Comparative Evaluation of Algorithm based Approach for Intrusion Detection using a Hybrid Model

Shardul S. Mahadik
Computer Department
K. J. Somaiya College of Engineering
Mumbai, India

Sukanya S. Parsekar
Computer Department
K. J. Somaiya College of Engineering
Mumbai, India

Sushant B. Choudhary
Computer Department
K. J. Somaiya College of Engineering
Mumbai, India

Pallavi S. Kulkarni
Computer Department
K. J. Somaiya College of Engineering
Mumbai, India

ABSTRACT

Adequate system security is the first step towards data integrity and protection, however even with the most advanced protection, modern computer and communication infrastructures are susceptible to various types of attacks. With traditional signature based systems losing proficiency, the Hybrid Intrusion Detection System (HIDS) approach proves the vitality of detecting intrusions and anomalies, simultaneously, by automated data mining over network traffic and signature generation. This paper will focus on analyzing different anomaly detection techniques used to detect zero day attacks and an automatic attack signature generation mechanism that can be complemented with the former. This will serve to be an elemental analysis of a few techniques, their working, and their pros and cons put together in a concise form.

General Terms:

System Security, Intrusion Detection, Anomaly Detection, Signature Generation

Keywords:

hybrid intrusion detection system, online oversampling principal component analysis, change point detection, shiryaev roberts, f-sign, automatic signature generation

1. INTRODUCTION

In Hybrid Intrusion Detection System, both the signature based and anomaly based systems are integrated in a sequential manner. In the training phase, a database of known attack signatures is constructed for the signature based system, and normal attack-free traffic is passed through the feature extractor of the Anomaly Detection System (ADS) to generate the episode rule database [1]. When the traffic data passes through a signature matching engine, it detects any known attacks but novel attacks can bypass it. If the traffic, which is passed to the ADS, cannot find any match with the normal traffic rules in the database then an anomaly is detected. Hence both the systems, used in series, improve the effectiveness of an IDS. Once an anomaly is detected, a signature generation mechanism generates a unique signature which can be used to match the activity with a signature based system in the future.

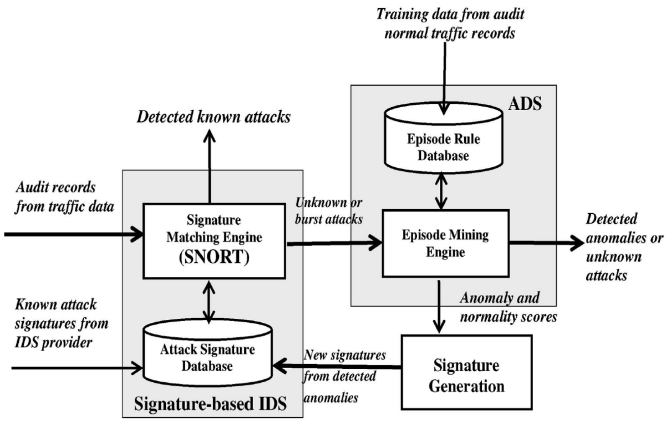

Fig. 1. Basic architecture of HIDS [1]

As the already existing signature based tools like SNORT [2], [3] are mature enough and have scope to be integrated easily with an ADS, this review focuses on analyzing a couple of algorithms relating to the anomalous episode detection based on two fundamentally different principles. Online Oversampling Principal Component Analysis (osPCA) [4] uses Principal Component Analysis (PCA) which is well known dimension reduction method, whereas the other relies on Changepoint Detection using Shiryaev-Roberts technique [5]. After detection of anomalous episodes, signatures need to be generated, which can be stored in the attack signature database for further use. F-Sign [6] which is an automatic attack signature generation mechanism complements to generate signatures specific and sensitive in nature. The scholarly material regarding these procedures is studied, analyzed and evaluated to gain detailed information about the same.

