ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 11, November 2017

# Anomaly Detection in Policy Authorization Activity Logs

[1] Zahedeh Zamanian, [2] Ali Feizollah, [3] Nor Badrul Anuar, [4] Miss Laiha Binti Mat Kiah
[1][3] Faculty of Computer Science and Information Technology,
University of Malaya, Kuala Lumpur, Malaysia

**Abstract - Security in corporations is a crucial issue. As number of users in these corporation increases, the chance for having intruder also increases. It is important to develop effective methods to deal with such threat. Luckily, users leave an electric footprint behind, as log files. Analyzing these log files results in examining users' activity and detecting an intruder. Recent works have proposed methods for detecting intruders inside corporations. However, these methods are complex for today's corporation. In this work, we proposed a lightweight and effective method to detect an intruder inside corporations using log files. The dataset in this work was provided from NextLabs, one of the high-profile companies in information security. The experiment using random forest algorithm shows that this method detects intruders with 97.18% accuracy.**

**Keywords: Intrusion detection system, anomaly detection, log file, inside intruder.**

## 1. INTRODUCTION

Information is the crown jewels of every business but integrity, availability and confidentiality of these information are predominant concerns of Companies and organizations. Information and Communications Technologies have moved forward in leaps and bounds in the last couple of years. This has created new opportunities for Corporations to running and expand their business across the globe. Whereas this event has resulted in increasing accessibility to the Internet and reduced costs for corporations, it has also resulted in vulnerability of organization to both insiders and outsiders threats [1]. Therefore, data protection and keep network secure becomes vitally important. As defined by International ISO/IEC 17799:2000, Confidentiality means ensuring that information is accessible only to those authorized to have access. Therefore, unauthorized access can be grouped in two classes:

- External penetrator: an agent from outside the organization who are not authorized to have access.
- Internal Penetrator: an agent that belong to the organization but surpasses his or her legitimate access rights [2].

In the CSI Computer Crime & Security Survey 2010, 13% and 11% of attacks were Unauthorized access or privilege escalation by insider and System penetration by outsider , respectively [3] . Therefore it is crucial for companies to realize threats which influence their assets and the areas which each threat could affect [4].

Analyzing log files is one form of defense mechanism against these kind of attacks [5]. Activity log or log file is a collection of event records which is occurring within a company's systems and networks. Logs are consisted of log entries; each entry contains information related to an event including the use of specific system resources, system status changes, and general performance issues. The significant role of log files can be recognized by wide usage of logs in different area including anomaly detection [6, 7], troubleshooting errors and debugging [8-10], performance issues [11] , system behavior understanding [12] , workload modeling [13] ,etc.

The typical content for a log file are

- Timestamp: The occurrence time associated with the event
- Source: System that generated the log file represented in IP address or hostname format
- Data: No standard format, it could represent source and destination IP address, source and destination ports, user names, program names, resource objects like file, directory, byte transferred in or out

Log files come from many different sources for instance Unix and windows System , Switches, Firewalls ,Routers, Wireless Access Points ,Virtual Private Network (VPN) Server, AntiVirus ( AV) Systems and Printers [14].A medium to large company tends to generate and collect sheer size of activity logs which typically contains hundreds and thousands of lines. Therefore, analyze and

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 11, November 2017

classify such huge sets of data manually, for anomaly detection or reporting purposes, is tedious and nearly impossible [15]. Therefore, there is a need for automated analysis tools that detect peculiar and malicious behavior that is unlikely to be spotted by a human. Chandola et al. state that "*Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These nonconforming patterns are often referred to as anomalies or outliers*." The anomaly detection provides very important and crucial information from a computer security perspective. It can detect malicious activity such as unauthorized use, penetrations, and other forms of computer abuse [16] . When data needs to be analyzed in order to find pattern or to predict known or unknown, data mining techniques are applied. These could be categorized to clustering, classification and machine based learning techniques. In addition, hybrid methods are also being used to get higher level of accuracy on detecting anomalies [17]. Depending on whether the data labels are provided for learning, these techniques can be divided as supervised, semi-supervised or unsupervised. It should be known that obtaining accurate labeled data that representative of all types of behaviors, is very expensive. This task is time consuming and done manually by human expert [18]. Moreover, due to privacy and ethical concerns companies are not interested to share their dataset especially the data that may contain insider threats [19]. As a result ,acquire and research with real-world data is challenging . Gheyas et al mentioned data source which was used for insider threat by researcher can be categorized as below[20] :

