*… It has high recognition accuracy. In literature [7], it uses the Naïve Bayesian classification algorithm to detect the P2P traffic, in literature [8] and literature [9], they use the SVM algorithm and C4.5 algorithms to detect the traffic of P2P, and their detection performance is superior to the use of the detection model established by Naïve Bayes classification algorithm.

This paper uses the BP neural network algorithm to establish the P2P traffic identification model, and the usesthe improved particle swarm optimization algorithm to improve BP neural network in order to avoid the shortcoming of falling into local minimum.

Multilayer feedforward BP network is a kind of neural network used most frequently at present. It has the advantages of general neural network, but it is still not very perfect. Therefore, in order to better understand the application of neural network to solve problems, some discussions are made on its advantages and disadvantages.

Firstly, BP neural network has the following advantages:

1) Nonlinear mapping ability: in fact, BP neural network has achieved the mapping function from the input to the output, and the mathematical theories have proved that the three-layer neural network can approximate any nonlinear continuous function with arbitrary precision, which makes it particularly suitable for solving complex problems of internal mechanism. That is to say, BP neural network has strong nonlinear mapping ability.

2) Self-learning and adaptive ability: in training, BP neural network can automatically extract the "rule of reason" between input data and output data by learning, and memorize the learning contents on the network weights. In other words, BP neural network has high self-learning and adaptive ability.

3) Generalization ability: the so-called generalization ability means that in the process of designing a pattern classifier, it is necessary to consider whether the network has classified the objects in a right way, and whether after training, the network can correctly classify those models either have never been seen or have noise pollution. Consequently, BP neural network has the ability to apply the study results to new knowledge.

4) Fault tolerant ability: BP neural network will not cause a great impact on the whole training when local or part neurons are damaged. And the system can still work normally under local injury. Therefore, BP neural network has a certain degree of fault tolerance ability.

In view of the advantages of BP neural network, a lot of domestic and foreign scholars have conducted researches on it, but with the gradual expansion of application scope, BP neural network has exposed more and more disadvantages and shortcomings. In this thesis, the problems mainly include:

1) The local minimization problem: from the mathematical perspective, traditional BP neural network is an optimization method for local search, which is to solve a complex nonlinear problem. The weights of the network are adjusted gradually along with local improvement, which will make the algorithm fall into local extremum, the weights converging to the local minimum, thus lead to the failure of network training. Since BP neural network is very sensitive to initial weights, adopting different weights to initialize network will often make it converge to different local minima, which is the primary reason why many scholars obtain different outcomes in the trainings

2) Low convergence speed of the BP neural network algorithm: since the BP neural network algorithm is essentially a gradient descent method, the objective function it needs to optimize is very complex, therefore, the "saw-tooth phenomenon" will appear inevitably, which makes the BP algorithm inefficient. In addition, as the objective function to be optimized is very sophisticated, it will have some flat areas when the output of neurons is getting close to 0 or 1. In these areas, the change of weight error is so small that the training process is almost stopped. In BP neural network model, in order to make the network carry out the BP algorithm, we cannot use the traditional one dimension search method to figure out the length of each iteration step, but must input the updating rule of step length on the network in advance. However, this method can also lead to inefficient algorithm. All of these are the causes of low convergence speed of the BP neural network algorithm.

Improved particle swarm optimization:Improved particle swarm optimization is an optimization algorithm based on the simulation of birds' swarm intelligence to solve the optimization problems of continuous variables. Because of its simple concept, few parameters and easy implement, it has been paid much attention by researchers at home and is widely used in many fields since being put forward.

2. Improve BP Neural Network

2.1. BP Neural Network

BP neural network algorithm has got a very mature and extensive development, apply the BP neural network algorithm into P2P traffic identification technology can effectively improve the recognition rate and recognition speed of the P2P traffic recognition system.

The P2P traffic identification and detection method based on BP neural network usually train the detection system with a large number data flow characteristics sample, whose class in known already. This can make the system has a strong generalization ability. The classification training process of the P2P traffic detection is shown in the Figure 1.

