FEATURES SELECTION FOR IDS IN ENCRYPTED TRAFFIC USING GENETIC ALGORITHM

Mehdi Barati[1], Azizol Abdullah[2], Ramlan Mahmod[3], Norwati Mustapha[4], and Nur Izura Udzir[5]

[1]Universiti Putra Malaysia (UPM), Selangor, Malaysia, ac.barati@yahoo.com

ABSTRACT. Intrusion Detection System (IDS) is one method to detect unauthorized intrusions into computer systems and networks. On the other hand, encrypted exchanges between users are widely used to ensure data security. Traditional IDSs are not able to reactive efficiently in encrypted and tunneled traffic due to inability to analyze packet content. An encrypted malicious traffic is able to evade the detection by IDS. Feature selection for IDS is a fundamental step in detection procedure and aims to eliminate some irrelevant and unneeded features from the dataset. This paper presents a hybrid feature selection using Genetic Algorithm and Bayesian Network to improve Brute Force attack detection in Secure Shell (SSH) traffic. Brute Force attack traffic collected in a client-server model is implemented in proposed method. Our results prove that the most efficient features were selected by proposed method.

Keywords: feature selection, genetic algorithm, IDS, encrypted traffic

INTRODUCTION

In a shared network environment, unprotected traffic is posed some security attacks. Moreover, network users including online banking customers, need to be sure about their data integrity and privacy. The best method to provide data privacy and integrity is encrypting these traffics. All users in the network environment must be able to communicate securely with each other as well as with servers. Therefore, in a network environment and Internet as the most popular network, it is critical that traffic containing sensitive information is encrypted. Due to increasing trend of using Cloud computing, IPv6, Voice Over IP (VOIP), online banking, and P2P application the encrypted traffic is a huge part of any network traffic. Attackers take advantage of this increasing usage of encrypted traffic in order to launch attacks hidden in encrypted traffic. Therefore, encrypted attacks bring new vulnerabilities in the security field. Most of traditional attacks can apply this technique to bypass any detection system. In addition, encapsulating one application protocol by another one is also used to evade detection policy (Dusi, Crotti, Gringoli, & Salgarelli, 2009). An Intrusion Detection System (IDS) is a system used to detect unauthorized intrusions into computer systems and networks. IDS technology has been used for many years to defend valuable resources. Many kinds of IDS have been proposed by researchers, however proposing IDS in encrypted environment is still a challenging task. That is because of the fact that Deep Pcket Inspection (DPI) technique with tremendous computational cost is unable to detect encrypted attacks, and also causes user privacy violation. Therefore traditional IDS became progressively inaccurate in front of encrypted traffic. According to (Koch, 2011), the next generation IDS has to address some significant requirements including encrypted traffic. Therefore, some sophisticated method should be proposed to detect intrusion in encrypted and tunneled traffic.

Recently, some IDS have been proposed for encrypted environment, but none of them have provided acceptable accuracy.

ENCRYPTED AND TUNNELING OVERVIEW

End-to-end encryption ensures the confidentiality and integrity of data. SSH is a secure login program and its handshake contains three basic steps. It stars by transport layer protocol to support secure connections and continues by authentication step to negotiate the encryption and key exchange methods. Finally, it establishes an encrypted connection to provide Shell. In addition to encryption, tunneling method by using encapsulation technique can provide a channel for data transition. It operates with a client–server model which the client is inside the network and is connected to the server outside the network boundary. Some tunneling protocols such as SSH deploy both encapsulation and encryption in their method. By growth of such encrypted networks in the future especially with the adoption of IPv6 that has built-in encryption capability, IDS in encrypted traffic is a significant issue in this field of research.

ENCRYPTED TRAFFIC IDS OVERVIEW

The first step to design appropriate IDS is to understand the current technology of IDS. IDSs can be classified in to two groups including Host-based IDS (HIDPS) and Network-based IDS (NIDPS). The most noteworthy limitation of NIDS is inability to detect attack in encrypted traffic. One common method applied in IDS is Machine Learning (ML) technique. It focuses on prediction, based on some features learned from the training data. The focus of this research is in this kind of method. Three fundamental approaches to implement IDS in encrypted traffic are reviewed as follow.

