Revealing the Feature Influence in HTTP Botnet Detection

Nur Hidayah M. S, Faizal M. A, Siti Rahayu Selamat, Rudy Fadhlee M. D, Wan Ahmad Ramzi W. Y

Department of System and Computer Communication, Faculty of Information and Communications Technology, Universiti Teknikal Malaysia Melaka (UTeM), Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.

***Abstract*:** Botnet are identified as one of most emerging threats due to Cybercriminals work diligently to make most of the part of the users' network of computers as their target. In conjunction with that, many researchers has conduct a lot of study regarding on the botnets and ways to detect botnet in network traffic. Most of them only used the feature inside the system without mentioning the feature influence in botnet detection. Selecting a significant feature are important in botnet detection as it can increase the accuracy of detection. Besides, existing research focusses more on the technique of recognition rather than uncovering the purpose behind the selection. Therefore, this paper will reveal the influence feature in botnet detection using statistical method. The result obtained showed the accuracy is about 91% which is approximately acceptable to use the influence feature in detecting botnet activity.

***Keywords*:** Feature Selection, Botnet, HTTP Botnet, Statistical Approaches

1. Introduction

Nowadays, a botnet is widely utilized as a part of different cyber-attacks which prompt to the genuine dangers to our own system resources and organization's properties. As the networks develop massively in size and complexity, botnet detection is an exceptionally difficult issue which makes attention of several researchers to distinguishing the expanding issue of the malicious activities. Lately, botnets have been identified as one of the most developing threats to the security of the Internet which become much attention. These threats will make genuine harm corporate or government organizations, for example, what happened on February 2017, an enormous DDoS attack on Luxembourg's government servers that reportedly lasted more than 24 hours, and more than a hundred websites are affected. Other than that, in January 2016, online banking website and mobile application of HSBC's were temporarily knocked offline by a DDoS attack [1]. DDoS attack is directly attacking a single site or system which is usually targeted to companies, institutions and even governments and security companies [2]. As the result, it will distribute spams through network backbones, facilitating the denial of legitimate access, service become unavailable and loss in financial. In addition, on November 2016, about 4.5% of 20 million fixed-line customer of Deutsche Telekom customer in Germany was hit by network outage due to Mirai botnet [3]. This type of botnet is intended to transform network devices into remotely controlled "bots" that can be utilized to mount huge-scale network assaults. Thus, the essential action should be consider to protect organization from any security malicious threats that occur inside the network.

Intrusion Detection System (IDS) is a network system that built for analyze network activity and recognize suspicious pattern which can be harm to the network. By introducing this IDS, the possible damage that occur inside the network may be reduce. Anomaly-based and signature-based are two types of approaches to detect intrusion activity. Anomaly-based is a technique that monitoring unusual network traffic activities such as high traffic volumes and traffic on unusual ports. Meanwhile, signature-based is one of detection methods that observes the system streams and finds the examples of existing botnets which can perform prompt recognition and requires a low-asset to prepare.

Moreover, choosing significant features is important before introducing IDS as a defense tool. This is because the success of detection are depend on a set of features that involved in detecting botnet activity. Besides, there are no more specific features in botnet detection that might be used for detecting botnet attack in the network. Each researcher utilizes distinctive names for a similar of subset while others utilize a similar name but different type. Notwithstanding the significance in choosing the significant features, there no specific research on a set of the features that might use in detecting HTTP botnet activity. The existing research focusses more on the technique of recognition rather than uncovering the purpose behind the selection. Moreover, most of researcher only used the feature inside the system without mentioning the influence feature in botnet detection.

Therefore, it is necessary to reveal influence feature in botnet detection in order to overcome the difficulty to recognize the botnet activity in the network. In conjunction with that, this paper will reveal influence feature in botnet detection using statistical approach and comparative analysis from earlier researcher. Then, a set of selected features will be recommended to be used in detecting botnet activity in the network. The remainder of this paper is presented as follows: Section II discusses about related study and section III presents the methodology use for this paper. Section IV presents the analysis of the results. Section V concludes the paper and presents future work directions.

2. Related Work

The discovering significance feature is very important in order to avoid the problem such as redundant and duplication dataset. This is because it can reduce the misclassification of data and produce the best set of feature from dataset then produce the better rate of detection. For example, in the study [4] find an optimal set of features in detecting breast cancer from microarray dataset. This set of features produce better performance in detection accuracy which is capable to detect the breast cancer using active learning with accuracy 94%. In contrast with author [5], uses the minimum subset of feature from the original set to get efficiency load and price forecasts of electric power. From their result, eliminating an ineffective input of feature shows the improvement of forecast accuracy. However, [6] selecting significant features by using fuzzy

rule framework. The author control the redundancy to make their system to avoid measurement error in a particular feature. The experiment result of this author proves that selected feature with redundant control provide the best performance compared to the dataset using free parameters.

