International Journal of Security and Its Applications
Vol. 10, No. 2 (2016), pp.237-246
http://dx.doi.org/10.14257/ijsia.2016.10.2.21

Network Traffic Classification using Genetic Algorithms based on Support Vector Machine

Jie Cao[1, 2,*,] Zhiyi Fang[1]

[1] College of Computer Science and Technology Jilin University,
Changchun, 130012, P.R. China
[2] College of Information Engineering Northeast Dianli University,
Jilin, 132012, P.R. China
[1]caojiell78@126.com, [2]fangzy@jlu.edu.cn

Abstract

In recent years，machine learning method has been applied to the extensive research on traffic classification. In these methods, SVM (Support vector machine) is a supervised learning which can improve generalization ability of learning machine effectively. However, the penalty parameter C *and kernel function parameter* γ *are generally given by test experience during training of SVM. How to determine the optimal parameters of SVM is a problem to be solved. We proposed a method to deriving the optimal parameters of SVM based on GA (Genetic algorithm).This method does not need to traverse all the parameter points. The method extracts a certain number population from random solutions, and ultimately produces SVM optimal parameters according to the specific rules of operation. Through the method, we derived the optimal parameters combination* C *and* γ *of SVM. The accuracy of network traffic classification is improved greatly.*

Keywords: *Traffic classification, Genetic Algorithms, Support vector machine*

1. Introduction

Network traffic classification is to classify the traffic flow that is mixed with various applications [1]. Network traffic classification has been one of the hot topics in both academic and industrial fields. In recent years, the technology of traffic classification based on machine learning method has been widely researched. The machine learning methods include supervised learning methods and unsupervised learning methods [2] [3]. Support vector machine (SVM) is proposed by Vapnik *et al*., and it is based on the theory of structural risk minimization principle of statistical learning theory [4]. SVM is an effective supervised learning method. It can improve generalization ability of learning machine as much as possible. It can make the test sets to get the smaller error by the discriminant function which is obtained by the limited data set. In addition, SVM is a convex quadratic programming, the local optimal solution is certainly the global optimal one [5]. So SVM is an excellent machine learning method based on data. Recently, there are some researches to solve the problem of traffic classification based on SVM. Alice Este *at al*. [6] described a simple optimization algorithm that allows the classifier to perform correctly with as little training samples. Yuan Ruixi *et al*. [7] selected feature from a network flow and adopted a discriminator selection algorithm to obtain the optimal feature subset Zhu Li *at al*. [8] proposed a discriminator selection algorithm to obtain the best combination of the features for classification. Francesco

Jie Cao is the corresponding author.

Palmieri *at al*. [9] presented the potentialities of wavelet analysis in characterizing, and hence recognizing through binary classification, card-sharing traffic flows.

In the above researches, how to determine the penalty parameter C and the kernel function parameter γ during the training of SVM were not involved. In our previous work, we have used the grid search method to find the optimal parameters C and γ according to the highest accuracy of classification. However, the grid search method has the deficiency that it is very time-consuming to find the optimal parameters in a wider range, needs to adjust the step gradually, and traverses all the parameters points in the grids. In order to solve these problems, we proposed the genetic algorithm (GA) which is a heuristic algorithm to derive the optimal parameters C and γ of SVM. This method can reduce the computational complexity, and improve the accuracy of classification greatly.

2. SVM Model

SVM can be used for pattern classification and nonlinear regression. The main idea of SVM is to establish a classification hyperplane as the decision surface, making the maximum distance between positive and counter examples [10]. SVM is an approximate implementation of structural risk minimization. This principle is based on the learning machine's generalization error rate which takes the sum of training error rate and VC dimension term as the boundary [11] [12]. In the separable mode for SVM, the first term is zero and the second item is minimized [13]. Therefore, SVM can provide good generalization performance, and this attribute is unique to the SVM.

