Minimal Triangle Area Mahalanobis Distance for Stream Homogeneous Group-based DDoS Classification

Yudha Purwanto[1], Kuspriyanto[2], Hendrawan[3], and Budi Rahardjo[4]

School of Electrical Engineering and Informatics, Bandung Institute of Technology
Jl. Ganesha no. 10, Bandung 40132, Indonesia
[1]omyudha@telkomuniveristy.ac.id, [2]kuspriyanto@lskk.ee.itb.ac.id, [3]hend@stei.itb.ac.id,
[4]br@paume.itb.ac.id

Abstract: An Intrusion Detection System (IDS) which implement a group-based classification algorithm, theoretically has the benefit of higher accuracy. Unfortunately, higher accuracy only achieved if the observed group is homogeneous from a certain distribution. Recently, a distributed denial of service (DDoS) attack consists of multiple botnets which produce multi types of traffic in one attack session. It makes the IDS suffers from decreasing accuracy as the increasing heterogeneity within the observed group. To address the problem, we propose homogeneous grouping algorithm based on triangle area Mahalanobis distance to support IDS which implement group-based data analysis. First, the Mahalanobis distance measurement was used to construct homogeneous groups. Then, the covariance matrix of each homogeneous group was classified using a decision tree classifier. Classification performance was evaluated using known KDDCup 99 dataset. The results pointed out that the used of homogeneous grouping algorithm improve the classification performance for natural and mixed random DDoS traffic.

Keywords: Intrusion detection system, classification, distributed denial of service, Mahalanobis distance, covariance, decision tree.

1. Introduction

The research in Distributed Denial of Service (DDoS) detection system is very important in the security system. It is because the DDoS is not only affected the target but almost all users of the network [1]. As growing advances in internet technology and scale, the DDoS attack scale is also getting larger which consists of multi botnets in one attack session [2]. This situation known as random multi botnet scenario in a DDoS attack as the multi botnet will result in a random type of DDoS traffic at targeted attack.

Traffic anomaly-based detection is a popular method developed in a DDoS detection system, besides a signature-based detection. The capability focus of IDS anomaly detection research has been branching out from anomaly detection to classification, prevention and response action [3, 4, 5]. There are many features and methods which have been proposed to achieve higher IDS accuracy. Aggregate traffic features are commonly proposed to detect traffic anomaly, such as [6, 7, 8, 9]. By the used of machine learning such as clustering method, the system can determine clusters of traffic such in [10, 11, 12], but the output has no concern about the types of an anomaly of the formed groups. The classification has better used in security system such in [13, 14, 15], a system can single out the specific anomalous packet or connection and determine the known types of anomaly. Most of the research was proposed in one-by-one data analysis, such as in [16, 17, 18].

A different approach has been proposed by research in [19], which proposed group-based data analysis by the used of a covariance matrix. From the theoretical and simulation analysis, it has proved that group-based data analysis will achieve a higher probability of correctly classifying data. From the experimental result utilizing the KDDCup 99 dataset [20], the classification accuracy of the IDS achieves 99,98 % with zero false positive rate.

Unfortunately, the accuracy of group-based data analysis decreased along with the increasing mixture of traffic in a group. From the theoretical analysis, the higher accuracy of group-based

Received: August 25th, 2016. Accepted: June 3rd, 2018
DOI: 10.15676/ijeei.2018.10.2.12

data analysis only achieved when the data in a group is completely coming from the same distribution. In practice, the traffic may come from either distribution independently [17]. There is no guarantee that the natural sequence of data arrival will be automatically homogenous from a certain distribution such as in [19] experiment. Thus, it is very dangerous when a random multi botnet DDoS attack was straightly analyzed in IDS. It is because a random multi botnet scenario will be resulting mixture types of traffic within observed groups. Preliminary study shows that group-based classification accuracy decreased along with the decreasing purity of the group in a covariance feature space [21].

Our research focused to address the problem. We proposed stream homogeneous grouping algorithm to construct homogeneous groups for a group-based DDoS classification system. Constructing a homogeneous group was a key to gain high accuracy in group-based analysis. The minimal triangle area Mahalanobis distance was used to assign stream data into homogeneous groups which will enter the group-based classification process. We evaluated our system compared to a well established group-based classification system with no homogeneous grouping algorithm in testing phase [19]. The evaluation results show a significant improvement. We present our research contributions as follow. First, we proposed stream homogeneous group-based classification framework for random multi botnet DDoS attack in complete intrusion prevention system (IPS) sequence. It started by generating known classes in the training phase, constructing homogeneous groups, classification of a monitored group, and taking certain prevention action according to observed attack. Second, we proposed the use of minimal triangle area Mahalanobis distance for homogeneous grouping algorithm. This approach provided a new tool for homogeneous grouping and proved to provide high homogeneous grouping precision for multi botnet scenario. Third, we contribute to the development of group-based classification system by the use of homogeneous grouping algorithm. From the theoretical and simulation results analysis, it has improved the group-based classification performance for the possible real traffic stream in practice.

The rest of this paper is organized as follows: Section 2 review several related research on traffic anomaly detection, followed by Section 3 which presenting theoretical background, research design and details of the evaluation process. Furthermore, Section 4 shows our experimental results and analysis. Finally, the conclusion and future research directions are presented in Section 5.

2. Related Work

Types of anomaly detection research, especially types of DDoS detection research has become an important task in the security system. It has to detect and prevent the system from high traffic rate which affecting all participants in the system. Several methods have already been proposed in previous research such as statistic, machine learning, information theory, etc, and also the combination of those. Most of them pursued high classification accuracy, by accurately differentiate anomalous and normal traffic.

Aggregate traffic features mostly used for anomaly detection research, as it faster and took lower amounts of memory than per-flow traffic analysis. There are several proposed features and methods to detect anomaly traffic. Statistical analysis mostly used in anomaly detection research such as traffic volume deviation analysis [6], aggregate traffic statistic for bivariate parametric detection and sequential probability ratio test [7], statistical inference and α-stable model [8], and non-parametric statistical analysis [9, 22, 23]. Similarity distance method in information theory also used in research such as TCP flag for entropy and Mahalanobis distance in [24].

High classification accuracy was achieved in [19] by the used of decision tree classification on homogenous groups with covariance matrix analysis. The research proposal was able to theoretically proved that group-based analysis has a higher classification accuracy. But for heterogeneous traffic, the accuracy of group-based data analysis decreased and became vulnerable to attack. It is dangerous as one vulnerability influences other vulnerability [25] which lead to higher response cost such as in [26]. It stated in [17] that single data classification will result in a higher accuracy as a group not come from one distribution. It proved by multivariate

analysis with Triangle Area Map (TAM) of features flow sequence. Triangle area first proposed to provide more discriminative features in [27, 18]. These research occupied the Euclidean distance in a triangle area of features. It then was researched further by the use of the Mahalanobis distance in [28, 17]. Research results show high accuracy even in heterogeneous data. But unfortunately, these single data analysis does not pay much attention in types of DDoS classification.