In 2017, [7] acquire the set of features per sample to predict the metastasis in endometrial cancer. The feature is used to test on training and testing cohort. The result of the training cohort attain 100% accuracy while in testing cohort it divide by node-positive cases (90% accuracy) and node-negative cases (80% accuracy). This output causes to the significant enhancements in the estimation of lymphatic metastases in endometrial cancer patients. The study by [8], suggest a method of adaptive feature combine with feature distinctive degree to verify standard compact descriptors for visual search. The result takes 10% mean average precision and increase with very low bit rate mode in top match rate. Moreover, [9] selecting an optimal features to decrease the rate error of classifier and increase the rate of estimation with additional accuracy. Contrast with author [10], he used the method of embedded feature in bug prediction. The result of his study shows the method decrease the prediction error of the regressors and increase their stability. All this existing research was been carried out the study about significant feature however the researcher does not focus on botnet detection. Thus, this section also will be explained about the definition of flow, feature selection and statistical approaches.

2.1 Flow Definition

According to [11] a set of IP packets which passing an observation point during a certain time interval in the network is known as a flow. In a particular flow, every packet has its own common properties. Each property has their own information or feature that can be used in detecting the presence of botnet activity. This statement supported by [12]. He stated that the flow-based feature is important in order to identify the presence of bot in the network. He also clarifies that botnet activity can be recognize quickly before the task during C&C finished. Besides, the high accuracy of detection can be provided by monitoring the traffic flows [13]. A study from [14] used the flows of traffic to recognize botnet activity and location of zombie computers. Meanwhile, [15] stated that the detection of a botnet can be identified by observing the flow-based traffic features. The flow features must be extracted from packet headers in order to choose the most suitable feature. Therefore, selecting a significant feature is essential as it depends on feature to produce a better result in botnet detection.

2.2 Feature Selection

Feature selection is the method of selecting a subset of the variable in the training set and only use his subset as features to provide the better prediction results. According to [16], feature selection is a technique to eliminate the redundant and unnecessary features. Eliminate the redundant and unnecessary features are important since some of the features may have a subset of another feature. Feature selection can be categorized into two categories known as filter model and wrapper model. The filter model use estimation purposes and mostly rely on the properties of underlying data meanwhile wrapper model discovers appropriate features through repetitive application of the classification algorithm with different sets of features. For this project, a wrapper model (forward selection) will be used. This is because according to [17], wrapper model is the best method to detect botnet as it is can determine the most effectiveness subset of features that yield accuracy of detection. This method also are inclined to over-fit the informations particularly when the measure of information is insufficient [18].

The difficulty to detect HTTP botnet as it hiding behinds the normal HTTP flow are the reason [19] to identify the abnormal flow in web traffic. He has used source port number, destination port number, source IP address, destination IP address and protocol detect the presence of abnormal activity. These five features are used by author differentiate abnormal web traffic from regular web requests. Moreover, [20] used six tuples as a features in his research by discovering the similar patterns of communication and behaviors to detect malicious activity in the network. Source port number, destination port number, source IP address, destination IP address, protocol and number of packets are the six tuples that have been used by this researcher to recognize the malicious activity. The author [21] study that SYN flag, FIN flag and PSH flag of TCP connections give significant information related to the existence of web-based botnet. He uses this three feature in neural network and conclude that only SYN flag and FIN flag are essential in detecting botnet. Their result shows that this feature are able to detect HTTP botnet although the message being encrypted. Besides, [22] monitor the host that generates the failure patterns for a short period to detect a new bot in the network. He used the source IP address, the destination IP address, the source port number, and the destination port number in order to recognize the existence of abnormal activity in the network. In addition, based on the statement [23] HTTP botnet do not keep alive the connection to C&C servers and will terminate the session then re-establish for new transactions. This will lead to the number of outgoing TCP connection attempts to become large. Thus, in order to detect this situation occur, the ratio of incoming to outgoing TCP packets per time interval and the ratio of TCP packets to the total number of packets per time interval need to be found. Meanwhile, [24] calculate the amount of the length of HTTP reply packet payloads for a set of bunched flows. The author recommends that the payloads established for legitimate requests are tend to be bigger conversely. The author also deliberates that if the sum of payloads underneath the value threshold of 2 KB to be suspicious.

