ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/322525657

Comparative Analysis for Detecting DNS Tunneling Using Machine Learning Techniques

Article *in* International Journal of Applied Engineering Research · December 2017

CITATIONS
4

READS
347

3 authors, including:

Mahmoud Sammour
Roads and Transportation authority Dubai - RTA
10 PUBLICATIONS **56** CITATIONS
SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project Ground level Ozone Hierarchical analysis View project

Project VANETs View project

All content following this page was uploaded by Mahmoud Sammour on 16 January 2018.

The user has requested enhancement of the downloaded file.

International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 22 (2017) pp. 12762-12766
© Research India Publications. http://www.ripublication.com

Comparative Analysis for Detecting DNS Tunneling Using Machine Learning Techniques

[1] **Mahmoud Sammour, [2] Burairah Hussin, [3]Mohd Fairuz Iskandar Othman**
Faculty of Information and Communication Technology, Universiti Teknikal Malaysia Melaka, Malaysia.

[1]Orcid: 0000-0002-6860-2804

Abstract

DNS tunneling is one of the issues that have concerned the information security community in the last decade. Such malicious activity resembles a legitimate threat for many corporations where there are a respected amount of network traffic that would be embedded with DNS tunneling. The threats that caused by such tunneling could be ranged from the full remote control into file transfer or even a full IP tunnel. Therefore, different approaches have been proposed for detecting the DNS tunneling such firewalls and intrusion detection systems. However, these approaches are limited to specific types of tunneling. Therefore, researchers have tended to utilize machine learning techniques due to its ability to analyze and predict the occurrence of DNS tunneling. Nonetheless, there are plenty of choices for employing specific machine learning techniques. This paper aims to provide a comparative study for three machine learning techniques including SVM, NB and J48. A benchmark dataset for the DNS tunneling has been used in the experiment in order to facilitate the comparison. Experimental results showed that SVM has the superior performance compared to the other classifiers in terms of detecting DNS tunneling by achieving 83% of f-measure.

Keywords: Domain Name System, Tunneling, Support Vector Machine, Naïve Bayes, Decision Tree, Classification

INTRODUCTION

Domain Name System (DNS) is one of the important protocols that has a vital role regarding web activities such as browsing and emailing. This can be represented by allowing applications to use names such as example.com instead of a difficult-to-remember IP address [1]. Many organizations do not consider any threats regarding the DNS because it is not related to the data transfer. Nonetheless, many companies could be vulnerable to numerous types of threats throughout the DNS [2]. This is due to the respected amount of traffic that would be subjected to the DNS threats.

Nowadays, many utilities are being available for conducting the tunneling over DNS, most of these utilities aim at gaining a free Wi-Fi access for sites that required restricted access via http [3]. However, serious threats could be happened along with gaining the free Wi-Fi access. Such threats can be represented as malicious activities that would be accommodated via the DNS tunneling. Using the DNS tunneling, a full remote control could be conducted via a channel for a compromised internet host. In addition, different activities could be conducted via the DNS tunneling such as operating system commands, file transfer or even a full IP tunnel. Feederbot [4] and Moto [5] are examples of DNS tunneling tools known to use DNS as a communication method.

All the latter mentioned threats have motivated the information security community to provide robust methods that have the ability to detect the DNS tunneling [6]. Various types of detection DNS tunneling methods have been proposed, such methods can be categorized into two main classes; Traffic Analysis and Payload Analysis. The first class aims to analyze the overall traffic where some significant features could be identified such as volume of DNS traffic, number of hostnames per domain, location and domain history. While the second class aims to analyze the payload of a single request in order to identify features such as domain length, number of bytes and content.

Analyzing the features that are related to the DNS tunneling has led the researcher community to utilize rule-based approaches in which both traffic and payload are being analyzed in terms of some features. Once a predefined condition has been occurred, the identification of DNS tunneling will be operated. However, with the complex and tedious task of manual curation of rules, researchers have tended to utilize Machine Learning Techniques (MLT). The key characteristic behind the machine learning lies on the statistical model that has the ability to identify significant rules automatically [7-9]. In addition, with the emergence of annotated datasets such as the JSON [10] which contains network connections with predefined labels (e.g. Tunneled or Legitimate), the focus on machine learning has been expanded. This is due to MLT requires historical data that is annotated. Hence, MLT would have the ability to train the model based on such data. Based on such training, a new data can be tested.

In fact, there are numerous type of MLTs such as Support Vector Machine (SVM), Naïve Bayes (NB), Decision Tree (DT), K-nearest Neighbor (KNN) and others. With this variety, it is a challenging task to identify the most suitable classifier, that would fit the process of detecting DNS tunneling. This paper aims to accommodate a comparative analysis regarding the process of DNS tunneling using three MLTs classifier including SVM, NB and DT.