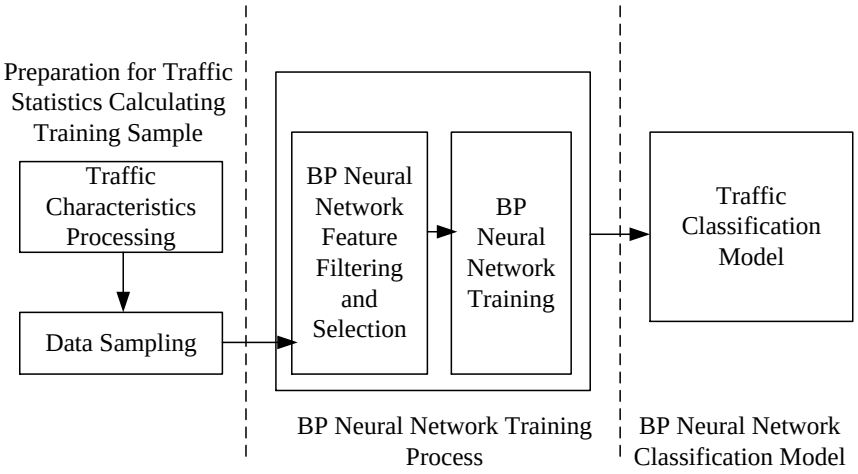


Figure 1. P2P Traffic Identification and Classification Training Process Based on BP Neural Network Algorithm

Actually, BP neural network is an iterative learning method of gradient descent algorithm. Due to stable learning gradient descent algorithm requires a small learning speed, therefore, the convergence rate of this method is very slow. Further, since when the BP neural network is in training, it will gradual close to the extreme value of the error along the error ramp at some point. Thus, different starting point will get different error extreme value and different solution. So the traditional BP neural network has the disadvantages of learning slow, weak anti-interference ability and easy to fall into local minimum value. Usually it requires using the genetic algorithm, particle swarm algorithm to optimize it. This paper uses an improved particles swarm optimization algorithm for the optimization of the general BP neural network algorithm.

2.2. Improved Particle Swarm Optimization (PSO)

By initializing a random swarm of particles, and in the later iterative process, and in the later iterative process, then enable each particle in the swarm search for the optimal location and the optimal particle in order to get the optimal solution, this is the Particle Swarm Optimization. Update the particle speed and position according to the method below:

$$v_i^{k+1} = wv_i^k + c_1 r_1 \left(p_{best} - z_i^k \right) + c_2 r_2 \left(g_{best} - z_i^k \right) \quad (1)$$

$$z_i^{k+1} = z_i^k + v_i^{k+1} \quad (2)$$

Where r_1 and r_2 is the random number between 0 and 1; c_1 and c_2 is the accelerating factor; k is the iterations of Particle Swarm Optimization. p_{best} is the optimal solution of the particle; g_{best} is the optimal particle among the swarm; v_i^k and z_i^k represent for the speed and position for the number i^{th} particle in the k^{th} iteration; w is the inertia weight factor[10].

Due to the conventional particle swarm optimization algorithm is easy to premature, and its convergence rate is very slow, this is because of that the conventional algorithm's optimization capability is determined by the initial parameter. So this paper adopts the operation of grouping, organizing, fission and mutation to overcome the shortcoming of the conventional algorithm.

The problems of premature and slow convergence rate can be partly overcome by grouping and organizing. In the paper below, I will introduce the operation of grouping and organizing.

In the conventional particle swarm optimization, each particle of the swarm tries to close to the global optimal solution, that is searching for the global optimal solution, so it is easy to fall into the problem of local optimal solution. After the operation of grouping, the ability of searching for the local optimal solution of the particles in the swarm will be reduced, so after the operation of grouping, we need to do organizing. So the improved algorithm will obtain the ability of searching for the global and local optimal solution.

This paper uses the method of extracting to do grouping, the principle of extracting is shown in the Figure. 2. When grouping, firstly, sort the particles of the swarm in accordance with the manner of fitness ascending, and divide the N particles into M subgroups. Where N and M must with a relationship of integer multiple. Extract a particle from each of the M subgroups, and then construct a subgroup with these N / M particles.

