Comparison of Properties between Entropy and χ^2 based Anomaly Detection Method

Shunsuke Oshima*, Takuo Nakashima† and Toshinori Sueyoshi‡
* *ICT Center for Learning Support*
Kumamoto National College of Technology
2627 Hirayama-Shinmachi, Yatsushiro, Kumamoto, Japan
oshima@kumamoto-nct.ac.jp
† *Dep. of Electronics and Intelligent Systems Engineering*
Tokai University
9-1-1 Toroku, Kumamoto, Japan
taku@ktmail.tokai-u.jp

‡ *Graduate School of Science and Technology*
Kumamoto University
2-39-1 Kurokami, Kumamoto, Japan
sueyoshi@cs.kumamoto-u.ac.jp

***Abstract*—As the typical anomaly detection methods using statistics, entropy and χ^2 based method has been researched and reported in terms of their properties for anomaly attacks. In this research, we compare the properties of both methods and discuss the accuracy of detection and the efficiency for different kinds of attacks. Our previous researches have clarified that the source IP address and destination port number are efficient statistical variables to view the anomaly packet property, which lead to detect correctly. In this paper, we propose EMMM method for entropy value and CSDM method of χ^2 value using multi statistical variables. The experiments to verify our proposed methods were conducted using source IP address, destination port number and arriving interval of packets. We could extract the following results. Firstly, EMMM method could decrease the value of False-Positive and False-Negative. Secondly, CSDM method could increase the F-metric, which is the evaluation standard for accurate detection. In the experiments using the same condition of parameters such as probability valuables and window width, CSDM method enlarges the F-metric compared to EMMM method.**

***Keywords*-DoS/DDoS detection, anomaly detection, Entropy, chi-square value, statistical approach**

I. Introduction

The computer systems connected to the Internet are exposed the threats of DoS/DDoS attacks aiming to destroy the server functions. The malicious users hijack the vulnerable PCs connected to the Internet at the first stage of DoS/DDoS attacks, and then generate the attacking PCs called as BOT. The port scan and IP address scan methods are utilized to find the vulnerable PCs. Huge packets are flooding over the LAN in the organizations in these scanning processes. In the second stage, the malicious users send the attacking instructions to a large number of BOTs aiming to attack and paralyze the server functions as the victim of the DoS/DDoS attacks. A huge number of packets are exchanged between BOTs and the servers in this second stage. In both stages, a huge number of packets are flowing over the Internet. We call these malicious packets as the anomaly packets observed on the BOT PC and servers. The method to detect these anomaly packets early is required to sustain the damage of the DoS/DDoS attacks. These anomaly detection methods are required with no learning data to prevent the new diverse anomaly accesses.

We have researched to develop the detection system to split the normal traffics and anomaly traffics based on the statistical methods. In this research, we focus on the statistical method such as the entropy method and χ^2 method. Our proposed method utilizes the arriving time deviation, and also utilizes the source IP address and destination port number. We discussed and compared the detection accuracy of both entropy and χ^2 methods under the same condition.

This paper is organized as follows. Firstly, the previous researches in terms of the χ^2 method and the problems of these researches are described in Section II. Section III describes the proposed method followed by the setting algorithm of the threshold to detect anomaly attacks in Section IV. The experimental method is shown in Section V, Section VI gives the detail of evaluation method in terms of the results, the experimental results are discussed in Section VII. Section VIII gives the summary and discussion for future works.

II. Related work

A. Entropy

The statistic method of anomaly detection methods firstly defines the independent variables based on the packet features such as the source IP address or the destination port number, secondly extracts the symbols in packet header, finally calculate the entropy value and χ^2 value based on the frequency of features. The entropy value increases when each symbol has the uniform frequency. On the other hand, the entropy value decreases when the specific symbol appears with concentration. Using these properties, various

978-0-7695-4458-8/11 $26.00 © 2011 IEEE
DOI 10.1109/NBiS.2011.40

researches have been proposed in relation to the anomaly detection system [1] [2] [3] [4].

We will show the calculation method of entropy value. Let an information source have n independent symbols each with the probability of P_i. Then the entropy H is defined as:

$$H = -\sum_{i=1}^{n} P_i \log_2 P_i \tag{1}$$

In the calculation procedure, firstly all incoming packets put in the time sequence order are split into non-overlapping window with window width (W [packets]). Secondly, the parameters in each packet are extracted. Each probability of the parameter is calculated $P_i = x_i/W$, where x_i is the frequency of the parameter.