3. Research Method

A. Group-based Classification

Group-based data analysis has been researched in [19], which stated that the probability of correctly classifying a group of data was higher than one-by-one data. Stream data come in sequence of data, $X = [x^1\ x^2 \dots x^t]; t = 1 - \infty$. For multi features anomaly detection, each x^t will be identified by $x^t = \left[x_{f1}^t\ x_{f2}^t\ x_{f3}^t\ \dots\ x_{fp}^t\right]^T$. In two class classification case $C(i); i = 2; C1$ and $C2$ both have classification profile as $(\mu C1, \sigma^2 C1)$ and $(\mu C2, \sigma^2 C2)$. When data X enter the system, the classifier will classify incoming x^t to $C1$ or $C2$ with the probability of correctly classify each x^t as $Pi(x)$, which are

$$P_1 = \int_{-\sim}^{\mu^*} \left(\frac{1}{\sqrt{2\pi}\sigma C1}\right) \exp[-\frac{1}{2}(\frac{x-\mu C1}{\sigma C1})^2]\ dx \tag{1}$$

$$P_2 = \int_{\mu^*}^{\sim} \left(\frac{1}{\sqrt{2\pi}\sigma C2}\right) \exp[-\frac{1}{2}(\frac{x-\mu C2}{\sigma C2})^2]\ dx \tag{2}$$

where the threshold is defined by

$$\mu^* = \mu C1 \,.\frac{\sigma C2}{\sigma C1 + \sigma C2} + \mu C2 \,.\frac{\sigma C1}{\sigma C1 + \sigma C2} \tag{3}$$

In group data analysis, the analysis was done over a group of m data which is $X_m = [x^1\ x^2\ x^3 \dots x^m]$. In two class classification case, $C(i); i = 2$; each $C1\ and\ C2$ are identified by the sample mean of each homogeneous group profile *($\mu C1$, $\frac{1}{m}\sigma^2 C1$)* and *($\mu C2$, $\frac{1}{m}\sigma^2 C2$)*. Traffic with self-similar and long-range dependent was assumed stationary with a sample variance of $\sigma^2 x(m) = \sigma^2 m^{-\beta}$ such as in [29, 30]. Thus, the probability of correctly classify X_m to $C1$ or $C2$ are

$$\mathrm{Q_1} = \int_{-\sim}^{\mathrm{z}^*} \left(\frac{1}{\sqrt{2\pi}\ \sigma/\sqrt[\beta]{\sqrt{\mathrm{m}}}}\right) \exp[-\frac{1}{2}(\frac{\mathrm{x}-\mu}{\sigma/\sqrt[\beta]{\sqrt{\mathrm{m}}}})^2]\ \mathrm{dx} \tag{4}$$

$$\mathrm{Q_2} = \int_{\mathrm{z}^*}^{\sim} \left(\frac{1}{\sqrt{2\pi}\ \sigma/\sqrt[\beta]{\sqrt{\mathrm{m}}}}\right) \exp[-\frac{1}{2}(\frac{\mathrm{x}-\mu}{\sigma/\sqrt[\beta]{\sqrt{\mathrm{m}}}})^2]\ \mathrm{dx} \tag{5}$$

where the threshold is defined by

$$z^* = \mu C1 \,.\frac{\sigma C2}{\sigma C1 + \sigma C2} + \mu C2 \,.\frac{\sigma C1}{\sigma C1 + \sigma C2} \tag{6}$$

However, according to [17], there was no guarantee that incoming traffic came from one distribution. In a random multi botnet attack, each botnet comes from different homogeneous profiles $C(i)$. So, in a group data X_m (group with m number of data) will consist of a mixture of each botnet data x_i and generate the heterogeneous group. According to i number of classes, the probability of correctly classify X_m by the used of group-based classification degrade to $\frac{1}{i^m}\ Q_i(X_m)$, as Q_i achieved from homogeneous profiles $C(i)$. The central limit theorem can be utilized to extract statistical profiles of a mixed group and create heterogeneous group profiles. But in supervised learning, it is hard to construct mixed group profiles as there will be i^m mixing combination. Thus, it is more sufficient to utilize homogeneous profiles $C(i)$.

To utilize group-based classification in a real natural environment, we have to minimize the i^m problem; which is a homogeneity problem; to achieve a high probability of classifying $Q_i(X_m)$ accurately. The heterogeneous group can be viewed as mixed types of distribution in a mixture model. In term of supervised learning, known distribution parameters in Bayesian rule felicitous utilized to perform homogeneous grouping in a mixture model. By distance calculation of X to each known group profiles $C(i)$; $y_j = dist(x, Ci)$; we could construct stream homogeneous groups G_i from X. Each distance measurement y_j will construct statistical Y_j profiles in $(\mu Y_j, \sigma Y_j)$. In the two distributions mixture case; distance profile of x being in group G_1 is $Y_1 = (\mu Y_1, \sigma Y_1)$, and x being in G_2 is $Y_2 = (\mu Y_2, \sigma Y_2)$. For each distance x to $C(i)$; which is y_j; we can estimate the probability of x being in any group G_i such as

$$P_1(G_1|y_j) = \frac{P(y_j|G_1)P(G_1)}{P(y_j|G_1)P(G_1) + P(y_j|G_2)P(G_2)} \quad (7)$$

$$P_2(G_2|y_j) = \frac{P(y_j|G_2)P(G_2)}{P(y_j|G_1)P(G_1) + P(y_j|G_2)P(G_2)} \quad (8)$$

where

$$P(y_j|G_i) = \int_{-\sim}^{\mu^*} \left(\frac{1}{\sqrt{2\pi}\sigma Y_j}\right) \exp[-\frac{1}{2}(\frac{y - \mu Y_j}{\sigma Y_j})^2]\, dy \quad (9)$$

As y_j getting smaller, the exponential component in $P(Y_j|G_i)$ will get lower and so the resulting $P_j(G_i|y_j)$ will get higher. From the expectation-maximization approach, x will be grouped to $G(i)$ based on minimal distance x to each $C(i)$ profile. On the other hand, $y'_j = arg\ min_i\ (dist(x, C(i)) = arg\ min_j(y_j)$ will always provide the highest probability value of rightly grouping x into $G(i)$; $P(G_i|y'_j)$. By the used of resulting group G_i, m number of data in G_i can construct homogeneous group X_m. Thus, each homogeneous group X_m will be classified as known $C(i)$ with the probability of accurately classify X_m is $Q_i(X_m)$, which surely higher than $\frac{1}{i^m}\ Q_i(X_m)$, as $i = 1$. The illustration of group-based classification is depicted in Figure 1.

