# Evaluation of Anomaly Based Intrusion Detection System with *N-Gram* and *Incremental Learning*

I Made Agus Adi Wirawan, Royyana Muslim Ijtihadie, dan Baskoro Adi Pratomo
Jurusan Teknik Informatika, Fakultas Teknologi Informasi
Institut Teknologi Sepuluh Nopember (ITS)
Jl. Arief Rahman Hakim, Surabaya 60111 Indonesia
*e-mail*: {roy, baskoro}@if.its.ac.id

*Abstract*—***The rapid development of information technology is inevitable wich made its necessity is growing every single day. Data transaction through internet has become the primary need of most software nowadays. Software like social media, cloud server, online game, e-government, remote application, etc. With the various use of the internet, it is obvious that we need a method that can guarantee its safety.***

***IDS which stands for Intrusion Detection System is the solution to protect the internet network. This system will decide whether a packet is safe or dangerous for the network depends on certain condition. Nowadays many IDS (Intrusion Detection System) has been developed, but most of them are developed based on signature or using rules, and some of them use anomaly. Anomaly is a method used to look for irregularities in the data.***

***The IDS concept that is implemented in this application is the anomaly based IDS, which the data analysis is carried on the transmitted data packets. This thesis using two methods, the n-gram method used to calculate the distribution of byte character on data packet while the mahalanobis distance methods used to calculated the distance between the normal data packets and the intrusion data packets.***

***Mahalanobis distance methods can distinguish between normal data packets and intrusion data packets by calculating the average and standard deviation of the data packets.***

***Kata Kunci*—*N-Gram*, Mahalanobis Distance, Incremental Learning.**

## I. INTRODUCTION

THE rapid development of information technology make it easier for people to exchange data either via the internet or intranet. Of course, with easy sharing of data that is very possible to attacks on these data, mainly through a computer network. Intrusion detection system or generically called IDS (Intrusion Detection System) [1] is the main weapon to secure a network where this system would have the task to identify and record whether a data packet is a form of attack or normal data packets.

Recently many IDS (Intrusion Detection System) has been developed, but most of them are developed based on signature or using rules, and some of them using anomaly. Anomaly is basically a search of data that deviates from the normal set of data. IDS based on anomaly [2] is more flexible, because it can recognize new attack patterns without having to update the database of attack patterns. IDS based on anomaly has an artificial intelligence that is able to detect and recognize an attack. IDS based on anomaly combines analysis and statistical methods to identify such deviations. The downside of this method is the possibility of misidentification in the processed data.

An intrusion detection system is necessary to distinguish between normal data packets and data packet in the form of intrusion where later on, the intrusion detection system will use a combination of analysis and statistical methods that identifies differences in normal data packets and data packet in the form of intrusion. In addition, intrusion detection systems can also learn about the new normal data packet to update the training data.

N-gram method can be used to create a simple and fast to compute data packets models, especially the calculation of characters distribution on some data packets. N-Gram is the most efficient method and effective in making a model of a data packet.

Mahalanobis distance method is useful to distinguish data packets based on the anomalies that occurred. Meanwhile, incremental learning method is used to learn the normal data packet, where later on this method will renew the average and standard deviation of the data packets models on existing training data.

This article uses the article [3] as a reference implementation of the method mahalanobis distance to classify the packets of data and methods of n-gram to model the data packets. And this article discusses the evaluation of the addition of incremental learning process on the anomaly-based intrusion detection. Does the addition of incremental learning process can improve the accuracy of intrusion detection or not.

## II. APPLICATION DESIGN

### A. *Anomaly Based Intrusion Detection System*

Anomaly-based intrusion detection systems were basically looking for a data that deviate from normal data set. Anomaly-based IDS combines analytical and statistical methods to identify such deviations [2]. IDS based on anomaly is more flexible, because it can recognize new attack patterns without having to update the database of attack patterns. IDS based on anomaly has an artificial intelligence that is able to detect and recognize an attack.

The drawback of the anomaly method is the possibility of misidentification in the processed data, there is also the possibility of errors in the normal data, which causes the application can not recognize the attack.