B. χ^2 Value

The χ^2 value is calculated using the following equation.

$$\chi^2 = \sum_{i=1}^{B} \frac{(O_i - E_i)^2}{E_i} \tag{2}$$

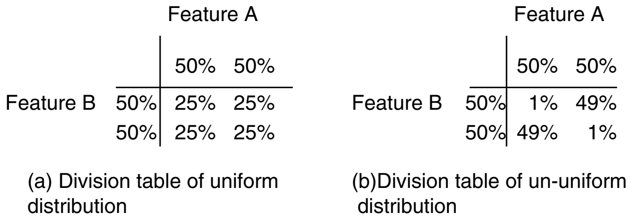
Where B is the number of observed values, E_i is the expected frequency of symbols, and O_i is the observed frequency. This value obeys the $B-1$ degrees of freedom and increases when the difference between the expected frequency and the observed frequency is large. We discussed the previous researches proposing the anomaly detection method using χ^2 value. In the researches [5] [6], firstly the moving averages are calculated based on the frequency of event occurrence in the audit log of the Solaris BMS. The researches showed that the χ^2 values based on the moving average managed to detect the anomaly attacks. In the research [7] [8], the method to detect the abnormal users was proposed from the information of the number of output frequency and pages of printer or the number of access to the URL of the specific home page. On the other hand, the research [1] showed that χ^2 values based on the source IP address of the arrival packets were effective to detect for DDoS attacks.

III. Proposed Method

In previous approach, the probability variable calculated in entropy and χ^2 method is one feature such as the source IP address. Packet header, however, has a lot of fields such as the destination port number, packet length, TTL and flags which have meaningful information to detect attackers. These multiple features are required to analyze in each method to increase the detection accuracy. In this paper, we propose the new entropy and χ^2 method to handle the multiple features.

A. Entropy-based Multidimensional Mahalanobis-distance Method(EMMM)

Firstly, we propose the Entropy based Multi dimensional Mahalanobis distance Method (EMMM), which enable to include the properties of multi features in packet header. Let the Entropy Vector is $\boldsymbol{H}_{t+1}$ at the time $t+1$, also the average entropy vector and variance-covariance matrix is $\bar{\boldsymbol{H}}_{\boldsymbol{t}}$ and $\boldsymbol{\Sigma}_t$ at the time t respectively. The Mahalanobis distance between two entropy vectors $\bar{\boldsymbol{H}}_t$, $\boldsymbol{H}_{t+1}$ is defined as follows.



		Feature A	
		50%	50%
Feature B	50%	25%	25%
	50%	25%	25%

(a) Division table of uniform distribution

		Feature A	
		50%	50%
Feature B	50%	1%	49%
	50%	49%	1%

(b)Division table of un-uniform distribution

Figure 1. An example of the division table for two features.

$$d_m = \sqrt{(\boldsymbol{H}_{t+1} - \bar{\boldsymbol{H}}_t)^T\,(\boldsymbol{\Sigma}_t)^{-1}(\boldsymbol{H}_{t+1} - \bar{\boldsymbol{H}}_t)} \tag{3}$$

We decide whether the packet is attack packet or not using the following equation.

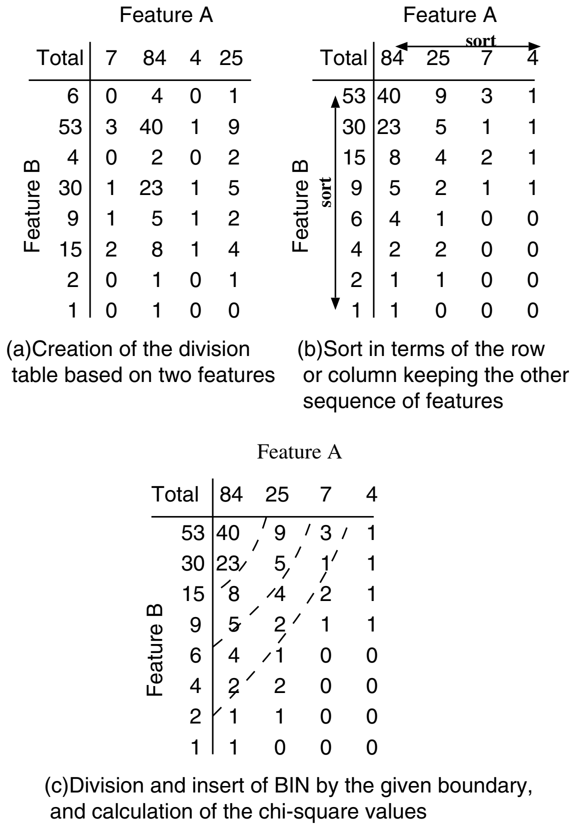
$$d_m > \theta \tag{4}$$

The setting mechanism of θ in equation (4) is described in section IV in detail.

In this paper, we adopt 1)source IP address and time deviation of arriving packets, 2)destination port number and time deviation of arriving packets as the probability variables. Mahalanobis distance is applied to these probability variables.