The classification function of SVM is shown as equation (1), x and y are two kinds of linear separable sample sets (x_i, y_i), k is a dot product kernel function, and classification plane is $\omega \cdot x + b = 0$. In short, SVM first transforms the input space to a high dimensional space with the nonlinear transform of the inner product function, then, presents the optimal classification plane in the space. At present, there is no unified model about the SVM method and its parameters, and the selection of kernel function and its parameters. Normally, the optimal parameters selection of SVM can only be found by experience, the experimental comparison and the large range searching.

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{n} \alpha_i^* y_i k(x_i \cdot x) + b^*\right) \tag{1}$$

3. SVM Parameters Optimization based on GA

During training of SVM, the penalty parameter C and the kernel function parameter γ are set normally by test experience. How to derive the optimal parameters C and γ, and get the high classification accuracy is the key problem of researching. Though we have used the grid search method to find the optimal parameters C and γ according to the highest accuracy of classification, that is global optimal solution. However, this method is very time-consuming to find the optimal parameters in a wider range. We proposed a method for deriving the optimal parameters of SVM based on GA [14]. This method can derive the global optimal solution and does not need to traverse all the parameter points. GA is a heuristic algorithm, and its essential feature is the population search strategy and the simple evolutionary operators [15]. GA has its unique advantages according to its ability of generalization, robustness, parallelism and simple operation [16]. It has the following features compared with other traditional optimization algorithms:

(1) The search process uses parameters encoding not directly acting on the variables.
(2) The algorithm only uses the information of the target function value as the search information.
(3) The algorithm is to carry out large-scale evolutionary optimization in the population, rather than to optimize on a single point. It has a good global search ability, and can avoid falling into local optimal solution.
(4) The selection, crossover and variation operators are all random operations. The search direction is guided by transition rules of probability, so that the direction of search process moves to the optimal solution field of search space.
(5) The algorithm has strong robustness. Even if there is any interference, the results are similar when solving multiple times.

Steps of parameters optimization based on GA are as follows:

(1) Determination of fitness function： $Accuracy = \sum_{i=0}^{n} \frac{TP_i}{TP_i + FP_i}$. True Positive (TP): the number of class X members that are correctly classified as belonging to it. False Positive (FP): the number of non-class X members that are incorrectly classified as belonging to class X .（Suppose X is a network traffic class. Class X and other non-class X mixed together.）
(2) Initialization: N individuals as the initial population are randomly generated, and the population is a set of feasible solutions to the fitness function. Set the initial value of the evolution algebra is 0, and the maximum evolution algebra is 100.
(3) Evaluation of the individual: According to the fitness function, the fitness of the each population is calculated to see whether it meets the fitness standards.
(4) Judgment of termination condition: To determine whether the algorithm satisfies the termination condition. If it satisfies the condition, go to the step (8). If it is not satisfies the condition, go to the step (5).
(5) Selection operation: The initial population is executed by the selection operation. The excellent individual is copied largely, and the unsound individual is copied little even to be eliminated.
(6) Crossover operation: Operation is implemented according to the crossover probability.
(7) Variation operation: Operation is implemented according to the variation probability.
(8) After selection, crossover and variation operation of the population, the next population which is composed of N individuals is generated, then go to the step (2), otherwise go to the step (4).
(9) By continuous evolution, the individual with the highest fitness on objective function is derived ultimately. This individual is outputted as the optimal solution, and the calculation is terminated.

According to the above steps, the process of the SVM parameters optimization based on GA is as shown in Figure 1.

Selection operator: Selection operation is a process which selects the individual with high fitness and low elimination fitness from the current population. The selection operation is based on the evaluation of the fitness of individuals in the population. The purpose is to make the optimal individual survive with high probability, so as to improve the computational efficiency and global convergence.

Crossover operator: Based on crossover operator, crossover operation generates a new individual through carrying on gene exchange of two individuals in the population with a certain probability. The purpose is to obtain the next generation of the best individual, and improve the searching ability of GA.

