Determining Attributes of Encrypted Data Traffic using Feature Selection Method

Tasmi, Herri Setiawan, Deris Stiawan, Husnawati, Sasut Analar Valiata

Abstract: Encrypted packages such as banking and e-commerce are widely used in various fields because it is advantages in terms of data security. However, the problem occurs when checking attributes package to determine if it is a safe packet instead of malware. The purpose of this study is to get the best attributes using feature selection processes by ranking.The results of this study found that from the two best methods of IG and One R, in average IG better than One R. If based on the results of the response, the data produced for the estimated data of TLS V1.0 IG method has better accuracy compared to the One R method, on the contrary inTLS V1.2 One R data is better than IG.

Keywords: Feature Selection, Feature Ranking, Encrypted Traffic

I. INTRODUCTION

In data communication over the network, it is possible to lose confidentiality, message integrity or authentication endpoints. There are three main aspects of the discussion about data security, namely; 1) Privacy or Confidentiality, which includes the confidentiality of information. The major aspect of privacy is how to protect information so as not to be seen or accessed by unauthorized people, 2) Message Integrity, which includes the integrity of information. The major aspect of integrity is how to keep the information intact or can be stated what is received should be the same as what is sent, 3) Authentication, which is related to the validity of the owner of the information [1].

Applications on the internet already use encrypted communication that has advantages in security toreduce the risk of data loss. Inresearch conducted by[2]aspects included in encrypted data,traffic includes Peer to Peer (P2P) applications, Security and Privacy (Security Socket Layer (SSL), Virtual Private Network (VPN), and Secure Shell (SSH)). One use of encrypted data types is Hypertext Transfer Protocol Secure (HTTPS) which uses secure socket layer (SSL). SSL is a type of sockets communications between transmission protocol/internet protocol (TCP / IP) transmission and the application layer. However, from the results of several previous studies, there had been difficulties in recognizing encrypted packet patterns because this type of package is more complex compared to other data types [3].

There are three[4]to classify encrypted packages, namely; 1) fine-grained classification of encrypted traffic, 2) scale datasets generation and labeling, and 3) overcoming countermeasures against traffic analysis.

There are many solutions that have been used in identifying packages types on the internet using machine learning (ML). Research conducted by [5] uses feature selection in classifying data packages. Research by [6] uses WSU_AUC and SRSF method to optimize data traffic, as well as research conducted by[7] and dissertations project[8] use feature selection to recognize anomaly packets on the network,and research by[9] use information gain for traffic classification.

One method of feature selection uses the ranking feature. Where the ranking feature is a method that generates a value for each attribute, each attribute is calculated and sorted by score, where the score is the degree of relevance that depends on the application.In research conducted by [10] and [11], FR is used for the attribute selection process. There are five methods used to find the ranking of attributes, namely Information Gain (IG), Gain Ratio (GR), Symmetrical Uncertainty (SU), Relief-F (RF) and Entropy-based One R[12].Basically, entropy is used to interpret the uncertainty of some attributes of a data set.The higher the entropy of an attribute then the higher the uncertainty value.

To find the value of entropy used Equation (1) as follows:

$$H(Y) = -\sum_{y\epsilon Y} p(y) log_2(p(y)) \qquad (1)$$

Where p (y) is a marginal probability function for a random variable of value Y.If the Y values measured in the S dataset are divided by the second feature values X, and entropy Y on the partition is affected by X less than Y's entropy before partitioning process, then there is a relationship between Y and X features, then the expression of entropy Y after observing X is shown in Equation (2)

$$H(Y|X) = -\sum_{x\epsilon X} p(x) \sum_{y\epsilon Y} p(y|x) log_2 p((y|x)) \qquad (2)$$

Conditional*p (y | x)* is the probability of y against x.

Information Gain*(IG)*

Entropy as invalid criteria in training data set S, it can be defined as the size is additional information about Y which is provided by X which represents the number where the entropy Y value decreases. This section is known as Information Gain (IG*)*

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y) \qquad (3)$$

Revised Manuscript Received on September 22, 2019.