B. χ^2-based Space Division Method(CSDM)

In the calculation of the χ^2 value using two independent variables, $n \times m$ division table is normally created, and each cell is considered as the observed value O_i. In this case, the created χ^2 values obey the χ^2 distribution with the $(n-1)\cdot(m-1)$ degree of freedom. This method, however, is difficult to be applied to the real traffic flowing over the Internet. It is known [9] that the frequency of each symbol revealing the feature in the packet header such as the source IP address and packet length obeys the power law. This character generates the zero value in the cell when the typical concentrates into the specific cell. We show two distribution tables in Figure 1. Both Figure 1 (a) and (b) are examples of the distributions of 50% in terms of two features. If the distribution obeys the pattern in Figure 1 (a), this division table will be applied to the χ^2 method. On the contrary, if the distribution obeys the pattern in Figure 1 (b), which occurs more likely in the large number of divisions or features, then this division table will not be appropriate to calculate the χ^2 values. We proposed the new mechanism to store data in BIN in Figure 2. The case of two features is explained in Figure 2, and the case of N features is expanded to the N dimension.



Feature A
Feature B
sort
sort

Total	7	84	4	25
6	0	4	0	1
53	3	40	1	9
4	0	2	0	2
30	1	23	1	5
9	1	5	1	2
15	2	8	1	4
2	0	1	0	1
1	0	1	0	0

(a)Creation of the division table based on two features

Total	84	25	7	4
53	40	9	3	1
30	23	5	1	1
15	8	4	2	1
9	5	2	1	1
6	4	1	0	0
4	2	2	0	0
2	1	1	0	0
1	1	0	0	0

(b)Sort in terms of the row or column keeping the other sequence of features

Total	84	25	7	4
53	40	9	3	1
30	23	5	1	1
15	8	4	2	1
9	5	2	1	1
6	4	1	0	0
4	2	2	0	0
2	1	1	0	0
1	1	0	0	0

(c)Division and insert of BIN by the given boundary, and calculation of the chi-square values

Figure 2. Calculation of the χ^2 values based on two features

The proposed calculation procedure of the χ^2 method is explained as follows.

(a) Let N to be the number of features. The N order metrics with the index of each symbol showing the specific feature is generated. Each symbol is extracted from the packet with the time sequence order in the window. The frequency of the matrix cell associated to the symbol is incremented, and the total frequency of each feature is counted up.

(b) After finishing of the total calculation of all packets in the window, the frequencies of each feature are sorted keeping the position of the row and the column. The symbols such as the source IP address in anomaly packets tend to distribute or concentrate. This property makes the position close to each other after the sorting process.

(c) The total evaluation value s is calculated from the order of each feature. The packets are divided in to each BIN based on the s. Finally, the χ^2 values are calculated using the observation frequency of the packets in each BIN.

We indicate the distribution procedure of BIN as follows. Firstly, the evaluation value s_j of each symbol j is calculated using the following equation.

$$s_j = \sum_{k=1}^{N} \frac{r_{j,k}}{m_k} \tag{5}$$

Where N is the number of features, $r_{j,k}$ is the order of the packet j in terms of the feature k, and m_k is the number of symbols of feature k. For example, we consider the condition that the number of symbol of the feature A and B is 4 and 8 respectively. When the order of packet j of the feature A and B is 2 and 3 respectively, we are able to get the $s_j = 2/4 + 3/8 = 0.875$. After this calculation, if the condition of $d_{i-1} < s_j \leq d_i$ is correct, the applicable packet push in the BIN$_i$, where d_i indicates the boundary of BIN$_i$. For example, when the boundary values have $d_1 = 0.5, d_2 = 0.75, d_3 = 1.0$, if s_j is smaller than or equal to 0.5, then the packet is put in the BIN$_1$; if that is smaller than or equal to 0.75 the packet is put in the BIN$_2$, if that is larger than 1.0, the packet is put in the BIN$_4$. In this case, $s_j = 0.875$ causes to put in the BIN$_3$.

The change of the calculation reduces the increase of the time and space complexity of the sort in terms of the large N in our proposed method. Firstly, the frequency of symbol for each feature in the window is sorted leading that the order of each symbol is decided. Secondly, the searching table for the order is created. The order of symbol is decided using the order searching table with picking the packet from the front of the window. These processing create the calculation of the evaluation value s_j, and make the count value of BIN increase. This new processing procedure reduce the N order matrix.

In this paper, we adopted the source IP address and the destination port number as the first feature which were proved the effectiveness for the anomaly detection method in the research [1]. As the second feature, we adopted the time difference Δt of the arrival time, which is the difference of the time interval of the arrival packets. The packets generated by the accesses of person such as Web access, ssh and telent does not normally consists of the constant Δt. We are able to distinguish the DoS packets and the normal packets in comparing the two dimensional distribution for both packets with the Δt, even if the distribution similar to the DoS attacks is consisted of the normal packets using the source IP address only.

