Attribute Selection Using Information Gain and Naïve Bayes for Traffic Classification

Ahmad Fali Okilias\(^1,2\), Tasmi\(^2\), Sri Desy Siswanti\(^3\), Mira Afrina\(^4\), Herri Setiawan\(^5\)

\(^1,2\) Department of Computer Engineering, Faculty of Computer Science, Universitas Sriwijaya, Indonesia
\(^3\) Department of Information Systems, Faculty of Computer Science, Universitas Sriwijaya, Indonesia
\(^2\) Computer Engineering Computer Science Department Indo Global Mandiri University
\(^5\) Engineering Informatics Computer Science Department Indo Global Mandiri University

*fali@ilkom.unsri.ac.id

Abstract. This paper presents the value of good attributes in carrying out the package classification process. Furthermore, attribute selection results can be of value in determining one type of data packet traffic. In obtaining the results (data) this study uses several stages of Data Capture, Feature Extraction and Feature Selection, but in this study only focuses on the process of feature selection using the gain and Entropy Information and Naïve Bayes algorithm. The testing process by dividing raw data into parts is 70 per cent for Training data and 30 per cent for testing data. The total data used is 6632, so for the training data obtained is 4642, while the data for testing is 1990. The results obtained are for the https data package type is 9 attributes, while for the HTTP data is 10 attributes. For further research can be added using another classification method as a comparison of data.

1. Introduction

The internet has become one of the needs in human life today because there (the internet) people can get all the information in all fields such as education, health, entertainment and others. With the rapid development, the quality of security and user settings is needed so that security is maintained. One way to keep information safe, website development already uses Secure Socket Layer (SSL) or Transport Layer Security (TLS).

The communication protocol in the internet world that functions to connect between users and servers is HTTP and HTTPS. The difference between HTTP (Hypertext Transfer Protocol) and HTTPS (Hypertext Transfer Protocol Secure) is when the process of sending data from a user to a server where an application that uses HTTP does not have a data encryption process. Based on its ability (HTTP), it opens the opportunity of those who are not responsible for the process of theft and reading data during the communication process. While HTTPS is equipped with additional Secure Socket Layer (SSL) or Transport Layer Security (TLS) security protocols to maintain data security during the communication process. There are three types of data security using Secure Socket Layer (SSL), namely: (i). Authentication, (ii) Confidentiality and (ii) Integrity.

Some previous studies have carried out the process of classifying traffic in recognizing patterns of a data plan. One of them is [1] traffic type classification which is one method used to optimize bandwidth with the aim of reliable and stable internet connection. In the paper Molina and his friends [2] mention the management network as a supporting medium in the case of identification of data packages with the approach of Operation, Administration, and Maintenance & Provisioning. Not only that from Zhang [3]...
the results of their research managed to recognize package patterns well, but the system was still passive so there was no traffic control in and out. Furthermore, by using the Deep Packet Inspection (DPI) method in the data packet classification process [4] use the DPI method to classify traffic types, comparing DPI tools in classifying types of internet traffic [5].

In the process, the results of classification of attribute selection have a very important role to get maximum results [6]. Various feature selection search techniques have been widely used in the machine learning process. One way in the attribute search process uses feature selection. The more complex the attributes are very influential in the computational process and also with the number of attributes that do not match it will produce bad accuracy [7]. The purpose of this process is to eliminate attributes that are not suitable so that the classification process can work optimally and the results of accuracy will be better [8]. In the pre-processing process, the feature selection method has been widely used, such as using information gain [9]. In IDS many have used information gain in the detection of an attack [10] [11].

2. Overview and Background Problems

In previous studies conducted by [4] using the DPI method in traffic classification and research comparing DPI tools in classifying types of internet traffic [5]. The focus of this study is to look for strong attributes to produce good accuracy in recognizing a data packet, this is in accordance with research conducted by [7] and Witten [8] which states that attributes that are not good will produce accuracy that is not good.