Other than that, [25] extract GET or POST requests in HTTP traffic and group by the similarity of the message. This author defines three features in order to identify botnet traffic which is periodic factor, the range of absolute frequencies and the time sequence factor. All this three feature is a measure related to the group by similar HTTP message. However, [26] study packet size to detect low-rate DDoS attack in the network traffic. He discovers the universal entropy of packet sizes by defining the smaller result in packet size are tend to be attacked. He also stated that the attack is detected when the distance between the probability distribution of packet sizes greater than the value of the threshold. The research study [27], observing the traffic of single host to detect HTTP botnet. Entropy of time gap and packet count are two feature that produced from traffic model. Based on their perception, benign HTTP activity bursty in nature and has a higher estimation of this both

features contrasted with bot movement which is considerably more occasional in nature.

Furthermore, [28] proposed a minimum set of influence features to detect fast attack. His study reveal three influence feature in detecting fast attack which is src_count, srv_count and dst_count. The result from experiment show that 3 connection per second to distinguish intrusion activity at attacker and 3 connection per second to distinguish intrusion activity at the victim. Meanwhile, [29] used the features based on a host which can detect botnet precisely. The result of their research produce the highest rate of botnet detection. In 2017, [30] obtain the set of features from NSL-KDD intrusion dataset by using discretized differential evolution and C4.5 machine learning algorithm. Their result shows a significant change in detection accuracy which is able to detect new attack with 88.73% accuracy. Therefore, understanding the relationship between the features that influence in detecting botnet activity is necessary in order to avoid features selected redundant in the botnet detection. Previous researcher only used the feature inside the system without mentioning the influence feature in botnet detection. Hence, this paper will expose feature that influence in botnet detection.

2.3 Statistical Approach

According to [26] statistical methods is a set of principles and procedures used by successful scientists in their pursuit of knowledge which involves data collection, data summarization and statistical analysis of related observed data. Some researcher uses this technique by using the estimated values of the parameters to detect botnet. This statement supported by [31] which highlight the calculation of statistical parameters such as maximum or minimum can be used to distinguish anomalous activity. Besides, [32] use a statistical approach to recognize and to reduce critical malicious patterns in malware families, which are vital features towards automated classification of identified and unidentified malware in large amount. Their study also introduce the novel formalization methodology defined as a statistical approach which automates the identification of critical malicious patterns for each malware family which is more consistent compared to related works.

Moreover, [33] proposed a method using statistical based features combine with a machine learning method in order to address the problem of dependency on the signature of network packets. The review objectives is to look at the utilization of straightforward measurable elements in rule-based system to identify network interruptions. From the studies, the authors extract seven statistical features from network traffic for detection of intrusive behavior which simplifies the effectiveness of their Intrusion Detection System (IDSs) against varied types of network attacks. However, [34] installed spam botnets to capture network traffic and characterize this network traffic in order to identify the main activities. Variations were observed in the behavior/features of different spamming botnets after intensive statistical analysis, which can further be explored to design various spam botnet detection techniques. As a result, all botnets contain some individual elements that can be discovered to improve the differences of botnet detection techniques.

In addition, [35] have proposed botnet-versus network setting and virulence estimation approach based on random sampling along with a novel statistical learning technique. The authors claimed that maximum likelihood approximation was used in their statistical method to botnet recognition. Meanwhile, [36] present a novel botnet detection system that is capable to detect stealthy P2P botnets. The authors classify all hosts within a monitored network that perform to be engaging in P2P communications and then derive statistical fingerprints of the P2P communications generated by these hosts. The author also leverages the obtained fingerprints to distinguish between hosts that are part of legitimate P2P networks and P2P bots [36].

Furthermore, [37] use logistic regression to calculate the probability that a packet contains malware. This approach is the best method compared to the current signature detection and anomaly detection methods. It is because logistic regression can replace all the signatures related to a single malware family with the same accuracy as signature detection. Other than that, logistic regression is more accurate with less false positives than the anomaly detection methods. Thus, statistical approach is the best way to detect HTTP botnet.

3. Methodology

In order to understand how to detect botnet in the network, the testbed is implemented. The network traffic from internal to external source will be captured by connecting one host running on Linux to mirror port. One mirror port has been configured on the network switch in order to monitor the network traffic on the firewall interface. The network card inside the host needs to be configured as well after configuring the network switch. The network card needs to set as a promiscuous mode. Therefore, the monitored host can capture any botnet activity inside the network traffic as shown in Figure 1. Meanwhile, the capturing process for HTTP botnet started with configuring HTTP botnet environment setup. Then, the captured packet will be run using crontab service and will be saved in *tcpdump.gz file. That file will be located at the root privileged in dumpit folder for the next step.