RELATED WORK

As mentioned earlier, researchers nowadays have concentrated on the machine learning techniques in order to handle the task of DNS tunneling detection. For instance, Allard et al. [9] presented an approach for detecting DNS tunneling using MLT. In their work, the authors have utilized two classifiers Decision Tree (DT) and Random Forest (RF) in order to be trained on the ciphered flows by conducting a statistical analysis on the inner protocol. Such analysis aims to identify the size and the inter arrival delays of the packets in the flow.

On the other hand, Aiello et al. [3] presented an approach for detecting DNS tunneling based on the SVM classifier. The proposed approach exploited the statistical features of the DNS queries and answers. This has been conducted by analyzing the content of the queries and the answers in order to identify malicious data hidden by the legitimate DNS.

The same authors have presented an extension method using the same classifier in [8]. Such extension aims to exploit different features related to the payload analysis. This can be represented by analyzing the inter arrival times and the packet size regarding to the protocol messages.

Buczak et al. [11] have examined different types of features for the task of DNS tunneling detection. The comparison has been performed using RF classifier. The features consist of number of answers provided in the response, time between two consecutive packets for a specific domain and time between two consecutive responses for a specific domain.

Aiello et al. [12] have focused on the statistical features of DNS tunneling by using two approaches including Principal Component Analysis (PCA) and Mutual Information (MI). Such approaches aim at examining the frequent occurrence of a particular pattern in order to identify the tunneling. In addition, the proposed features extraction approaches have been combined with a KNN classifier. The proposed method showed an efficient performance in terms of traffic profiling in the existence of DNS tunneling.

Homem et al. [13] have presented an entropy method in order to classify the DNS connections into tunneling and legitimate. The proposed method analyzes the internal packet structure in order to characterize the information entropy of different network protocols with their corresponding tunnels. Consequentially, the proposed method utilized a prediction approach based on the entropy distribution averaging.

Van Thuan Do et al. [7] have addressed the detection of DNS tunneling within the mobile network. The authors have utilized the conventional SVM classifier with some features such as source, destination and length of the queries in order to identify the tunneling.

PROPOSED METHOD

The proposed method is composed of three main phases including Dataset, Feature Extraction and Classification. First phase aims to identify a benchmark dataset of DNS tunneling in order to facilitate the comparison among the classifiers. Second phase aims to exploit some features related to both payload and traffic analysis. Third phase aims at carrying out three classification methods including SVM, NB and J48. Fig. 1 depicts the framework of the proposed method.

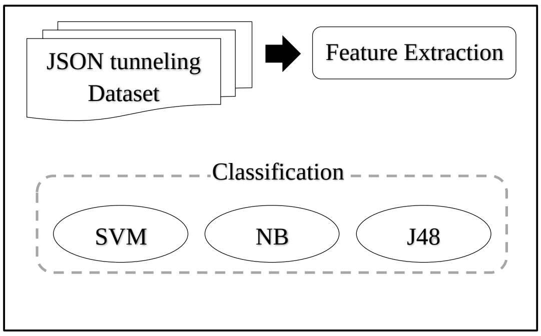

Figure1: Proposed method

JSON Tunneling Dataset

In order to facilitate the machine learning technique for the process of DNS tunneling detection, it is necessary to identify an annotated dataset that contains DNS connections with their corresponding class label (i.e. tunneled or legitimate). For this purpose Homem & Papapetrou [10] have simulated the use of different protocols in terms of their own DNS tunnel. The authors have utilized a script in Python programming language to capture the network traffic using multiple protocols including HTTP, HTTPS, FTP and POP3. For the first two protocols, the data has been generated by visiting five random websites. Whereas, for the FTP, five random files have been downloaded. Eventually, for the POP3 protocol, the data was simulated by requesting and downloading five random emails from a mail server in which some of these email contains plain text and the other contains randomly generated files as attachments. Table 1 shows the dataset details.

Table 1: Dataset details

Network Protocol	Number of Samples
HTTP	52
HTTPS	53
FTP	53
POP3	53
Total	211

Feature Extraction

One of the vital tasks in the machine learning techniques lies on employing robust set of features that has the ability to discriminate the occurrence of DNS tunneling. For this purpose, this study aims to focus on the information entropy regarding to its direct correlation with the actual data bytes composing packets. Information entropy has the ability to examine the variations among the regular content size of different protocols. In this vein, it is possible to address the unusual size for the content in different requests. The entropy is being computed as the probability of a specific byte occurrence p(x_i) multiplied by the logarithm of the probability of that occurrence, this would be summed up for all byte occurrence as in the following equation:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \times log\, p(x_i) \qquad (1)$$

Besides the information entropy, there are other features that were provided for each connection such as DNS request length, IP packet sender length, IP packet response length, encoded DNS query name length, request application layer entropy, IP packet entropy and query name entropy.

