Computer and Information Science
www.ccsenet.org/journal.html
Vol. 2, No. 4
November 2009

Intrusion Detection Method Using Protocol Classification and Rough Set Based Support Vector Machine

Xunyi Ren
College of Computer Science, Nanjing University of Post & Telecommunications
Nanjing 21003, China
E-mail: renxy@njupt.edu.cn

Ruchuan Wang
College of Computer Science, Nanjing University of Post & Telecommunications
Nanjing 21003, China
State Key Lab. for Novel Software Technology, Nanjing University, Nanjing 210093, China

Hejun Zhou
College of Computer Science, Nanjing University of Post & Telecommunications
Nanjing 21003, China

Abstract

In order to improve the efficiency of support vector intrusion detection, we first do protocol Classification for the intrusion data, then refine its characteristic by rough set reduction. By using these procedures, we propose an intrusion detection method using protocol classification and rough set based support vector machine. The method is divided into training and testing processes. In the process of training, we first do protocol classification for the training data, and then do rough set refinement. The refined characteristics are stored as the pre-defined process, and finally the usage of support vector machine for data reduction training, the training model will be stored in accordance with the agreement. In the testing process, the data is classified according to protocol classification and then start the characteristics reduction procedure according to protocol classification. Finally, make a decision using the Support Vector Machines that corresponding to the agreement. The experimental results based on KDDCUP'99 data show that the method is the method is faster and the detection accuracy is comparable compared with the SVM without using protocol classification and using all characteristic.

Keywords: Intrusion detection, Support Vector Machine, Rough set

1. Introduction

Support Vector Machine, refer to Vapnik, 1995, Burges, 1996, P.121-167, is based on structured risk minimization and statistic theory. It overcomes the shortcoming such as difficult to handle of small samples, high dimension, over-matching, local minimization problems etc, that exists in the conventional methods like natural network. Therefore, it is a new high performance learning method, and it has been widely applied in intrusion detection face reorganization, voice processing and so on.

Intrusion detection is essentially a classification problem. It can be viewed as a classification process for test samples of training models. However, the construction of intrusion detection model needs to do learning for thousands of samples; there are tens of characteristics for every sample. Moreover, samples have the property of different structure. If we put the entire characteristic into intrusion detection, SVM will have to solve complex a quadratic programming problem. Therefore, the method is inefficient.

Actually, certain dependent relationship exists in the high dimension characteristics, therefore, how to find this dependence, and then compress the data so as to reduce the dimension, are significant for shorten SVM training time, detection time, and choosing the optimal parameter (Mukkamala, Janoski &.Sung, 2001, P. 1702-1707, Sung, 405-411, Lin & .Cunningham, P.190-198). In (Frohlich, Chapelle & B.Scholkopf, 2003, P.142-148), the genetic algorithm is adopted to optimize the model and characteristic chosen. In (Roberto, Guofei & Wenke, ICDM'06), n-grams is chosen to choose the host computer character and construct a combined SVM detector. In (Sung, P.405-411), the weighted SVM W is adopted to order and choose the characteristic, and by deleting the low influenced characteristics so as to find the most efficient two kinds of methods.

These kinds of methods have made considerable progress; however, these methods are always distilling the characteristic from all the data. Actually, the intrusion detection usually uses the leak of the protocol, and for every kind of protocol, the intrusion data has different characteristic. For different protocols, if different characteristic is used, the method will more powerful, and hence it will be helpful for improving the learning efficiency of the model.

Rough sets (Pawlak, 1982, P.341-356) is a frequently used method for distilling the characteristics, it is efficient in decreasing the dimension of data. In this paper, we propose to combine the protocol classification and rough sets methods, and so as to produce a intrusion detection method that is based on protocol classification and rough set SVM. By classifying the data based on data protocol, and reduction, we can give the training and detection model. Using the KDDCUP'99 intrusion data, we verify the method.