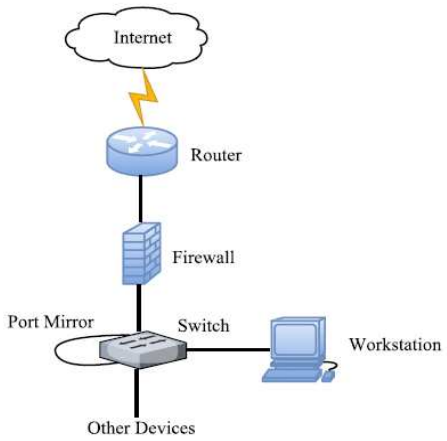

Figure 1. Network Design

Besides, the process of the feature selection can reveal feature that influence in botnet detection. After pre-processing data is done, the data will be analyzed by using a feature selection algorithm. Figure 2, shows the process of feature selection. From the figure, the capture packet which

involve normal and botnet dataset has been collected and going through data preprocessing.

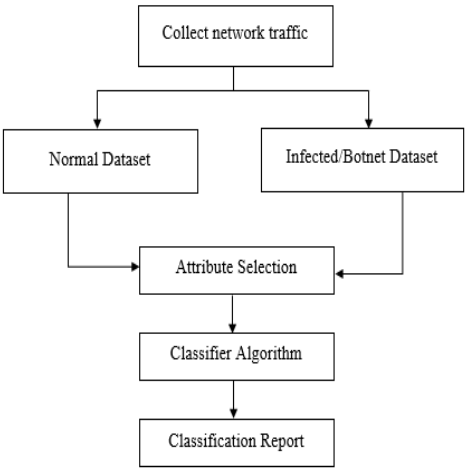

Figure 2. Process of Feature Selection

After that, data (57 of the feature) will be analyzed by using feature selection to select the features that can be used in machine learning classifier. Then, the dataset will be trained by using three classifier algorithms which are Naïve Bayes, Decision Tree and Random Forest. Table 1, shows the description of the classification algorithm.

Table 1. Result accuracy of Classifier [38]

Classification Algorithm	Description
Naïve Bayes	Based on the Bayes rule of conditional probability. It makes use of all the characteristics contained in the data, and analyses them individually as though they are similarly significant and independent of each other.
Decision Tree	A predictive machine-learning model that chooses the objective estimation of a new sample based on several characteristic values of the available data.
Random Forest	A troupe learner technique that produces numerous individual learners and totals the outcomes. The best parameter at every node in a decision tree is produced using an arbitrarily chosen number of components.

The comparisons of accuracy percentage between this three classifier have been done in order to identify the best and suitable result that can be used for the next phase. Table 2, shows the result of classifier with different operators. From the figure, it concludes that optimize operator and the forward operator give a better performance in accuracy of detection. As mentioned earlier, the researcher decided to select forward selection because its yield accuracy of detection. Lastly, the classification will be generated with the selected features that involved for botnet detection.

Moreover, the Likelihood Ratio test and Wald test were the test that give a trivial contribution to the model in detecting botnet attack. The feature that influence in botnet detection can reveal by exploring the significant contribution of the feature. When the result of feature selected gave a good influenced to the model then, the feature can be used. In addition, other researchers who focused on botnet detection can refer the result of the feature influence in this paper for future studies. Figure 3 shows the process of feature influence.

Table 2. Result accuracy of Classifier

Operator	Classifier		
	Random Forest	Naïve Bayes	Decision Tree
Backward	91.12%	16.35%	91.92%
Forward	91.65%	91.08%	91.90%
Optimize	91.65%	91.08%	91.90%

Each of selected feature (7 of the feature) will testing by using a Likelihood Ratio test and Wald test model. After that, all the influence features will be analyzed in order to prove whether the model gave a decent impact to the model or not in expecting the result.

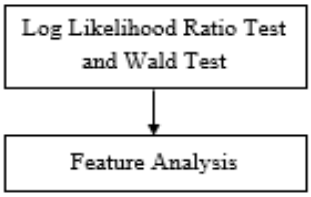

Figure 3. Process of Feature Influence

The Likelihood Ratio Test (1) and Wald Test (2) were two techniques that involve in assessing the contribution of the feature [39]. This two test can be used to specify whether the features have a good prediction of the result or outcome variable.