Feature selection is one step that is mostly done in the process of pre-processing and machine learning. The function of the attribute selection process is: first to determine the ranking of each attribute; the second is to eliminate attributes that are not relevant. In the study by Slocum [12] states that basically entropy is used to interpret the uncertainty of some attributes of a dataset. The higher the entropy of an attribute, the higher the uncertainty value, while the research [13] to find the entropy value using equation (1) as follows:

\[
(Y) = - \sum_{y \epsilon Y} (y) \log_2 (p(y))
\]  
Where \(p(y)\) is the marginal probability function for the value of a random variable \(Y\). If the values of \(Y\) measured in dataset \(S\) are divided by the values of the second feature \(X\), and the entropy \(Y\) for the partition is affected by \(X\) less than entropy \(Y\) before the partition process, then there is a relation between \(Y\) and \(X\) features. So that the entropy equation \(Y\) after observing \(X\) is shown in equation 2.

\[
H(Y|X) = - \sum_{x \epsilon X} p(x) \sum_{y \epsilon Y} p(y|x) \log_2 p((y|x))
\]

\(p(y|x)\) is a conditional probability of \(y\) against \(x\). Entropy as an invalid criterion in training data set \(S\), it can be defined in size is an additional information about \(Y\) provided by \(X\) which represents the number where the entropy value of \(Y\) decreases. This section is known as Information Gain (IG).

\[
IG = H(Y) - H(Y|X) = H(X) - H(X|Y)
\]

From equation 3 the information obtained about \(Y\) after observing \(X\) is the same as the information obtained about \(X\) after observing \(Y\). The focus of this study uses Feature selection (ranking) in determining strong attributes in traffic classification and further research will use these attributes to detect data package.

3. Proposed Design

The result of data capture (raw data) has many attributes that will be searched for strong attributes to recognize the type of data packet. One of the attribute determination techniques using the ranking method with the aim of finding the order of attributes by means of the score of each attribute. In this study using information gain method to find the sequence of attributes and then be validated using the classification method. The stages proposed in this study for the feature ranking process and data validation are divided into five stages: (i) capture data, (ii), Feature Extraction, (iii) Data Cleaning and Implementation of Feature Selection, (iv) Data validation and (v) analysis as shown in Figure 1.
The first part that is done is collecting data obtained from the results of capture traffic on the computer science faculty network Universitas Sriwijaya (fasilkom unsri). Then the second stage of preprocessing of the data obtained. This stage is the process of cleaning the data that is used to remove duplicate data, see inconsistent data, and eliminate data errors that are obtained with good accuracy. At the retrieval stage, the dataset produces 19 attributes which are the result of data processing and one attribute as a variable class, namely the types of data packages. The third part is the process of determining the attributes that are relevant by using ranking feature methods, namely information Gain (IG). Ranking feature method to find rankings that will be validated using the Naive Bayes classification method. The results of the classification will be analyzed based on the accuracy of each attribute. In this section, the dataset is divided into two parts, namely 70 per cent training and 30 per cent testing. These results are used for the learning process to get predictions of strong attributes in determining the pattern of a data plan.

4. Implementation and Result
The dataset used in the research is that there is a real capture result on the campus network of the Unsri campus. Sniffing package is an application that is used to monitor traffic on a network passing. This refers to Patel [14] which states packet sniffing is part of a hardware or software used to monitor all traffic on a network. While according to Spangler [15] the information obtained from packet sniffing can be used by an administrator to identify strange packages and also to keep transmitting data on the network to remain efficient.