1) Proposed decision method of the BIN boundary: The boundary of BIN depends on the number of features N, the number of symbols of the IP address or the port number in the features, window width W and the number of BIN leading to the difficulty of decision of the boundary. In addition, the condition of BIN requires that the number of packets in each BIN exceeds five in the normal condition. We propose the boundary $d_i(t)$ of BIN to be calculated based on the following equation.

$$d_i(t) = d_i(t-1) \cdot \left(1 - \eta \frac{\bar{p_i}(t-1) - E_i}{E_i}\right) \tag{6}$$

where the expected frequency E_i is constant based on W/B, and then η is the coefficient of the window variation. This equation constantly make the average packet frequency $\bar{p_i}(t)$ in each BIN divided by the boundary value $d_i(t)$ close to the expected value. The average packet frequency $\bar{p_i}(t)$ is defined as the exponential moving average as follows.

$$\bar{p_i}(t) = \lambda p_i(t) + (1-\lambda)\bar{p_i}(t-1) \tag{7}$$

where λ is called the smoothing coefficient positioning between 0 and 1.

In the researches [1] [5] [6], the expected frequency E_i is dynamically calculated from the constant BIN width. This calculation uses the exponential moving average based on the occurrence of previous events to deal with the traffic variation in the term such as one day or one week. On the contrary, the expected frequency E_i is fixed, and dynamically varies the bin boundary $d_i(t)$ in this paper. Therefore, the proposed equation guarantees that the expected frequency exceeds five and traces the dynamic variation of traffics. The boundary of BIN is dynamically calculated based on the number of symbols, window width W, the number of BIN B and so on leading that the calculation process to decide the BIN width does not need at the parameter changes.

IV. Threshold value

The threshold values of χ^2 method and mahalanobis distance are statistically important parameters. We describe the policy and mechanism to decide the threshold value. In addition, the real values of threshold used in the experiments are explained.

A. The threshold of EMMM

The threshold of EMMM should be determined by the definition of Mahalanobis distance. It is generally known that the n order squared Mahalanobis distance d_m^2 obeys the χ^2 distribution with the degree of freedom n. Using these distribution property, we introduce the threshold value θ in Mahalanobis distance as the corresponding value of 2σ or 3σ in the normal distribution. Let $\chi^2(\alpha, n)$ be the upper probability α of χ^2 distribution with the degree of freedom n, and $P(Z > p)$ be the probability of standard normal distribution with the standard deviation is p or more. The threshold θ_1, θ_2 associated to the 2σ of the normal distribution are represented 1st order and 2nd order Mahalanobis distance respectively. Both values are defined as follows.

$$\begin{aligned}\theta_1 &= \sqrt{\chi^2(P(Z>2)+P(Z<-2),1)} = 2.000\\ \theta_2 &= \sqrt{\chi^2(P(Z>2)+P(Z<-2),2)} = 2.486\end{aligned}$$

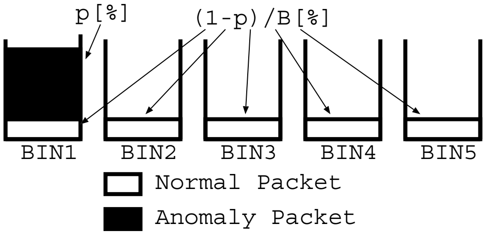


Figure 3. Condition of BIN with the $p[\%]$ anomaly packets (DoS)

The correctness of this equation will be confirmed by $\sqrt{\chi^2(2 \cdot P(Z > p), 1)} = p$. In this paper, we use θ_1 and θ_2 are applied as the thresholds.

B. The threshold of CSDM

When χ^2 value makes the maximum value, features of all packets in window width (W) will be the same and all packets are sorted in one BIN. In our proposed CSDM, the BIN width are automatically arranged to be the same interval using the equation (6) to enter the same number of packets on each BIN. Therefore, the expected frequency E of these equation will be set $E = W/B$, and maximum χ^2 value χ^2_{max} is expressed as follows.

$$\chi^2_{max} = \frac{(W-E)^2}{E} + \frac{(0-E)^2}{E} \cdot (B-1) \tag{8}$$

From this equation, the size of χ^2_{max} changes largely depending window width and E. The reason, however, depends on the decision process to be one threshold. In this paper, we normalize to range [0:1] by calculation that χ^2 value is devided by χ^2_{max}. This manipulation could set the threshold value not to depend on the window width and the number of BIN.

We assume that the $p[\%]$ of all packets in window width W are DoS attacks. The calculation value χ_p^2 is revealed the following equation (9).

$$\begin{aligned}\chi^2_{\mathrm{p}} &= \frac{(pW + \frac{1-p}{B}\cdot W - E)^2}{E}\\ &\quad + \frac{\left(\frac{1-p}{B}W - E\right)^2}{E} \cdot (B-1)\end{aligned} \tag{9}$$

χ_p^2/χ^2_{max} does not depend on window width (W) and the number of BIN (B) as follows.