Variation operator: Variation operation is the phenomenon of the some allele mutation on chromosome. And it is another way to generate a new individual. The main purpose of variation is to maintain the diversity of the population, to prevent premature convergence phenomenon. In addition, it can make the GA has a local random search ability.

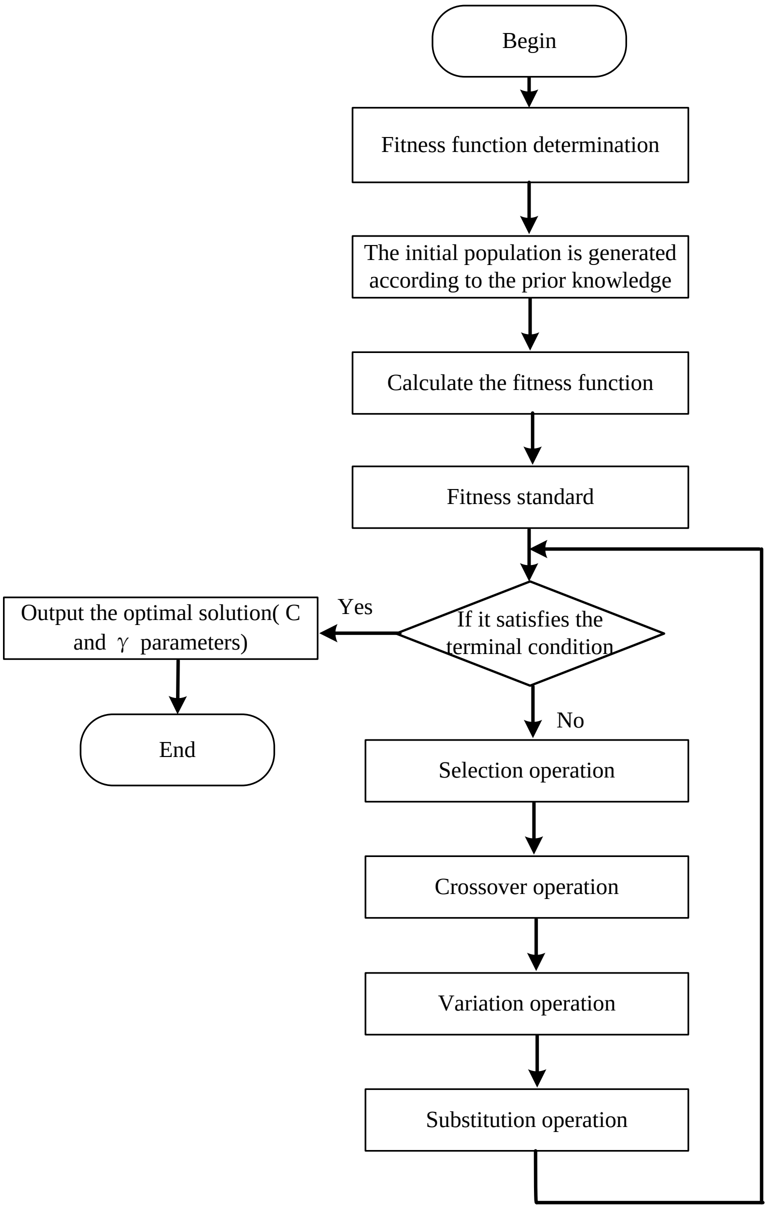
Begin
Fitness function determination
The initial population is generated according to the prior knowledge
Calculate the fitness function
Fitness standard
If it satisfies the terminal condition
Yes
Output the optimal solution(C and γ parameters)
End
No
Selection operation
Crossover operation
Variation operation
Substitution operation

Figure 1. The Process of the SVM Parameters Optimization based on GA

4. Evaluation based on SVM

4.1 Data Set

We used the Andrew Moore datasets. The datasets consist of 10 separate sub data sets each from a different period of the 24-hour day [17]. In our experiments, we extracted samples not more than 3000 from every subset randomly. Because the samples of Games and Interactive flows were very few, we deleted these traffic flows. The composition of our data set is presented in Table 1.