2. RELATED WORK

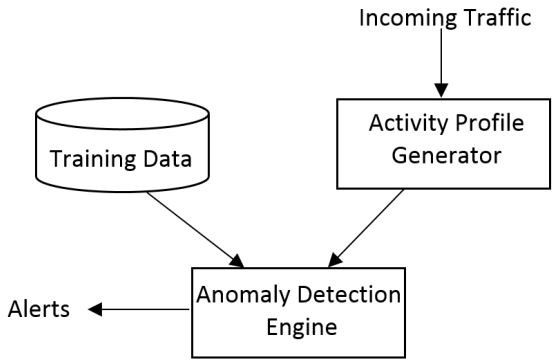

Fig. 2. Outline of ADS

An anomaly based IDS tends to observe network traffic and compare it against an established baseline. This technique parts the normal traffic from the anomalous or abnormal traffic. This division of traffic takes place based on a number of parameters such as the bandwidth, port numbers, protocols, connecting devices and so on. An activity profile is generated for the network based on the information which is then examined. Whenever such network traffic, which is anomalous, or significantly different, than the expected baseline is identified, and alert is generated and sent to the system administrator or the user. Therefore, a novel attack or 0-day attack can be handled efficiently. In the past, many outlier detection methods have been proposed [7], [8], [9], [10], [11], [12] which fall into three major categories.

—Statistical approaches assume that the data follows some standard or predetermined distributions, and finds deviations [7], [9].

—Distance based: The distances between each data point of interest and its neighbours are calculated. Threshold helps to identify outliers [11], [12].

—Density Based: Based on the local density of each data instance, the local outlier factor (LOF) determines the degree of outlierness. Has the ability to estimate local data structure via density estimation not write or print anything outside the print area [8], [10].

2.1 Online Oversampling Principal Component Analysis

PCA is a well-known unsupervised dimension reduction method, which determines the principal direction of the data distribution. It obtains principal directions of data by constructing the data covariance matrix and calculating its dominant eigenvectors. It uses the Leave One Out (LOO) strategy to calculate the score of outlierness of each point by computing difference in the principle direction due to the point. This score of outlierness (st) can be used to determine whether the added data is anomalous or not and is calculated as shown

$$s_t = 1 - \left| \frac{\langle \tilde{\mathbf{u}}_t, \mathbf{u} \rangle}{\|\tilde{\mathbf{u}}_t\| \|\mathbf{u}\|} \right|$$

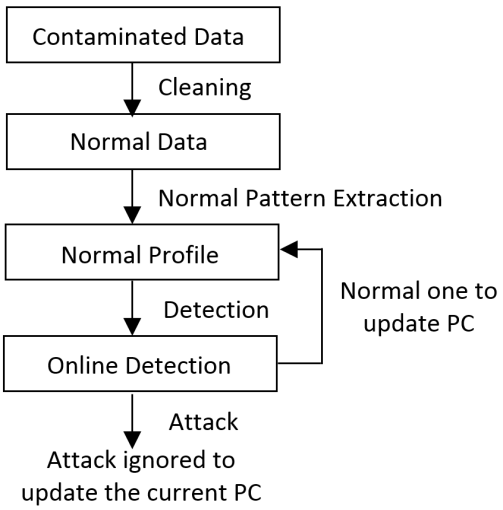

Fig. 3. osPCA Framework

2.1.1 Oversampling PCA. For practical anomaly detection problems, the size of the data set is typically large, and thus one requires osPCA for large scale anomaly detection problems that will duplicate the target instance multiple times, in order to amplify the effect of outlier rather than that of normal data. If the target instance is an outlier, this oversampling scheme allows us to overemphasize its effect. E.g. 10% of the size of data set

2.1.2 Online Updating. The major concern of this technique is the computation cost of calculating or updating the principal directions in large-scale problems. In such a case, the online updating technique can be used, wherein principal direction of the previous calculations is used, and the resultant principal component is approximated. This concept of regression reduces the computation complexity from $O(np^2)$ of the osPCA with power method [13] to O(p) for the online updating system, where n and p are the size and dimensionality of data respectively.

2.1.3 Architecture. In practicality, the training data may contain some noise or contamination which has to be removed, else it will affect the first principal component adversely. For each target instance, st is calculated and if it crosses a threshold, an anomaly is said to be detected, else the principal component is updated via online updating scheme and next instance is tested.