$$\begin{aligned}\frac{\chi_p^2}{\chi^2_{max}} &= \frac{\left(pW + \frac{1-p}{B}\cdot W - \frac{W}{B}\right)^2 +}{\left(W - \frac{W}{B}\right)^2 + \left(\frac{W}{B}\right)^2 \cdot (B-1)}\\ &\qquad \underline{\left(\frac{1-p}{B}\cdot W - \frac{W}{B}\right)^2 \cdot (B-1)}\\ &= \frac{(pB + (1-p) - 1)^2 +}{(B-1)^2 + (B-1)}\\ &\qquad \underline{((1-p)-1)^2 \cdot (B-1)}\end{aligned}$$

$$= \frac{p^2(B-1)+p^2}{(B-1)+1} = p^2 \tag{10}$$

This equation means that if we want to detect anomaly packet more than $p[\%]$, the threshold should be set p^2 to calculate the $\chi^2/\chi^2_{\max}$. In this paper, we set $p = 0.5$ to detect DoS attacks consisting the half of all packets meaning that the threshold of $\chi^2/\chi^2_{\max}$ is set $p^2 = 0.25$.

V. Experimental method

The dataset of this experiment adopts the DARPA1999 [10], which is created by the MIT in 1999. DARPA1999 has a lot of diverse anomaly attacks, and gives the pcap format file including the attacking packets with no processing format. This dataset is profitable to our proposed method utilizing the multi features in packet header as the probability function. In addition, this dataset provide the label which could distinguish whether the packet is anomaly or normal packet causing that the dataset is profitable to evaluate the detection accuracy using F-measure.

DARPA1999 provide the dataset for five weeks. In this research, we used the outside packets having a lot of diverse anomaly attacks. Firstly, the second week (March 8th to March 12th in 1999) dataset, which has no attacking data, is used to learn as the expected value in χ^2 calculation. Secondly, the fourth week (March 29th to April 2nd) and fifth week (April 5th to April 9th) dataset have a lot of diverse anomaly attacks. For this reason, we used this dataset to evaluate the accuracy of our proposed detection method.

In this paper, we selected the source IP address(`srcip`) and the destination port number(`dstport`) as the probability variable, since two parameters are effective to detect anomaly packets in previous research [1]. In addition, we selected other two features, packet interval time with the same source IP address (`dt-srcip`) and packet interval time with the same destination port number(`dt-dstport`), which could able to enhance the detection accuracy.

In previous research [11], we had the good detection accuracy using the source IP address with the window width $W = 10,000$ and using the destination port number with the window width $W = 2,000$. In this paper, we focus on these two window width, and compare the EMMM and CSDM. In the CSDM, we set the number of BIN $B = 5$, and the BIN arranging parameters; $\eta = 0.05$ and $\lambda = 0.2$.

The maximum values of EMMM and CSDM have a large fluctuation depending on the type of probability variables and window width. The normalization of EMMM is conducted to set the maximum value is 1, and that of CSDM is conducted to set that the χ^2 values are divided by the equation (8).

VI. Evaluation equation

Evaluation metrics for the detection methods have adopted the False-Positive (FP) and False-Negative (FN). Both metrics have the tradeoff meaning that if a detection method could reduce the FP, it will increase the FN, and vice versa. In this paper, F-measure, which is combined both metrics FP and FN, is adopted to evaluate totally each method. The definition of F-measure is defined as follows.

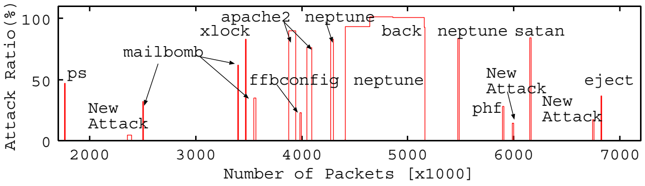


Figure 4. Distribution of attacking packets

$$F = \frac{1}{\frac{1}{2}(\frac{1}{R}+\frac{1}{P})} = \frac{2RP}{R+P} \tag{11}$$

Where R is the Recall and P is the Precision defined as following equations.

$$R = \frac{tp}{tp+fn} \tag{12}$$

$$P = \frac{tp}{tp+fp} \tag{13}$$

Where tp is the number of True-Positive (TP); the number of the detection for the attacking packets, and fp and fn is the number of FP and FN respectively. Both metrics R and P have the value between 0 and 1. If metrics is close to 1, the accuracy of detection method is good. F-measure is also has the value between 0 and 1, and $F = 1$ enable to detect accurately.

VII. Results of experiments

Figure 4 shows the all attacking packets labeled in DARPA1999 dataset with the label name. The vertical axis indicates the attack ratio [%]; the number of attacking packets is divided by the number of all packets in the window. The number of anomaly packets in DARPA1999 has the large range between 50 and 50,000. In this research, we selected and illustrated the anomaly packets exceeding the 2,000 packets classified into 19 attacks. We have the two reasons to choose the 2,000 packets. Firstly, the window width is 10,000 in both EMMM and CSDM. Secondly, we set the threshold to detect 20[%] or more anomaly attacks based on the total packets in CSDM.