- Real-world system log data [21]
- Real data injected with synthetic anomalies [22]
- Game-theoretic approach (GTA) [23]
- Social media data Simulated data drawn from stochastic models[24]
- Simulated data drawn from stochastic models which are developed from real data

In academic literature, mostly behavior -based modeling has been presented to detect insider threat. This model can be grouped to system behaviors and user behaviors. The system behaviors are generated by hosts and networks and relate to the host activities and network status. In contrast, the user behaviors can be defined as direct interaction between the user and the system such as typing patterns. It should be known that there is relation between these two types and can affect each other [25].

## 2. RELATED WORK

Anomaly detection has been an important research problem in security analysis, therefore development of methods that can detect malicious insider behavior with high accuracy and low false alarm is vital [26]. In this problem layout, McGough *et al* [27] designed a system to identify anomalous behavior of user by comparing of individual user's activities against their own routine profile, as well as against the organization's rule. They applied two independent approaches of machine learning and Statistical Analyzer on data. Then results from these two parts combined together to form consensus which then mapped to a risk score. Their system showed high accuracy, low false positive and minimum effect on the existing computing and network resources in terms of memory and CPU usage.

Bhattacharjee *et al* proposed a graph-based method that can investigate user behavior from two perspectives :(a) anomaly with reference to the normal activities of individual user which has been observed in a prolonged period of time and (b) finding the relationship between user and his colleagues with similar roles/profiles .They utilized CMU-CERT dataset in unsupervised manner .In their model , Boykov Kolmogorov algorithm was used and the result compared with different algorithms including Single Model One-Class SVM, Individual Profile Analysis, k-User Clustering and Maximum Clique (MC).Their proposed model evaluated by evaluation metrics Area-Under-Curve (AUC) that showed impressive improvement compare to other algorithms [28]. Log data are considered as high-dimensional data which contain irrelevant and redundant features. Feature selection methods can be applied to reduce dimensionality ,decrease training time and enhance learning performance[29] .

In [30] Legg *et al* offered an automated system that construct tree structured profiles based on individual user activity and combined role activity. This method helped them to attain consistent features which provide description of the user's behavior. They reduced high dimensionality of this feature set by using principal component analysis(PCA) and compute anomaly scores

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 11, November 2017

based on Mahalanobis distance anomaly metrics. Their system was tested on synthetic dataset which ten malicious data injected. Their system performed well for identifying these attacks .

In a similar line, Agrafiotis *et al* [31]applied same model as offered by Legg *et al* but they used real -world data set from multinational organization .Moreover ,their approach abided the ethical and privacy concerns. Their result showed high accuracy and low false alarm. Although finding a sequence is a common choice for modeling activities and events through time but catching nomalous sequence in a dataset is not an easy task. One of the algorithm that has ability to recognize temporal pattern and widely has been used is Hidden Markov Models (HMM).

Rashid *et al* [32] proposed a model based on HMM to identify insider threat in CERT dataset. They tried to model user's normal behavior as a week-long sequence. Their modeled showed accurate result with low false alarm. Although author mentioned using shorter time frame for instance a day long sequences could build a more accurate model of employee's daily behavior. Moreover, their system was trained based on first 5 weeks so not able to detect insider threats amongst short-term users such as contractors whose are a real threat.

## 3. EXPERIMENTAL DETAILS

The experiment conducted in this paper aims at detecting anomalies in users' log files. The dataset was acquired from NextLabs Corporation, which includes 1,000 records of logs. It is a collection of log data (time, date, id), user data (email, host id, host IP), resource data (file name, file id), policy name, and policy decision. The records are labelled as normal or anomaly. This dataset is used as training and testing data in machine learning algorithm.