Tasmi,Computer Engineering, Computer Science Department, Universitas Indo Global Mandiri Indonesia

Herri Setiawan,Engineering Informatics, Computer Science Department, Universitas Indo Global Mandiri Indonesia

Deris Stiawan,Computer Engineering, Computer Science Department, Universitas Sriwijaya Indonesia

Husnawati,Computer Engineering, Computer Science Department, Universitas Indo Global Mandiri Indonesia

Sasut Analar Valiata,Master of Engineering Informaticst of Computer System, Department, Universitas Sriwijaya

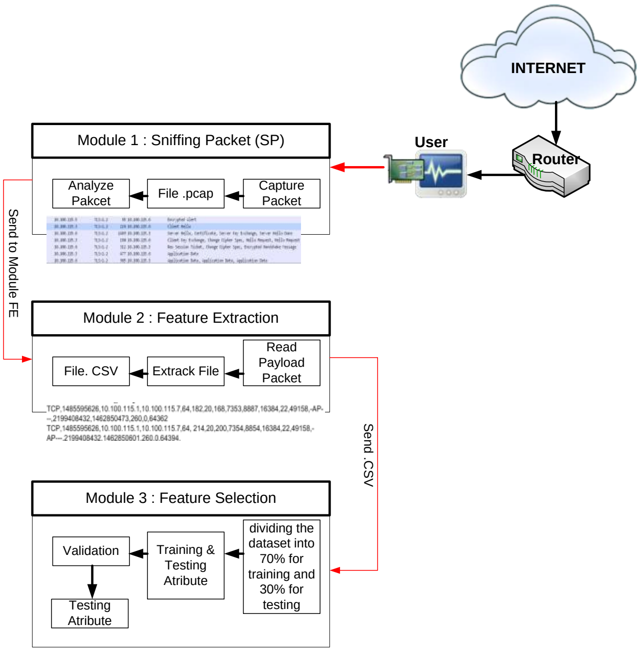
From Equation (3) the information obtained about Y after observing X is the same as the information obtained about X after observing Y.

One R

One-R is an algorithm that will generate a rule by creating a rule for each attribute and then selecting the rule with the smallest error.

II. Research Method

At this stage, it will produce strong attributes to be used as a pattern in recognizing a package. In this study,feature selection is used to find influential features and override features that have no effect. Many algorithms are used in the feature selection process include ranking features. This study used two copy-based ranking feature methods, namely: (1) Information Gain (IG), (2) One-R (OR). The procedure around the Attribute crackingshown in Fig. 1.


INTERNET
Router
User
Module 1 : Sniffing Packet (SP)
Analyze Pakcet
File .pcap
Capture Packet
Send to Module FE
Module 2 : Feature Extraction
File. CSV
Extrack File
Read Payload Packet
Send .CSV
Module 3 : Feature Selection
Validation
Training & Testing Atribute
dividing the dataset into 70% for training and 30% for testing
Testing Atribute

```
TCP,1485595626,10.100.115.1,10.100.115.7,64,182,20,168,7353,8887,16384,22,49158,-AP-
--,2199408432,1462850473,260,0,64362
TCP,1485595626,10.100.115.1,10.100.115.7,64, 214,20,200,7354,8854,16384,22,49158,-
AP---.2199408432.1462850601.260.0.64394.
```

Fig. 1 Research framework

There are three steps to get the encrypted packet attribute using Feature Ranking. The first stage is sniffing packet which is used to get the dataset. The second stage is processing, namely: feature extraction used to get information from the attributes of the dataset.

Stage three is feature selection which is used to find influential attributes and remove features that have no effect. The process is carried out in three parts; first dividing the dataset into 70 percent for training and 30 percent for testing, second validation by classification methods and the last part is testing attribute as knowledge of encrypted packages.

Dataset

The process of sniffing in real-time, where data will be directly inspected and classified as based on the application protocol. In the process of capturing this, it will be conditioned on several computers that do browsing activities and then will be captured, the length of capturing time is adjusted to the conditions of the length of the user activity process.

Procedure for Capturing Data

```
input
capturing data ← parsing packet
interface ← eth device
        begin
$while capturing\_ data \leftarrow 1$
$if interface \leftarrow 1$
                      $if protocol \leftarrow TCP$
                      $print packet\_data$
                                 $endif$
$else\ Break;$
$end\ if$
 $end\ while$
```

Feature Extraction

The feature extraction process is needed to analyze data packages for search patterns and signatures from traffic. The feature extraction process is the conversion process in the data package, data that is still in the form of raw data resulting from capturing is converted into clear text so that the future analysis process will be easier.