### *B. N-Gram Payload Model*

Basically, the N-Gram models [4] is a probabilistic model that was originally designed by a mathematician from Russia in the early 20th century and later developed to predict the next item in the sequence of items. The items can include letters or characters, words, or the other according to the application. One of them, n-gram models based on the words used to predict the next word in a particular word order. In the sense that an n-gram is simply a collection container with each word has a length of n word. For example, an n-gram size 1 is referred to as unigram; size 2 as Bigram; size 3 as trigrams, and so on.

In character generation, N-gram consists of a substring along the n characters of a string, the other definition of n-grams are pieces of a number of n characters of a string. N-gram method is used to take the n character pieces letters of a word that continuously read from the source text to the end of the document. For example: the word "TEXT" can be decomposed into the following several n-gram:

*uni-gram* : T, E, X, T
*bi-gram* : TE, EX, XT
*tri-gram* : TEX, EXT
*quad-gram* : TEXT, EXT_
and so on.

While in the generation of the word, n-gram method is used to retrieve *n* fragments of a series of words (sentences, paragraphs, reading) that continuously read from the source text to the end of the document. For example: the phrase "i can see the light." It can be decomposed into the following several n-gram:

*uni-gram* : i, can, see, the, light
*bi-gram* : i can, can see, see the, the light
*tri-gram* : i can see, can see the , see the light_
and so on.

One advantage of using the n-gram as compared to using a whole word as a whole is that the n-gram is less sensitive to error in writing contained in a document.

### *C. Simplified Mahalanobis Distance*

Mahalanobis distance [5] is a statistical method to calculate the distance between points P and distribution D. The principle of Mahalanobis Distance is counting the distance in multidimensional space between an object of observation with the center of all the observations. In this article Mahalanobis Distance is used to calculate the distance between the distribution of byte characters from the new payload to the existing models on training data. The farther the distance, the more likely this is not a normal payload.

Mahalanobis distance of a new payload can be calculated if the system already has the training data, to further calculate the average and standard deviation of the existing models in the training data. To calculate the average of the existing models in the training data can be seen in equation 1, while for calculating the standard deviation of the existing models in the training data can be seen in equation 2. Once the average and standard deviation of the existing models in the training data have been calculated, then the distance mahalanobis of new payload can be calculated using equation 1. the format of the raw data contained on the Mahalanobis disatance can be seen in Table 1.

$$d(x,\overline{y}) = \sum_{i=0}^{n-1}(|x_i - \overline{y_i}|/(\overline{\sigma_i} + \alpha)) \qquad (1)$$

Where $d$ denotes the mahalanobis distance, $x_i$, the value of $i$-th variable from new *payload*, $\overline{y_i}$, the average of $i$-th variable from data training's model, $\overline{\sigma_i}$ , the standard deviation of $i$-th variable from data training's model, $\alpha$, the smoothing factor.

| | Variable | | | | | | |
|---|---|---|---|---|---|---|---|
| Object | $X_1$ | $X_2$ | ... | $X_i$ | ... | $X_{p-1}$ | $X_p$ |
| 1 | . | . | ... | . | ... | . | . |
| 2 | . | . | ... | . | ... | . | . |
| 3 | . | . | ... | . | ... | . | . |
| . | . | . | ... | . | ... | . | . |
| . | . | . | ... | . | ... | . | . |
| . | . | . | ... | . | ... | . | . |
| K | $X_{k1}$ | $X_{k2}$ | ... | $X_{ki}$ | ... | $X_{k,p-1}$ | $X_{k,p}$ |
| . | . | . | ... | . | ... | . | . |
| . | . | . | ... | . | ... | . | . |
| . | . | . | ... | . | ... | . | . |
| N | $X_{N1}$ | $X_{N2}$ | ... | $X_{Ni}$ | ... | $X_{N,p-1}$ | $X_{N,p}$ |
| Average | $\overline{X_1}$ | $\overline{X_2}$ | ... | $\overline{X_i}$ | ... | $\overline{X_{p-1}}$ | $\overline{X_p}$ |
| Standar deviation | $S_1$ | $X_1$ | ... | $X_1$ | ... | $X_1$ | $X_1$ |

*Table 1 The data format in the Mahalanobis Distance*

The equations used to find the average is:

$$\overline{X_i} = \frac{1}{N}\sum_{k=1}^{N} X_{ki} \qquad (2)$$

Where $\overline{X_i}$ denotes the average of $i$-th variable, $N$, the amount of model, $X_{ki}$, the value of $i$-th variable.