Encryption Protocol based IDS

In encryption protocol based IDS, the misuse of the encryption protocols is detected. When an encryption protocol does not behave in an expected pattern, it can be assumed that the protocol is being misused. Remote timing attack (Brumley & Boneh, 2003) is a good illustration of this kind of attack. Some presented IDSs including (Joglekar & Tate, 2004), (Sperotto, Sadre, Boer, & Pras, 2009), (Fadlullah & Taleb, 2010), and (Hellemons & Hendriks, 2012) have been proposed to detect this kind of attack against encrypted protocols. Their method is proper to detect attack on encrypted protocol itself, but it is not able to detect malicious activities hidden inside the encrypted tunnel.

Modification based IDS for Encrypted Traffic

Network infrastructure is modified to detect the intrusion in this approach. The encrypted traffic is duplicated and sent to both receiver and IDS by using an additional encryption layer to keep the confidentiality. This approach uses a framework that retrieves decrypted network traffic for DPI technique. (Li & Zhang, 2009) and (Goh, Zimmermann, & Looi, 2010)propose IDS by integrating NIDS into an encrypted network. Their approach is not suitable to be implemented in all environments due to inability of decryption in some scenarios.

Traffic Analysis based IDS for Encrypted Traffic

Analyzing encrypted traffic statistically to gain information from frequently observed patterns is the last approach. The key idea is that flow based information are enough to infer the nature of the application protocol that generated those traffic. (Borders & Prakash, 2004) proposed one method for detecting HTTP tunnels and named as Web tap. They emphasized on detecting spyware and backdoors over HTTP traffic. Another paper (Yamada, Miyake, Takemori, Studer, & Perrig, 2007) presented an approach using only data size and timing information for encrypted web accesses. Information from each web client which is encrypted is extracted and access frequencies are calculated. Then, malicious activities are detected using the rules generated from the accesses frequency. Attacks are detected without decryption using the intrusion signatures for SSH in (Foroushani, Adibnia, & Hojati, 2008) and anomaly detection for SSL in (Augustin & Balaz, 2011). Their results demonstrated detection with undeniable false alarm. In (Dusi et al., 2009) a technique for blocking unwanted tunneled traffic proposed by characterizing legitimate uses of application-layer protocols. Their method is capable to detect this traffic by fingerprinting allowed protocols even in encrypted tunneled protocol including SSH. Results demonstrated that Tunnel Hunter is able to detect legitimate tunnels with high accuracy except P2P tunneled traffic. Due to high false alarm rate, their systems are not efficient enough for a production environment. Therefore, a hybrid method using GA and Bayesian Network is proposed to find the most efficient feature set for encrypted and tunneled traffic IDS and improve the detection accuracy.

OUR PROPOSED MODEL

Feature Selection

Feature selection is one basic step in preprocessing of data mining. The reason is that number of features is usually very huge in real problems. Therefore, a frequent challenging issue in IDS is to select the most efficient feature subset to improve the detection accuracy and speed as well as reduce dimensionality of data. Using all features from the dataset can cause large memory and disk usage, and make the detection phase very slow. Therefore, the aim of feature selection is choosing the most representative features which may have the most discriminative power over a dataset (Tsai, Eberle, & Chu, 2013). Various features are used in IDS which are generated by using packet headers, payload, or protocol handshaking. A hybrid method using Genetic Algorithm is applied for feature selection in the current research.

Genetic Algorithm

Genetic Algorithm (GA) is a kind of search method suited for optimizing which maintains a population of potential solutions. It continues the search until find the best solution or termination situation appears. For the first time, GA was used by John Holland (Holland, 1975) in the computer world for developing system. A collection of solutions for one problem are represented by chromosomes composed of genes. Each chromosome represents the properties of one solution and algorithm tries to select the best chromosome or actually solution using fitness function. Indeed, fitness function presents the amount of similarity of chromosome to the answer of search problem. Algorithm creates new chromosomes from one or two parents by crossover mechanism. Moreover, in order to prevent selecting any local results by search algorithm, it uses another mechanism as mutation. New chromosomes with high fitness function added to the initial population. In other words, chromosomes that are fitted to the answer survive (Renner & Ekárt, 2003). A simple procedure of GA is summarized by Figure 1.