Table 1. The Number of Data Set's Samples

Traffic flows	Number of Samples	Traffic flows	Number of Samples
WWW	2999	P2P	2391
Mail	2999	DataBase	2943
Ftp-Control	2990	FTP-Data	2997
FTP-PASV	2989	MultiMedia	576
Attack	1793	Services	2220

4.2 Pretreatment

In our previous work, we adopted wrapper algorithm and determined the number of features corresponding the highest classification accuracy of each flow, such as table 2. Based on this feature subset, we will derive the optimal parameters C and γ of SVM with GA.

Table 2. The Number of Feature Subset

Traffic flows	The number of feature subsets	Traffic flows	The number of feature subsets
WWW	243	P2P	238
Mail	225	DataBase	175
Ftp-Control	247	FTP-Data	98
FTP-PASV	228	MultiMedia	222
Attack	227	Services	86

4.3 The Simulation Experiments and Analysis

In order to reduce the computational complexity, we extracted 500 data samples randomly from each traffic flow. Parameters optimization processed with a total of 5000 data samples. With libSVM [18], we took half the data as the training set and the other half as the test set. The optimal parameter combination of (C,γ) which is derived by GA algorithm as shown in Table 3. The fitness of main flow's accuracy on GA is as shown in Figures 2-8. By deriving the optimal parameter combination, we tested the optimized SVM based on GA. The average classification accuracy of optimized SVM is higher than that of traditional SVM.

As shown in Figure 9, accuracy of WWW is increased from 87.96% to 99.2%, accuracy of Mail is increased from 88.17% to 99%, accuracy of Ftp-control is increased from 88.01% to 97.08%, accuracy of Ftp-pasv is increased from 88% to 99.12%, accuracy of Attack is increased from 92.8% to 99%, accuracy of P2P is increased from 90.48% to 95.08%, accuracy of Database is increased from 88.32% to 99.76%, accuracy of Ftp-data is increased from 88.34% to 99.88%, accuracy of Services increased from 91.5 to 99.72%. In addition to the accuracy of MultiMedia is close to the traditional SVM, the accuracy of the other traffic is significantly improved. And the overall average classification accuracy is increased from 90.13% to 98.31%.

Table 3. The Optimal Parameters C and γ based on GA

Traffic class	C	γ	Accuracy (%)
WWW	5.1307	0.5166	99.2
Mail	5.2648	0.6386	99.00
Ftp-Control	65.5170	4.9964	97.08
FTP-Pasv	50.2177	0.1931	99.12
Attack	69.9928	0.0100	99.00
P2P	20.3203	3.5438	95.08
DataBase	72.6331	0.3274	99.76
FTP-Data	65.9683	0.5349	99.88
MultiMedia	10.6126	3.7757	95.24
Services	16.4908	3.0433	99.72