A. Results of experiments based on the source IP address

Figure 5 and 6 show the EMMM and CSDM values respectively based on the source IP address. Both figures include three figures; EMMM values (1) based on the source IP address(`srcip`), (2) based on the interval time with the same source IP address(`dt-srcip`), and (3) based on both metrics. Both EMMM and CSDM could detect anomaly packets such as apache2 and neptune. On the other hand, these methods could not detect application level anomaly packets such as mailbomb, since mailbomb packet

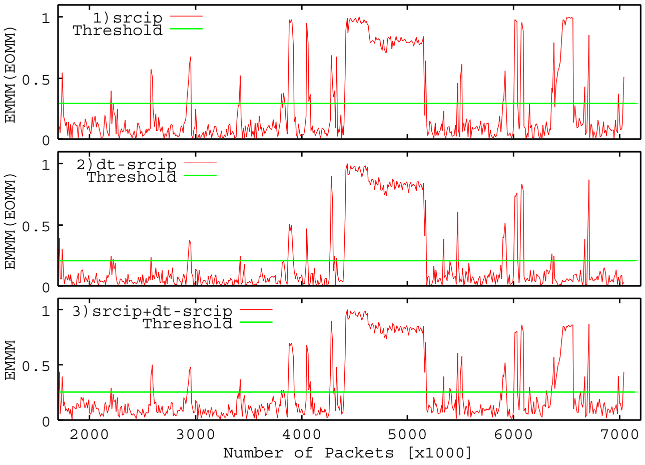


Figure 5. EMMM values based on the source IP address ($W = 10,000$)

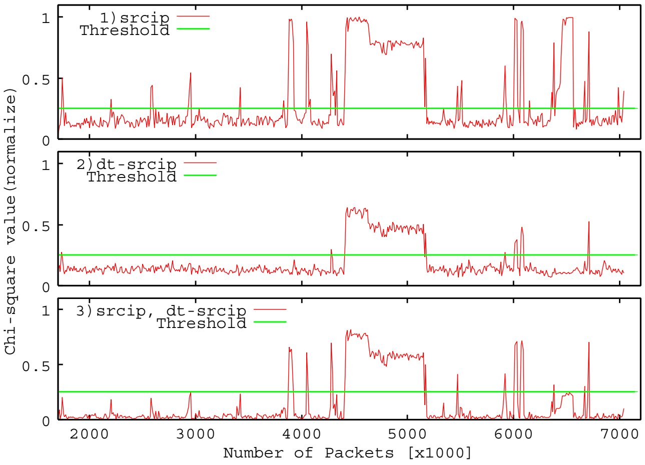


Figure 6. CSDM values based on the source IP address ($W = 10,000$)

Table I
METRIC VALUES OF EMMM BASED ON THE SOURCE IP ADDRESS ($W = 10,000$)

Random variable	fp	fn	R	P	F
1) `srcip`	52	9	0.904	0.620	0.736
2) `dt-srcip`	31	12	0.872	0.726	0.792
3) `srcip+dt-srcip`	58	7	0.926	0.600	0.728

sequence is the similar to the SMTP packet sequence. These application level anomaly attacks is difficult to detect using the statistic method. The some large values in both EMMM and CSDM do not indicate the anomaly attacks, since the human activating protocol such as ssh, ftp and telnet appear in such area.

Table I and II show the experimental results of EMMM and CSDM respectively. F-measure of `srcip+dt-srcip` has good accuracy compared to the one of srcip and `dt-srcip` in CSDM due to the decrease of FP. The information of time difference could distribute the concentrated access of telnet an ssh in case of the source IP address, and detain the CSDM values. On the other hand, F-measure of `dt-srcip` has the best accuracy compared to the one of `srcip+dt-srcip` in EMMM. The degradation of F-measure of `srcip+dt-srcip` is caused by FP. The number of symbols makes the large value in the case of time difference, which generates the large variation of entropy value, and finally causes the large number of FP. To reduce the number of symbols in the case of time difference, the aggregation methods such that the neighbor symbols are aggregated to one symbol are required in this EMMM.

Table II
METRIC VALUES OF CSDM BASED ON THE SOURCE IP ADDRESS ($W = 10,000$)

Random variable	fp	fn	R	P	F
1) `srcip`	56	8	0.915	0.606	0.729
2) `dt-srcip`	10	19	0.798	0.882	0.838
3) `srcip+dt-srcip`	13	12	0.872	0.863	0.868

In addition, the relation between two probabilistic variable in both EMMM and CSDM has the difference. For example, the relation between `srcip` and `dt-srcip` for one packet does not affect to the calculation of entropy value in EMMM. On the other hand, CSDM keeps the relation of each packet, and calculates the χ^2 values from the second order matrix. EMMM will lose the relation between two features and affect to the accuracy of detection.