2. Classify Support Vector Machine (Vapnik, 1995, Burges, 1998, P.121-167)

Suppose $\{(x_1,y_1),(x_2,y_2),\cdots,(x_n,y_n)\}$ is a group of sample data, with $x_i \in \Re^d, y_i \in \{-1,+1\}$. We want to find the optimal partition plane $y = W \cdot X + b$, which is equivalent to solve the convex quadratic Burges programming problem:

$$\begin{array}{ll} \min\limits_{w,b,\xi_i} imize & \frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i \\ st & y_i(w^T X_i + b) + \xi_i - 1 \ge 0, \xi_i \ge 0, 1 \le i \le N \end{array} \tag{1}$$

Where w is the normal vector of hyperplane, b is the deviation, while C a punish function parameter in the case of incomplete integral, and ξ_i is a relaxation parameter in the case of relaxing the constraint conditions. By introducing the Lagrange multiplier:

$$L_p = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} a_i y_i (w^T X_i + b) + \sum_{i=1}^{n} a_i - \sum_{i}^{n} a_i \xi_i \tag{2}$$

Then do partial differential for L_p:

$$\begin{cases} \frac{\partial Lp}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{n} a_i x_i y_i \\ \frac{\partial Lp}{\partial b} = 0 \Rightarrow \sum_{i=1}^{n} a_i y_i = 0 \\ \frac{\partial Lp}{\partial \xi_i} = 0 \Rightarrow 0 \le a_i \le c \end{cases} \tag{3}$$

In order to obtain a_i, we convert the original problem in to a dual problem, and introduce kernel function $K(\bullet,\bullet)$:

$$\begin{array}{ll} \max imize & Q_D = \sum_i a_i - \frac{1}{2}\sum_{i,j} a_i a_j y_i y_j K(X_i, X_j) \\ st & 0 \le a_i \le C, \sum_{i=1}^{n} a_i y_i = 0, 1 \le i \le N \end{array} \tag{4}$$

By solving (4) we can obtain a_i, then submit it into (3) we have: $w = \sum_{i=1}^{n} a_i X_i Y_i$. As the quadratic programming problem satisfy the KKT condition, so we have $b = y_j - \sum_{i=1}^{n} y_i a_i^* K(X_i, X_j)$, with a_i^* is a coefficient larger than 0. As only when $a_i > 0$, it has effect on the value of Q_D. Therefore, we call the support vector corresponding to $a_i > 0$ as the support vector of X_i. Then we can get the decision function $f(X) = \text{sgn}(\sum_{i=1}^{n} y_i a_i^* K(X_i, X) + b)$

3. Rough Sets

Rough sets are proposed by Z.Pawlak in 1982, it is a data mining method which can be used to study the incompleteness of data, and uncertainty of knowledge. The basic idea of data reduction by using rough set theory can be outlined as follows: it find the decision regulation by the dependence relationship between the sample attribute and the decision attribute; then judge the importance by the degree of influence of attribute to the decision. By these procedures, the unimportant attribute can be removed, so as to achieve the classification ability of reduce the data characteristic and preserve the data nature.

Definition 1. Information system is a four number set I=<S,A,V,F>, with U is the nonempty sample set, and A is the attribute set, and V is the attribute value region, and F is the map, which can give a value from V for every sample attribute A in S.

For the training sample, there is some classification marks, such as the 42 dimensional intrusion sample of KDDCUP'99 is "normal", "abnormal" and so on. These attributes are called decision attributes. By introducing the decision attribute, we can obtain the decision graph by the information system.

Definition 2. The decision graph of information system is a four number set T=<S,A&{d}, V, F>, with A be the sample attribute, and its value a is called as condition attribute, and d is the decision attribute.

Definition 3. Indiscriminate relationship can be described as follows, in the decision graph DT, with $B \subseteq A$, for any sample in S, we have $F(a) = F(a^{'})$, then such a relationship is called the inseparable relationship between A and its subset B (B-indiscriminate relation), denoted by $IND_I(B)$, where $IND_I(B)$ refers to the indiscriminate relationship of attribute, i.e. the sample can not be discernible from attribute B. The decision indiscriminate relationship can be constructed based on the concept.