2.2 Changepoint Detection

"Changepoint" is the time instance at which the state of the process changes from "normal" to "abnormal". Changepoint detection is a technique tries to identify changes in the probability distribution of a time series. It can therefore be considered a subtopic in anomaly detection. Its goal is to identify whether or not a change has occurred, or whether several changes might have occurred and thus assessing the time of change. There are mainly two types of changepoint detection methods.

—Online: Sequential analysis ("online") approach is used.

—Offline: Offline algorithms may employ clustering based on maximum likelihood estimation.

2.2.1 The Shiryaev Roberts Procedure. The Shiryaev Roberts Procedure [14], [15], [16] is a Sequential Changepoint detection

method for anomaly detection The likelihood ratio based SR procedure has appealing optimality properties, particularly it is exactly optimal in a multi-cyclic setting geared to sense a change occurring at a far time horizon. This process successfully detects the attack with very slight delays. In this, the mean and standard deviation is calculated for both legitimate and attack traffic and thus outcome of attack leads to a considerable increase in the mean and standard deviation of the connections birth rate [5].

2.3 F-Sign

F-Sign is intended to generate simple signatures that can be used by intrusion detection systems for filtering malware in real-time. F-Sign lessens risk of false positive detection errors by generating signatures that are both exact and sensitive. This continuous up gradation of signature storehouses help in handling Zero-day attacks proficiently.

2.3.1 Common Function Library Construction. Appropriate benign files are given to a function extractor. Each of which is then processed in order to extract all known functions. After the mining is done, function matching module filters known functions leaving only novel functions behind which are then inserted in into the common function library (CFL). Accuracy and up gradation of the CFL is of utmost importance since generating good signatures greatly depends on it. [6].

2.3.2 Signature Generation. The procedure of signature generation initiates with identification of functions using IDA (Interactive Disassembler) or State Machine. The recognized functions are then matched with the CFL to remove all the common functions present and generate candidates for further processing. Once this is done, the final candidate for generation of the unique signature is selected using one of the 2 methods; intelligent candidate selection using entropy score or random selection. Candidate with the highest entropy have large amounts of information, thus they are best suited to generate signatures [6].

3. COMPARISON

Both, the osPCA algorithm and Changepoint Detection using SR Procedure, are based on fundamentally different principles. An attempt has been made to compare them based on various parameters.

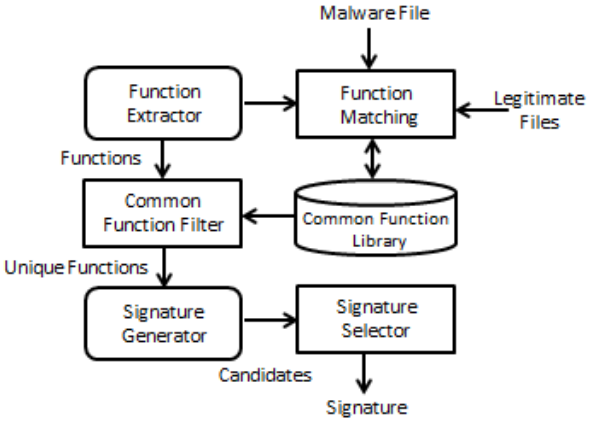

Fig. 4. F-Sign Process

Table 1. Comparison of Anomaly Detection Algorithms

OSPCA	SR Procedure
Principal Component Analysis as means of anomaly detection.	Sequential Changepoint detection method for anomaly detection.
Principal Component of data is calculated with and without the target instance (LOO strategy)	Mean and standard deviation is calculated for both legitimate and attack traffic
Computationally inexpensive as regression is used	Computationally inexpensive
osPCA can be used practically	SR procedure is not practically used
Preferable in online, streaming data, or large scale problems	Preferable in detecting attack with very small delays
Less number of false alarm or low positive rate	Relatively overflowed with false alarms
Support for multidimensional data	No Support for multidimensional data
Computational complexity:O(n)	Computational complexity:$O(n^4)$
Time Complexity: O(n)	Time Complexity: $O(n^2)$
Memory Requirements: O(n)	Memory Requirements: O(n)