The data collection process was carried out four times. The results obtained can be seen in Figure 2 with the line graph, that the data in the fourth experiment produced good data because the lost package was only about 0.05 per cent, TCP packet was 77.53 per cent, and the UDP package was 22.41 per cent. In the first experiment, the UDP package was bigger than TCP data which was 10.67 per cent. This is because the router provides services to users of several UDP protocols such as DHCP and DNS. In the second experiment the number of packages 226,516, TCP = 87.11 per cent and UDP = 12.78 per cent, then TCP is 74.33 per cent greater than UDP. The next experiment, the number of packets decreases and TCP is greater than UDP, so next until the fourth experiment there is a decrease too.
Figure 2 shows the results of capture based on applications accessed by the user, the algorithm built only focuses on TCP and UDP protocols. While some packages such as ICMP, SIP, SNMP and others are not processed. The next part: first is to determine the result attributes of captured traffic in real-time. Second is to determine the relevant attributes from the capture that are obtained from the traffic on the Faculty. Ranking feature method that is information Gain (IG) is used to find rankings that will be validated using the Naive Bayes classification method.

Table 1. The result of Ranking Attribute by Using IG

| No | Attribute        | Information Gain | Entropy | Rank |
|----|------------------|------------------|---------|------|
| 1  | Protocol         | 0.0000           |         | 18   |
| 2  | Source_IP        | 0.07236785       | 2       |      |
| 3  | Dst_IP           | 0.07414856       | 1       |      |
| 4  | TTL              | 0.00397622       | 8       |      |
| 5  | Capture_length   | 0.00055618       | 11      |      |
| 6  | Header_Length    | 0.0000           | 16      |      |
| 7  | Total_Lengt      | 0.00055618       | 10      |      |
| 8  | Identification   | 0.02781127       | 3       |      |
| 9  | Checksum_Header  | 0.00025073       | 12      |      |
| 10 | Frag_Offset      | 0.01413719       | 4       |      |
| 11 | Source_Port      | 0.00002238       | 13      |      |
| 12 | Dst_Port         | 0.00002238       | 14      |      |
| 13 | Flags            | 0.0063704        | 7       |      |
| 14 | Ack              | 0.01102643       | 5       |      |
| 15 | Seq              | 0.00676296       | 6       |      |
| 16 | Window           | 0.0039587        | 9       |      |
| 17 | Urgent_Pointer   | 0.0000           | 17      |      |
| 18 | Checksum_Protokol| 0.00000774       | 15      |      |

The results of the classification will be analyzed based on the accuracy of each attribute. This stage is the process of cleaning the data used to eliminate data errors so that the results of good accuracy are obtained. The next step is to divide the dataset into two parts, namely 70 per cent training and 30 per cent testing. This is used for the learning process to get predictions of strong attributes in determining the pattern of a data plan.

The results of feature selection testing to get the rank attribute and entropy value are shown in table 1 and figure 3. From the test results, each attribute is calculated entropy value and looking for relations between attributes by using equation 2. Then by using equation 3 looks for the highest attribute rank. The results obtained are the first rank is attribute DST_IP with entropy value of 0.07414856, ranked two
and three, the attribute Source_IP and identification are obtained with each entropy value of 0.07236785 and 0.02781127. From the results of the uncertainty value in an attribute and also the ranking results obtained that three attributes that do not have a good correlation with the class. These three attributes are protocol, Header_Length and Urgent_Pointer. This is because the value of the three attributes has the same value. The next step is to do the attribute validation process using the Naïve Bayes classification method. This testing process by dividing raw data into 2 parts, namely: 70 per cent for training data and 30 per cent for testing data. The total data used is 6632, so for training data, there are 4642, data for testing are 1990. Figure 4 is the result of training and testing will be tested to get the validation value of each attribute that will be used as new knowledge in the process next.

Fig 3. Results of Ranking and Entropy Attribute Value

Based on the results of the experiment, shows that the Naïve Bayes algorithm with ranking-based attribute selection can produce accuracy in each attribute. From the attribute selection results obtained as many as eight attributes whose accuracy is above 60 percent, namely: (i) TTL 97.2942 percent, (ii) Capture_length 60.4554 percent, (iii) Total_length 60.4554 percent, (iv) Identification 71.2501 percent, (v) Checksum_Header 60.3045 percent, (vi) Frag_Offset 73.2959 percent, (vii) Flags 79.0799 percent, and (viii) Window 93.8767 percent. While from the results of the experiment there are three attributes that have 0.0000 per cent accuracy namely (i) Protocol, (ii) Header_Length and (iii) Urgent_Pointer.