One of problem in EMMM, the equation of entropy value consists of the simple summation of probabilities. Table III shows the comparison between entropy and χ^2 values under different conditions. The "Normal" in this table means the values under normal packets. Since the number of packets obeys the power law behavior, the packet number under "Normal" case is calculated. The entropy value under "Normal" case is 1.371 meaning that the decision whether the packet is normal or anomaly is decided using this entropy value. On the other hand, χ^2 method calculates the expected frequency using the packets in BIN under normal condition. The DoS and DDoS attacks could be detected by the differences between normal condition and attacking condition on both entropy and χ^2 value. On the other hand, if the attacking pattern is (DoS+DoS) meaning that the two DoS attacks are overlapping; χ^2 value has a large value even if the entropy value is almost same in the case of normal. This indicates that entropy method is difficult to detect the multiple overlapped DoS attacks.

B. Experimental results based on the destination port number

We conducted the same experiments relating the destination port number. Figure 7 and 8 show the result of EMMM and CSDM respectively. For both experiments, we show the results of three patterns; 4) `dstport`, 5) `dt-dstport` and 6) `dstport+dt-dstport`. In addition, Table IV

Table III
COMPARISON BETWEEN ENTROPY AND χ^2 VALUES UNDER DIFFERENT CONDITIONS

Condition	Number of Packets BIN1	 BIN2	 BIN3	 BIN4	H	χ^2
Normal	3	9	24	64	1.371	0.000
DoS	0	0	0	100	0.000	56.25
DDoS	25	25	25	25	2.000	213.6
DoS+DoS	2	6	46	46	1.387	26.56
DoS+DDoS	10	10	10	70	1.357	25.17

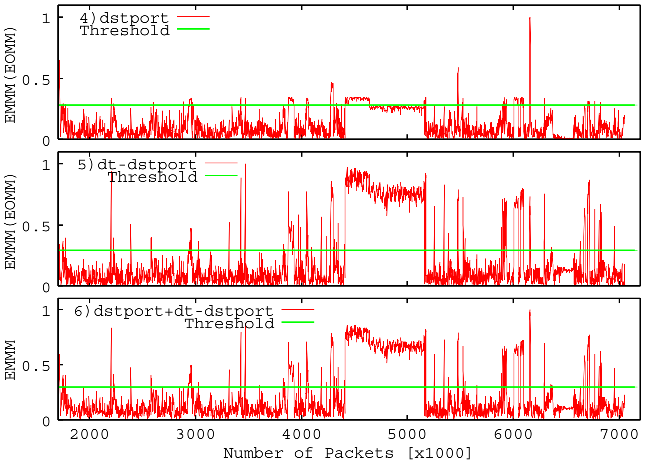


Figure 7. EMMM values based on the destination port number ($W = 2,000$)

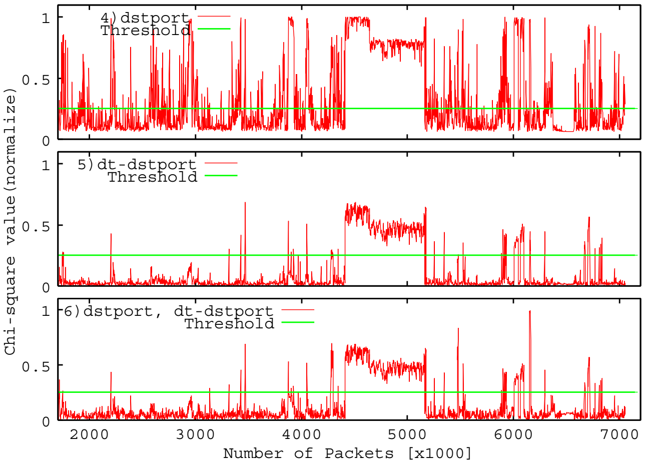


Figure 8. CSDM values based on the destination port number ($W = 2,000$)

and V illustrate the number of FP and FN, and F-measure in EMMM and CSDM respectively. In the experiments of 4)destport, the number of FP is small but the number of FN is very large in EMMM. The small window width such as $W = 2,000$ in these experiments cause the instability of entropy value. This fluctuation of entropy value leads to the instability of EMMM leading to the increase of FN. On the other hand, experiment 6) `dstport+dt-dstport` make FN reduce, however FP increase. As the same of the experiments of srcip, the increase of the kind of symbols on the time difference causes the increase of variation of entropy values. In addition, the relation between the probability variables affected the results.