4. CONCLUSION

Anomaly Detection based on osPCA with online updating algorithm is suitable when working with large data sets and multidimensional data sets. Also it is suitable for applications where there are constraints on the computation and memory as it is based on approximation, whereas the Changepoint detection algorithm is suitable in scenarios where minimized delay is needed during detection. However, these algorithms have high false positive rates when compared to their signature based counterparts. Thus, these anomaly detection systems in alliance with an automatic signature generator will minimize the false alarm rate, guaranteeing very small detection delays. F-Sign which is an automatic signature generation method can be used to enhance the performance of the hybrid system by generating signatures (from malicious code) that are both exact and sensitive, used for filtering malware in real time. Taking into consideration all the pros and cons of the aforementioned algorithms, the authors infer that anomaly detection when complemented with a signature-based IDS will eliminate its drawbacks, improving the system's overall performance, thereby accomplishing a perfect criteria for Hybrid IDS

5. ACKNOWLEDGMENT

The authors gratefully acknowledge the contributions of Prof. Prasanna Shete and Prof. Suchita Patil and thank them for their endless support and motivation. They also thank the college for providing the necessary infrastructure and platform for doing this research.

6. REFERENCES

[1] K. Hwang, M. Cai, Y. Chen, and M. Qin, "Hybrid intrusion detection with weighted signature generation over anomalous internet episodes," *Dependable and Secure Computing, IEEE Transactions on*, vol. 4, no. 1, pp. 41–55, 2007.

[2] B. Caswell and J. Beale, *Snort 2.1 intrusion detection*. Syngress, 2004.

[3] M. Roesch *et al.*, "Snort: Lightweight intrusion detection for networks.," in *LISA*, vol. 99, pp. 229–238, 1999.

[4] Y.-J. Lee, Y.-R. Yeh, and Y.-C. F. Wang, "Anomaly detection via online oversampling principal component analysis," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 25, no. 7, pp. 1460–1470, 2013.

[5] A. G. Tartakovsky, A. S. Polunchenko, and G. Sokolov, "Efficient computer network anomaly detection by changepoint detection methods," *Selected Topics in Signal Processing, IEEE Journal of*, vol. 7, no. 1, pp. 4–11, 2013.

[6] A. Shabtai, E. Menahem, and Y. Elovici, "F-sign: Automatic, function-based signature generation for malware," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 41, no. 4, pp. 494–508, 2011.

[7] V. Barnett and T. Lewis, *Outliers in statistical data*, vol. 3. Wiley New York, 1994.

[8] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: identifying density-based local outliers," in *ACM sigmod record*, vol. 29, pp. 93–104, ACM, 2000.

[9] D. M. Hawkins, *Identification of outliers*, vol. 11. Springer, 1980.

[10] W. Jin, A. K. Tung, J. Han, and W. Wang, "Ranking outliers using symmetric neighborhood relationship," in *Advances in Knowledge Discovery and Data Mining*, pp. 577–593, Springer, 2006.

[11] N. L. D. Khoa and S. Chawla, "Robust outlier detection using commute time and eigenspace embedding," in *Advances in Knowledge Discovery and Data Mining*, pp. 422–434, Springer, 2010.

[12] E. M. Knox and R. T. Ng, "Algorithms for mining distancebased outliers in large datasets," in *Proceedings of the International Conference on Very Large Data Bases*, pp. 392–403, Citeseer, 1998.

[13] Y.-J. Lee, Y.-R. Yeh, and Y.-C. F. Wang, "Anomaly detection via oversampling principal component analysis," in *New Advances in Intelligent Decision Technologies*, pp. 449–458, Springer, 1998.

[14] S. Roberts, "A comparison of some control chart procedures," *Technometrics*, vol. 8, no. 3, pp. 411–430, 1966.

[15] A. N. Shiryaev, "The problem of the most rapid detection of a disturbance in a stationary process," *Soviet Math. Dokl.*, vol. 2, pp. 795–799, 1961.

[16] A. N. Shiryaev, "On optimum methods in quickest detection problems," *Theory of Probability and its Applications*, vol. 8, no. 1, pp. 22–46, 1964.