Fig 4 Attribute Validation Results with Naïve Bayes
5. Conclusion

From the results of the discussion, it can be drawn the conclusion that using the ranking method (information Gain) and Naïve Bayes results in ranking attributes in the case selection feature. The results of feature selection using the feature method show that the highest entropy value is 0.0741864 for the DST_IP attribute. For the validation attribute, the highest value in TTL attribute is 97.2942 per cent. The attribute validation results using the Naïve Bayes method produce TTL attributes get the highest value of 97.2942 per cent for training and 97.0525 per cent for data testing. This research will then try to use the knowledge from the ranking results and will also try to use several types of ranking methods and also use other classification methods in order to get good accuracy in the detection process.

Reference:

[1] I. G. Siqueira, L. B. Ruiz, and a. a. F. Loureiro, “Coverage area management for wireless sensor networks,” Int. J. Netw. Manag., no. October 2005, pp. 17–31, 2007.
[2] M. Molina, I. Paredes-oliva, W. Routly, and P. Barlet-ros, “Operational experiences with anomaly detection in backbone networks,” Comput. Secur., vol. 31, no. 3, pp. 273–285, 2012.
[3] H. Zhang, D. Yao, N. Ramakrishnan, and Z. Zhang, “Causality reasoning about network events for detecting stealthy malware activities,” Comput. Secur., vol. 58, pp. 180–198, 2016.
[4] A. F. Oklilas and Tasmi, “Monitoring and Identification Packet in Wireless With Deep Packet Inspection Method,” Int. Conf. Recent Trends Phys. 2016 IAES Int. Conf. Electr. Eng. Comput. Sci. Informatics, vol. 365, p. 011001, 2017.
[5] T. Bujlow, V. Carela-Español, and P. Barlet-Ros, “Independent comparison of popular DPI tools for traffic classification,” Comput. Networks, vol. 76, pp. 75–89, 2015.
[6] V. Bolón-Canedo, N. Sánchez-Maroto, and A. Alonso-Betanzos, “Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset,” Expert Syst. Appl., 2011.
[7] E. Alpaydın, “Introduction to machine learning,” MIT Press, 2012.
[8] I. H. Witten, E. Frank, and M. a Hall, Data Mining: Practical Machine Learning Tools and Techniques (Google eBook). 2011.
[9] B. Azhagusundari and A. S. Thanamani, “Feature Selection based on Information Gain,” Int. J. Innov. Technol. Explore. Eng., 2013.
[10] S. Mukherjee and N. Sharma, “Intrusion Detection using Naive Bayes Classifier with Feature Reduction,” Procedia Technol., 2012.
[11] Akashdeep, I. Manzoor, and N. Kumar, “A feature reduced intrusion detection system using ANN classifier,” Expert Syst. Appl., 2017.
[12] M. Slocum, “DECISION MAKING USING ID3 ALGORITHM,” vol. 8, no. 2, pp. 1–12, 2012.
[13] J. Novaković, P. Strbac, and D. Bulatović, “Toward optimal feature selection using ranking methods and classification algorithms,” Yugosl. J. Oper. Res., vol. 21, no. 1, pp. 2334–6043, 2011.
[14] H. Patel, “Network Traffic Analysis Using Packet Sniffer Pallavi Asrodia *, Hemlata Patel **;,” vol. 2, no. 3, pp. 854–856, 2012.
[15] R. Spangler, “Packet sniffer detection with antisniff,” " Wisconsin, Dep. Comput. ..., no. May, 2003.