Intrusion detection System Based on improved SVM Incremental Learning

Hongle Du, Shaohua Teng , mei Yang
Faculty of Computer
Guangdong University of technology
Guangzhou, China
Dhl5597@163.com, shteng@gdut.edu.cn

Qingfang Zhu
College of Mathematics
Luoyang normal university
Luoyang, China
Ann55@163.com

Abstract—**When collecting Network connection information, we can not obtain a complete data set at once, which result in SVM training insufficiently and high error rate of prediction. To solve this problem, this paper proposes a new method that combines Support Vector Machine with clustering algorithm, based on analyzing the relation between boundary support vectors and KKT condition. In the method, firstly, presents incremental support vector machine learning algorithm based on clustering, and describes the running process of the algorithm detailedly; then give the intrusion detection model based on incremental SVM learning; finally, the performance of the model is tested by computer simulation with KDD CUP1999 data set. The experimental results show it has higher detection accuracy rate and improves the speed of SVM training and classification, as keeping the generalization ability of incremental SVM learning algorithm.**

Keywords-Support vector machine;Clustering algorithm; Incremental learning; Intrusion detection system

I. INTRODUCTION

Support vector machine (SVM) is a new machine learning based on the statistical learning theory (SLT) developed by Vapnik [1]. It shows lots of unique advantages to solve the problem about the small sample, nonlinear and high dimensional pattern recognition problems based on structural risk minimization principle. In recent years, SVM has been successfully applied into a variety of fields, such as handwritten digit recognition, face identification, and text categorization. But to solve the massive data sets and online learning problems, SVM will be no advantage. For example, if data set is too big to read into memory at once, SVM will be no way to train it. And if we can not obtain complete data sets at beginning, we have to use online learning. To solve these two problems, incremental SVM learning algorithm is presented [2-9].

Syed proposed the incremental SVM learning algorithm at first [2]. The algorithm only keeps down the support vectors and discards all non-support vectors. In fact, these discarded samples also include the information of classification of data set. And with the adding of new datasets, non-support vectors and support vectors may be transformed into each other. The classification accuracy rate will be seriously affected if we discard samples too fast. Especially in the initial case of smaller samples, then the follow learning may be unstable and result in shake.

Intrusion Detection can be seen as a classification problem, according to the network information to classify network behavior: normal behavior or abnormal behavior. Thus the intrusion detection problem is transformed into a pattern recognition problem. The reference [10] presented the intrusion detection model based on support vector machines, and discussed the work process of the model. The references [11-13] introduce a fuzzy membership function to the penalty in the quadratic problem of proposed by Weston and Watkins, and also achieved good results.

Above methods all need to get the complete training data set in order to have good prediction accuracy; however it is very difficult to collect the complete training data set from network stream. So the reference [14] use the method that was proposed by Syed [2] into intrusion detection and reduces the training time and predicting time, but it reduces more accuracy of prediction at the same time. In this paper, incremental learning algorithm of SVM based on clustering (CISVM) is proposed, taking into account the boundary support vectors may change into support vectors after adding new samples. The method, firstly, cluster the training data set using unsupervised clustering (Abbreviate UC) algorithm and gets cluster particles; then reconstruct the new training set with the centers of cluster particles and retrain it; then, discard the samples that meet the KKT condition and add the samples that contrary to the KKT condition into the support vector set using the UC (Unsupervised Clustering) algorithm; finally, we can get new training data set again and retrain it again. And the algorithm can keep the information of classification, and reduce training and classification time, at the same time improve classification accuracy rate.

This paper is organized as follow. Section 2 analyzes the process of the SVM and current incremental SVM learning. Section 3 gives the new algorithm for SVM incremental learning. Section 4 constructs the model of intrusion detection based on incremental SVM learning algorithm. Section 5 presents the simulation experiments and results. Section 6 gives the conclusions.