Definition 4. The indiscriminate relationship of decision is refer to the following fact, in $IND_I(B)$, we have $F(x,d) = F(x^{'},d)$, denoted by $IND_I(B,d)$.

Definition 5. The decision reduction refers to that in DT; we seek the smallest attribute set such that $IND_I(B,d) = IND_l(A,d)$ holds.

Though the decision reduction is a NP hard problem, there exist many fast reduction algorithms; this topic is beyond the discussion of this paper. Decision graph can be established by the decision graph discriminate matrix.

Definition 6. Suppose M is the decision graph discriminate matrix constructed based on DT, the element Mij on the (i, j) position is defined as follows,

$$M_{ij} = \begin{cases} \{a \mid a \in A \wedge f(x_i) \neq f(x_j)\} & f(x_i,d) \neq f(x_j,d) \\ 0 & f(x_i,d) = f(x_j,d) \end{cases}$$

By classifying the data protocol, and construct a decision graph for every group of data, then reduce the decision graph using the reduction algorithms, then we can obtain the different data set reduced from different data protocol.

4. The SVM intrusion detection method using protocol classification and rough Set

In the former investigation of rough set data reduction, the protocol is indiscriminate and the reduction is for all the data. There are two shortcomings in these approaches: firstly, all the data is strongly different structured, study the data using SVM, we need to introduce a new computation method for distance. On the other side, intrusion usually takes the leak of the different structured data. The indiscriminate protocol is just a broad detection method, it does not consider the different characteristic in different data, and hence these methods are not aimed. We propose the SVM intrusion detection method using protocol classification and rough sets, it is able to remove the shortcomings in the original methods, and is able to improve the detection time and the accuracy.

Classifying the protocol, using the rough set to reduce the data, then do training to the reduced data, i.e. the corresponding SVM input. The obtained training model is the SVM detector corresponding to different protocol. The SVM intrusion detection method using protocol classification and rough sets can be described as the following Fig 1.

In Fig 1, the real line illustrates the training process, the training data is classified according to protocol. Three different kinds of intrusion data is divided, denoted by TCP, UDP, and ICMP. Then carrying out the rough sets study for these three kinds of intrusion data, the studying procedure is denoted by T, U, and I. The reduced characteristic after study is used as the SVM study input; on the other hand, the reduced regularization is stored as the pre-definition process, denoted by reduction T, reduction U and reduction I. Three SVM study apparatus will become three detector models after study; they are stored as three detectors T, U and I. In Figure 1, the dash line denotes the detection process of the test data. The test data first classified by the protocol, then the reduction procedures are started based on different protocol data, the reduced data is inputted into corresponding detector, and the test results come from the detector. The

SVM intrusion detection method using protocol classification and rough sets can be described as the following algorithm:

Step 1: Input the training data, start protocol classification, the data is divided into TCP, UDP, AND, ICMP according to different data protocol; and they are stored in database.

Step 2: Start the rough sets study machine, reduce three kinds of data separably,then obtain their own reduced characteristic set T, U and I. Then construct a SQL sentence based on the characteristic set, which is stored as the pre-definition process. Finally, the reduced training data is inputted into the corresponding SVM study machine.

Step 3: Start the SVM study machine T, U and I, then obtain their own decision function by study. $f(X) = \mathrm{sgn}(\sum_{i=1}^{n} y_i a_i^* K(X_i, X) + b)$, stored as detector U, T, and I.

Step 4: For the input data X to be detect, first do protocol classification, then start the pre-defined rough sets reduction process according to classification.

Step 4: Input the reduced data into the corresponding SVM detector, the output the detection results through the SVM detector, normal is denoted by +1, and abnormal is denoted by -1.