Classification

This phase aims to carry out three different classifiers including SVM, NB and J48 for the task of DNS tunneling detection. The goal behind applying such classification methods lies on the comparative analysis that would take a place in order to identify the most accurate classifier. Such classifier can be illustrated as follows:

J48

J48 or so-called Decision Tree (DT) is one of the machine learning classification methods that has a flow-chart-like tree structure [14]. Every node on such tree represents an attribute value, every branch resembles an outcome of the test, and tree leaves resembles the classes. The key characteristics behind the J48 lies on the rule-based nature of such classifier in which the attribute values are being converted into rules that distinguish the occurrence of DNS tunneling [15].

Naïve Bayes

Naive Bayes is one of the feature-based classification method in which the occurrence of a single feature is being addressed in accordance with the class labels (i.e. tunneling or legitimate) [16]. The key strength behind NB lies on its ability to independently perfrom the classification with the absence or presence of specific feature. The probability of NB can be computed as:

$$p(c_j \mid w_i) = \frac{p(c_j)\, p(c_j \mid w_i)}{p(w_i)} \qquad (2)$$

Where *P(c)* is the probability of a class label, *P(w)* is the probability of a specific feature instance, and *P(c/w)* is the probability of a specific feature instance in accordance to the class label.

Support Vector Machine

SVM is one of the classification methods that perform the classification based on dividing the data space into multiple class labels using a hyperplane [17]. Hyperplane is considered to be a margin that should be optimally identified. An accurate adjustment of the hyperplane leads to accurate classification accuracy. Identifying the optimal hyperplane can be computed as follow:

$$f(\vec{x}) = sgn\big((\vec{x} \times \vec{w}) + b\big)$$
$$= \begin{cases} +1: & (\vec{x} \times \vec{w}) + b > 0 \\ -1: & Otherwise \end{cases} \qquad (3)$$

RESULTS

Due to this study is utilizing MLT therefore; the traditional evaluation methods including precision, recall and f-measure will be used to evaluate the classification methods [13]. To compute such metrics, the contingency table will be used as shown in Table 2.

Table 2. Contingency table

	Predicted DNS connection		
	Predicted Actual	Normal	Tunnel
Actual DNS connection	Normal	*True Positive (TP)*	*False Positive (FP)*
	Tunnel	*False Negative (FN)*	*True Negative (TN)*

False Negative (FN): is the number of actual tunneling connections that have been predicted as normal.

False Positive (FP): is the number of actual normal connection that have been predicted as tunneling.

True Negative (TN): is the number of correctly un-predicted connections.

True Positive (TP): is the number of correctly predicted connections.

In this vein, the precision, recall and f-measure can be calculated based on the following equations.

$$Precision = \frac{TP}{TP + FP} \qquad (4)$$

$$Recall = \frac{TP}{TP + TN} \qquad (5)$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \qquad (6)$$

Precision is the ratio between the number of correctly classified connections over the total number of connections. While Recall is the ratio between the number of correctly classified tunneling connections and the total number of tunneling connections. Finally, F-measure is considered to be the total accuracy.

Based on the latter illustration for the evaluation method, Table 3, 4 and 4 show the results of NB, J48 and SVM classifiers respectively.

Table 3. NB results

Class	Precision	Recall	F-measure
FTP	0.77	0.90	0.83
HTTPS	0.78	0.73	0.75
HTTP	0.64	0.63	0.63
POP3	0.97	0.94	0.96
Average	0.79	0.80	0.79

Table 4. J48 results

Class	Precision	Recall	F-measure
FTP	0.88	0.90	0.89
HTTPS	0.72	0.70	0.71
HTTP	0.61	0.62	0.61
POP3	0.92	0.94	0.93
Average	0.78	0.79	0.78

Table 5. SVM results

Class	Precision	Recall	F-measure
FTP	0.78	0.88	0.82
HTTPS	0.85	0.73	0.79
HTTP	0.67	0.74	0.71
POP3	1.00	1.00	1.00
Average	0.82	0.83	0.83

In general, it is obvious that the POP3 tunneled protocol has been effectively identified by all the classifiers (96% NB, 93% J48, 100% SVM). This is due to the information entropy that successfully identified the occurrence of tunneling within such protocol. On the other hand, the HTTP tunneling was the less identified protocol by all the classifiers. This is due to various paradigms of embedding tunneling within the DNS while browsing.