Procedure Feature Extraction

```
input
read data ← sniffing interface
write data ← write to file csv
protocol ← TCP
begin
    whilereading data ←1
for each row on reading data
for each column on reading data
if protocol position (row, column) is occupied
                  if protocol ← 1 then
                           write data
end if
              else return 0; break
              end if
                       else return 0; break
end for
end for
          end while
```

III. Experimental Results and Discussion

Sniffing Packet

In this study, the initial stage is to produce a dataset,so that it builds an algorithm using the libpcap library for the packet sniffing process as. This algorithm defines the interface used, time sniffing and output produces "*file.out*".

In the process of snatching packet to obtain raw data done by process "*dev_handle = pcap_open_live (dev_in, 65536, 1, 0, errbuff)*" where dev_in is the name of the interface used for the sniffing process, 65536 is the maximum amount of data to be synchronized, 1 used to activate the device in the form of promiscuous mode, 0 defines the timeout time, errbuff is a buffer that is used to store error messages. Then the package obtained from sniffing will be converted into numeric form based on the following features (a) IP header (b) TCP header (c) UDP header, the time interval used inthe 20s, 40s, 60s, 80s, so raw 100s the data obtained is not too widespread.

The results of testing the packet sniffing based on the protocol communication process are shown in table 1. The results obtained from the sniffing packet based on communication protocols and applicationsused, where the data retrieval process is done in the 20s, 40s, 60s, 80s, and 100s. So that the data obtained is more accurate. The results obtained from the test show that the data generated is linear (number of data packets with time), so it can produce the number of the file to be processed.

Table. 1 Packet Sniffing Results based on the Communication Protocol

Time	QT Packet	File Size (MB)	TCP	UDP	Packet Drop
20s	6632	1,25	96,00%	2,01%	1,99%
40s	15835	3,00	88,82%	11,06%	0,13%
60s	22276	4,00	58,20%	2,16%	39,63%
80s	30314	5,63	73,51%	1,27%	25,22%
100s	38147	6,56	95,56%	1,68%	2,76%
Time					
	File Size (MB)	TCP	UDP	Packet Drop	File Size (MB)
20s	1,41	45,79%	2,18%	52,03%	1,41
40s	3,08	56,06%	2,34%	41,60%	3,08
60s	3,82	57,30%	1,50%	41,20%	3,82
80s	5,52	91,39%	1,29%	7,32%	5,52
100s	7,24	95,09%	1,44%	3,47%	7,24

When testing is done, the first step is to determine the capture time which is used as the interrupt program function, then determine the interface to be used, namely eth0 as outbound and eth1 inbound.

The package drops of 52.03 percent occurred at the time of the 20s for the dataset II, 1.99 percent for the dataset I trial, because of the difference in conditions at the time of the router sniffing process. In the trial dataset I, the router has provided several services such as DHCP and DNS. Whereas in the dataset II experiment, the timing of sniffing along with the router gave some useful service to the user, while the stable one was generated for testing for the time of 100s and the packet dropped by 3.47 percent and 2.76 percent.

The results of testing packet sniffing applications obtained the number of the packet for the trial time of the 20s, 40s, and 100s of data produced in the second experiment is greater than the number of packages in the first experiment as presented in Fig. 2. The 60s and 80s were more than the data in the second experiment. This difference in number occurs because of the addition of several applications accessed by users such as SMTP, FTP,and SSH.

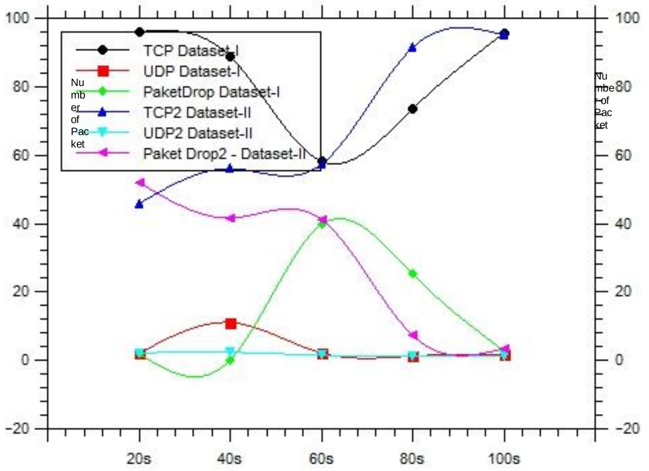


Fig. 2 Results of TCP and UDP Package Sniffing

Feature Ranking

The next stage in this research is the feature selection process. The purpose of feature selection is to clear out irrelevant features. Many previous studies have used machine learning for the feature selection process to get strong attribute values.