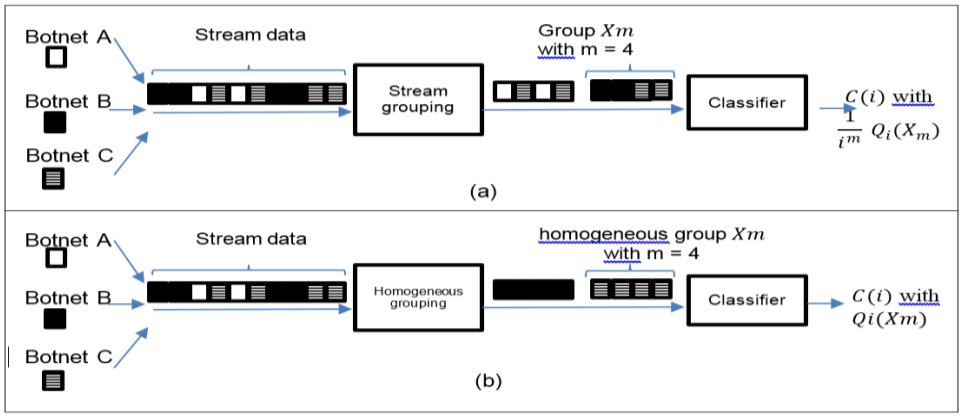

Figure 1. Illustration of group-based classification (a) without homogenous grouping, and (b) with homogeneous grouping.

B. Research Design

Our research capability focus was on types of DDoS classification in IPS. Research result could lead to types of attacks prediction, thus, IPS could take certain prevention or response action according to certain types of attack. It is very important as passive action in IDS is not

enough to stop the attack [31]. We focus on group-based classification as there is still left a problem for heterogeneous group data.

We add stream homogeneous grouping algorithm in covariance features classification to overcome the problem of group-based classification in random multi botnet environment. Our stream homogeneous group classification, consisting of four main processes as shown in Figure 2. The first process was traffic features selection, which selects proper features for system input. In the second process, we construct a homogeneous group by adopting the micro-clustering approach in Denstream [12]. The stream traffic was grouped according to the minimal distance to existing group profile and processed further when the group size is adequate for group-based classification input. The distance measurement was done by triangle area Mahalanobis distance (MD) which provide benefit as MD is scale-invariant and consider correlation during calculation [32]. Triangle area Mahalanobis distance which has been proposed in [17, 27, 18], proved to provide significant improvement of detection accuracy. Our previous report in [33] has reported the ability of triangle area Mahalanobis distance to discriminate traffic types among different distance measurement methods. The next process was group-based classification. We have occupied second-order covariance matrix decision tree prediction for group-based classification algorithm as in [19] to achieve high classification performance. The last process was prevention action to choose certain action according to certain detected types of attack.

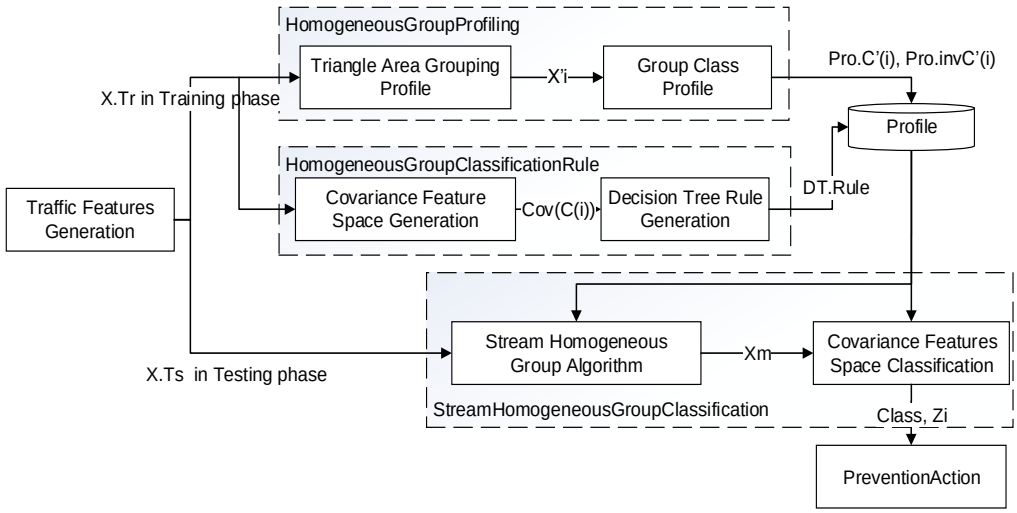

Figure 2. A framework of stream homogeneous group-based classification system.

B.1. Training phase

Suppose there is data traffic X_{Tr} which consist of successive data, $X_{Tr} = [x_1\ x_2\ x_3 \dots x_t]$; with x was data with multi features fp with p number of features; $x = [x_{f1}\ x_{f2}\ x_{f3}\ \dots\ x_{fp}]^T$. We have implemented supervised learning with multi features input from the dataset, and obtain $C(i) = [x^i{}_1\ x^i{}_1 x^i{}_2 \dots\ x^i{}_{t(i)}]$; where *C(i)* was classes of data traffic, *i* was the number of classes, and $t(i)$ was the number of data in each class $C(i)$. There are two processes in training phase which are homogeneous group profiling algorithm and group-based classification rule construction. The group profiling algorithm input $X'_{Tr} = [x'_1\ x'_2 \dots\ x'_t]$; with $x' = [x'_{f1}\ x'_{f2}\ x'_{f3}\ \dots\ x'_{fg}]^T$ was data with triangle area features g which was obtained by permutation of each features fp ; $x_{fa}/\ x_{fb}$ *; a ≠ b; a, b* $\in p$; of *x*. Each group profile *C'(i)* has drawn from the statistical mean of each training data from each class label i. In group-based classification rule construction, data in each class $C(i)$ was grouped according to size m. Thus, we obtain a sequence of $X_{Tr}(i) = [x^i{}_1\ x^i{}_1 x^i{}_2 \dots\ x^i{}_m]$ as the input to construct group-based classification rule. Classification rule was generated from every covariance of $X_{Tr}(i)$ and corresponding i class label. The algorithm is shown in Figure 3 and Figure 4.

