Evaluasi Sistem Pendeteksi Intrusi Berbasis Anomali dengan *N-Gram* dan

*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 wether 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.

# PENDAHULUAN

T

HE 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 neccessary 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.

# DESAIN DAN PERANCANGAN

## Anomaly Based Intrusion Detection System

*Anomaly is basically looking for a data set of data deviates from normal. 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 berbais the anomaly has an artificial intelligence that is able to detect and recognize an attack.*

*The downside of this method is the possibility of anomaly misidentification in the processed data, there is also the possibility of errors in the data that causes the normal application can not recognize the attack.*

## N-Gram Payload Model

Basically, the model N-Gram [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 item. The items can include letters / characters, words, or the other according to the application. One of them, a model-based n-gram word is used to predict the next word in a particular word order. In the sense that an n-gram container is simply a set of words with each having a length of n word. For example, an n-gram size 1 is referred to as unigram; size 2 as Bigram; 3 sizes as trigrams, and so on.

In character generation, N-gram consists of all n characters substring from a string in the definition of another n-gram is a piece of n characters from a string. N-gram method is used to pick up the pieces of character font number n of a word continuity is read from the source text so that the end of the document. For example: the word "TEXT" can be decomposed into several n-gram following:

*uni-gram* :T, E, X, T

*bi-gram* : TE, EX, XT

*tri-gram* : TEX, EXT

*quad-gram* : TEXT, EXT\_

dan seterusnya.

While in generating word, n-gram method is used to pick up the pieces sejmlah n words from a set of words (sentences, paragraphs, reading) that continuity is 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 several n-gram following:

*uni-gram* :saya, dapat , melihat, cahaya, itu

*bi-gram* : saya dapat, dapat melihat,

melihat cahaya, cahaya itu

*tri-gram* : saya dapat melihat, dapat melihat

cahaya, melihat cahaya itu\_

and so on.

While in generating word, n-gram method is used to pick up the pieces sejmlah n words from a set of words (sentences, paragraphs, reading) that continuity is 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 several n-gram following:

## Simplified Mahalanobis Distance

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

Mahalanobis distance of a new payload can be calculated if the system already has the training data. Next 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. As for calculating the standard deviation of the existing models in the training data can be seen in equation 2. Once completed calculate the average and standard deviation of the existing models in the new training data can calculate the distance mahalanobis of new payload using equation 1. the format of the raw data contained on the Mahalanobis disatance can be seen in Table 1.

(1)

dimana,

= jarak mahalanobis

= variable ke-i dari *payload* baru

= rata-rata variable ke-i dari model data *training*

= standar deviasi variable ke-i dari model data *training*

= *smoothing factor*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Variabel (karakteristik) | | | | | | |
| Object |  |  | … |  | … |  |  |
| 1 | . | . | … | . | … | . | . |
| 2 | . | . | … | . | … | . | . |
| 3 | . | . | … | . | … | . | . |
| . | . | . | … | . | … | . | . |
| . | . | . | … | . | … | . | . |
| . | . | . | … | . | … | . | . |
| K |  |  | … |  | … |  |  |
| . | . | . | … | . | … | . | . |
| . | . | . | … | . | … | . | . |
| . | . | . | … | . | … | . | . |
| N |  |  | … |  | … |  |  |
| Average |  |  | … |  | … |  |  |
| Standar deviation |  |  | … |  | … |  |  |

Tabel 1 Format data kasar didalam Mahalanobis Distance

Persamaan untuk mencari rata-rata, yaitu:

(2)

dimana,

= rata-rata variabel ke-i

= jumlah object model

= nilai variabel ke-i

Persamaan untuk mencari nilai standar deviasi, yaitu:

(3)

dimana,

= standar deviasi variabel ke-i

= nilai dari variabel ke-i

= rata-rata variabel ke-i

= jumlah object model

## Incremental Learning

*Incremental Learning* is a process to update 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 added new sample data.

To calculate the Mahalanobis distance required version of Incremental Learning average and standard deviation of each ASCII character for each new sample is 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 necessary number of samples have been counted previously [6]. To calculate the average value of the new can be seen in Equation 4.

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

The equation for calculating a new average of the models were observed, namely:

(4)

dimana,

=rata-rata baru

= nilai dari variabel yang baru

=jumlah sampel sebelumnya

Persamaan untuk menghitung standar deviasi baru dari model yang diamati, yaitu:

(5)

dimana,

= standar deviasi variabel ke-i

= nilai dari variabel ke-i

= nilai dari variabel yang baru

= jumlah object model

## Alur Kerja Sistem Secara Umum

In general, the system is built consists of three main processes, namely the process of training data sets, sniffing and attack identification process. The workflow system in general can be seen in Figure 1.