978-0-7695-3816-7/09 $26.00 © 2009 IEEE
DOI 10.1109/AICI.2009.254

II. SUPPORT VECTOR MACHINE

A. Fuzzy support vector machine

Fuzzy support vector machine is SVM introduced fuzzy membership function. The fuzzy membership function is constructed different function based on the impact of the decision-making surface of different samples. It can avoid the impact of the noise samples and outliers' samples, and thus improve the classification accuracy rate of SVM [13, 14]. The essential of SVM is to construct an optimal separating hyperplane through training data set. Given training sample set with label of categories

$$T=\{(x_1,n_1,y_1),(x_2,n_2,y_2),\cdots,(x_l,n_l,y_l)\}$$,

$x_i \in R^n$, $y_i \in \{1,-1\}$,and n_i is membership function of sample. The main purpose of SVM is to construct a separating hyperplane to differentiate two different types of samples, meanwhile insuring classification of the maximize margin and the minimize error rate. We can solve the below quadratic optimization problem to get the decision function.

$$\min \frac{1}{2}<w,w>+\frac{1}{l}\sum_{i=1}^{l} n_i(\varepsilon_i-\nu\rho)$$

$$s.t.\ \ y_i(<w,x_i>+b)\geq\rho-\varepsilon_i \qquad (1)$$

$$\varepsilon_i\geq 0,\ \rho\geq 0,\ i=1,2,\cdots,l$$

Transform the Optimal problem of classification into its dual form through introducing Lagrange multipliers:

$$\max_a W(a)=\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l}\alpha_i\alpha_j y_i y_j K(x_i,x_j)$$

$$\text{s.t.}\ \sum_{i=1}^{l}\alpha_i y_i=0,\frac{\nu}{l}n_i\leq\alpha_i\leq\frac{n_i}{l}\ \ i=1,2,\cdots,l \qquad (2)$$

Where $K(x_i,x_j)$ is kernel function, and let $K(x_i,x_j)=<\phi(x_i),\phi(x_j)>$ .It is using non-linear mapping $\varphi:R^k\mapsto F$ to make training sample set map into feature space from input space. They are linear classification in the feature space.

We obtain the following decision function: $f(x)=\mathrm{sgn}(\sum_{x_i\in sv} a_i^* y_i K(x,x_i)+b^*)$. (3)

The above decision function show the effectiveness of SVM classification is support vectors. In other words, samples that $a_i \neq 0$ play a major role and other non-support vectors have no any effect in the decision function.

B. Incremental SVM learning

Incremental learning is proposed to solve two type machine learning problems: one is that computer's memory is not enough and training time is too long if the training data set is too large; another is we can't obtain the maturity data set at beginning and have to use online learning, that may improve learning precision in the using process with increasing of samples. The key of incremental learning is which learning information should be retrained in the previous training and how to deal with newly adding data set. Syed proposed the incremental SVM learning algorithm at first [2].The algorithm, first, train training data set and obtain classifier and all support vectors; then we obtain new training data set through merging support vectors and new adding data set; finally, we train new training data set. References [2,3] have a common characteristic is that they discard all samples that is not support vector, but they lost some classifications information at the same time; Based on analyzing the relations between KKT(Karush-Kuhn-Tucker) condition and samples distribution, reference [4] proposed an equivalence algorithm because previous data set and new adding data set have equivalence effect on constructing decision function; Taking into account the samples near the hyperplane may become support vector after adding new training set, reference [6] introduce a redundant incremental learning algorithm: so add redundant samples near the hyperplane in next training; reference[7] proposed active set iteration method that mainly solves the large-scale learning problems.