$$CovX = [\ \sigma(x_{f1}, x_{f2})\ \sigma(x_{f1}, x_{f3}) \dots\ \sigma(x_{f1}, x_{fp})\ \sigma(x_{f2}, x_{f3}) \dots\ \sigma(x_{fp-1}, x_{fp})\ ] \quad (10)$$

$$\sigma\ (x_{fa}, x_{fb}) = \frac{1}{m}\sum_{i=1}^{m}\left(x_{fa} - E(x_{fa})\right)(x_{fb} - E(x_{fb})) \quad (11)$$

$$E(x_{fa}) = \frac{\sum_1^m x_{fa}}{m-1} \quad (12)$$

$$\overline{C(i)} = \left[E(x_{fa}), E(x_{fb}), \dots\ E(x_{fp}))\right]^T; \quad (13)$$

$$x' = P_2^p = \frac{x_{fa}}{x_{fb}}\ ; a\ \neq b; a, b \in \text{p} \quad (14)$$

$$g = \left[\frac{p!}{(p-2)!}\right] \quad (15)$$

$$x' = \left[x'_{f1}\ x'_{f2}\ x'_{f3}\ \dots\ x'_{fg}\right]^T \quad (16)$$

$$E(x') = \frac{\sum_1^t x'}{t-1} \quad (17)$$

$$\overline{C'(i)} = \left[E(x'_{f1})\ E(x'_{f2})\ E(x'_{f3})\ \dots\ E(x'_{fg})\right]^T; \quad (18)$$

$$MD(x', E(x')) = \sqrt{\frac{(x\prime - E(x\prime))^T (x\prime - E(x\prime))}{covX\prime}} \quad (19)$$

Algorithm GroupClassificationRuleConstruction
Input $C(i)$; i; m
Output DT.Rule

```
for n = 1 to i do
    Cov(n) = [ ];
    for t = 0 to t(i)/m do
        y = m * t
        X_t(i) ← [x^i_{y+1} ... x^i_{y+m}] of C(n)
        Cov(t) = Cov(X_t(i))
        Cov(t) = linspace(Cov(t))^T
        Cov(n) = [Cov(n); [Cov(t)]]
    end for
end for
for all Cov(n) construct DecisionTree
return DT.Rule
```

Figure 3. Group-based classification rule construction algorithm in training phase.

Algorithm HomogeneousGroupProfiling
Input *i*; $X'i$; $C(i)$; *t(i)*; *g*
Output *Pro.C'(i), Pro.invCovC'(i)*

```
for n = 1 to i do
    for m = 1 to g do
        for t = 1 to t(i) do
            E(x'^n_m) ← (Σ_1^t x'^n_t) / (t-1)
        end for
    end for
    C'(n)] ← [E(x'^n_1), E(x'^n_2), ... E(x'^n_g)]^T
    Pro.C'(n) ← C'(n)
    Pro.CovC'(n) ← Cov(C'(n))
    Pro.invCovC'(n) ← inv(Pro.CovC'(n)
end for
return Pro.C'(i), Pro.invCovC'(i)
```

Figure 4. Group profiling algorithm in training phase.

B.2. Testing phase

In term of distance grouping, each data x in X_{Ts} was firstly grouped in any $C(i)$ according to $min(MD(x', C'(i)))$. When the group size of $C(i)$ is equal to defined group size m, all data in group $C(i)$ are classified according to group-based classification. From classification output, prevention process has acquired types of attack information so that prevention system can single out certain action according to distinct types of attack. Algorithm involved in testing phase is depicted in Figure 5 and Figure 6.

The evaluation was done by several performance indices. We separate the grouping performance out of IDS detection performance. The homogeneous grouping was evaluated by grouping accuracy (Acc). It is because the grouping was done in term of multiclass classification. And the classification output was evaluated by classification accuracy (Acc), detection rate (DR), false positive rate (FPR) and detection precision rate (PR). The classification Acc measures the multiclass classification performance. And the DR, FPR, and PR were measured in term of IDS detection performance which is in binary classification (normal or anomaly). Suppose the real label of $X = [x_1\ x_2\ x_3 \dots x_t]$ is $= [y_1\ y_2\ y_3 \dots y_t]$ according to i types of traffic. And the classification predicted label of X is $Z = [z_1\ z_2\ z_3 \dots z_t]$. Thus, we can measure the performances as formulated as Formula 20 to 23.

Algorithm StreamHomogeneousGroupClassification Input *X'*; $Pro.C'(i)$, $Pro.invCovC'(i)$, m : defined group size Output *Class, Z(i), NewClassAlert*

for $n = 1$ to i do
  $MD(x', Pro.C'(i))$
end for
$[x,\ dist] \leftarrow Min(MD(x', Pro.C'(i)))$
$C(i) \leftarrow x$
if $dist > theshold\ C(i)$ then
  $C_o(i) \leftarrow x$
  update $C_o(i), size.C_o(i)$
else
update $C(i),\ size.C(i)$
end if
if $size.C(i) = m$ then
  $Cov(i) = Cov(C(i))$
  match $Cov(i)$ using DT.rule
  return $Class$
  $Zi \leftarrow IP_{source} in\ C(i)$
  return $Z(i)$
  update HomogeneousGroupProfiling $C(i)$
  empty $C(i)$
end if
if $size.C_o(i) = m$ then
  return $NewClassAlert$
end if

Figure 5. Algorithm for group-based classification.

$$Acc_{classification} = \frac{\sum_1^t (z_t = y_t)}{t} \tag{20}$$

$$PR_{detection} = \frac{\sum_1^t (z_t^{i \neq normal} = y_t^{i \neq normal})}{\sum_1^t (z_t^{i \neq normal} = y_t^{i \neq normal}) + \sum_1^t (z_t^{i \neq normal} = y_t^{i = normal})} \tag{21}$$

$$DR_{detection} = \frac{\sum_1^t (z_t^{i \neq normal} = y_t^{i \neq normal})}{\sum_1^t y_t^{i \neq normal}} \tag{22}$$

$$FPR_{detection} = \frac{\sum_1^t (z_t^{i \neq normal} = y_t^{i=normal})}{\sum_1^t y_t^{i=normal}} \quad (23)$$

Algorithm PreventionAction Input $Class, Z(i)$ Output prevention.
1. for all $Z(i)$ do 2. if $Class$ = Smurf then 3. prevention.Smurf 4. drop packets in $Z(i)$ 5. s_attack ← IP_source $Z(i)$ 6. else if $Class$ = Back then 7. prevention.Back 8. deny packets in $Z(i)$ 9. else if $Class$ = Neptune then 10. …. 11. end for