5. Experiment

5.1 The tested data

KDDCUP'99 is obtained in the real net work. It can be used to simulate the 5 classes including 23 different kinds of data arising from attack, these data can be used as experimental data in data mining. The 10% subset of the data has 494021 records, and each record has 41 characteristics, which incorporate the continuous, discrete and text data. We can put a note at the end of each record to show whether the data is normal. Therefore, such kind of data set is a classical multi-protocol multi-attack

different structured data set. By classifying the protocol for the normal and attack situations, the results are illustrated in Figure 2 as follows

Statistical results show that TCP protocol records are 190064, and ICMP protocol records are 283602, and UDP protocol records are 20354. In the TCP protocol classification, there are all different kinds of attack, and DoS attack most frequently. In the UDP and ICMP protocol, the R2L and U2L attack almost never appear. For the UDP protocol, the abnormal data includes DoS and Probe. For the ICMP protocol the DoS attack has 280 thousands records. The abnormal data is mainly DoS data.

After protocol classification, we begin to do test from selected training data and test data, the test results are outlined as follows,

(1). TCP test data: Choosing 30000 records from the TCP data set, where the normal data is 12802 items; and abnormal data is 17198 items (DOS has 16560 items, Prob has 422 items, R2L has 188 items, U2L has 8 items).

(2). UDP test data: Choosing 10173 records from the UDP data set, where the normal data has 9586 items, and abnormal data has 587 items (DoS has 489 items, Prob has 98 items).

(3). ICMP test data: Choosing 28353 records from the ICMP data set, where the normal data has 128 items, and abnormal data has 28225 items (DoS has 28105 items, and Prob has 120 items).

Taking 70% data randomly from the test data set for training; then leaving other 30% for test.

5.2 The reduction of the test data

Reducing the data by means of Rosetta tool Komorowski, 1997, P.403-407, and form different reduction set from the 41 reduced characteristic. The characteristic set reduced from TCP, UDP and ICMP are outlined in the following Figure 3, Figure 4 and Figure 5.

Choosing two groups of characteristic set, for example, take the first and the eighth from TCP, and take the sixth and the 30th from UDP, and take the sixth and the eighth from ICMP. By reducing the characteristic for the corresponding training data and test data, we can obtain the training data and the test data after characteristic reduction. Compare with the characteristic with the ones given in Sung P.405-411, we can see our approach has less characteristic and easier to deal with, and finally the test result shows that the our method can preserve high accuracy and much faster.

5.3 Data training and detection

In the test, we choose RBF function $f(x_i, x_j) = \exp(-(x_i - x_j)/2\sigma^2)$ as the SVM kernel, and adopt 5-Fold Cross Validation, embedded in the LibSVM software by Chihjen. The test is in three steps, firstly, we use grid search (grid.py command) to compute the optimal punish parameter C and σ^2, then obtain the training model by train the training data using the optimal parameter. and finally test using the trained data. Take the example using 21000 TCP training data and 9000 test data, the parameter search is outlined in Figure 6. The optimal value is $C = 512, \sigma^2 = 0.03125$. By using these two parameters to train the 21000 TCP data, we obtain train.txt.model. Then we use this model to do test for these 90000 data. Finally, we obtain the training time, the detection time, and the accuracy.

For comparison reasons, the intrusion data and detection is divided into three situations. The first is to do test on the classified data by the complete characteristic. The second is to do test on the classified data by the reduced characteristic. The third is to do test on the unclassified data. The final test results are outlined in Figure 1, Figure 2 and Figure 3.

Comparing Figure 2 and Figure 3, we can discover that the training time and the detection time is shorten by using protocol classification, moreover, the detection accuracy is not damaged.

Comparing Figure 1 and Figure 2, we can see that using characteristic reduction and not using characteristic reduction has similar accuracy, however, the detection time and training time is saved obviously by using the characteristic reduction. Therefore, our conclusion is as follows, protocol classification along with characteristic reduction need less time, while using the complete characteristic need much more time, further more time is needed if protocol classification and characteristic reduction are all not carried out.

6. Concluding remarks

In this paper, we propose to use internet protocol classifying the intrusion data, and use rough sets to reduce unclassified data, and then do training for the reduced data, and finally produce a training model. In the test procedure, we first do protocol classification for the data, then do test for the model after training. We do some tests on the KDDCUP'09 data under three cases, the test results show that the new method produce more accuracy results, and need less training and test time. By theoretical analysis, the reason is as follows: as we have adopted protocol classification, which eliminate the difficulty caused by the unstructured protocol character, this reduces the time needed in dealing with data. On the other hand, as intrusion is due to the hole of protocol, so it is more targeted and the accuracy is not damaged because of the characteristic decrease. The future work will be implement a intrusion detection system based on the algorithm proposed in this paper. This will not only consider the protocol classification, but also need to consider that real internet intrusion actually a unsupervised character classification. Furthermore, it also needs multi-class classification technique research.