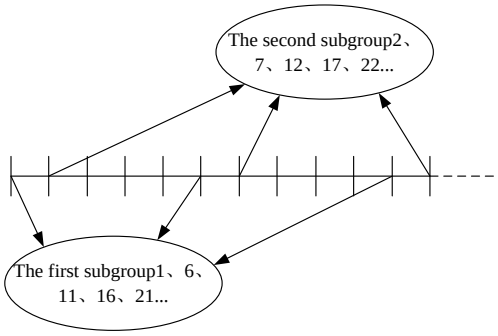


Figure 2. Extraction Method Clustering Principle

Use the method of extracting can ensure both the particles of excellent and poor performance are included in these subgroups, and these particles have a uniform distribution in each subgroup. Use this method can ensure the diversity of each subgroup and has a positive effect to improve the global search ability.

In order to improve the local search ability of the algorithm, to do the operation of organizing is necessary. Organizing is to combine the subgroups together. Take the optimal solution get by the operation of grouping as the initial optimal solution of the operation of organizing iteration [11].

Do the operation of fission and mutation is in order to improve the particle swarm optimal algorithm's ability of searching for the global optimal solution, and avoid the reduction of algorithm's convergence speed and accuracy.

The purpose of fission is to initialize the particles that have a poor performance, and then improve the diversity of the entire swarm. Enlarge the searching area and prevent the algorithm falling into the local optimal solution. Meanwhile, because of the operation of fission is not to initialize all of the particles, so the main structure of the current swarm will be guaranteed. In order to avoid a decrease in the algorithm convergence speed and accuracy, fission will be done only in the first half process of iteration. In the first half process of iteration, if within a certain number of times, the optimal solution obtained substantially the same, then sort the swarm in accordance to fitness, and initialize these particles of poor performance; while do grouping iteration, get the initializing particles within the range randomly, while do organizing iteration, get the initializing particles within the range of grouping optimal solution [12].

Mutation will be done only in the second half of iteration, by searching in the area nearby to avoid the optimal solution in the neighboring area, and this can improve the accuracy of the searching. In the second half of iteration, if within a certain number of times, the optimal solutions we can obtain is substantially the same, then sort the swarm in accordance to fitness, and do mutation to these particles of poor performance, the detail method is shown like below:

$$z_{id} = \begin{cases} \begin{cases} z_{id} + 1, r_{id}^2 < \rho_2 \\ z_{id} - 1, r_{id}^2 \geq \rho_2 \end{cases}, r_{id}^1 < \rho_1 \\ z_{id} \qquad\qquad\quad , r_{id}^1 \geq \rho_1 \end{cases} \tag{3}$$

Where ρ_1 and ρ_2 is the value of mutation rate; r_{id}^1 and r_{id}^2 is the random number corresponding to z_{id}, and their range is between 0 and 1; z_{id} is the variable in d^{th}dimension of the i^{th} particle.

Update the speed of grouping iterating operation by the following process [13]:

$$\begin{aligned} v_i^{k+1} &= w v_i^k + c_1 r_1 \left(p_{best} - z_i^k \right) \\ &+ c_2 r_2 \left(l_{best} - z_i^k \right) + c_3 r_3 \left(g_{best} - z_i^k \right) \end{aligned} \tag{4}$$

Where l_{best} is the optimal solution of the subgroup after grouping.

Update the weighting factor in the operation of grouping iteration by the following method:

$$w = w_0 - \frac{0.8(k-1)}{(T_1 + T_2 - 1)} \tag{5}$$

Where T_1 is algebra of subgroup's iteration; T_2 is the algebra of organized-group's iteration; w_0 is the initial inertia weight factor.

Use the method below, we can update the inertia weight factor in the operation of organizing iteration:

$$w = w_0 - \frac{0.8(T_1 + k - 1)}{(T_1 + T_2 - 1)} \tag{6}$$

And the initial speed of the particles can be generated by the method below:

$$v_i^1 = \text{round} \begin{bmatrix} 2rand(1, D) \cdot \mathbf{V}\max(1:D) \\ -\mathbf{V}\max(1:D) \end{bmatrix} \tag{7}$$

Where $\text{round}(\ )$ is a function used for round to the nearest integer; v_i^1 is the initial speed of the i^{th} particle; $rand(1, D)$ is applied for a vector of D-Dimension, and each of the element in this vector is a random number within (0,1); $\mathbf{V}\max(1:D)$ is the vector with the highest speed.