Apart from the protocols, SVM has outperformed the two other classifiers by achieving an f-measure of 83% compared to 78% achieved by J48 and 79% achieved by NB. This is due to the superior performance of SVM in terms of handling multiple number of class labels.

CONCLUSION

This paper has conducted a comparative analysis for three machine learning techniques in terms of DNS tunneling detection. The proposed techniques consists of SVM, NB and J48. A benchmark dataset for the DNS tunneling has been used in the experiment. Results showed that SVM has been outperformed the other classifiers by achieving the highest f-measure. For the future researchers, combining multiple classifiers may lead to enhance the performance of classifying the tunneling.

ACKNOWLEDGEMENT

The authors would like to thank the UTeM Zamalah Scheme. This research is supported by Universiti Teknikal Malaysia Melaka (UTeM) under UTeM Zamalah Scheme.

REFERENCES

[1] G. Farnham and A. Atlasis, "Detecting DNS tunneling," *InfoSec Reading Room,* 2013.

[2] R. Rasmussen, "Do you know what your dns resolver is doing right now," *Security Week. DOI= http://www. securityweek. com/do-you-know-what-your-dnsresolver-doing-right-now,* 2012.

[3] M. Aiello, M. Mongelli, and G. Papaleo, "Basic classifiers for DNS tunneling detection," in *2013 IEEE Symposium on Computers and Communications (ISCC)*, 2013, pp. 000880-000885.doi:10.1109/ISCC.2013.6755060.

[4] C. Dietrich, "Feederbot-a bot using DNS as carrier for its C&C," ed, 2011.

[5] C. Mullaney, "Morto worm sets a (DNS) record," *Symantec Official Blog,* 2011.

[6] J. LIU and G.-y. QIU, "The firewall penetrating techniques based on the inverse connection, http-tunnel and sharing dns," *Journal of Zhengzhou University of Light Industry (Natural Science),* vol. 5, p. 014, 2007.

[7] P. E. Van Thuan Do, B. Feng, and T. van Do, "Detection of DNS Tunneling in Mobile Networks Using Machine Learning," *Information Science and Applications 2017: ICISA 2017,* vol. 424, p. 221, 2017.

[8] M. Aiello, M. Mongelli, and G. Papaleo, "DNS tunneling detection through statistical fingerprints of protocol messages and machine learning," *International Journal of Communication Systems,* vol. 28, pp. 1987-2002, 2015.

[9] F. Allard, R. Dubois, P. Gompel, and M. Morel, "Tunneling activities detection using machine learning techniques," DTIC Document2010.

[10] I. Homem and P. Papapetrou, "Harnessing Predictive

Models for Assisting Network Forensic Investigations of DNS Tunnels," 2017.

[11] A. L. Buczak, P. A. Hanke, G. J. Cancro, M. K. Toma, L. A. Watkins, and J. S. Chavis, "Detection of Tunnels in PCAP Data by Random Forests," in *Proceedings of the 11th Annual Cyber and Information Security Research Conference*, 2016, p. 16.doi.

[12] M. Aiello, M. Mongelli, E. Cambiaso, and G. Papaleo, "Profiling DNS tunneling attacks with PCA and mutual information," *Logic Journal of IGPL,* p. jzw056, 2016.

[13] I. Homem, P. Papapetrou, and S. Dosis, "Entropy-based Prediction of Network Protocols in the Forensic Analysis of DNS Tunnels," 2016.

[14] N. B. Amor, S. Benferhat, and Z. Elouedi, "Naive bayes vs decision trees in intrusion detection systems," in *Proceedings of the 2004 ACM symposium on Applied computing*, 2004, pp. 420-424.doi.

[15] R. C. Barros, M. P. Basgalupp, A. C. De Carvalho, and A. Freitas, "A survey of evolutionary algorithms for decision-tree induction," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* vol. 42, pp. 291-312, 2012.

[16] D. M. Farid, L. Zhang, C. M. Rahman, M. Hossain, and R. Strachan, "Hybrid decision tree and naive Bayes classifiers for multi-class classification tasks," *Expert Systems with Applications,* vol. 41, pp. 1937-1946, 2014.

[17] A. Puri and N. Sharma, "A NOVEL TECHNIQUE FOR INTRUSION DETECTION SYSTEM FOR NETWORK SECURITY USING HYBRID SVM-CART," 2017.

View publication stats