The above methods do not take into account the samples near hyperplane or not enough resulting in higher classification error after incremental learning. As shown in Figure 1, non-support vectors may change into support vector after adding new sample set in order to minimized error rate and maximize classification margin. So we can't discard all non-support vectors. Reference [8] point out that support vector set will change after incremental learning if the new samples include some classification information that the previous training set don't include. And point out that the conversion is relation with the KKT condition. The new samples meet KKT conditions will not change the previous support vector set and the new samples contrary to KKT condition will change the previous support vector set. Karush-Kuhn-Tucker theorem, also known as KKT conditions, gives the optimal solution conditions for an optimization problem. Reference [9] gives three situations in violation of the KKT conditions, we can sum up as: $|f(x_i)|<1$.

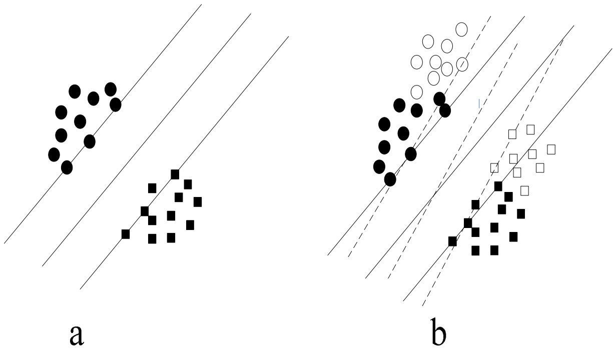

Figure 1. classification hyperplane's change after adding new samples

III. Incremental SVM Learning Based on UC Algorithm

A. Design idea

Based on above analyzing, this paper presents a new incremental learning algorithm combined SVM with clustering algorithm. In this algorithm, firstly, we deal with training set using clustering algorithm (Class label is looked as one attribute of data set. And the samples that belong to the same cluster have same class label if radius $r > 1$). Then we can get clusters $O(n_i, o_i, y_i)$(Where n_i is the number of sample point; o_i is the center of the cluster; y_i is class label).Next, we construct new training data set using centers of clusters and n_i is fuzzy membership function. Then we train new training data set with FSVM and obtain new support vectors. There are two strategies to deal with new adding data set: one is to add new adding samples into the support vector set that is get in the first step using clustering algorithm; another is only to add samples that contrary to KKT condition using UC algorithm and thrown away the samples that are satisfied with KKT condition .Then we can get new training set and training it using FSVM again. In this paper, we will compare two treatment methods with experimental results and the corresponding analysis.

B. UC algorithm

Reference [9] presents a simple UC algorithm. Compared with tradition K-means algorithm, the algorithm doesn't need to pre-specify the number of classification. There is a high speed of clustering, so we deal with the training data set using the UC algorithm in this paper. Given the training set $T = \{x_1, x_2, \cdots, x_l\}, x_i \in R^{n+1}$ .Here, we look the label of the SVM training set as a dimension, so the number of dimensions is n+1. C_number is the number of cluster. Then algorithm can be described as follows:

1. Read one record x_i from training set. If $C\_number = 0$, Create a new cluster center o_1 and set $o_1 = x_i$; otherwise go to step 2;

2. Compute distance d_i between sample x_i and each o_k , $d_i = \sqrt{(x_{j1} - o_{k1})^2 + (x_{j2} - o_{k2})^2 + \cdots + (x_{jm} - o_{km})^2}$,set $d_m = \min\limits_{i=1,2,\cdots,n} (d_i)$,where n is number of cluster particles and m is index of cluster particle;

3. If $d_m < r$,then x_i is added into cluster o_m and reset o_m : $o_m = \dfrac{o_m * n_m + x_j}{n_m + 1}$;and reset $n_k = n_k + 1$;Otherwise recreate a cluster and set $C\_number = C\_number + 1$, $o_{C\_number} = x_j$, $n_{C\_number} = 1$;