How the system works more detail, namely, reading the file data sets, model making training data, save the model training data, packet capture, packet processing and comparing the distance mahalanobis packet with models and decision making on the results of the comparison. Creating a data model training is the process whereby data packets and read files stored on an array of objects. Files that can be read only files with extension \* .cap, \* .pcap, \* .tcpdump by utilizing the library Jpcap [7]. The next process is the capture packet from the network interface with the help of library Jpcap then store it on an array of objects. Next is the process of comparing the distance mahalanobis packet to the data model. Packet reflect only the packet which has a destination port less than 1024. When there is a packet that qualify will do the calculation process, starting from the calculation of average and standard deviation of the packet. And then the calling function to calculate the distance Mahalanobis Distance mahalanobis packet with training data model. After getting mahalanobis distance, and then compared with a threshold value that is predetermined. Threshold value each port has different magnitudes. If the distance mahalanobis packet exceeds the threshold value, then the packet which can be considered a package that is not normal.

The above process is repeated until stopped by the user application and write the results in the comparison process to a log file.



Gambar 1 Diagram alir kerja sistem secara umum

# UJI COBA DAN EVALUASI

## Skenario Pengujian

Tests performed 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. Window size is dimasksud is the number of data packets captured, if the number of data packets captured it meets next window size data packets are processed to determine the distance mahalanobis each package.

## Data Pengujian

Test data which will be processed by two fold cross validation method [8]. This test method is used because the system created has properties similar to applications based on machine learning. Learning stage referred to in this system go round the stage for the determination of threshold detection. Departing from this, the two fold cross validation method is chosen so that the need to obtain and test data threshold is met. With two fold cross validation method, the test data set will be divided into 2 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 previous data sets into the test data will be used as training data and training data set used as test data. The following Table 2 is presented test data where the parameters distinguishing between the test data is the size of the window. Window size is dimasksud is the number of data packets captured, if the number of data packets captured already meet the new window size data packets are processed to determine the distance mahalanobis each package. The test data used is file data packets catches of data packets on the external network DARPA [9] at week 4, amounting to 5 file data packets.

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

(6)

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **No** | **Ukuran *window*** | **Port TCP** | | | | **Port UDP** |
| 1 | 10000 | 21 | 23 | 25 | 80 | 53 |
| 2 | 15000 | 21 | 23 | 25 | 80 | 53 |
| 3 | 20000 | 21 | 23 | 25 | 80 | 53 |

Tabel 2 Data uji

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 |

Gambar 2 Model Confussion Matrix untuk pengujian

## Hasil Pengujian

### Pengujian akurasi tanpa menambahkan proses incremental learning

Armed with the threshold is then tested against the data testing week 5, day 1 without incremental learning process. The test results from 10,000 data packets, which are 5328 5308 connection consisting of normal packages and 20 packages of intrusion.

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **PREDICTED** | | |
| INTRUSI | NORMAL |
| **ACTUAL** | INTRUSI | 6 | 355 |
| NORMAL | 14 | 4935 |

Tabel 3 Confussion matrix uji coba 1

Based on the number above, then obtained a rating based on formulas related to confussion matix presented in Table 4.

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Jenis Penilaian** | **Nilai** | **Persentase** |
| 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% |

Tabel 4 Hasil penilaian percobaan 1

### Pengujian akurasi dengan menambahkan proses incremental learning

Armed with the threshold is then tested against the data testing week 5, day 1 with additional incremental process leraning. The test results from 10,000 data packets, which are 5328 connection consisting of normal packets 1577 and 3751 intrusion package.

After all the data is tested, then the data can be processed result by the confusion matrix. In Table 5 presented the classification of the amount of each class based on the test results.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | | **PREDICTED** | | |
| INTRUSI | NORMAL |
| **ACTUAL** | INTRUSI | 187 | 174 |
| NORMAL | 3564 | 1403 |

Tabel 5 Confussion matrix uji coba 2

Based on the number above, then obtained a rating based on formulas related to confussion matix presented in Table 6.

|  |  |  |  |
| --- | --- | --- | --- |
| **No** | **Jenis Penilaian** | **Nilai** | **Persentase** |
| 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% |

Tabel 6 Hasil penilaian percobaan 2

# KESIMPULAN

## 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.

## Saran

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.

DAFTAR PUSTAKA

|  |  |
| --- | --- |
| [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. [Diakses 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. [Diakses 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/. [Diakses 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. [Diakses 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. [Diakses 23 Mei 2016]. |
| [10] | Kohavi, “Confusion Matrix,” 1999. [Online]. Available: http://www2.cs.uregina.ca/~dbd/cs831/notes/confusion\_matix/confusion\_matrix.html. [Diakses 24 June 2016]. |