Proposed Feature Selection using GA

As abovementioned, GA is used as an efficient search algorithm to find the optimum solution. In this research, this algorithm together with Bayesian Network classifier are used for feature selection of IDS in encrypted traffic. In addition, using flow-based features instead of packet features can improve the classification of encrypted traffic (Alshammari & Zincir-Heywood, 2011). Indeed a flow is a sequence of packets from a source to a destination passing an observation point in the network during a certain time interval. Figure 2 illustrates our hybrid flow-based feature selection model that consists of Genetic Algorithm and

Bayesian Network classifier. In chromosome (feature subset) selection phase, fitness function is applied by GA to find the best feature subset. In this model, the classification accuracy is used as fitness function. It means that in each generation the selected features are implemented in classification to calculate the accuracy. If the termination condition (specific number of iteration or accuracy) is not met, it creates a new generation by using crossover and mutation operation. It continues this trend until find the best chromosomes or feature set in term of accuracy.

1. Random initial population is created.

2. Each chromosome is evaluated using the fitness function.

3. If the termination condition (specific number of iteration or suitable fitness of chromosome) is met, the best chromosome is returned.

4. Best chromosomes are chosen based on fitness values and new chromosomes are created by implementing genetic operators including Crossover and Mutation.

5. New population is created by adding new chromosomes and deleting the least fit chromosomes.

6. Actions starting from step 2 are repeated until the termination condition is met.

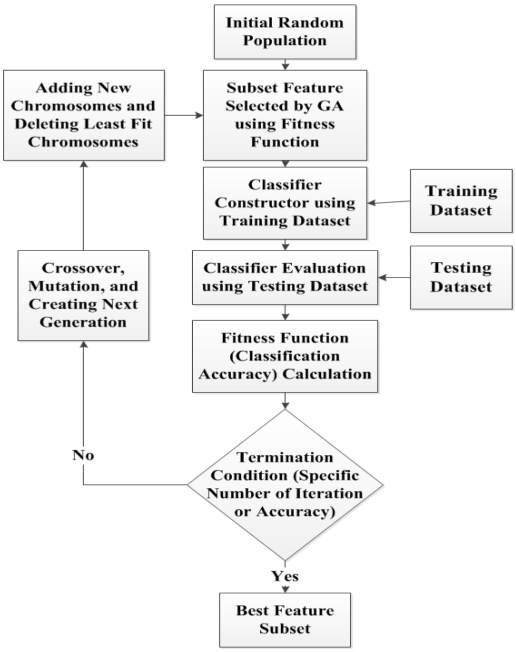

Initial Random Population
Adding New Chromosomes and Deleting Least Fit Chromosomes
Subset Feature Selected by GA using Fitness Function
Classifier Constructor using Training Dataset
Training Dataset
Crossover, Mutation, and Creating Next Generation
Classifier Evaluation using Testing Dataset
Testing Dataset
Fitness Function (Classification Accuracy) Calculation
No
Termination Condition (Specific Number of Iteration or Accuracy)
Yes
Best Feature Subset

Figure 1. Simple procedure of GA

Figure 2. Proposed Feature Selection

Data Collection

To evaluate the proposed method efficiency, the traffic regarding a Brute Force attack was launched by implementing a client-server model in SSH. Figure 3 demonstrates the data collection scenario in an encrypted SSH environment. SSH protocol was deployed in our research because of the fact that it uses both tunneling and encryption techniques in establishing connection. The attack is implemented by a client to make a secure connection to server and traffic was captured using a traffic collector on server side. In addition to traffic related to Brute Force attack, traffic in non-attack scenario was also captures and stored in the server.