4. If all samples are dealt with, then stop; otherwise go to step 1.

C. Combined SVM with UC algorithm

In order to keep more classification information of original samples and have a high accuracy of classification, the clustering radius should be smaller. If $r = 0$, that is to say, we do not cluster. But in order to keep more boundary support vectors and improve the speed of training, we should make the radius larger. In this paper, we set $r = 1$. Given training sample set with label of categories $A = \{(x_i, y_i), i = 1, 2, \cdots, l\}$ and new adding data set $B = \{(x_i, y_i), i = 1, 2, \cdots, p\}$, $x_i \in R^n$, $y_i \in \{1, -1\}$. l, p are number of samples. The CISVM algorithm can be described as follows:

1. Deal with the data set A using the algorithm in section 2.2. Then we can get the clustering center set $O = \{(n_1, o_1), (n_2, o_2), \cdots, (n_p, o_p)\}$;

2. Reconstruct training set with clustering center set O . Firstly, we separate the last characteristic of the vectors o_i and get o_i ' and y_i . Here y_i is the label of classification. Then we can get the new training set $T = \{(n_1, o_1', y_1), (n_2, o_2', y_2), \cdots, (n_p, o_p', y_p)\}$ .Where n_i is the fuzzy membership function. Finally, retrain the new training set T using FSVM and get classifier φ and the support vector set SVs ;

3. There are two methods to deal with new adding set B . The first method is that we look the set SVs as the clustering center set and deal with set B using clustering algorithm of section 2.2. Then we can get new clustering center set O' ; Second method, according to KKT condition, discard the samples that meet the KKT condition and deal with the samples that contrary to KKT using the first method.

4. Do with O' using the method of step 2; then we can get the new classifier φ' and new the support vector set SVs' .

IV. The Model of IDS Based on CISVM

The essential of Network intrusion detection is a classification problem, according to the network link information to classify the network behavior: normal behavior or abnormal behavior. Thus the intrusion detection problem is transformed into a pattern recognition problem. In this field, some scholars do in-depth study. In the reference [15], intrusion detection was seen as a process of distinguishing between "self" and "non-self", and Forrest and others scholars propose the model of intrusion detection based on immune technology. Ghosh use neural networks to extract features and classify [16]. W. Lee discusses the question of achieving intrusion detection from the point of data mining techniques [17]. Above methods requires a large number or a complete samples set to achieve the desired performance and need a longer training time. The support vector machine is an effective tool to solve the problem with small samples set, and it's been widely applied to various fields of pattern recognition. In the reference [10, 11], SVM was used in intrusion detection system, and made a very good experiment results.

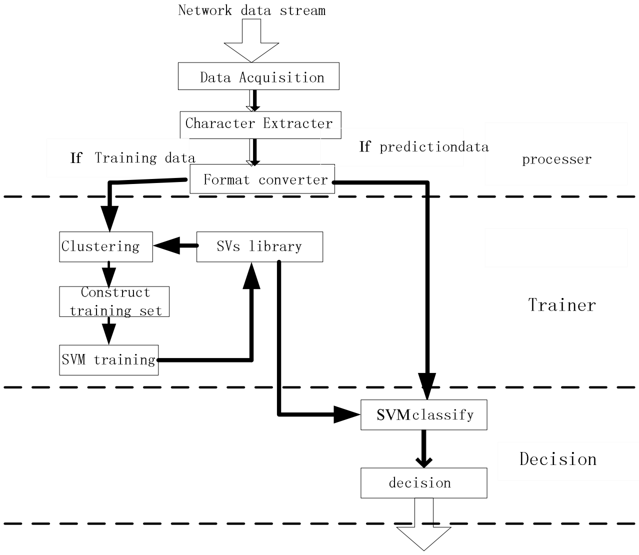

Figure 2. the model of IDS based on CISVM

Network intrusion detection system based on SVM is mainly composed of three parts: data-preprocessing, trainer and decision and response. As shown in Figure 2, the data-preprocessor mainly extracts the network data from the large network data stream, and converts the data format. Support vector machines only identify the data of digital type and the data must be the same dimension, therefore, it must extract effective information of network connection and at the same time transform the original network data into digital vectors. Trainer train SVM with the data that combine the new adding set and the support vector set (If this is the first time, the support vector set is null.) using the method proposed in section 3 and store the results of training to support vector library. Decision-making classify the network connection behavior according to information of support vector database, and then makes the appropriate decision.