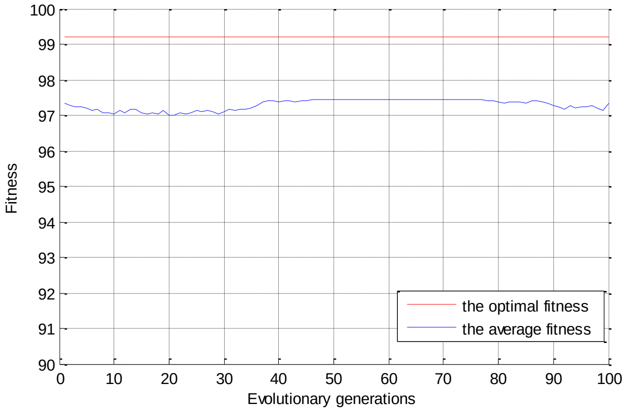

Figure 2.Fitness (%) of WWW's Accuracy on GA

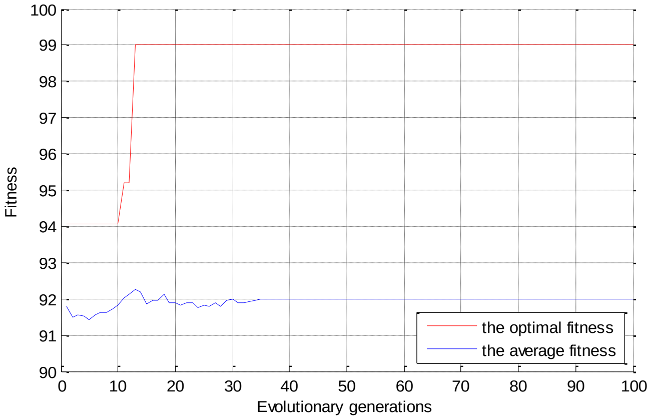

Figure 3. Fitness (%) of Mail's Accuracy on GA

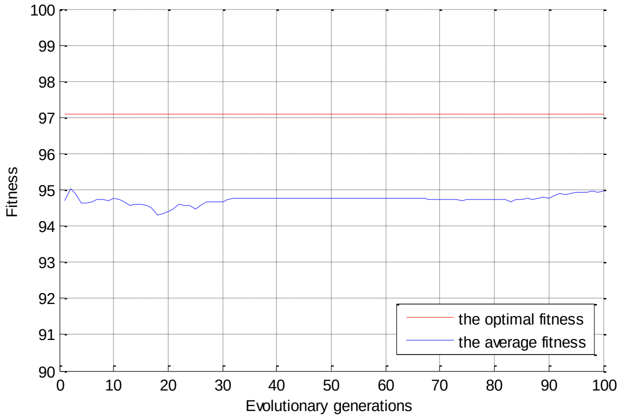

Figure 4. Fitness (%) of Ftp-Control's Accuracy on GA

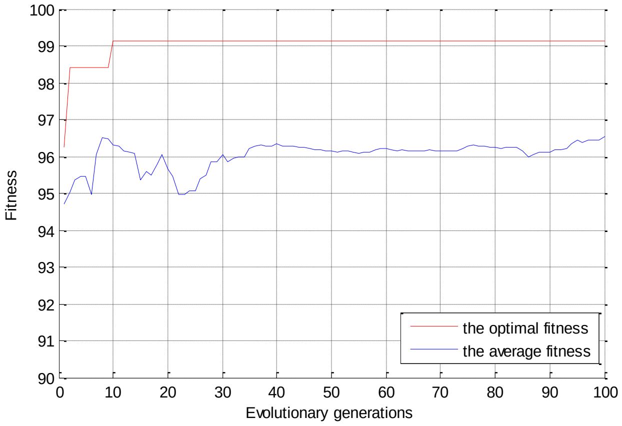

Figure 5. Fitness (%) of FTP-Pasv's Accuracy on GA

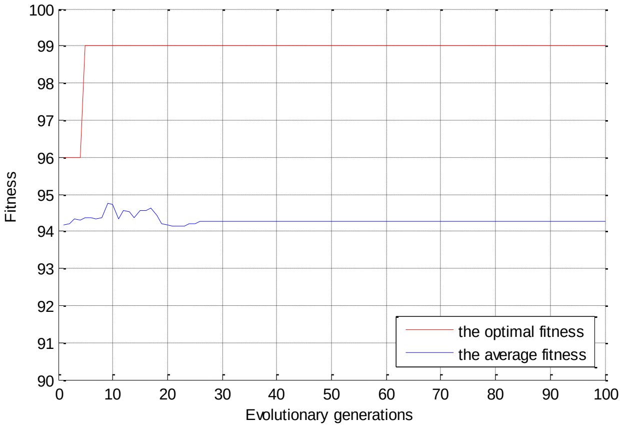

Figure 6. Fitness (%) of Attack's Accuracy on GA

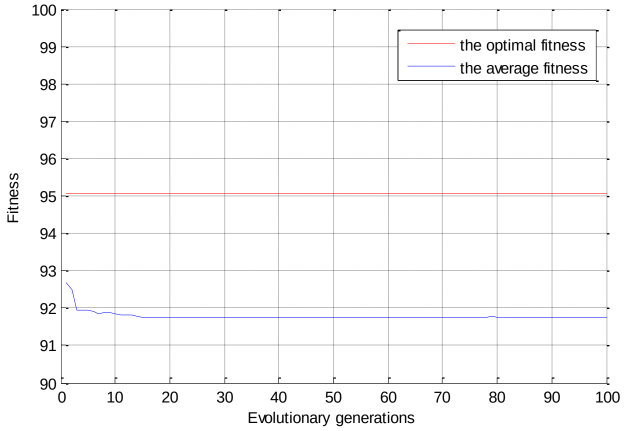

Figure 7. Fitness (%) of P2P's Accuracy on GA

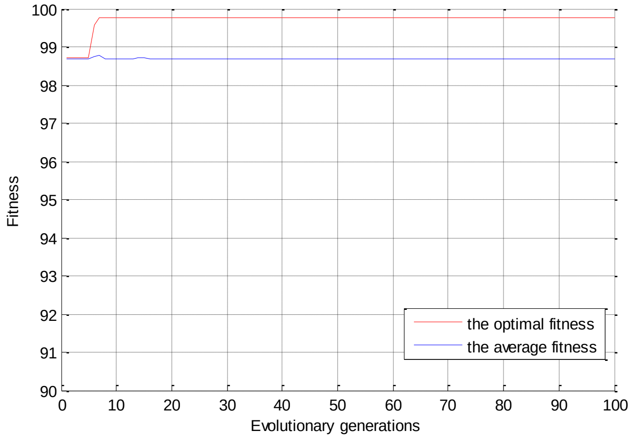

Figure 8. Fitness (%) of DataBase's Accuracy on GA

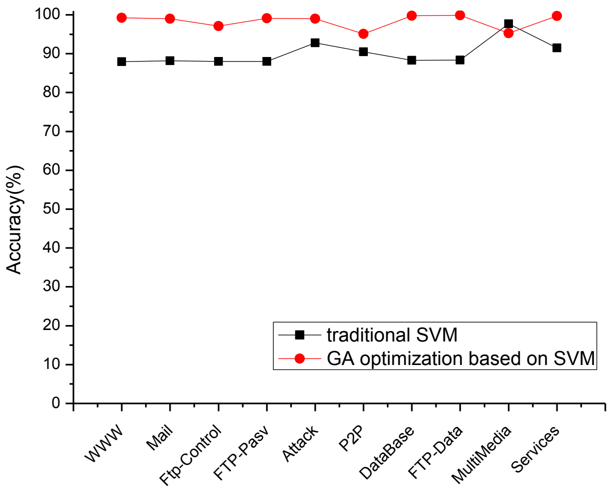

Figure 9. Accuracy on Traditional SVM and GA Optimized SVM

5. Conclusion

In this paper, we proposed a method to derive the optimal parameters of SVM based on GA. This method can find the global optimal solution and does not need to traverse all the parameter points in grid. By this method, the accuracy of traffic classification is greatly improved. Our next work is to research more swarm intelligence algorithms to improve the traffic classification accuracy of SVM.

References

[1] Alberto Dainotti and Antonio Pescapé, "Issues and future directions in traffic classification", IEEE Network, vol.26, **(2012)**, pp.35-40.

[2] Joao V. Gomes, Pedro R. M. Inacio, Manuela Pereira , and Mario M. Freire , "Detection and classification of peer-to-peer traffic: A survey", ACM Computing Surveys, vol.45, no.3, **(2013)**.

[3] Jawad Khalife, Amjad Hajjar, and Jesus Diaz-Verdejo, "A multilevel taxonomy and requirements for an optimal traffic- classification model", International Journal of Network Management, vol.24, no.2, **(2014)**, pp.101-120.

[4] N. Vapnik, "The Nature of Statistical Learning Theory", Technometrics, vol.38, no.4, **(1996)**, pp.400.