The equations used to find the standar deviation is:

$$S_i = \sqrt{\frac{\sum_{k=1}^{N}(X_{ki} - \overline{X}_i)^2}{N-1}} \qquad (3)$$

Where $S_i$ denotes the standard deviation of $i$-th variable, $X_{ki}$, the value of $i$-th variable, $\overline{X}_i$, the average of $i$-th variable, $N$, *the* amount of model.

### *D. Incremental Learning*

*Incremental Learning* is the process to renew the average value and standard deviation of the existing models in the training data when adding a new payload. This process is necessary to improve the accuracy of each model when a new sample data is added.

It takes an average and standard deviation of each character ASCII to calculate the Incremental Learning version of Mahalanobis distance for each new sample being calculated. To calculate the average of a character can be seen in equation 2. Furthermore, in order to renew the average value of the existing models in the training data, the number of samples that have been calculated previously is required [6]. To calculate the new average value can be seen in equation 4.

Meanwhile, to calculate new standard deviation takes an average of $x_i^2$ on the previous model. To calculate the new standard deviation can be seen in the equation 5.

The equation for calculating a new average of the observed models, that is:

$$\overline{\boldsymbol{x}} = \frac{\overline{x} \times N + x_{N+1}}{N+1} = \overline{\boldsymbol{x}} + \frac{x_{N+1} - \overline{x}}{N+1} \qquad (4)$$

Where $\overline{\boldsymbol{x}}$ denotes the new average, $\boldsymbol{x}_{N+1}$, the value of new variable, $\boldsymbol{N}$, amount of the previous model.

The equation for calculating a new standard deviation the observed models, that is:

$$\boldsymbol{S_i} = \sqrt{\frac{(n+1)\times(\sum_{i=1}^{n} x_i{}^2 + x_{n+1}{}^2) - (\sum_{i=1}^{n} x_i + x_{n+1})^2}{(n+1)n}} \qquad (5)$$

Where $S_i$ denotes the standar deviation of $i$-th variable, $x_i$, the value of $i$-th variable, $x_{n+1}$, the value of new variable, $n$, the amount of model.

### *E. Work Flow System in General*

In general, the system built with three main processes, the process of training data sets, the process of capturing data packets and the process of identifying the attack. The system workflow in general can be seen in Figure 1.

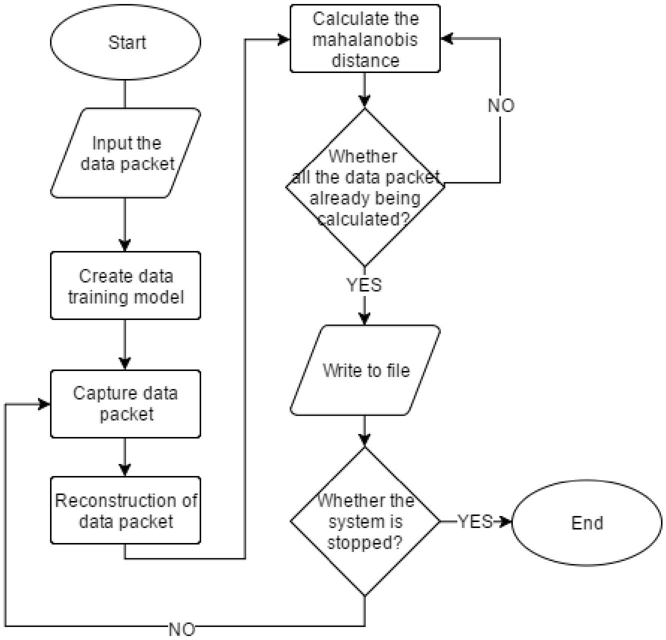
Start
Input the data packet
Create data training model
Capture data packet
Reconstruction of data packet
Calculate the mahalanobis distance
Whether all the data packet already being calculated?
NO
YES
Write to file
Whether the system is stopped?
YES
End
NO

*Figure 1 Work flow diagram of the system in general*

In more detail, how the system works is, read the file data sets, create data training model, save the data training model, capturing data packets, processing data packets and comparing the distance mahalanobis between data packets with model and also take a decision on the results of the comparison between the distance mahalanobis with the threshold. Creating data training model is the process of reading file data packets and store them in an array of objects. Only files with * .cap, * .pcap, * .tcpdump extension is readable, by utilizing the Jpcap library [7]. The next process is capturing data packets from the network interface with the help of Jpcap library then store it on an array of objects. Next is the process of comparing the distance mahalanobis between data packets with data model. The next process is calling mahalanobis distance function to calculate the mahalanobis distance between the data packets with the data training model, then compare it with the threshold value that is predetermined. Threshold value of each port has different magnitudes, if the distance mahalanobis of data packets exceeds the threshold value, then the data packet is categorized as data packets that are not normal.