The process to improve the conventional particle swarm optimization is shown in the Figure 3 below [14]:

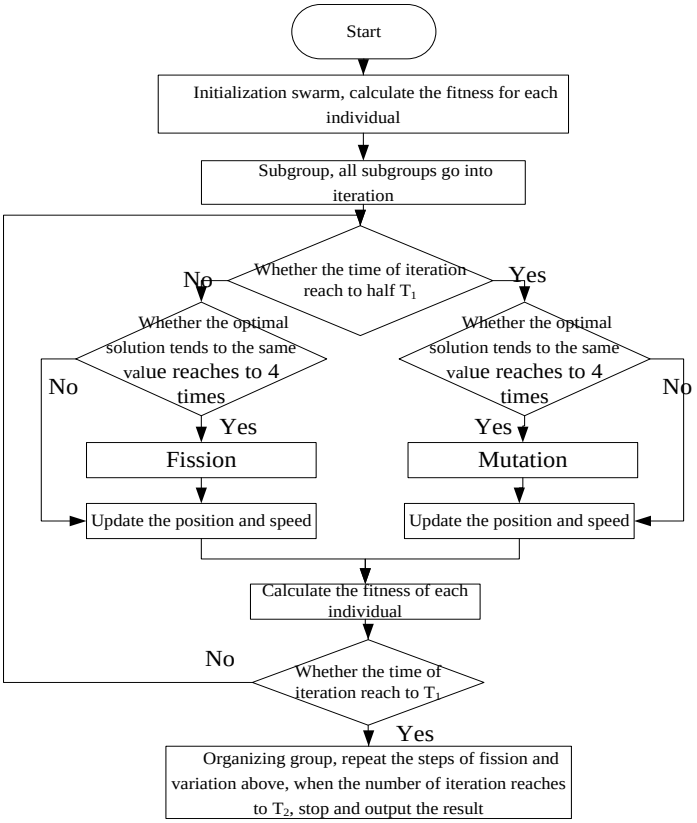


Figure 3. Improved Particle Swarm Algorithm Process

3. Research on Experiment

Collect the P2P data under the actual complex network environment with the sniffing software, and then preprocess the data to get P2P traffic's 5 major characteristics for identification, they are TCP traffic ratio, proportion of upstream traffic, total number of data packet, average length of data packet and the ratio between the number of connections and the number of IP.

Preprocess the traffic data from these 4 Web applications of Bittorrent、EMule、PPlive and PPStream, extract 1000 data among these to do experimental analysis. With these data, 500 of them are used to train the detecting model, and the other 500 data are used to test the model.

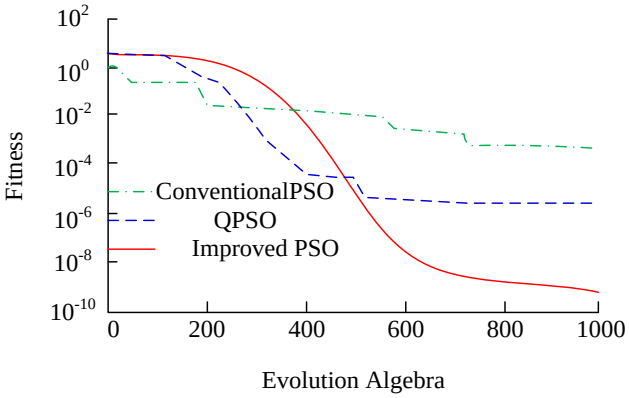
Build the P2P traffic detection model with the software of MATLAB. The number of the neurons in the input layer of BP neural network is set to 5, because there are 5 main traffic characteristics for P2P traffic detection. The number of the neurons in the output layer of BP neural network is set to 4, and uses the purelin type transfer function. The learning rate of the BP neural network is set to 0.01, the maximum number of training is set to 2000, and the learning accuracy is set to 10^{-4}.

Meanwhile, to test the improved particle swarm optimization algorithm researched in this paper, it uses the classical function of Ackley function, Griewank function and Sphere function as the fitness function, and use the conventional particle and the particle swarm optimization optimized by quantum computing as the control group. The dimensions and searching range of these three classical testing functions are shown in the table 1.