The research proposed to research [13]that used ranking features and Support Vector Machines (SVM) [14]for packet detection anomaly. The ranking feature is one method that produces values for each attribute, each attribute is calculated and sorted by score, where the score is the degree of relevance that depends on the application. This study used two methods that are used to find the ranking of attributes, namely Information Gain (IG) and OneR, which are entropy-based, basically, entropy is used to interpret the uncertainty of some attributes of a data set. The higher the entropy of an attribute, the higher the uncertainty value.The order of each attribute based on the entropy value and ranking used shown in Table 2.

Table. 2 Attribute Ranking Results

No	Attribute	IG		OneR	
		Entropy	Ranking	Entropy	Ranking
1	Protocol	0,0000	19	96,0375	18
2	IP_Source	0,0782	2	96,3780	8
3	IP_Dest	0,1162	1	96,0375	19
4	TTL	0,0015	11	96,3780	7
5	Data_Capture	0,0394	6	96,4420	5
6	Lenght_Header	0,0000	17	96,9540	2
7	Total_Lenght	0,0394	5	96,4420	4
8	Iden__Header	0,0294	8	96,6340	3
9	Checksum_Header	0,0003	12	96,4420	9
10	Fragment_Offset	0,0248	9	96,3140	10
11	P_Source	0,0678	4	96,3140	11
12	P_Dest	0,0678	3	96,3140	12
13	Flags	0,0315	7	96,0375	17
14	Ack	0,0012	13	96,2292	16
15	Seq	0,0007	14	96,2292	15
16	Window	0,0029	10	96,2292	14
17	Urg_Pointer	0,0000	18	96,2292	13
18	Checksum_Protocol	0,0013	15	96,4210	6
19	Service	0,0000	16	98,7000	1

In Table 2 the order of each attribute based on the entropy value and ranking used are presented, such as the service attribute required by the OneR method with an entropy value of 98,700, while the IP_Dest attribute is needed for the IG method with an entropy value of 0,1162, but on the contrary on the OneR method, the IP_Dest attribute is not very influential. Three attributes, namely, Urg_Pointer, Protocol and Length_Header, are stated to be the lowest value by the IG ranking method with an entropy value of 0,000. Protocol, IP_Dest, and Flags have an entropy value of 96.0375 which is the attribute that has the lowest value in the OneR ranking method.

For testing attributes of the dataset, the distribution will be divided into 70 percent for training and 30 percent used for testing. Based on the attribute ranking results, it will be validated by the classification method to find the best attributes. Attributes that have high presentation values from the results of training will be tested to get new knowledge.

Each attribute obtained from ranking will be validated using the classification method Bayesian Network. Comparison of accuracy values is presented in Fig. 4 using the *Naïve Bayes Classifier* method with several feature ranking methods, where the accuracy of classification with the Information Gain (IG) method obtains accuracy below 50 percent, whereas, OneR gets results above 50 percent.

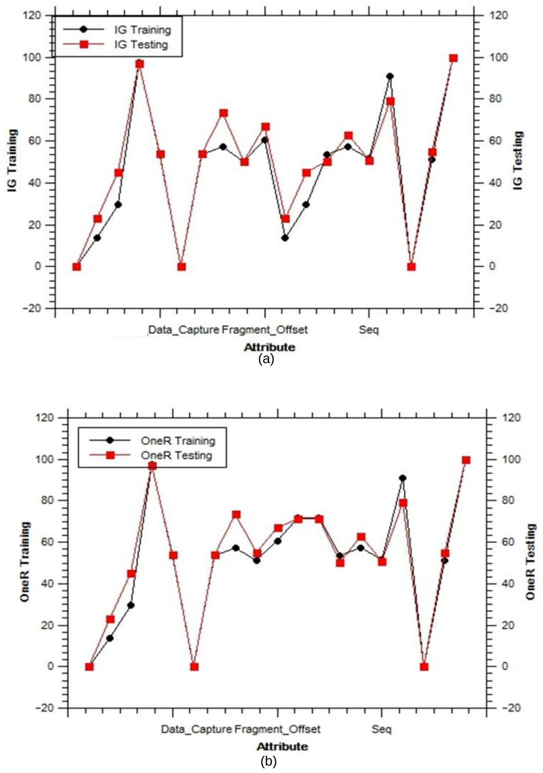


Fig. 4Ranking Attribute

Fig. 4 displays the results of training and testing ranking using the IG and OneR methods which are validated using Bayesian Network. The results obtained by the highest accuracy of the classification results from all methods are TTL attributes 97.2944 percent, window attributes 91.0311 percent obtained for the IG method, while the OneR windows attribute values are 87.5586 percent.