The whole system can be divided into two processes: training and predicting. The main work, in this paper, is in training stage. Compared with the tradition SVM, the training stage take into account the relation of new adding data set and support vector set and deal with them using the algorithm presented in section 3. And not to add new adding data set and support vector set simply. AS show in Figure 2, when adding new data set, we deal with it using UC algorithm and get new training set. And then retrain it. The method takes into account the affection of boundary support vectors. Because of reducing the scale of training set, it improves the speed of training. Predicting stage is a process that is using support vector machines to classify the network data that have been processed with above method. According to the discriminate function (3) we can get the sort of networks behavior and submit the results to the decision-making system and make the appropriate decision.

V. Experiments Results and Analysis

We have generated UC algorithm implementation in C++ and improved SVM based on LIBSVM [18]. To evaluate the effectiveness and the performance of the CISVM, we choose the benchmark data sets from the KDD CUP1999 [19] for experiments. In all experiments, we use personal computer (P4 3.0GHZ, 512M RAM), and the operating system is Windows XP.

A. Selection experiment data sets

KDDCUP1999 data are standard data sets for Intrusion Detection, including the training data sets and the test data sets. The training data sets include 494,022 records and test data sets include 311030 records. There are 24 types of attacks in training data sets, and add new 14 kinds of attacks in the test data sets. They can be divided into four major categories: Probing, Denial of Service (DoS), User-to-Root (U2R) and Remote-to-Local (R2L). Each TCP connection record includes 41 attributes. These attributes are the types of values, and others are types of characters, but SVM can only deal with numerical vector. Therefore, before training we must make the input data numerical and normal. This study uses a simple substitution of symbols with numerical data types. The protocol-type, service and flag replaced by digital attributes. For examples, three kinds of Protocol-type (TCP, UDP and ICMP) are instead with 1, 2, and 3. 71 kinds of service are substituted with 1,2,…, 71 .Label of attack instead of 1 or -1,where normal record is 1 and abnormal record is -1.Last, normalized the input data set with Libsvm. Because original data sets are too large, so we only select Correct set.

In order to reduce the training time and ensure representation of the chosen data, use the same interval to selected data sets. We select four training sets: train1 is that get one record every 10 from first and all are 31103 records; train2 is that get one record every 10 from second and all are 15552 records; traint3 is that get one record every 40 from third and all are 15552 records; train4 is that get one record every 10 from fourth and all are 15552 records. Test is that get one record every 10 from fifth and all are 31103 records.

B. Experimental results and analysis

Where CN is the number of cluster, and SV is the number of support vectors after training, and AN is the number of records that are correctly predicted on the test set using SVM, and T-time is the time spend in training SVM, and P-time is the time spend in predicting test set, and AV is accuracy rate of classification. $r=0$ is directly trained without clustering. Table1 is the classification error rate of SVM affected by different radius of clustering. From table 1, when $r=1$, the number of cluster only is 27.98% of the samples. That is to say, we decrease seriously training set and reduce the SVM training time, and meanwhile the accuracy rate only reduces 0.177%. As radius increases, training time gradually reduce. But accuracy rate also reduce at same time. In the follow experiments, we set $r = 1$.