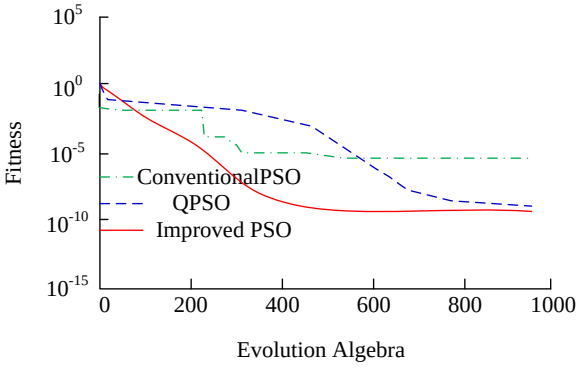
Table 1. Test function Parameter Settings

Testing Function	Dimensions	Searching Range
Ackley Function	50	[-32,64]
Griewank Function	50	[-50,100]
Sphere Function	50	[-600,600]

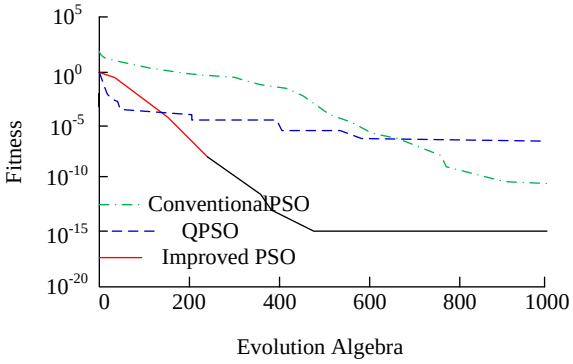
Set the maximum iteration of each particle swarm optimization to 1000, the accelerating factor c_1 and c_2 are all set to 2, the initial inertia weight factor is set to 1.2, both the value of mutation rate ρ_1 and ρ_2 are set to 0.5, the number of particles N is set to 100, the number of the subgroup M is set 10, the iteration of subgroup is set to 30; do simulating study on these particle swarm optimization, whose performance to be tested, with the 3 classical test function, and get the results of fitness are shown in the Figure. 4 below:



（a）Ackley testing function



(b)Griewank testing function



(c)Sphere testing function

Figure 4. Adaptive Value Curve of Three Kinds of Particle Swarm Optimization Algorithm

From the Figure 4, it can find that, no matter the convergence accuracy or convergence rate, the improved particle swarm optimization algorithm researched in this paper is better than the other two particle swarm optimization.

The testing results of the conventional PSO, the PSO optimized by quantum computing and the improved PSO proposed in this paper are shown in table 2.

Table 2. P2P Traffic Identification Test Results

P2PApplication	Identification rate(%)		
	Based on Conventional PSO-BPNeural Network	Based on QPSO-BPNeural Network	Based on PSO-BPNeural Network
Bitcomet	83.74	86.31	97.47
thunder	82.58	89.50	95.25
QQLIVE	82.80	89.93	96.18
PPStream	81.24	88.56	95.64

From the testing results, it can be find that, for the application of these 4 kinds of P2P network, the results get by the model proposed in this paper is 96.14% higher than the PSO-BP model, and is 9.4% higher than the QPSO-BP model.

4. Conclusion

(1) Research on the traffic detection and identification based on P2P has become a hot issuenowadays. This paper builds a P2P traffic detecting model with the BP neural network, and optimizes the network with the improved PSO to avoid the shortcoming of easy falling into local minimum.

(2) In the conventional PSO, all of the particles in the swarm try to close to the global optimal solution, therefore, there will exists the problem of easy falling into the local optimal solution. By the operation of grouping, it can enable the particles in the swarm to do the distributed search within a wide range for the optimal solution, in order to overcome the problem of easy falling into the local optimal solution. After grouping, particles' ability of searching for the optimal solution will be reduced. So, after grouping, we need to do the operation of organizing, in order to makes the improved algorithm obtains the ability of searching for the local and global optimal solution.