Figure 6. Rule-based algorithm for prevention action

B.3. Test & Data Acquisition

Stream homogeneous group classification has evaluated using KDD Cup 99 corrected dataset [20], which has been used in most anomaly detection and classification research. It was used as a training and testing dataset since this dataset considered mostly in related work. From 41 traffic features in KDD Cup 99 corrected dataset, not all features have meant for the DDoS attack. A high number of features made high computation cost but not always linearly result in high classification [34]. We have done features selection by filter method using mutual information calculation on continuous features, such as in [35]. Mutual information is one of the fast calculations of mutual dependence between features in different classes which measures arbitrary dependencies between random variables. In this research, we have defined features set (fp) which consists of 13 highest mutual information features to classify types of DDoS attack, listed in Table 1. The 13 features were chosen by wrapper method from training phase which provided lowest decision tree classification error, lowest 10 fold cross validation lost, and a minimal number of features.

Table 1. List of features from features selection.

Feature number	Feature name	Mutual information
23	*count*	0.792814
24	*srv_count*	0.706286
33	*dst_host_srv_count*	0.536724
32	*dst_host_count:*	0.19473
36	*dst_host_same_src_port_rate*	0.015505
28	*srv_rerror_rate*	0.010674
41	*dst_host_srv_rerror_rate*	0.010673
27	*rerror_rate*	0.010671
40	*dst_host_rerror_rate*	0.010643
34	*dst_host_same_srv_rate*	0.009792
29	*same_srv_rate*	0.009571
12	*logged_in*	0.006167
10	*hot*	0.001107

For training and testing phase, we have used 60% randomly selected data of each traffic type in the dataset (normal and DDoS dataset) into training phase ($X.Tr$), and the rest 40% dataset used in testing phase ($X.Ts$). The dataset in training and testing were disjoint. We have reproduced training scenario from research [19], in which training dataset is arranged in homogeneous groups $C(i)$ as normal ($i = 1$), Neptune ($i = 2$), Smurf ($i = 3$), Back ($i = 4$), Teardrop ($i = 5$), Pod ($i = 6$) and Land ($i = 7$). The number of data involving in training phase can be seen in Table 2.

Table 2. Number of DDoS data from dataset.

Traffic Type	Number of data in dataset	Number of data in training phase			
		m=150	m=100	m=50	m=10
normal	60592	36300	36300	36300	36300
neptune	58001	34800	34800	34800	34800
smurf	164091	98400	98400	98400	98400
back	1098	600	600	600	600
teardrop	12	-	-	-	-
pod	87	-	-	50	50
land	9	-	-	-	-

Evaluation has divided into two main scenarios. First was evaluation in artificial random multi botnet with mix types of DDoS attack in a group. And second was natural DDoS attack stream which came from the natural sequence of data from the dataset. In the artificial random DDoS scenario, $X.Ts$ construct a heterogeneous group of n data which averagely consist of predefined composition mix $C(i)/n$ percentages of i traffic types. For example, if the composition mix [40%, 25%, 15%, 5%, 5%, 5%, 5%], then a sequence of 100 data will consist of averagely 40% normal, 25% Neptune, 15% Smurf, 5% Back, 5% Teardrop, 5% Pod, and 5% Land, in unordered data from randomly chosen $X.Ts$. In natural DDoS attack traffic test, it was hard to construct 40% natural flows of DDoS because the natural mix of the botnet may not preserve. To keep representing a natural mix of DDoS attack, we have occupied coarse 100% dataset (including 60% of unarranged $X.Tr$) to preserve a natural mix of DDoS attack in the testing phase. In term of group-based classification analysis, we have analyzed four different group-size of 10, 50, 100 and 150 data and two different mixes of traffic. This scenario reflecting the importance of group size to the classification accuracy.

4. Results and Analysis

A. Stream Homogeneous Grouping Algorithm

We have tested our proposed Stream Homogeneous Group Classification (SHGC) compared to natural sequence grouping of traffic arrival for Covariance Feature Space Classification (CFSC) from [19]. This was done to analyze the improvement of group-based classification in the possible practical traffic stream. The traffic stream was represented by the used of unordered mixed types of traffic and natural sequence of data in the dataset.

This section describes research result and analysis by implementing minimal distance threshold in stream homogeneous grouping algorithm (SHGA) to construct a homogeneous group. In artificial random multi botnet testing, **w**e have done two scenario, first represent low rate DDoS with mix composition of [70% 10% 10% 10% 0% 0% 0%], and second represent high rate DDoS with mix composition of [10% 30% 30% 30% 0% 0% 0%]. Each scenario was done by generating 10.000 data in ten experiments.

The use of minimal Mahalanobis in triangle features area in our proposed grouping algorithm have shown to provide high grouping accuracy. From statistical Mahalanobis distance in our previous research [33], seems that minimal Mahalanobis distance of each DDoS traffic type to

normal profile has settled encouraging tool in stream grouping. Our proposed grouping algorithm achieved encouraging grouping accuracy, such as in Table 3 for low rate DDoS scenario. Even in high rate DDoS mix, the accuracy was above 98,2% on average. In the natural mix of traffic, the mix type of traffic in a group was naturally depending on KDDCup 99 data sequence, which was originally constructed from DARPA 98 dataset. Average grouping accuracy achieved 99,144% for a natural mix of different group size. The overall accuracy is above 98,2% as shown in Figure 7.

Table 3. A sample of grouping confusion matrices in low rate DDoS scenario.

Actual	Predicted at m= 50				Predicted at m= 100			
	normal	*back*	*smurf*	*neptune*	*normal*	*back*	*smurf*	*neptune*
normal	6870	0	16	0	6844	0	15	0
back	8	992	0	0	7	993	0	0
smurf	0	0	1000	0	0	0	1000	0
neptune	15	0	0	984	19	0	0	979

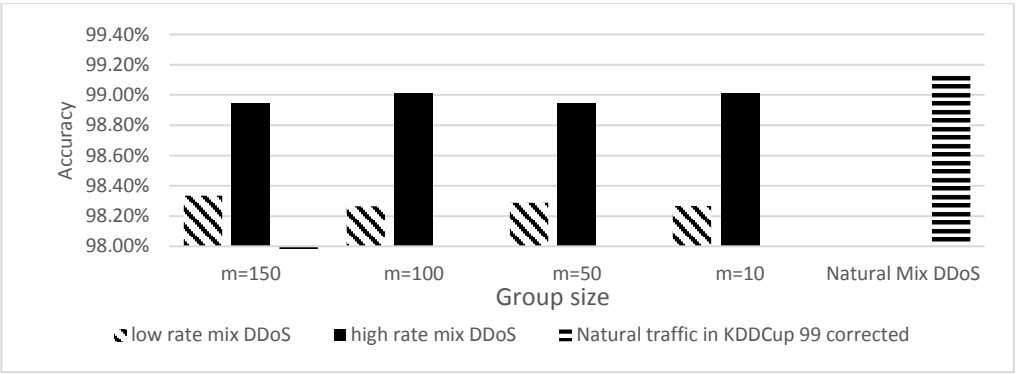

Figure 7. Average grouping accuracy.

