

Development Trend of Insider Anomaly Detection System

MinKyu Kim *, KiHwan Kim *, HoonJae Lee**

*Dept. Ubiquitous It, Dongseo University, Busan, Rep. of KOREA

**Div. Of Information and Communication Engineering, Busan, Rep. of KOREA

alsrb1030@naver.com, ghksdl90@naver.com, hjlee@dongseo.ac.kr

Abstract— Recently, industrial and national infrastructure suffered economic losses due to internal leaks caused by insider leaks and key data leaks. As a result, many companies applying not only physical external and internal penetration methods, but also software, machine learning, and other methods to detect people's abnormal behaviour. This paper surveys trends and forecasts of the intrusion detection techniques by categorizing into basic software and machine learning technique.

Keywords— Insider Threat Detection, Anomaly Detection, Machine Learning, Software, Algorithm

I. INTRODUCTION

In recent years, information leaked from an enterprise or country has caused massive economic losses [1]. As a result, major facilities began to separate the external network from the internal network in order to prevent internal data from being leaked to the outside [2]. It was believed that most people would be safe from external threats [3]. However, comparing the cases attacked by outsiders with the cases attacked by insiders, it can be seen that the cases attacked by insiders cause more losses [4]. Malicious behaviour caused by an insider can be seen as a user's carelessness or corporate spy.

In this paper, we refer to industrial spies as "malicious users" and write users who have caused accidental accidents as "careless users". Malicious users can easily create new companies by stealing information or sell data that has been leaked to competitors. In the case of careless users, damage is caused by mistakenly downloading spam or accidentally connecting to the internal network without knowing that malicious code exists on USB. In order to prevent damage from information leakage, corporate and national facilities are constructing their own anomaly detection systems or using solutions sold by security companies to prevent attacks by insiders [5].

II. ANOMALY DETECTION TECHNIQUES AND RELATED STUDIES

A. Rule-based anomaly detection

An anomaly detection system is a system that enables network administrators to regularize the behaviour of network users and to detect violations of rules and data leakage.

The disadvantage of the rule-based detection system is that it cannot actively detect and defend against attacks that fall outside of the restricted rules [7].

Malicious users continue to exploit the disadvantages of rule-based detection systems and add new changes when the system environment is updated to prevent legitimate rules from being properly detected [8].

Recently, a malicious user extracts the characteristics of the target data and finds the characteristics of the malicious user by using the text mining technique without specifying the type in advance. However, countermeasures were needed for criminals attempting to attack a little differently from the defined rules [9].

B. Machine Learning

Machine learning means that a computer learns data from a variety of cases. Machine learning has a disadvantage that learning takes a lot of time. But it has the advantage of being able to actively respond to new types. Table 1 lists the types of anomaly detection systems that utilize machine learning [10, 11].

TABLE 1. TYPES OF ANOMALY DETECTION SYSTEMS USING MACHINE LEARNING [10,11]

Type	Positive	negative
Neural Networks	You can generalize unsafe data and discover new and unknown threats.	The training process takes a long time and is not suitable for real-time detection.
Decision Tree	High detection rate even in large data.	Building a decision tree is computationally intensive.
Fuzzy Logic	Reasoning is Approximate rather than precise. Effective, especially against port scans and probes.	High resource consumption Involved.
Genetic Algorithm	Biologically inspired and employs an evolutionary algorithm, it is possible to derive optimal classification	Optimization response time cannot be guaranteed.

	rules and select good parameters.	
Support Vector Machine	High training rate and decision rate, insensitiveness to dimension of input data.	The training takes a long time.
Bayesian Network	Encodes probabilistic relationships among the variables of interest	If prior knowledge is wrong, good data cannot be obtained.

In recent years, there have been various devices besides smart phones and computers, and the amount of data that insiders make and use has increased exponentially and the types of data have diversified.

Rule-based detection methods have become more difficult for people to manage all their data as the threats become more complex and diverse.

To cope with this, most security companies prefer to detect using machine learning rather than rule-based anomaly detection.

Therefore, there has been an increase in the demand for active detection and defence according to the new type.