(3) The operation of fission is to initialize the particles with poor performance, and improve the diversity of the swarm; the searching range will also be enlarged. This can avoid falling into the local optimal solution. Mutation is by searching within the area nearby, to avoid losing the optimal solution within neighboring region and increase the searching accuracy.

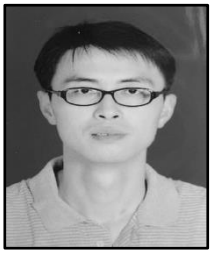
(4) Using the classical testing function to test, and no matter the convergence accuracy or convergence rate, the performance of the improved PSO researched in this paper is better than the other 2 kinds of PSO.

(5) Using the Matlab to build the P2P traffic detection model, and do traffic detection experiment on these 4 kinds of Web P2P application of Bittorrent, EMule, PPlive and PPStream. The testing result illustrates that, to these 4 kinds of P2P Web application, the average recognition rate of the identification model proposed in this paper is the best.

References

[1] X. Sun, R. Torres and S. Rao, "Preventing DDOS attacks on Internet servers exploiting P2P systems. Computer Networks, vol. 54, no. 15, **(2010)**, pp. 2756-2774.
[2] R. Zhao, "The research and implementation of P2P traffic identification based on feature string", Chengdu: University of Electronic Science and Technology of China, **(2009)**.
[3] H. Bleul, E. P. Rathgeb and S. Zilling, "Advanced P2P multiprotocol traffic analysis based on application level signature detection", IEEE Computer Society, **(2006)**, pp. 1-6.
[4] Z. B. Guo and Z. D. Qiu, "Identification of BitTorrent traffic for high speed network using packet sampling and application signatures", Journal of Computer Research and Development, vol. 45, no. 2, **(2008)**, pp. 227-236.
[5] F. Constantinou and P. B. I. Mavrommatis, "Identifying known and unknown peer-to-peer traffic", IEEE Computer Society, **(2006)**, pp. 93-102.
[6] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule and K. Salamatian, "Traffic classification on the fly", ACM SIGCOMM Computer Communication Review, vol. 36, no. 2, **(2006)**, pp. 23-26.
[7] X. Shu, J. H. Yang, D. F. Zhang and G. G. Xie, "Compare and analysis of clustering algorithms oriented traffic identification system", Computing Technology and Automation, vol. 27, no. 3, **(2008)**, pp. 1-6.
[8] P. Xu, Q. Liu and S. Lin, "Internet traffic classification using support vector machine", Journal of Computer Research and Development, vol. 46, no. 3, **(2009)**, pp. 407-414
[9] P. Xu and S. Lin, "Internet traffic classification using C4.5 decision tree", Journal of Software, vol. 20, no. 10, **(2009)**, pp. 2692-2704.
[10] H. Jia, W. Yaowu and L. Suhua, "Dynamic reactive power optimization based on particle swarm optimization algorithm", Power System Technology, vol. 31, no. 2, **(2007)**, pp. 47-50.
[11] W. Fangjie, Z. Chengxue and D. Zhiyuan, "Application of modified particle swarm optimization in reactive power optimization", Power System Technology, vol. 31, no. 24, **(2007)**, pp. 35-39.
[12] J. Kennedy and R. C. Eberhart, "A discrete binary version of the particle swarm algorithm", IEEE International Conference on Systems, Man and Cybernetics, Orlando, USA：IEEE, **(1997)**, pp. 4104-4108.
[13] Z. Wen and L. Yutian, "Adaptive particle swarm optimization and its application in reactive power optimization", Power System Technology, vol. 30, no. 8, **(2006)**, pp. 19-24.
[14] Z. Bo and C. Yijia, "A multi-Agent particle swarm optimization algorithm for reactive power optimization", Proceedings of the CSEE, vol. 25, no. 5, **(2005)**, pp. 1-7.

Authors



Song Rong, hisis nationality is Han. He was born in March, 1978. His birth place is Nanchang, Jiangxi Province. He is an associate professor of Jiangxi Vocational and Technical College of Communication. His research direction is computer science.

Li Xiating, her nationality is Han. She was born in June, 1979. Her birth place is Yichun, Jiangxi Province. She is an associate professor of Jiangxi Vocational and Technical College of Communication. Her research direction is computer science.