TABLE I. COMPARISON EXPERIMENTAL RESULTS OF DIFFERENT RADIUS

r	CN	SV	AN	T-time(s)	P-time(s)	AV(%)
R=0	31103	7891	29932	689	282	96.235
R=1	8704	6374	29877	105	236	96.058
R=2	6945	6210	29859	71	206	96.001
R=3	6420	6068	29805	64	195	95.827
R=4	6111	5952	29787	61	188	95.769

TABLE II. COMPARISON WITH TRADITION ALGORITHM AND CISVM

Data set	CN	SV	AN	T-time(s)	P-time(s)	AV (%)
train1	31103	7875	29932	689	282	96.235
C_train1	8704	6374	29859	71	236	96.058
Atrain2	46655	11335	30050	1587	404	96.612
C_Atrain2	12324	9272	30043	218	332	96.593
Atrain3	62207	14767	30082	2158	526	96.717
C_Atrain3	15741	12034	30060	326	431	96.647
Atrain4	77759	18087	30147	2834	641	96.926
C_Atrain4	19038	14718	30131	499	528	96.875

TABLE III. COMPARISON WITH TISVM AND CISVM

Data set	algorithm	CN	AN	SV	T-time(s)	P-time(s)	AV (%)
Adding train2	TISVM	23343	30039	11252	778	398	96.579
	CISVM2	12145	30042	8352	145	278	96.589
Adding train3	TISVM	26804	30083	14620	862	519	96.721
	CISVM2	14598	30090	9769	192	304	96.742
Adding train4	TISVM	30172	30101	15718	1030	626	96.779
	CISVM2	18865	30110	11068	236	381	96.807

Table2 is the comparison of results between directly adding new samples into training set and results of the first method of CISVM. $Atrain2, Atrain3, Atrain4$ are the new adding training set of $train2, train3, train4$ respectively, and $c\_train2, c\_train3, c\_train4$ are the new adding set of $train2, train3, train4$ using clustering proposed in this paper. Accuracy rate is higher as the number of samples increasing. But training time and predicting time are growing explosively. After adding train4,the algorithm of CISVM decrease nearly 90% training time and 30% classifying time in the case of loss 0.051% classification error rate.

Table3 is the comparison of results between tradition algorithm that was presented by Syed[2] and the second method of CISVM. Comparing with the tradition incremental learning, CISVM reduce the number of training samples and the number of support vectors. So it decreases the training time and predicting time. At the same time, the classification accuracy rate is improved.

Figure3 is comparison accuracy rates the method (CISVM1 and CISVM2) in this paper with TISVM (tradition incremental SVM) with the same test set. X-axis is adding $train2, train3, train4$ in turn. Y-axis is the prediction accuracy rates with same test set. As show in Figure 3, CISVM have a high accuracy than TISVM. And CISVM can reduce the scale of training data set, so training time and prediction time are short.

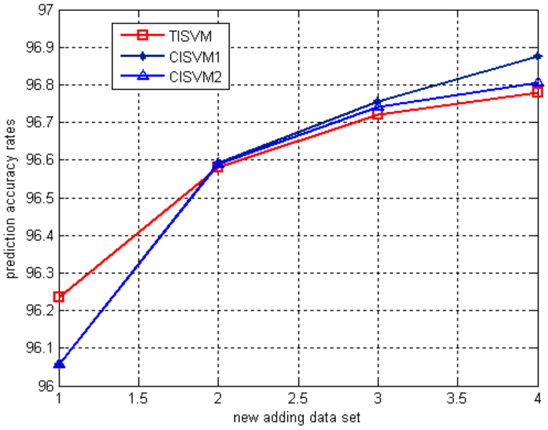

Figure 3. comparison accuracy rates

VI. CONCLUSION

In this paper, we analyze samples point near classification hyperplane detailedly. Combined with KKT conditions, we analyze current incremental SVM learning algorithm, then propose CISVM algorithm through introducing clustering algorithm into SVM. The algorithm takes account of the relation between KKT conditions and samples point near the hyperplane. It not only decreases the training time and predicting time but also improves the classification accuracy rate. The experiment results show that CISVM has high performance than traditional methods. Further research work is incremental SVM learning of multi-class classification.

VII. ACKNOWLEDGEMENT

This work was supported by Guangdong Provincial Natural Science Foundation (Grant No. 06021484), Guangdong Provincial science & technology project (Grant No. 2005B16001095, Grant No. 2005B10101077), Yuexiu Zone Guangzhou city science & technology project (Grant No. 2007-GX-075).