Table IV
METRIC VALUES OF EMMM BASED ON THE DESTINATION PORT NUMBER ($W = 2,000$)

Random variable	fp	fn	R	P	F
4) `dstport`	77	407	0.304	0.698	0.424
5) `dt-dstport`	119	150	0.744	0.785	0.764
6) `dstport+dt-dstport`	133	145	0.752	0.768	0.760

Table V
METRIC VALUES OF CSDM BASED ON THE DESTINATION PORT NUMBER ($W = 2,000$)

Random variable	fp	fn	R	P	F
4) `dstport`	490	112	0.809	0.491	0.611
5) `dt-dstport`	79	188	0.679	0.834	0.748
6)`dstport+dt-dstport`	82	175	0.701	0.833	0.761

The experimental results of CSDM in table V shows that the F-measure of 6) has the largest value, which is the same result of srcip. Compared to 4), 6) reduces the large number of FP. The small number of FP on the experimental results of 5) affected the calculation of CSDM, which keeps the relation between probability variables, leading to the decrease of the number of FP.

VIII. CONCLUSION

We proposed anomaly detection method based on the entropy based method called EMMM and χ^2 based method called CSDM utilizing the multiple features. In both methods, we measured the values of F-measure of anomaly detection combining the probabilistic variables such as source IP address and time difference, destination port number and time difference.

From the results of experiments, EMMM and CSDM could detect the neptune and satan attacks, but they could not detect the application level attacks such as mailbomb. EMMM sometime miss detect the normal ftp, telnet and ssh tending to concentrate in terms of srcip or dstport. In addition, EMMM caused the large fluctuation of entropy values in the case of the large number of symbols such the time difference, and led the increase of FP, finally degraded the F-measure. These results show that direct utilization of the time difference is not appropriate in the entropy method; we should cluster some area of symbols.

On the other hand, CSDM could revise the F-measure to reduce the number of FP such as telnet keeping the relation in one packet. These results show that CSDM is prefer to use the multiple probabilistic variables including the time difference compared to EMMM. In the results of experiments using two probabilistic variables with the same condition such as window width, CSDM has the large

F-measures than that of EMMM. As the results, CSDM effectively works at the case of using two probabilistic variables.

In the next step, we should arrange the number of symbols and the value of the probabilistic variable. In our method, a lot of probabilistic variable such as packet length, TTL and flag are available to each method. We should verify both EMMM and CSDM using such a lot of different variables. We especially focus on the efficiency of time differences in both EMMM and CSDM under the same condition.

References

[1] L. Feinstein, D. Schnackenberg, R. Balupari, and D. Kindred, "Statistical approaches to ddos attack detection and response," *Proceedings of DARPA Information Survivability Conference and Exposition*, vol. 1, pp. 303–314, Apr. 2003.

[2] K. Lee, J. Kim, K. H. Kwon, Y. Han, and S. Kim, "Ddos attack detection method using cluster analysis," *Expert Systems with Applications*, vol. 34, pp. 1659–1665, Apr. 2008.

[3] G. Nychis, V. Sekar, D. G. Andersen, H. Kim, and H. Zhang, "An empirical evaluation of entropy-based traffic anomaly detection," in *Proceedings of the 8th ACM SIGCOMM Conference on Internet measurement*, Vouliagmeni, Greece, Oct. 2008, pp. 151–156.

[4] A. Wagner and B. Plattner, "Entropy based worm and anomaly detection in fast ip networks," in *Proceedings of the 14th IEEE International workshops on Enabling Technologies, Infrastructure for Collaborative Enterprise*, Linköping, Sweden, Jun. 2005, pp. 172–177.

[5] N. Ye and Q. Chen, "An anomaly detection technique based on a chi-square statistic for detecting intrusions into information systems," in *Quality and Reliability Engineering International 17*, 2001, pp. 105–112.

[6] N. Ye, S. M. Emran, Q. Chen, and S. Vilbert, "Multivariate statistical analysis of audit trails for host-based intrusion detection," *IEEE Transactions on Computers*, vol. 51, no. 7, pp. 810–820, Jul. 2002.

[7] R. Goonatilake, A. Herath, S. Herath, S. Herath, and J. Herath, "Intrusion detection using the chi-square goodness-of-fit test for information assurance, network, forensics and software security," *Papers of the Fourteenth Annual CCSC Midwestern Conference and Papers of the Sixteenth Annual CCSC Rocky Mountain Conference*, vol. 23, no. 1, pp. 255–263, Oct. 2007.

[8] B. Zhou, Q. Shi, and M. Merabti, "Intrusion detection in pervasive networks based on a chi-square statistic test," in *Proceedings of the 30th Annual International Computer Software and Applications Conference(COMPSAC06)*, Chicago, Illinois, Sep. 2006, pp. 203–208.

[9] M. Falutsos, P. Faloutsos, and C. Faloutsos, "On power-law relationships of the internet topology," in *Proc. of ACM SIGCOMM*, Aug. 1999, pp. 251–262.

[10] MIT, "Darpa intrusion detection evaluation data set," http://www.ll.mit.edu/mission/communications/ist/index.html, Lincoln Laboratory.

[11] S. Oshima, T. Nakashima, and T. Sueyoshi, "Anomaly detection using chi-square values based on the typical features and the time deviation," in *25th International Conference on Advanced Information Networking and Applications(AINA2011)*, Mar. 2011, p. will be prepared.