(i) Log-Likelihood Test

The Likelihood Ratio Test is the comparison between model with and without a particular predictor. The predictor the model gave a decent affected in foreseeing the result when the reduction values of the likelihood ratio model without the predictor if the predictor was incorporated inside the model.

The equation of the log-likelihood ratio test are:

$x^2 = 2$ [Log Likelihood(New(with predictor) – Log Likelihood(Baseline(without predictor)) (1)

(ii) Wald Test

Wald test has a special distribution known as the chi-square distribution which is used to estimate the statistical consequence of each coefficient of b in the model. Wald test tells us whether the b coefficient for that predictor is significantly different from zero [39]. The indicator is making a huge commitment to the expectation of the result when the coefficient is altogether not the same as zero by utilizing condition (2).

$$\text{Wald} = b / SE_b \quad (2)$$

In this manner, if the estimation of the Wald test inside the model is greater, then the indicator give a significance commitment to the model in anticipating the result.

4. Result

In this paper, only 7 are selected from 57 of feature that can be used for detecting botnet activity in the network as shown in Table 3 below. All these seven features have their own characteristic which is important in detecting HTTP botnet.

Table 3. Feature Selection for Botnet Detection

Feature	Description
avg_segm_size_b2a	The normal segment size detected during the lifetime of the connection calculated as the value reported in the actual data bytes field divided by the actual data packets reported.
initial_window_bytes_a2b	The entire number of bytes sent in the first window
unique_bytes_sent_b2a	The quantity of restrictive bytes sent.
max_win_adv_a2b & max_win_adv_b2a	The higher number of window advertisement have been discovered. When both sides negotiated window scaling, then it is the maximum window-scaled advertisement seen.
min_segm_size_a2b	The smallest number of segment size detected during the lifetime of the connection.
max_segm_size_a2b	The higher number of segment size detected during the lifetime of the connection.

Besides, revealing the feature influence and purpose behind the collection of the feature is a good opportunity as most of previous researchers did not focused about it. The reason behind revealing the influencing feature may help to assess the significant contribution of the feature used to detect botnet. Furthermore, by understanding that, it may help to give knowledge about the relationship of the feature in contributing a role to detect the botnet activity. The result was discussed based on the statistical value from likelihood ratio test and Wald test.

4.1 Avg_segm_size_b2a Feature

This feature has been selected for detecting botnet activity in the network. By using a statistical value from the likelihood ratio test, it validates that avg_segm_size_b2a influence in botnet detection. Table 4 demonstrate the value of the likelihood ratio statistic after the element is incorporated inside the model. Hence, the value of the baseline model (without the indicator) can be figured by referring to the equation of the Log Likelihood test. At the point when just the consistent was incorporated, -2LL = 404980.373, however avg_segm_size_b2a has been included this has been decreased to 402052.506. The features have a critical affected at expecting the result (botnet) based on the reduction value of the likelihood ratio.

Table 4. Avg_segm_size_b2a Summary

-2 Log Likelihood	Wald
402052.506	2173.349

Table 4 also demonstrates the value of Wald is 2173.349. According to [26] if the value was different from zero, it showed that the indicator gave a decent impact to the model in expecting the result. The chosen feature gave a decent affected to the model in expecting the result since the value of the outcome was significantly different from zero.

4.2 Initial_window_bytes_a2b Feature

Initial_window_bytes_a2b also gave a significant influence in botnet detection. Table 5 demonstrate the value of the likelihood ratio statistic after the element is incorporated inside the model. Meanwhile, when just the consistent was incorporated, -2LL = 402052.506 and this has been decreased to 400666.766. This diminishment implies that the features have a critical affected at expecting the result (botnet).

Table 5. Initial_window_bytes_a2b Summary

-2 Log Likelihood	**Wald**
400666.766	7445.696

The Wald value for the new model which is 7445.696 as shown in above. Since, the result significantly different from zero it means that the feature selected gave a decent affected to the model in in expecting the result.

4.3 Unique_bytes_sent_b2a Feature

Table 6 demonstrates the value of the likelihood ratio statistic after the element is incorporated inside the model. In this way, the value of the baseline model (without the predictor) can be figured. At the point when just the consistent was incorporated -2LL = 400666.766 and this value has been decreased to 400651.702. This diminishment implies that the elements have a critical affected at expecting the result (botnet). Meanwhile, the value of the Wald test for this feature is 0.322.