B. Group-based Classification Analysis

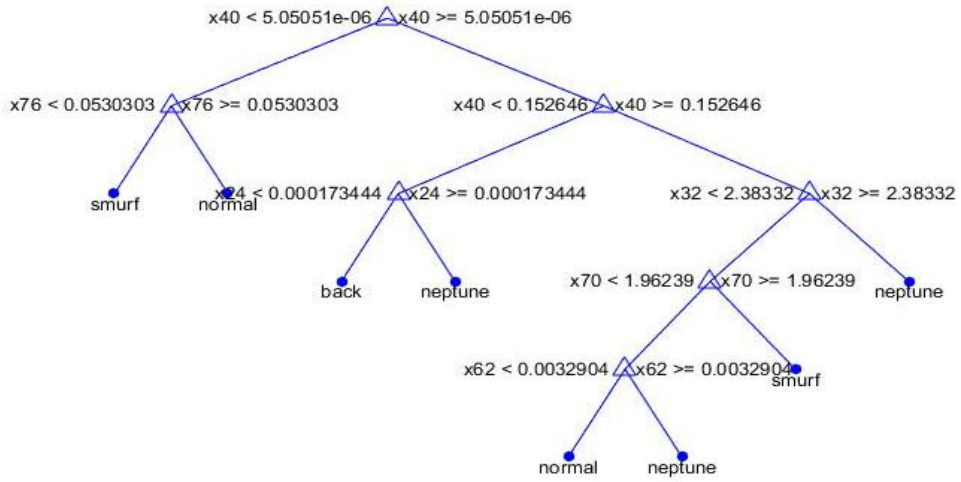

Figure 8. Decision tree at group-size of 100.

We have done the group-based classification analysis in four different group windowing size. Unfortunately, not all type of traffic in the dataset has a larger number of data than predefined group size. In order to construct a precise classification rule, the traffic with the number of data less than the group size has not been included in each training and testing phase, such as in Table 2. By the used of classification rule construction algorithm, we obtain four different decision rules according to four group sizes. From the decision rule we conclude that lower group size has produce more data training. Thus, we obtain more rigid decision rule for lower group size. An example of decision rule for group size of 100 was shown in Figure 8.

From the testing result, the used of homogeneous grouping has significantly improved group-based classification. IDS detection parameter best achieved in group size of 100, which maintain PR averagely above 99.2% and FPR below 0.37% in every testing scenario. From the multi-class confusion matrix, the used of homogeneous grouping in SHGC has clearly improved group-based classification accuracy as shown in Figure 9. This made the packet loss of normal traffic which dropped by prevention action will significantly be improved along with the lower FPR value achieved by SHGC.

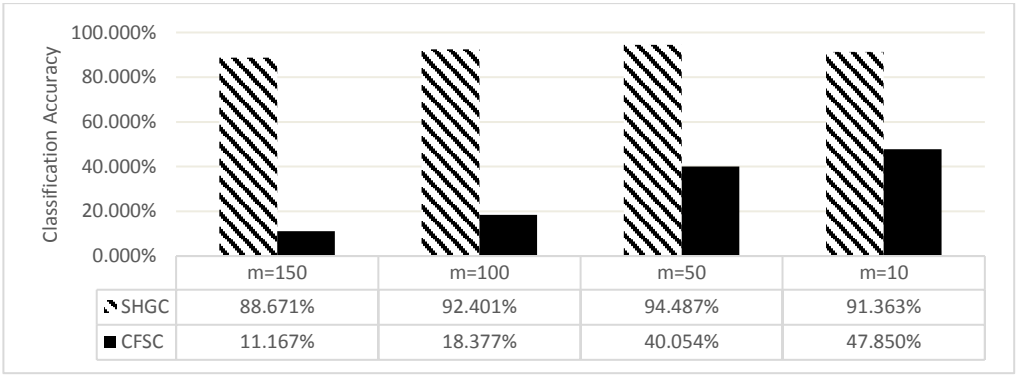

Figure 9. Average classification accuracy on low rate DDoS mix.

In the dataset, multi botnet DDoS mixing patterns was not much present. The nature of DDoS traffic that streamed continuously has made the mixing patterns within the group rarely found. With a low mixing rate on traffic, SHGC still achieve better classification performances. Thus, SHGC provides better IDS detection performance as shown in Table 4.

Table 4. IDS performance of SHGC compared to CFSC from a natural stream simulation.

Performance indices	m = 150		m = 100		m= 50		m= 10	
	SHGC (%)	CFSC (%)	SHGC (%)	CFSC (%)	SHGC (%)	CFSC (%)	SHGC (%)	CFSC (%)
Acc	98.257	91.042	98.845	91.450	98.908	94.191	98.854	97.812
FPR	6.240	5.888	1.867	3.076	2.613	2.320	4.095	4.129
DR	99.478	90.210	99.038	89.966	99.320	93.245	99.654	98.339
PR	98.326	98.263	99.492	99.082	99.292	99.330	98.898	98.873

From the overall measurement result, smaller group size has shown better classification performance than a larger group size. The classification accuracy was so affected by the noise data in the group itself. As the covariance matrix provide correlation among features, even a small portion of noise data in a feature will have avalanche impact for other features. For example in 13 features, one feature will take effect to 12 elements in the covariance matrix. For m

sequential data; $X = [x_1\ x_2\ x_3 \ldots x_t]$; probability to accurately classify sequence of m data x is $P(t) = C_m^t\ P^t(1-P)^{m-t}$, which P is in equation (1). From this equation, it is clearly seen that the higher number of m will result in lower probability of P. Thus, in group-based classification, the lower the group size m will have a higher probability of homogeneous data and produce a higher probability of accurately classifying all m data. For example, a normal type with an accuracy of 98% means that there are two noise data occurrence in 100 data. Thus, in a group size of 10 will certainly have the probability of noise value lower than size 100 which is about 2%.