This work uses random forest algorithm to train, test, and generate a model for anomaly detection in log files. Random forest is an ensemble learning algorithm. The basic premise of the algorithm is that building a small decision-tree with few features is a computationally cheap process. If we can build many small, weak decision trees in parallel, we can then combine the trees to form a single, strong learner by averaging or taking the majority vote. In practice, random forests are often found to be the most accurate learning algorithms to date.

The random forest algorithm uses the bagging technique for building an ensemble of decision trees. Bagging is known to reduce the variance of the algorithm. However, the natural question to ask is why does the ensemble work better when we choose features from random subsets rather than learn the tree using the traditional algorithm? Recall, that ensembles are more effective when the individual models that comprise them are uncorrelated. In traditional bagging with decision trees, the constituent decision trees may end up to be very correlated because the same features will tend to be used repeatedly to split the bootstrap samples. By restricting each split-test to a small, random sample of features, we can decrease the correlation between trees in the ensemble. Furthermore, by restricting the features that we consider at each node, we can learn each tree much faster, and therefore, can learn more decision trees in a given amount of time. Thus, not only can we build many more trees using the randomized tree learning algorithm, but these trees will also be less correlated. For these reasons, random forests tend to have excellent performance.

## 4. RESULTS AND DISCUSSION

This section presents results of the experiment. The data were divided into 70% training and 30% test data. The training data are used to train the algorithm. Then, the learned model is tested using test data. The test data is fed to the model as new data to measure how the algorithm is trained. The results are measured in terms of accuracy, which is number of correctly classified data over all of data. Table 1 shows results of the experiment.

***Table 1. Experiment Result***

| Accuracy | Error |
|---|---|
| 97.18% | 2.82% |

The random forest algorithm achieved 97.18% accuracy, and 2.82% of error. The error is wrongly classified data as normal or anomaly. Among various features in the dataset, policy decision is more important than others. It shows whether a user is allowed access to a resource. As illustrated in Figure 1, the random forest recognizes such importance.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Vol 4, Issue 11, November 2017

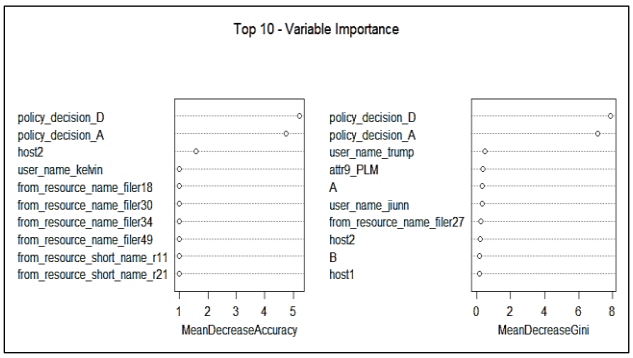
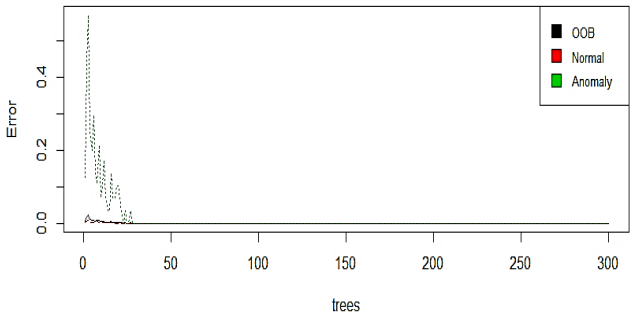



***Figure 1. Variable Importance in Random Forest***

This figure shows importance of features using two methods. In both methods (mean decrease accuracy and mean decrease gini), policy decision has higher rank compare to other features.It is also possible to see




progress of the algorithm as it trains. Figure 2 shows progress of random forest in

terms of number of trees and associated error.

***Figure 2. Random Forest Progress Graph***

Figure 2 shows that the algorithm starts with high rate of error and gradually the error decreases as it learns the data. Eventually, it reaches to almost zero error rate.