[5] Lei Ding, Fei Yu, Sheng Peng and Chen Xu, "A classification algorithm for network traffic based on improved support vector machine", Journal of Computers, vol.8, no.4, **(2013)** , pp1090-1096.

[6] Este Alice, F. Gringoli, and L. Salgarelli, "Support vector machines for tcp traffic classification", Computer Networks, vol.53, no.14, **(2009)**, pp. 2476–2490.

[7] Ruixi Yuan , Zhu Li , and Xiaohong Guan, "An SVM-based machine learning method for accurate Internet traffic classification", Information Systems Frontiers, vol.12, no.2, **(2010)** , p149-156.

[8] Li Zhu, and Yuan Ruixi, "Accurate Classification of the Internet Traffic Based on the SVM Method", IEEE International Conference on Communications, **(2007)**, pp.1373-1378.

[9] Francesco Palmieri, Ugo Fiore, Aniello Castiglione and Alfredo De Santis, "On the detection of card-sharing traffic through wavelet analysis and Support Vector Machines", Applied Soft Computing Journal, vol.13, no.1, January **(2013)** pp.1373-1378.

[10] Gabriel Gómez Sena, and Pablo Belzarena, "Statistical traffic classification by boosting support vector machines", Proceedings of the 7th Latin American Networking Conference, **(2012)**, pp.9-18

[11] Francesco Palmieri, Ugo Fiore, Aniello Castiglione, and Alfredo De Santis, "On the detection of card-sharing traffic through wavelet analysis and Support Vector Machines", Applied Soft Computing, vol.13, no.1, **(2013)** , pp.615-627.

[12] Buyun Qu, Zhibin Zhang, Xingquan Zhu, and Dan Meng, "An empirical study of morphing on behavior-based network traffic classification", Security and Communication Networks, vol.8,no.1, **(2015)** , pp.615-627.

[13] Xinlu Zong, Chunzhi Wang, and Hui Xu, "Density-based adaptive wavelet kernel SVM model for P2P traffic classification", International Journal of Future Generation Communication and Networking, vol. 6, no.6,**(2013)** , pp.615-627.

[14] Taeshik Shon, Jongsub Moon, "A hybrid machine learning approach to network anomaly detection", Information Sciences, vol.177, no.18, **(2007)** , pp.3799-3821.

[15] H. Hasan Orkcü, Hasan Bal, "Comparing performances of backpropagation and genetic algorithms in the data classification", Expert Systems with Applications, vol.38, no.4, **(2011)** , pp.3799-3821.

[16] Ba-Vui Le, Jae Hun Bang, and Lee Sungyoung, "Hierarchical emotion classification using genetic algorithms", Proceedings of the 4th Symposium on Information and Communication Technology, **(2013)**, pp.158-163.

[17] Moore A, and Papagiannaki K, "Toward the accurate identification of network application", Proceedings of PAM, Boston, MA, U.S.A. ,**(2005)** March-April.

[18] Chih-Chung Chang, and Chih-Jen Lin, "LIBSVM: A Library for support vector machines", ACM Transactions on Intelligent Systems and Technology, vol.2, no.3, **(2011)**.

Authors

Jie Cao, She received the MS degree in information engineering from Northeast Dianli University, Jilin, China, in 2007. She is currently a lecturer of information engineering in Northeast Dianli University. And now she is a Ph.D. candidate in Computer science and Technology of Jilin University. Her research interests include the areas of computer network, computer system architecture , and data mining.

Zhiyi Fang, He received the PhD degree in computer science from Jilin University, Changchun, China, in 1998, where he is currently a professor of computer science. He was a senior visiting scholar of the University of Queensland, Australia, from 1995 to 1996, and the University of California, Santa Barbara, from 2000 to 2001. He is a member of China Software Industry Association (CSIA) and a member of Open System Committee of China Computer Federation (CCF). His research interests include distributed/parallel computing system, mobile communication, and wireless networks.