References

[DB/OL].http://www.csic.ntu.edu.tw/~cjlin/papers/libsvm.pdf.

Burges C. (1998). A tutorial on support vector machines for pattern recognition, *Data Mining and Knowledge Discover.* No. 2 .P. 121-167.

Chihjen L. LIBSVM: a library for SVMs (Version 2.6)

Frohlich H., Chapelle O., & Scholkopf B. (2003). Feature selection for support vector machines by means of genetic algorithm. *In: Proceedings of 15th IEEE International Conference on Tools with Artificial Intelligence.* No.3-5. P. 142 – 148.

http://kdd.ics.uci.edu/databases/kddcup99/task.htm.

Komorowski J.O., & ROSETTA. (1997). A rough set toolkit for analysis of data. *Fifth International Workshop on Rough Sets and Soft Computing.* Tokyo,Japan. P. 403-407.

Lin Y., &.Cunningham A. A New Approach to Fuzzy-Neural System Modeling. *IEEE Transactions on Fuzzy System*s, No.3. P.190-198.

Mukkamala S., Janoski G., & Sung H.(2003). Intrusion Detection Using Neural Networks and Support Vector Machines. *Proceedings of IEEE International Joint Conference on Neural Networks,* P.1702-1707.

Pawlak Z. (1982). Rough sets. International Journal of Information and Computer Sciences. No.11, P.341-356.

Roberto P., Guofei, G., & Wenke L. (2006). Using an Ensemble of One-Class SVM Classifiers to Harden Payload-based Anomaly Detection Systems. *In Proceedings of the Sixth International Conference on Data Mining (ICDM'06).*

Sung A. (1998). Ranking Importance of Input Parameters of Neural Networks. *Expert Systems with Applications*. No. 15. P.405-41.

Vapnik V. (1995). The nature of statistical learning theory. *New York: Springer-Verlag.*

Table 1. Experimental result with protocol difference but without reduction

protocal	best c, σ 2	Training time	Test time	Test correct rate
TCP	c=512, σ 2=0.03125	5.1s	1.6s	99.7889%
UDP	c=128, σ 2 =0.03125	2.3s	1.3s	99.9344%
ICMP	c=8.00, σ 2=0.078125	2.7s	1.4s	99.9765%

Table 2. Experimental results with protocol difference and reduction

protocal	Feature set	Best c, σ 2	Training time	Test time	Test correct rate
TCP	1	c=2048, σ 2=0.5	4.8	0.9	99.7287%
	8	c=32768, σ 2=0.125	4.2	1.2	99.765%
UDP	6	c=2048, σ 2=0.5	1.6	0.8	99.8689%
	30	c=2048, σ 2=8.0	1.8	0.9	99.9017%
ICMP	6	c=512, σ 2=5.0	2.1	1.0	99.9882%
	8	c=32, σ 2=0.078125	2.4	0.8	99.9765%

Table 3. Experimental results without protocol difference and reduction

Data set	best c, σ 2	Training time	Test time	Test correct rate
30000	c=512, σ 2=0.5	8.6	6.2	99.5%
10173	c=128, σ 2=0.125	4.7	3.7	99.8%
28353	c=64, σ 2=0.125	5.3	5.2	98.6%