The above process is repeated until the application is terminated by the user and writes the results of the comparison process to a log file.

## III. TESTING AND EVALUATION

### *A. Testing Scenario*

The performed test is divided into two parts, namely:

1. Testing accuracy without adding incremental learning process; and
2. Testing accuracy by adding incremental learning process.

Testing parameter used is the size of the window. The Window size is meant here is the number of data packets captured, if the number of data packets captured already meet the window size then the next data packet is processed to determine mahalanobis distance of each data packet.

### *B. Data Testing*

The test data will be processed using the two-fold cross validation method [8]. This test method is used because the created system has similar properties to machine learning based applications. The learning phase in this system is referred to the determination of detection threshold. From here, the two fold cross validation method is chosen so that the need to obtain threshold and testing data is met. With two fold cross validation method, the test data set will be divided into two equal lots. Then one data set will be used as training data as well as determine the threshold of detection applications. After completion, the activity is repeated with training data which previously was a test data. Presented in Table 2 is the test data where the distinguishing parameters between the test data is the size of the window. The test data used is the result of a data packets file captured on the DARPA’s external network [9] at week 4, which amounted to 5 file data packets.

Data to be recorded later is the mahalanobis distance from the normal data packet and packet data in the form of intrusion. Both of these variables will be processed into a threshold value by adding the smallest mahalanobis distance of packet data in the form of intrusion with the greatest distance mahalanobis of normal data packets and then divided by 2. Here in the equation 6 described how to get detection threshold value.

$$Threshold = \frac{min\ of\ attack\ distance + max\ of\ normal\ distance}{2} \qquad (6)$$

| No | Window size | TCP Port | | | | UDP Port |
|---|---|---|---|---|---|---|
| 1 | 10000 | 21 | 23 | 25 | 80 | 53 |
| 2 | 15000 | 21 | 23 | 25 | 80 | 53 |
| 3 | 20000 | 21 | 23 | 25 | 80 | 53 |

*Table 2 Data testing*

To calculate the accuracy, use the confusion matrix method [10]. This method is used because it is relatively easy to use and can generate values besides testing accuracy, like a true positive rate and false positive rate. Confussion matrix to be used for testing the accuracy is the confusion matrix with size 2x2. Following the model shown in Figure 2 along with the confusion matrix then explained the definition of each class.

| | | PREDICTED | |
|---|---|---|---|
| | | INTRUSI | NORMAL |
| ACTUAL | INTRUSI | A | B |
| | NORMAL | C | D |

*Figure 2 Confussion matrix models for testing*

### C. Testing Result

#### 1) Testing accuracy without adding incremental learning process

With the given threshold then the 5th week 1st day testing data is tested without incremental learning process. From 10,000 data packets tested, there are 5328 connection consisting of 5308 normal data packets and 20 intrusion data packets.

After all the data is tested, the results can be processed to produce the confusion matrix. Table 3 shows the classification of the amount of each class based on the test results.

| Detection without adding incremental | | PREDICTED | |
|---|---|---|---|
| | | INTRUSI | NORMAL |
| ACTUAL | INTRUSI | 6 | 355 |
| | NORMAL | 14 | 4935 |

*Table 3 Confussion matrix of 1st testing*

From the number above, a rating can be obtained based on formulas related to the confusion matrix presented in Table 4.

| No | Types of assessments | Value | Percentage |
|---|---|---|---|
| 1 | Akurasi (AC) | 0.9307 | 93.07% |
| 2 | *True positive rate* (TP) | 0.0166 | 1.66% |
| 3 | *False negative rate* (FN) | 0.9834 | 98.34% |
| 4 | *False positive rate* (FP) | 0.0028 | 0.28% |
| 5 | *True negative rate* (TN) | 0.9972 | 99.72% |
| 6 | Presisi(P) | 0.3 | 30.0% |

*Table 4 Rating result of 1st testing*

#### 2) Testing accuracy with adding incremental learning process

With the given threshold then the 5th week 1st day testing data is tested without incremental learning process. From 10,000 data packets tested, there are 5328 connection consisting of 1577 normal data packets and 3751 intrusion data packets.