Table 6. Unique_bytes_sent_b2a Summary

-2 Log Likelihood	**Wald**
400651.702	0.322

4.4 Max_win_adv_a2b Feature

The result from the analysis indicate that max_win_adv_a2b gave a decent affected to the model in in expecting the botnet detection. Table 7 demonstrate the value of the likelihood ratio statistic after the element is incorporated inside the model. When just the consistent was incorporated, -2LL = 400651.702 and this value has been decreased to 398849.06. This diminishment implies that the elements have a critical affected at expecting the result (botnet). Table 7 also demonstrates the value of Wald is 0.854.

Table 7. Max_win_adv_a2b Summary

-2 Log Likelihood	Wald
398849.06	0.854

4.5 Min_segm_size_a2b Feature

Table 8 demonstrates the estimation of the likelihood ratio statistic after the element is incorporated inside the model. In this way, the estimation of the baseline model (without the indicator) can be figured. At the point when just the consistent was incorporated, -2LL = 398849.06 and this value has been decreased to 378687.144. This diminishment implies that the elements have a critical affected at expecting the result (botnet).

Table 8. Min_segm_size_a2b Summary

-2 Log Likelihood	Wald
378687.144	13961.988

The Wald value for the new model which is 13961.988 as shown in Table 8. Since, the result significantly different from zero which means that the feature selected gave a decent affected to the model in in expecting the result.

4.6 Max_segm_size_a2b Feature

Max_segm_size_a2b also gave a significant influence in botnet detection and the validation was made by using the same test with other influence feature. Table 9 demonstrate the value of the likelihood ratio statistic after the element is incorporated inside the model. When just the consistent was incorporated, -2LL = 378687.144 and this value has been decreased to 378445.002. This diminishment implies that the features have a critical affected at expecting the result (botnet). Meanwhile, the value of Wald test is 0.124.

Table 9. Max_segm_size_a2b Summary

-2 Log Likelihood	Wald
378445.002	0.124

4.7 Max_win_adv_b2a Feature

The result from the analysis indicate that max_win_adv_b2a gave a decent affected to the model in in expecting the botnet detection. Table 10 demonstrate the value of the likelihood ratio statistic after the element is incorporated inside the model. When just the consistent was incorporated, -2LL = 378445.002 and this value has been decreased to 377280.474. This diminishment implies that the elements have a critical affected at expecting the result (botnet). Table 10 also demonstrates the value of Wald is 0.354.

Table 10. Max_win_adv_b2a Summary

-2 Log Likelihood	Wald
377280.474	0.354

Therefore, from the discussion above it conclude that from seven feature selected, only three features that gave a decent affected to the model in expecting the result. The influence features are avg_segm_size_b2a, initial_window_bytes_a2b and min_segm_size_a2b with the value of Wald test is different from zero, which is 2173.349, 7445.696 and 13961.988 respectively. These three features can be used to identify the threshold selection.

5. Conclusions and Future Work

Choosing significant features are important as it give a contribution in terms of accuracy of detection. Most of researcher only focused on the method of recognition instead of revealing the reason behind the selection. The previous researcher only used the feature inside the system without mentioning the influence feature in botnet detection. Besides, in order to avoid redundant features, understanding the relation between influence features may reduce the potentials of choosing unnecessary feature which might give an effect in detecting botnet activity. Thus, this paper will reveal the feature that influence in botnet detection by using statistical approach and comparative analysis from earlier researcher. For future work, we would like to implement an experiment based on selecting features and produce the suitable value of threshold in detecting of HTTP botnet activity in the network.

6. Acknowledgement

This work has been supported under Universiti Teknikal Malysia Melaka research grant FRGS/1/2015/ICT04/FTMK/02/F00292 and KPT MyBrain15. The authors would like to thank to Universiti Teknikal Malaysia Melaka and all members of INSFORNET research group for their incredible supports in this project.

References

[1] Warwick Ashford, 2017. Lloyds Bank hit by massive DDoS attack. Retrieved from http://www.computerweekly.com/news/450411443/Lloyds-Bank-hit-by-massive-DDoS-attack [Accessed on March 8, 2017].

[2] Khattak, S., Ramay, N.R., Khan, K.R., Syed, A.A. and Khayam, S.A., "A taxonomy of botnet behavior, detection, and defense," IEEE Communications Surveys & Tutorials, Volume 16, No. 2, pp. 898-924, 2014.

[3] Eric Auchard, 2016. "German internet outage was failed botnet attempt: report". Retrieved from http://www.reuters.com/article/us-deutsche-telekom-outages-idUSKBN13N12K [Accessed on February 13, 2017].

[4] Begum, S., Bera, S.P., Chakraborty, D. and Sarkar, R., "Breast cancer detection using feature selection and active learning," In Computer, Communication and Electrical Technology, pp. 43-48, CRC Press, 2017.