The ranking results OneR for P_Source attributes were 71.9003 % and P_Dest was 71.9003 %, but the P_Source attribute was 13.6770 % and P_Dest was 29.4205 % for the IG method, it means that these two attributes do not require IG method. From the results of training data and testing calculation from the data set, there are three attributes that have a value of 0,000 percent, namely Urg_pointer, Lenght_Header and Protocol, which indicates these attributes do not contribute to the type of encrypted packet traffic, while 16 attributes have values greater than 0 meaning that these attributes are relevant to the characteristics of encrypted traffic

IV. CONCLUSIONS

In this research, the selection of features and classification methods is done to find the attribute rank and delete irrelevant features, by calculating the value of each attribute and sorted by score. The results showed that the two ranking methods (IG and OneR) obtained from the TTL attribute had the highest accuracy value of 97.2944% and three attributes (protocol, Urg_Pointer Lenght_Header) had an accuracy of 0,000%. This is considered to have no effect on both. In the next ranking method, the highest rank will be determined as new knowledge in detecting encrypted packets. In the future, several classification methods can be implemented to find the best accuracy value.

ACKNOWLEDGMENT

Thank you so much for the great support from many people. And a big thanks to Mr. Deris Stiawan, P.hD for their guidance, advice, and efforts, which are very important at all stages of this research. Also, very grateful to the leadership of the Indo Global Mandiri University and Yayasan for their attention and support this research.

REFERENCES

1. P. Panwar and D. Kumar, “Security through SSL,” vol. 2, no. 12, pp. 178–184, 2012.
2. 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.
3. H. Nazief, T. Sabastian, A. Presekal, and G. Guarddin, “Development of University of Indonesia Next Generation Firewall Prototype and Access Control With Deep Packet Inspection,” *ICACSIS*, pp. 47–52, 2014.
4. Z. Cao, G. Xiong, Y. Zhao, Z. Li, and L. Guo, “A Survey on Encrypted Traffic Classification,” *Appl. Tech. Inf. Secur.*, pp. 73–81, 2014.
5. G. Urvoy-keller, “Application-based Feature Selection for Internet Traffic Classification,” *2010 22nd Int. Teletraffic Congr. (lTC 22)*, 2010.
6. H. Zhang, G. Lu, M. T. Qassrawi, Y. Zhang, and X. Yu, “Feature selection for optimizing traffic classification,” *Comput. Commun.*, vol. 35, no. 12, pp. 1457–1471, 2012.
7. M. Mantere, M. Sailio, and S. Noponen, “Network Traffic Features for Anomaly Detection in Specific Industrial Control System Network,” pp. 460–473, 2013.
8. A. Bhandari and M. O. F. Science, “Feature Selection for Anomaly Detection,” *thesis*, no. June, 2015.
9. A. F. Oklilas, Tasmi, S. D. Siswanti, M. Afrina, and H. Setiawan, “Attribute Selection Using Information Gain and Naïve Bayes for Traffic Classification Attribute Selection Using Information Gain and Naïve Bayes for Traffic Classification,” 2019.
10. R. Ruiz, J. C. Riquelme, and J. S. Aguilar-Ruiz, “Heuristic Search over a Ranking for Feature Selection,” *Comput. Intell. Bioinspired Syst.*, pp. 742–749, 2005.
11. 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.
12. M. Slocum, “DECISION MAKING USING ID3 ALGORITHM,” vol. 8, no. 2, pp. 1–12, 2012.
13. A. Jung, “Online Feature Ranking for Intrusion Detection Systems,” *arXiv*, vol. 2, no. May, 2018.
14. Husnawati, G. F. Fitriana, and S. Nurmaini, “The development of hybrid methods in simple swarm robots for gas leak localization,” *IEEE*, pp. 197–202, 2017.