REFERENCES

[1] Vapnik V. The Nature of Statistical Learning Theory[M]. New York ：Springer—Verlag, 1995

[2] Syed N. Liu H. Sung K. Incremental learning with support vector machines [A].Proceedings of the Workshop on Support Vector Machines at the International Joint Conference on Artificial Intelligence (IJCAI-99) [C]. Stockholm, Sweden: M organ Kaufmann, 1999.876—892.

[3] Yangguang Liu,Qinming He,Qi Chen.Incremental Batch Learning with Support Vector Machines[C].Proceedings of the 5th World Congress on Intelligent Control and Automation.Hangzhou,China.2004 (2):1857-1861

[4] Mitra P. Murthy C. A. and Pal S. K. Data condensation in large databases by incremental learning with support vector machines[C]. International Conference on Pattern Recognition, 2000:708-711.

[5] Xiaodan Wang, Chunying Zheng, Chongming Wu. A New Algorithm for SVM Incremental Learning[C].ICSP2006 Proceedings.Beijing, China.2006

[6] WEN-JIAN WANG. A redundant incremental learning algorithm for SVM[C]. Proceedings of the Seventh International Conference on Machine Learning and Cybernetics[C], Kunming, China. 2008.734-738.

[7] Tao Liang. Fast Incremental SVM Learning Algorithm Based on Active Set Iteration[J].Chinese Journal of System Simulation,2006,18(11):3305-3308

[8] ZHOU Weida, ZHANG Li，Jiao Licheng.An Analysis of SVMs Generalization Performance [J]. Chinese journal of electronic .2001，29(5):590—594

[9] Li XL，Liu JM，Shi ZZ. A Chinese Web page classifier based on support vector machine and unsupervised clustering . Chinese Journal of Computers，2001，24 (1):62—68.

[10] Rao xian, Dong Chunxi, Yang Shaoquan. An Intrusion Detection System Based on Support Vector Machine [J]. Chinese journal of software, 2003, 14(4):798-803

[11] Li Kunlun, Huang Houkuan, Tian Shengfeng. Fuzzy Multi-Class Support Vector Machine and Application in Intrusion Detection[J].Chinese journal of computers, 2005, 28(2):274-280

[12] Huang, H. P.Liu, Y.H. Fuzzy Support Vector Machines for Pattern Recognition and Data Mining [J]. International Journal of Fuzzy Systems, 2002 4(3):826-835

[13] Hongle Du. Shaohua Teng. Qingfan Zhu. Intrusion detection Based on Fuzzy support vector machines[C]. 2009 International Conference on Networks Security, Wireless Communications and Trusted Computing[C], Wuhan, China.2009 (2)636-639

[14] Liu Ye, Wang Zebing, Feng Yan, Gu Hongying. DoS Intrusion Detection Based on Incremental Learning with Support Vector Machines[J].Chinese Computer Engineering,2006,32(4):179-181

[15] Forrest S. Perrelason AS. Allen L. Cherukur R. Self_Nonself discrimination in a computer[C]. Proceedings of the 1994 IEEE Symposium on Research in Security and Privacy. Oakland, CA: IEEE Computer Society Press, 1994.202~12.

[16] Chen You, Shen Huawei, Li Yang, Cheng Xueqi. An Efficient Feature Selection Algorithm Toward Building Lightweight Intrusion Detection System [J]. 2007，30（8）：1398-1408

[17] Lee W, Stolfo SJ, A data mining framework for building intrusion detection model[C]. Proceedings of the 1999 IEEE Symposium on security and Privacy. Oakland, CA: IEEE Computer Society Press, 1999, 120~132.

[18] Chang CC，Lin CJ. LIBSVM：A library for support vector machines.200 1.http://www.csie.ntu edu.Tw/cjlin/libsvm

[19] http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html