## CONCLUSION

Security in large companies has become a crucial issue. This work proposed a lightweight and effective method to detect inside intruders for corporations. We used random forest algorithm for detection purpose. The dataset was provided from NextLabs Corporation. The result shows that the algorithm achieved 97.18% of accuracy.

## ACKNOWLEDEGEMENT

The work described in this paper was supported by the Collaborative Agreement with NextLabs (Malaysia) Sdn Bhd (Project title: Anomaly detection in Policy Authorization Activity Logs).

## REFERENCES

[1] R. Prasad, “Insider Threat to Organizations in the Digital Era and Combat Strategies,” in Indo-US conference and workshop on "Cyber Security, Cyber Crime and Cyber Forensics, Kochi, India, 2009.

[2] P. A Diaz-Gomez, G. Vallecarcamo, and D. Jones, "Internal Vs. External Penetrations: A Computer Security Dilemma," The cybersecurity dilemma: hacking, trust and fear between nations, 2017.

[3] Robert Richardson, “CSIComputer Crime and Security Survey,” Computer Security Institute, 2010/2011.

[4] S. Bauer, and E. W. N. Bernroider, “From Information Security Awareness to Reasoned Compliant Action: Analyzing Information Security Policy Compliance in a Large Banking Organization,” SIGMIS Database, vol. 48, no. 3, pp. 44-68, 2017.

[5] R. Vaarandi, M. Kont, and M. Pihelgas, "Event log analysis with the LogCluster tool." pp. 982-987.

[6] J. D. Parmar, and J. T. Patel, “Anomaly Detection in Data Mining: A Review,” International Journal, vol. 7, no. 4, 2017.

[7] J. Breier, and J. Branišová, “A Dynamic Rule Creation Based Anomaly Detection Method for Identifying Security Breaches in Log Records,” Wireless Personal Communications, vol. 94, no. 3, pp. 497-511, June 01, 2017.

[8] K. Kinshumann, K. Glerum, S. Greenberg, G. Aul, V. Orgovan, G. Nichols, D. Grant, G. Loihle, and G. Hunt, “Debugging in the (very) large: ten years of implementation and experience,” Commun. ACM, vol. 54, no. 7, pp. 111-116, 2011.

[9] S. Kobayashi, K. Fukuda, and H. Esaki, "Mining causes of network events in log data with causal inference." pp. 45-53.

[10] Q. Lin, H. Zhang, J.-G. Lou, Y. Zhang, and X. Chen, "Log clustering based problem identification for online service systems," in Proceedings of the 38th International Conference on Software Engineering Companion, Austin, Texas, 2016, pp. 102-111.

[11] K. Nagaraj, C. Killian, and J. Neville, "Structured comparative analysis of systems logs to diagnose performance problems." pp. 26-26.

[12] H. Li, W. Shang, Y. Zou, and A. E. Hassan, "Towards just-in-time suggestions for log changes," Empirical Software Engineering, vol. 22, no. 4, pp. 1831-1865, August 01, 2017.

[13] F. Abbors, D. Truscan, and T. Ahmad, "Mining Web Server Logs for Creating Workload Models," Software Technologies: 9th International Joint Conference, ICSOFT 2014, Vienna, Austria, August 29-31, 2014, Revised Selected Papers, A. Holzinger, J. Cardoso, J. Cordeiro, T. Libourel, L. A. Maciaszek and M. van Sinderen, eds., pp. 131-150, Cham: Springer International Publishing, 2015.

[14] A. Chuvakin, K. Schmidt, and C. Phillips, "Chapter 2 - What is a Log?," Logging and Log Management ,The Authoritative Guide to Understanding the Concepts Surrounding Logging and Log Management, pp. 29-49, Boston: Syngress, 2013.

[15] J. Breier, and J. Branišová, "Anomaly Detection from Log Files Using Data Mining Techniques," Information Science and Applications, K. J. Kim, ed., pp. 449-457, Berlin, Heidelberg: Springer Berlin Heidelberg, 2015.

[16] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," ACM Comput. Surv., vol. 41, no. 3, pp. 1-58, 2009.