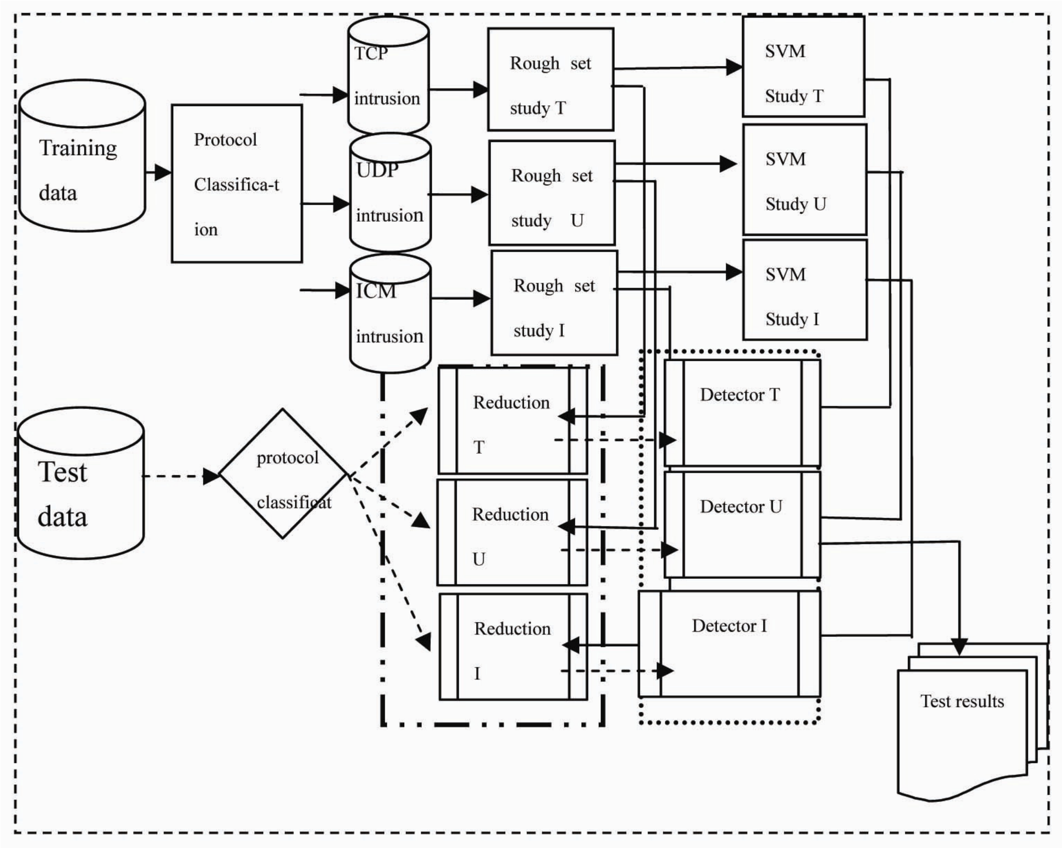

Figure 1. The SVM intrusion detection method using protocol classification and rough sets

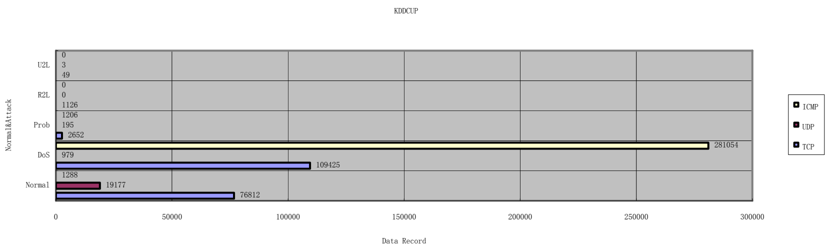

Figure 2. KDDCUP’99 intrusion data protocol classification.

NO	Feature set after Reduction	support	length
1	3,4,6,24,23,24,27,28,31,32,33,36,38	100	13
2	3,4,6,18,23,24, 27,28,31,32,33,36,38	100	13
3	3,4,6,14,23,24, 27,28,31,32,33,36,39	100	13
4	3,4,6,23,24, 27,28,31,32,33,35,36,39	100	13
5	3,4,6,23,24, 27,28,31,32,33,35,36,38	100	13
6	3,4,6,18,23, 24,27,28,31,32,33,36,39	100	13
7	3,4,6,10,23, 24,27,28,31,33,35,36,37,39	100	14
8	3,4,6,23, 24,27,28,31,32,34,35,36,37,39	100	14
9	3,4,6,10,23, 24,27,28,31,33,35,36,37,38	100	14
10	3,4,6,10,12,18,23, 24,27,28,31,32,34,35,36,37,38	100	17
11	3,4,6,10,12,14,23, 24,27,28,31,32,34,35,36,37,38	100	17

Figure 3. TCP reduced characteristic.