The poor classification rule is also contributing to lower classification performance, as all the data in the group will be classified in the same predicted class. If the classification rule ends up in a wrong prediction, the higher the group size will impact in higher wrongly labeled data. From the decision tree algorithm, the rule-based decision in the decision tree is in top-down structures. This made the decision is sensitively influenced by the rule structure. The top-down rule tree structure made each covariance element unequal in the decision process. As the covariance element is prone to noise data, the noised value of an element in a higher rule structure resulted in the wrong consecutive rule path. Moreover, the limitations of the classifier were caused by the error in the generation of decision tree rule in the training phase. From 10-fold cross-validation on X.Tr, we found 1,28%, 1.03%, 1.41% and 1.68% error for group size of 10, 50, 100 and 150.

In term of computational complexity, replenishment of stream homogeneous grouping in group-based covariance feature space classification has made overall complexity relatively the same. On one hand, triangle area Mahalanobis distance computed to construct homogeneous grouping has a computational complexity of $O(p^4 * i)$, where p is the number of features and i is the number of known classes. Thus, the overall computation of our proposed stream classification has a computational complexity of $O(p^4 * i) + O(p^2 + h)$, where h is decision tree depth. However, p, i, and h are fixed deterministic in practice. Thus, the overall computational complexity is $O(1)$, which is relatively the same as CFSC which has the overall computational complexity of $O(p^2 + h) = O(p^4) = O(1)$.

5. Conclusion

This paper studies the problem of detecting DDoS attack in IDS which implements stream homogeneous grouping in a homogeneous group covariance matrix classification. Firstly, we have theoretically analyzed the difficulty of the traditional group-based classification system, which fails to accurately classify random multi botnet DDoS attack. Stream homogeneous grouping algorithm generated high grouping accuracy of the homogeneous group by the used of minimal triangle area Mahalanobis distance among known classes. It theoretically and practically improved the group-based classification performance compared to with no homogeneous grouping in possible practical stream traffic.

The used of minimal Mahalanobis distance has proved to provide homogeneous grouping in the natural DDoS attack in KDD Cup 99 and the synthetic mix of multi botnet DDoS attack. The homogeneous grouping accuracy remained high even the mix of the multi botnet was exploited. The lower group size achieved better grouping performance as it has a lower probability of noise in a group. The homogeneous group was also proven to significantly improve the classification accuracy, instead of straightly process the traffic group with no homogeneous grouping. By the use of decision tree algorithm, the homogeneous group covariance matrix has shown more encouraging IDS performance which are 98% of minimal classification accuracy, 99% of minimal detection rate and 7% maximum of false positive rate. Furthermore, this made the packet loss of normal traffic which dropped by prevention action significantly decrease along with the achieved lower FPR value.

For part of the future research, seems that decision tree utilized in our classification system was slightly prone to noise data. The use of another classification algorithm which more resistant to noise will be beneficial to achieve higher accuracy. There exist the possibility of research in evolving and new types of attack as it has already captured by the existence of a new group.

6. Acknowledgement

Directorate of Research and Community Service partially supported this research, the General Directorate of Research and Development Strengthening, the Ministry of Research, Technology, and Higher Education of the Republic of Indonesia under the research contract FY 2018 No. 014/PNLT3/PPM/2018.

7. References

[1]. Ponemon Institute, "The Cost of Denial-of-Services Attacks," Ponemon Institute© Research Report, 2015.

[2]. Arbor Networks Technical Repor, "Worldwide Infrastructure Security Report Volume XII," Arbor Networks , 2016.

[3]. Varun Chandola, Arindam Banerjee and Vipin Kumar, "Anomaly Detection : A Survey," *ACM Computing Surveys,* vol. 41, no. 3, 2009.

[4]. Monowar H. Bhuyan, D. K. Bhattacharyya and J. K. Kalita, "Network Anomaly Detection : Methods, Systems, and Tools," *IEEE Communications Surveys & Tutorials,* vol. 16, no. 1; pp. 303-336, 2014.

[5]. Yudha Purwanto, Kuspriyanto, Hendrawan and Budi Rahardjo, "Traffic Anomaly Detection in DDoS Flooding Attack," in *International Conference on Telecommunication Systems, Services, and Applications*, Bali, 2014.

[6]. Yu Chen, Kai Hwang and Wei-Shin Ku, "Collaborative Detection of DDoS Attacks Over Multiple Network Domains," *IEEE Transactions on Parallel and Distributed Systems,* vol 18, no. 12, pp. 1649 - 1662 , 2007.

[7]. Gautam Thatte, Urbashi Mitra and John Heidemann, "Parametric Methods for Anomaly Detection in Aggregate Traffic," *IEEE/ACM Transactions on Networking,* vol. 19, no. 2, pp. 512 - 525, 2011.

[8]. Federico Simmross-Wattenberg, Juan Ignacio Asensio-Perez, Pablo Casaseca-de-la-Higuera, Marcos Martin-Fernandez, Ioannis A Dimitridis and Carlos Alberola-Lopez, "Anomaly Detection in Netowrk Traffic Based on Statistical Inference and α-Stable Modeling," *IEEE Transactions on Dependable and Secure Computing,* vol. 8, no. 4, pp. 494 - 509, 2011.

[9]. Ming Yu, "An Adaptive Method for Source-end Detection of Pulsing DoS Attacks," *International Journal of Computer Science Issues,* vol. 7, no 5, pp. 279-288, 2013.

[10]. Li Liu, Pengyuan Wan, Yingmei Wang and Songtao Liu, "Clustering and Hybrid Genetic Algorithm Based Intrusion Detection Strategy," *TELKOMNIKA Indonesian Journal of Electrical Engineering,* vol. 12, no. 1, pp. 762-770, 2014.

[11]. Gao Meng, Li Dan, Wang Ni-hong and Liu Li-chen, "A Network Intrusion Detection Model Based on K-means Algorithm and Information Entropy," *International Journal of Security and Its Applications,* vol. 8, no. 6, pp.285-29, 2014.

[12]. Zachary Miller, William Deitrick and Wei Hu, "Anomalous Network Packet Detection using Data Stream Mining," *Journal of Information Security,* vol. 2, pp. 158-168, 2011.

[13]. Zahra Jadidi, Vallipuram Muthukkumarasamy, Elankayer Sithirasenan and Kalvinder Singh, "Performance of Flow-based Anomaly Detection in Sampled Traffic," *Journal of Networks,* vol. 10, no. 9, pp. 512-520, 2015.

[14]. Yuji Waizumi, Hiroshi Tsunoda, Masahi Tsuji and Yoshiaki Nemoto, "A Multi-Stage Network Anomaly Detection Method for Improving Efficiency and Accuracy," *Journal of Information Security,* vol. 3, pp. 18-24, 2012.