After all the data is tested, the results can be processed to produce the confusion matrix. Table 5 shows the classification of the amount of each class based on the test results.

| Detection with adding incremental learning | | PREDICTED | |
|---|---|---|---|
| | | INTRUSI | NORMAL |
| ACTUAL | INTRUSI | 187 | 174 |
| | NORMAL | 3564 | 1403 |

*Table 5 Confussion matrix of 2nd testing*

From the number above, a rating can be obtained based on formulas related to the confusion matrix presented in Table 6

| No | Types of assessments | Value | Percentage |
|---|---|---|---|
| 1 | Akurasi (AC) | 0.2984 | 29.84% |
| 2 | *True positive rate* (TP) | 0.518 | 51.8% |
| 3 | *False negative rate* (FN) | 0.482 | 48.2% |
| 4 | *False positive rate* (FP) | 0.7175 | 71.75% |
| 5 | *True negative rate* (TN) | 0.2825 | 28.25% |
| 6 | Presisi(P) | 0.0499 | 4.99% |

*Table 6 Rating result of 2nd testing*

## IV. CONCLUSION

### A. Conclusion

From the results of trials that have been done, some conclusions can be drawn as follows:

1. Mahalanobis Distance method cannot be used to classify between normal data packets and packet data in the form of intrusion of the HTTP protocol. Distance generated during training using normal data packets and packet data in the form of intrusion resulted in a value of 0. So the normal data packets and data packets intrusion cannot be distinguished.
2. The system that made to detect intrusion using the Mahalanobis Distance without the "incremental learning" can detect intrusion by the percentage of the truth about 93%, but with the additional process of incremental learning can only detect intrusion by the percentage of the truth about 20%. From these results, in addition to the process of incremental learning reduce the level of accuracy of detection of intrusion.

### B. Suggestion

The advice given is the use of the method mahalanobis distance with the "incremental learning" to distinguish normal data packets with data packets attacks less accurate than without the addition of the incremental learning. This is because by adding the process of incremental learning, average and standard deviation of the model will be updated, but the threshold which is used to detect intrusion is not renewed, which resulted in threshold that is not accurate for detecting intrusion. There needs to be the implementation of other methods that can help improve the accuracy of intrusion detection.

## REFERENCES

[1] SANS Institute, "Understanding Intrusion Detection System," *SANS Institute Reading Room*, pp. 1-9, 2001.

[2] "Intrusion detection system," [Online]. Available: https://en.wikipedia.org/wiki/Intrusion_detection_system. [Accessed 22 June 2016].

[3] S. J. S. Ke Wang, "Anomalous Payload-based Network Intrusion Detection".

[4] A. Hanafi, "Pengenalan Bahasa Suku Bangsa Indonesia Berbasis Teks Menggunakan Metode N-gram. IT TELKOM," 2009.

[5] "Mahalanobis distance," [Online]. Available: https://en.wikipedia.org/wiki/Mahalanobis_distance. [Accessed 22 June 2016].

[6] D. E. Knuth, "The Art of Computer Programming," *Fundamental Algorithms. Addison Wesley,* vol. 1, 1973.

[7] K. Fuji, "a Java library for capturing and sending network packets," Jpcap, 15 May 2007. [Online]. Available: http://jpcap.gitspot.com/. [Accessed 23 May 2016].

[8] V. Galleys, "Cross Validation," 2006. [Online]. Available: http://www.cse.iitb.ac.id/~tarung/smt/papers_ppt/ency-cross-validation.pdf. [Accessed 24 June 2016].

[9] MIT Lincoln Laboratory, "MIT Lincoln Laboratory: Cyber system & technolog: DARPA Intrusion Detection," MIT Lincoln Laboratory, [Online]. Available: https://www.ll.mit.edu/mission/communications/cyber/CSTcorpora/ideval/docs/index.html. [Accessed 23 Mei 2016].

[10] Kohavi, "Confusion Matrix," 1999. [Online]. Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion_matix/confusion_matrix.html. [Accessed 24 June 2016].