[5] Abedinia, O., Amjady, N. and Zareipour, H., "A New Feature Selection Technique for Load and Price Forecast of Electrical Power Systems," IEEE Transactions on Power Systems, Vol. 32, Issues 1, pp.62-74, 2017.

[6] Chung, I.F., Chen, Y.C. and Pal, N., "Feature selection with controlled redundancy in a fuzzy rule based framework," IEEE Transactions on Fuzzy Systems, Vol. PP, Issue 99, pp. 1-1, 2017.

[7] Ahsen, M.E., Boren, T.P., Singh, N.K., Misganaw, B., Mutch, D.G., Moore, K.N., Backes, F.J., McCourt, C.K., Lea, J.S., Miller, D.S. and White, M.A., "Sparse feature selection

for classification and prediction of metastasis in endometrial cancer," BMC genomics, 18(3), p. 233, 2017.

[8] Zhu, C., Jia, H., Lu, T., Tao, L., Song, J., Xiang, G., Li, Y. and Xie, X., "Adaptive feature selection based on local descriptor distinctive degree for vehicle retrieval application," In Consumer Electronics (ICCE), 2017 IEEE International Conference on Las Vegas, NV, USA, pp. 66-69, IEEE, January, 2017.

[9] Radha, P. and Divya, R.," Multiple time series clinical data with frequency measurement and feature selection," In Advances in Computer Applications (ICACA), IEEE International Conference on Coimbatore, India, pp. 250-254, IEEE, October, 2017.

[10] Osman, H., Ghafari, M. and Nierstrasz, O., "Automatic feature selection by regularization to improve bug prediction accuracy," In Machine Learning Techniques for Software Quality Evaluation (MaLTeSQuE), IEEE Workshop on Klagenfurt, Austria, pp. 27-32, IEEE, February, 2017.

[11] B. Claise, 2008. "Specification of the IP flow information export (IPFIX) protocol for the exchange of IP traffic flow information". Retrieved from https://tools.ietf.org/html/rfc5101 [Accessed on March 8, 2017].

[12] Zhao, D., Traore, I., Sayed, B., Lu, W., Saad, S., Ghorbani, A. and Garant, D., "Botnet detection based on traffic behavior analysis and flow intervals," Computers & Security, 39, pp. 2-16, 2013.

[13] Stevanovic, M. and Pedersen, J. M., "An efficient flow-based botnet detection using supervised machine learning," In Computing, Networking and Communications (ICNC), 2014 International Conference on Honolulu, HI, USA, pp. 797-801, IEEE, February, 2014.

[14] W. Tarng, K. Ou, M. Chen, "The analysis and identification of P2P botnet's traffic flows," International Journal of Communication Networks and Information Security, Vol. 3, No.2, pp. 138-148, 2011.

[15] Saad, S., Traore, I., Ghorbani, A., Sayed, B., Zhao, D., Lu, W., Felix, J. and Hakimian, P., "Detecting P2P botnets through network behavior analysis and machine learning," In Privacy, Security and Trust (PST), 2011 Ninth Annual International Conference on Montreal, QC, Canada, pp. 174-180, IEEE, July, 2011.

[16] Bolon-Canedo, V., Sanchez-Marono, N. and Alonso-Betanzos, A., "Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset," Expert Systems with Applications, Vol. 38, No. 5, pp. 5947-5957, 2011.

[17] Beigi, E.B., Jazi, H.H., Stakhanova, N. and Ghorbani, A.A., "Towards effective feature selection in machine learning-based botnet detection approaches," In Communications and Network Security (CNS), 2014 IEEE Conference on San Francisco, CA, USA, pp. 247-255, IEEE, October, 2014.

[18] Brown, G., Pocock, A., Zhao, M. J. and Luján, M., "Conditional likelihood maximisation: a unifying framework for information theoretic feature selection," Journal of Machine Learning Research, pp. 27-66, January, 2012.

[19] Chen, C.M., Ou, Y.H. and Tsai, Y.C., "Web botnet detection based on flow information," In Computer Symposium (ICS), 2010 International on Tainan, Taiwan, Taiwan, pp. 381-384, IEEE, December 2010.

[20] Zeidanloo, H.R., Manaf, A.B., Vahdani, P., Tabatabaei, F. and Zamani, M., "Botnet detection based on traffic monitoring," In Networking and Information Technology (ICNIT), 2010 International Conference on Manila, Philippines, pp. 97-101, IEEE, June, 2010.