[15]. Ahmed Ahmim and Nacira Ghoualmi-Zine, "A New Fast and High Performance Intrusion Detection System," *International Journal of Security & Its Applications,* vol. 7, no. 5, pp. 67-80, 2013.

[16]. GuiPing Wang, ShuYu Chen and Jun Liu, "Anomaly-based Intrusion Detection Using Multiclass-SVM with Parameters Optimized by PSO," *International Journal of Security and Its Applications,* vol. 9, no. 6, pp. 227-242, 2015.

[17]. Zhiyuan Tan, Aruna Jamdagni, Xiangjian He, Priyandarsi Nanda and Ren Ping Liu, "A System for Denial of Service Attack Detection Based on Multivariate Correlation Analysis," *IEEE Transactions on Parallel and Distributed Systems,* vol. 25, no. 2, pp. 447 - 456, 2014.
[18]. Chih-Fong Tsai and Chia-Ying Lin, "A Triangle Area Based Nearest Neighbors Approach to Intrusion Detection," *Pattern Recognition,* vol. 43, pp. 222-229, 2010.
[19]. Shuyuan Jin, Daniel So Yeung and Xizhao Wang, "Network Intrusion Detection in Covariance Feature Space," *Journal of the Pattern Recognition Society 40, Elsevier,* vol. 40, no. 8, p. 2185–2197, 2007.
[20]. K. C. 1999, "Available on: http://kdd.ics.uci.edu/databases/kddcup," 28 October 1999. [Online]. Available: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html. [Acessed on 28-3-2016].
[21]. Trinita Samosir, Yudha Purwanto and Tito Waluyo, "A Sliding Window Technique for Covariance Matrix to Detect Anomalies on Stream Traffic," in *International Conference on Control, Electronics, Renewable Energy, and Communications (ICCEREC)*, Bandung, 2015.
[22]. Li Yu and Zhiling Lan, "A scalable non-parametric method for detecting performance anomaly in large scale computing," *IEEE Transactions on Parallel and Distributed System,* vol. 27, no. 7, pp. 1902-1914, 2016.
[23]. Richard Gow, Fethi A. Rabhi and Srikumar Venugopal, "Anomaly Detection in Complex Real World Application Systems," *IEEE Transactions on Network and Service Management,* vol. 15, no. 1, pp. 83-96, 2018.
[24]. Dolgormaa Bayarjargal and Gihwan Cho, "Detecting an Anomalous Traffic Attack Area Based on Entropy Distribution and Mahalanobis Distance," *International Journal of Security and Its Applications,* vol. 8, no. 2, pp. 87-94, 2014.
[25]. Jaka Sembiring, Mufti Ramadhan, Yudi S. Gondokaryono and Arry A. Arman, "Network Security Risk Analysis using Improved MulVAL Bayesian Attack Graphs," *International Journal on Electrical Engineering and Informatics,* vol. 7, no. 4, pp. 735-753, 2015.
[26]. Yudha Purwanto, Kuspriyanto, Hendrawan and Budi Rahardjo, "Cost Analysis for Classification-based Autonomous Response Systems," *International Journal of Network Security,* vol. 20, no. 1, pp. 121-130, 2018.
[27]. Aruna Jamdagni, Zhiyuan Tan, Priyadarsi Nanda, Xiangjian He and Ren Liu, "Mahalanobis Distance Map Approach for Anomaly Detection of Web-based Attacks," *Journal of Network Forensics,* vol. 2, no. 2, pp. 25-39, 2009.
[28]. Aruna Jamdagni, Zhiyuan Tan, Xiangjian He, Priyadarsi Nanda and Ren Ping Liu, "RePIDS: A multi tier Real-time Payload-based Intrusion Detection System," *Computer Networks,* vol. 57, no. 3, pp. 811-824, 2013.
[29]. Will Leland, Murad Taqqu, Walter Willinger and Daniel Wilson, "On the Self-Similar Nature of Ethernet Traffic," *IEEE/ACM Transactions on Networking,* vol. 2, no. 1, pp. 1-15, 1994.
[30]. O. I. Sheluhin, S. M. Smolskiy and A. V. Osin, Self-Similar Processes in Telecommunications, Chicester: John Wiley & Sons Ltd., 2007.
[31]. Kleanthis Malialis, Sam Devlin and Daniel Kudenko, "Distributed Reinforcement Learning for Adaptive and Robust Network Intrusion Response," *Connection Science,* vol. DOI: 10.1080/09540091.2015.1031082, 2015.
[32]. Prasanta Chandra Mahalanobis, "On The Generalised Distance in Statistics," in *Proceedings of the National Institute of Sciences of India, vol II, no. 1*, 1936.
[33]. Yudha Purwanto, Kuspriyanto, Hendrawan and Budi Rahardjo, "Statistical Analysis on Aggregate and Flow Based Traffic Features Distribution," in *International Conference on Wireless and Telematics*, Manado, 2015.
[34]. Richard Ernest Bellman, On the Theory of Dynamic Programming, Princeton, NJ: Princeton University Press, 1952.

[35]. R. Battiti, "Using The Mutual Information for Selecting Features in Supervised Neural Net Learning," *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537-550, 1994.

Yudha Purwanto completed his undergraduate degree at STTTelkom, Bandung and master degree at Electrical Engineering, Institut Teknologi Bandung. He currently works as a lecturer at Telkom University in Bandung. His research interests are future communication network, mobile application, and security system especially in network and digital forensics.

Kuspriyanto completed his undergraduate degree at Electrical Engineering, Institut Teknologi Bandung in 1974. He received his Master and Doctoral degree from Universit des Sciences et Techniques de Montpellier (USTL) France. He is currently a Full Professor at the Department of Electrical Engineering, Institut Teknologi Bandung, Indonesia. His current research interests include real time computing systems, computer architecture, and robotics. Contact at kuspriyanto@lskk.ee.itb.ac.id.

Hendrawan is an Associate Professor in School of Electrical Engineering and Informatics, Institut Teknologi Bandung, Indonesia. He completed undergraduate degree at Electrical Engineering, Institut Teknologi Bandung, Master and Doctoral degree in Telecommunications and Information Systems from University of Essex, UK. His current research interests are network management system, and future communication network. Contact at hend@stei.itb.ac.id.

Budi Rahardjo completed undergraduate degree at Electrical Engineering, Institut Teknologi Bandung. And received his Master and Doctoral degree from Manitoba University, Canada. His current research interests include security system, security forensics and cryptography. Contact at rahard@lskk.itb.ac.id.