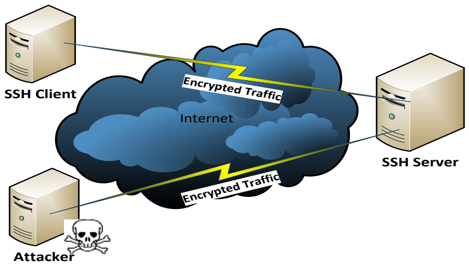


Figure 3. Data Collection Scenario

Both traffic collections are imported to feature extraction stage. Packets in the traffic are assembled and converted into a set of flows in this section. 42 important features regarding the flows were extracted by analyzing these flows statistically. Labeling of both traffic collections by attack and non-attack labels is also done in data preprocessing step.

Implementation

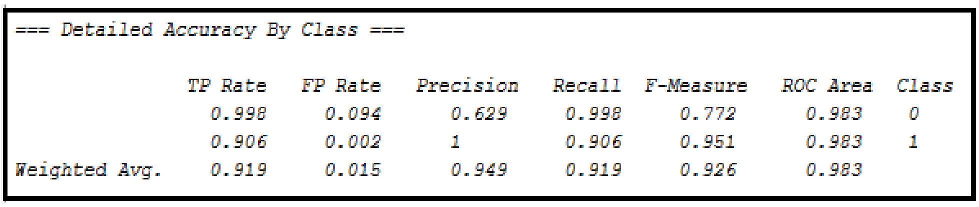
Traffic collections captured in data collection step are implemented in our proposed method for feature selection purpose. As abovementioned, a hybrid method using GA and Bayesian Network was used in feature selection step. In GA implementation, population size and the number of generation were set to 20 and 5 respectively. Moreover, to provide a valuable results 10 fold cross validation was used in our method.

Results

42 flow based features were used in our research and 12 most efficient features were selected as the best features by our model. Table 1 presents selected features and their descriptions. These features produced the best results in classification phase in terms of Precision, Recall, and Receiver Operating Characteristic (ROC) area respectively compared to other study. Precision is the fraction of retrieved instances that are relevant although Recall is the fraction of relevant instances that are retrieved. ROC curve is created by plotting True Positive Rate versus False Positive Rate. The area under the curve also is used to evaluate a classifier. According to Figure 4, the average Precision, Recall, and ROC area values are 0.949, 0.919, and 0.983 respectively. The proposed model also produced a deniable False Positive Rate (FPR) around 1.5%. Comparing to previous study in SSH traffic IDS, it can be revealed that our results are very promising. Selected features in this research are also able to contribute encrypted traffic classification problems.

Table 1. Selected Features using GA

Feature	Description	Feature	Description
total_fpackets	Total Forward Packets	min_bpktl	Minimum backward Packet Length
fpsh_cnt	Forward Push Flag Count	min_fiat	Minimum Forward Inter-arrival Time
furg_cnt	Forward Urgent Flag Count	max_fiat	Maximum Forward Inter-arrival Time
burg_cnt	Backward Urgent Flag Count	std_fiat	Standard deviation of Froward Inter-arrival Time
mean_fpktl	Mean Forward Packet Length	min_active	Minimum Active Time
max_fpktl	Maximum Forward Packet Length	min_biat	Minimum backward Inter-arrival Time



```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.998     0.094     0.629      0.998    0.772      0.983    0
                 0.906     0.002     1          0.906    0.951      0.983    1
Weighted Avg.    0.919     0.015     0.949      0.919    0.926      0.983
```

Figure 4. Classification Results

CONCLUSION AND FUTURE WORK

Due to growing trend of encrypted networks traffic usage, detection of intrusion hidden in encrypted or tunneled traffic is a challenging task. This research emphasizes on IDS in encrypted traffic and provides a review of recent proposed IDSs in this context. Proposed systems are not appropriate enough for the defined requirement due to the shortcomings already shown. A weak point of all these systems are the low efficiency. For a comprehensive development of encrypted traffic IDS in productive environments, false alarms should be minimized. In this study, efficient features for IDS in encrypted traffic are selected using GA. The idea of using the most efficient features instead of all features for classification is shown promising by results. Implementing another kind of attacks other than Brute Force attack and also evaluating the proposed method using a larger dataset to prove the robustness of the system are left as future work.