NO	Feature set after Reduction	support	length
1	3,5	100	2
2	5,24,34	100	3
3	5,6,36	100	3
4	5,31,36	100	3
5	5,34,35	100	3
6	5,33,36	100	3
7	5,32,33	100	3
8	5,6,32	100	3
9	5,8,36	100	3
10	5,32,34	100	3
11	5,8,32	100	3
12	5,23,33	100	3
13	5,30,34	100	3
14	5,31,32	100	3
15	5,34,36	100	3
16	5,33,34	100	3
17	5,29,34	100	3
18	5,8,29,35	100	4
19	5,8,23,35	100	4
20	1,5,23,34	100	4
21	5,6,30,35	100	4
22	3,33,35,36	100	4
23	5,23,34,40	100	4
24	5,6,29,35	100	4
25	5,8,30,35	100	4
26	5,8,30,33	100	4
27	5,23,31,34	100	4
28	5,24,30,33	100	4
29	5,29,31,33	100	4
30	5,6,23,34	100	4
31	5,30,31,33	100	4
32	5,33,35,36	100	4
33	5,30,31,35	100	4
34	5,6,29,31	100	4
35	5,29,31,35	100	4
36	5,24,33,35	100	4
37	5,24,29,33	100	4

Figure 4. UDP reduced characteristic.

NO	Feature set after Reduction	support	length
1	5,32	100	2
2	5,33	100	2
3	3,23,32,33	100	4
4	8,23,32,33	100	4
5	3,24,32,33	100	4
6	3,24,32,33	100	4
7	8,24,32,33	100	4
8	8,24,33,36,37	100	5
9	8,24,33,34,37	100	5

Figure 5. ICMP reduced characteristic.

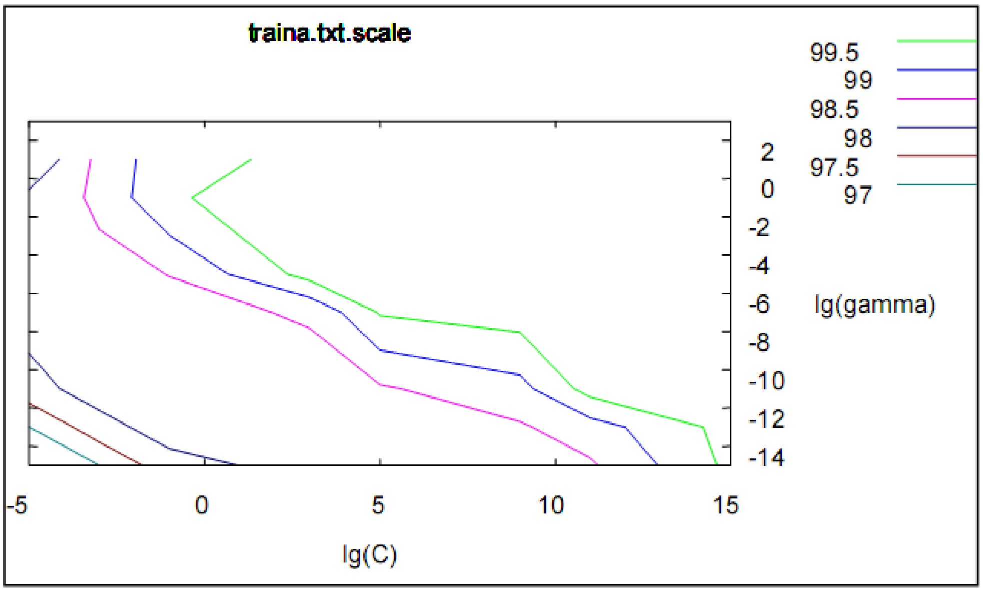

Figure 6. Parameter search of TCP training data