[21] Venkatesh, G.K. and Nadarajan, R.A., "HTTP botnet detection using adaptive learning rate multilayer feed-forward neural network," In IFIP International Workshop on Information Security Theory and Practice, Springer Berlin Heidelber, pp. 38-48, June, 2012.

[22] Huang, C.Y., 2013. "Effective bot host detection based on network failure models," Computer Networks, Vol. 52, No. 2, pp. 514-525, 2013.

[23] Cai, T. and Zou, F., "Detecting HTTP botnet with clustering network traffic," In Wireless Communications, Networking and Mobile Computing (WiCOM), 2012 8th International Conference on Shanghai, China , pp. 1-7, IEEE, September, 2012.

[24] Eslahi, M., Rohmad, M.S., Nilsaz, H., Naseri, M.V., Tahir, N.M. and Hashim, H., "Periodicity classification of HTTP traffic to detect HTTP Botnets," In Computer Applications & Industrial Electronics (ISCAIE), 2015 IEEE Symposium on Langkawi, Malaysia , pp. 119-123, IEEE, April, 2015.

[25] Xiang, Y., Li, K. and Zhou, W., "Low-rate DDoS attacks detection and traceback by using new information metrics," IEEE Transactions on Information Forensics and Security, Vol. 6, Issue. 2, pp.426-437, 2011.

[26] Ott, R.L. and Longnecker, M.T., 2010. An introduction to statistical methods and data analysis. Cengage Learning.

[27] B. Soniya and M. Wilscy, "Using entropy of traffic features to identify bot infected hosts," In Intelligent Computational Systems (RAICS), 2013 IEEE Recent Advances, Trivandrum, India, pp. 13-18, December 2013.

[28] Abdollah, M. F. "Fast Attack Detection Technique for Network Intrusion Detection System" (Doctoral dissertation, Ph. D. Thesis. Universiti Teknikal Malaysia Melaka, Malaysia), 2009.

[29] Huseynov, K., Kim, K. and Yoo, P., "Semi-supervised Botnet Detection Using Ant Colony System," In 31th Symposium on Cryptography and Information Security, Kagoshima, Japan, January, 2014.

[30] E. Popoola, A. Adewumi, "Efficient feature selection technique for network intrusion detection system using discrete differential evolution and decision tree," International Journal of Network Security, Vol.19, No.5, pp. 660-669, Sept. 2017.

[31] Ghanaei, V., Iliopoulos, C.S. and Overill, R.E., "A Statistical Approach for Discovering Critical Malicious Patterns in Malware Families," In The Seventh International Conferences on Pervasive Patterns and Applications,(Patterns 2015): IARIA, 2015.

[32] D. Lavrova, A. Pechenkin, "Applying correlation and regression analysis to detect security incidents in the internet of things," International Journal of Communication Networks and Information Security, Vol. 7, No. 3, pp. 131-137, 2015.

[33] Rastegari, S., Lam, C.P. and Hingston, P., "A Statistical Rule Learning Approach to Network Intrusion Detection," In IT Convergence and Security (ICITCS), 2015 5th International Conference on Kuala Lumpur, Malaysia, pp. 1-5, IEEE, August, 2015.

[34] Sousa, R., Rodrigues, N., Salvador, P. and Nogueira, A., "Analyzing the behavior of top spam botnets," In Communications (ICC), 2012 IEEE International Conference on Ottawa, ON, Canada, pp. 6540-6544, IEEE, June, 2012.

[35] Rrushi, J., Mokhtari, E. and Ghorbani, A.A., "A statistical approach to botnet virulence estimation," In Proceedings of the 6th ACM Symposium on Information, Computer and Communications Security, pp. 508-512, ACM, March, 2011.

[36] Zhang, J., Perdisci, R., Lee, W., Sarfraz, U. and Luo, X., "Detecting stealthy P2P botnets using statistical traffic fingerprints.," In Dependable Systems & Networks (DSN), 2011 IEEE/IFIP 41st International Conference on Hong Kong, China, pp. 121-132, IEEE, June, 2011.

[37] Hughes, K. and Qu, Y., "A theoretical model: Using logistic regression for malware signature based detection," In the 10th International Conference on Dependable, Autonomic, and Secure Computing (DASC-2012), 2012.

[38] Feizollah, A., Anuar, N.B., Salleh, R., Amalina, F., Ma'arof, R.U.R. and Shamshirband, S., "A study of machine learning

classifiers for anomaly-based mobile botnet detection," Malaysian Journal of Computer Science, 26(4), 2014.
[39] Field, A., 2009. Logistic regression. Discovering statistics using SPSS, pp.264-315.