REFERENCES

Alshammari, R., & Zincir-Heywood, a. N. (2011). Can encrypted traffic be identified without port numbers, IP addresses and payload inspection? *Computer Networks*, *55*(6), 1326–1350. doi:10.1016/j.comnet.2010.12.002

Augustin, M., & Balaz, A. (2011). Intrusion detection with early recognition of encrypted application. *Intelligent Engineering Systems (INES)*, 245–247. Retrieved from http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5954752

Borders, K., & Prakash, A. (2004). Web tap: detecting covert web traffic. *Proceedings of the 11th ACM conference on Computer and Communications Security, Washington DC, USA*, 110–120.

Brumley, D., & Boneh, D. (2003). Remote Timing Attacks are Practical. *12th USENIX Security Symposium*.

Dusi, M., Crotti, M., Gringoli, F., & Salgarelli, L. (2009). Tunnel Hunter: Detecting application-layer tunnels with statistical fingerprinting. *Computer Networks*, *53*(1), 81–97. doi:10.1016/j.comnet.2008.09.010

Fadlullah, Z., & Taleb, T. (2010). DTRAB: combating against attacks on encrypted protocols through traffic-feature analysis. *IEEE/ACM Transactions*. *18*(4), 1234–1247. Retrieved from http://dl.acm.org/citation.cfm?id=1959381

Foroushani, V. A., Adibnia, F., & Hojati, E. (2008). Intrusion detection in encrypted accesses with SSH protocol to network public servers. *2008 International Conference on Computer and Communication Engineering*, 314–318. doi:10.1109/ICCCE.2008.4580619

Goh, V., Zimmermann, J., & Looi, M. (2010). Experimenting with an intrusion detection system for encrypted networks. *International Journal of Business*, *5*, 172–191. Retrieved from http://inderscience.metapress.com/index/M768686PP3007869.pdf

Hellemons, L., & Hendriks, L. (2012). SSHCure: a flow-based SSH intrusion detection system. ... *Networks and Services*, 86–97. Retrieved from http://www.springerlink.com/index/6J666T627857V614.pdf

Holland, J. (1975). Adaptation in natural and artificial systems. *Ann Arbor: The University of Michigan Press*.

Joglekar, S. S. P., & Tate, S. R. S. (2004). ProtoMon: embedded monitors for cryptographic protocol intrusion detection and prevention. *Information Technology: Coding*, *11*(1), 81–88 Vol.1. doi:10.1109/ITCC.2004.1286430

Koch, R. (2011). Towards Next-Generation Intrusion Detection. *Cyber Conflict (ICCC), 2011 3rd International*, 151–168.

Li, L., & Zhang, Z. (2009). An intrusion detection model orienting towards encrypted conversation. *2009 2nd IEEE International Conference on Computer Science and Information Technology*, 541–545. doi:10.1109/ICCSIT.2009.5234889

Renner, G., & Ekárt, A. (2003). Genetic algorithms in computer aided design. *Computer-Aided Design*, *35*(8), 709–726. doi:10.1016/S0010-4485(03)00003-4

Sperotto, A., Sadre, R., Boer, P. De, & Pras, A. (2009). Hidden Markov Model modeling of SSH brute-force attacks. *Integrated Management*, Retrieved from http://www.springerlink.com/index/09u3281x23682j2w.pdf

Tsai, C.-F., Eberle, W., & Chu, C.-Y. (2013). Genetic algorithms in feature and instance selection. *Knowledge-Based Systems*, *39*, 240–247. doi:10.1016/j.knosys.2012.11.005

Yamada, A., Miyake, Y., Takemori, K., Studer, A., & Perrig, A. (2007). Intrusion Detection for Encrypted Web Accesses. *21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07)*, 569–576. doi:10.1109/AINAW.2007.212