[17] S. Agrawal, and J. Agrawal, "Survey on anomaly detection using data mining techniques," Procedia Computer Science, vol. 60, pp. 708-713, 2015.

[18] P. Gogoi, B. Borah, and D. K. Bhattacharyya, Anomaly Detection Analysis of Intrusion Data Using Supervised & Unsupervised Approach, 2010.

[19] F. L. Greitzer, D. A. Frincke, and M. Zabriskie, "Social/ethical issues in predictive insider threat monitoring," Information Assurance and Security Ethics in Complex Systems: Interdisciplinary Perspectives, pp. 132-161, 2010.

[20] I. A. Gheyas, and A. E. Abdallah, "Detection and prediction of insider threats to cyber security: a systematic literature review and meta-analysis," Big Data Analytics, vol. 1, no. 1, pp. 6, August 30, 2016.

[21] A. Ambre, and N. Shekokar, "Insider threat detection using log analysis and event correlation," Proc Comp Sci, vol. 45, 2015.

[22] H. Eldardiry, E. Bart, J. Liu, J. Hanley, B. Price, and O. Brdiczka, "Multi-domain information fusion for insider threat detection," 2013 IEEE Security and Privacy Workshops, San Francisco: IEEE, 2013.

[23] D. Liu, X. Wang, and J. Camp, "Game-theoretic modeling and analysis of insider threats," International Journal of Critical Infrastructure Protection, vol. 1, pp. 75-80, 2008.

[24] M. Kandias, V. Stavrou, N. Bozovic, L. Mitrou, and D. Gritzalis, "Can we trust this user? Predicting insider's attitude via YouTube usage profiling." pp. 347-354.

[25] J. Peng, K.-K. R. Choo, and H. Ashman, "User profiling in intrusion detection: A review," Journal of Network and Computer Applications, vol. 72, pp. 14-27, 2016.

[26] K. W. Kongsg, #229, rd, N. A. Nordbotten, F. Mancini, and P. E. Engelstad, "An Internal/Insider Threat Score for Data Loss Prevention and Detection," in Proceedings of the 3rd ACM on International Workshop on Security And Privacy Analytics, Scottsdale, Arizona, USA, 2017, pp. 11-16.

[27] A. S. McGough, D. Wall, J. Brennan, G. Theodoropoulos, E. Ruck-Keene, B. Arief, C. Gamble, J. Fitzgerald, A. v. Moorsel, and S. Alwis, "Insider Threats: Identifying Anomalous Human Behaviour in Heterogeneous Systems Using Beneficial Intelligent Software (Ben-ware)," in Proceedings of the 7th ACM CCS International Workshop on Managing Insider Security Threats, Denver, Colorado, USA, 2015, pp. 1-12.

[28] S. D. Bhattacharjee, J. Yuan, Z. Jiaqi, and Y.-P. Tan, "Context-aware graph-based analysis for detecting

anomalous activities," in Multimedia and Expo (ICME), 2017 IEEE International Conference on, 2017, pp. 1021-1026.

[29] R. Sheikhpour, M. A. Sarram, S. Gharaghani, and M. A. Z. Chahooki, "A Survey on semi-supervised feature selection methods," Pattern Recognition, vol. 64, no. Supplement C, pp. 141-158, 2017/04/01/, 2017.

[30] P. A. Legg, O. Buckley, M. Goldsmith, and S. Creese, "Automated insider threat detection system using user and role-based profile assessment," IEEE Systems Journal, vol. 11, no. 2, pp. 503-512, 2015.

[31] I. Agrafiotis, A. Erola, J. Happa, M. Goldsmith, and S. Creese, "Validating an Insider Threat Detection System: A Real Scenario Perspective," in 2016 IEEE Security and Privacy Workshops (SPW), 2016, pp. 286-295.

[32] T. Rashid, I. Agrafiotis, and J. R. C. Nurse, "A New Take on Detecting Insider Threats: Exploring the Use of Hidden Markov Models," in Proceedings of the 8th ACM CCS International Workshop on Managing Insider Security Threats, Vienna, Austria, 2016, pp. 47-56.