Machine learning tends to collect more data and extract automatically without renting to quickly respond to insider threats and continue to try out tutoring.

Table 2 compares the advantages and disadvantages of rule-based learning and machine learning methods.

TABLE 2. COMPARISON OF THE ADVANTAGES AND DISADVANTAGES OF RULE-BASED AND MACHINE LEARNING METHODS

Method	Advantages	Disadvantages
Rule-based learning	Rules are easy to configure.	It does not actively respond to attacks outside the rules. A lot of expert work is needed to enter the rules.
Machine learning	It is possible to learn machine itself without expert. During learning, you can also find internal attacks of new methods.	If the machine learns the wrong training data, the abnormal behavior detection does not work properly.

III. RESEARCH TREND OF INSIDER IDEAL BEHAVIOR

Recently, CERT (Computer Emergency Response Team) Dataset provided by Carnegie Mellon University was used as the experimental data used in the experiments of various papers used for detection of internal threat attacks.

The version of Cert Dataset is available from R1 to R 6.2, and as the version goes up, it is the latest version with the old data. Experimental data were collected over 18 months, with R4.2 collected for 1000 people and R6.2 collected for 4000 people. Table 3 is a scenario for a CERT data set.

TABLE 3. CARNEGIE MELLON UNIVERSITY CERT SCENARIO

Malicious Scenario
User who did not previously use removable drives or work after hours begins logging in after hours, using a removable drive, and uploading data to wikileaks.org. Leaves the organization shortly thereafter.
User begins surfing job websites and soliciting employment from a competitor. Before leaving the company, they use a thumb drive (at markedly higher rates than their previous activity) to steal data.
System administrator becomes disgruntled. Downloads a keylogger and uses a thumb drive to transfer it to his supervisor's machine. The next day, he uses the collected keylogs to log in as his supervisor and send out an alarming mass email, causing panic in the organization. He leaves the organization immediately.
A user logs into another user's machine and searches for interesting files, emailing to their home email. This behavior occurs more and more frequently over a 3 month period.
A member of a group decimated by layoffs uploads documents to Dropbox, planning to use them for personal gain.

The internal file of Insider Threat Dataset stores logs and useful information. Table 4 is the directory or file used in the CERT Dataset.

TABLE 4. CERT DIRECTORIES OR FILES AT CARNEGIE MELLON UNIVERSITY

Name	Files Inside Datasets	Function
LDAP	employee_name, user_id, email, role	There is a file that lists the employees of the month.
device.csv	id, date, user, pc, activity	Record the file access behavior for the device.
email.csv & email-supplemental.csv	id, date, to, from. Note that the email addresses	Records e-mail transactions between employees.
http.csv	id, date, user, pc, url	URLs visited by each employee are recorded.
logon.csv	id, date, user, pc, activity	Record the logon and logoff activity of each employee.

There is an example of an outlier detection algorithm that predicts the degree of threat of an internal member's behavior by using the above-mentioned CERT Data. This algorithm generates 60 variables by generating the variables used in the internal intrusion detection, which are used in CERT Dataset, and by dividing the time unit into day by working time and nonworking time. In the abnormal category, about 90% of the observations were concentrated, and the internal intrusion detection experiment was performed with only three tables. The algorithm used was Gaussian density estimation, Parzen window estimation, PCA (principal component analysis), and KMC (K-Means clustering). The characteristics of each algorithm are shown in the Table 5 [12].

TABLE 5. TYPES OF OUTLIER DETECTION ALGORITHMS USING CERT DATA

Name	Characteristic
Gauss	A method of calculating an ideal score using the probability that new data will be generated at an estimated probability density after assuming a given data as a normal distribution and estimating averaging variance to obtain a probability density function.
Parzen	A method for estimating density using kernel function as one of estimation methods of non - parametric density.
PCA	A technique for finding new key components that maximizes the variance of a given data.
KMC	A clustering algorithm that assigns each observation to a cluster with the closest centroid to itself.

As a result of experiments in the above environment, the results of the algorithm of Table 5, K = 3, 5, 7 value of KMK algorithm and the ensemble method of Parzen and PCA were derived.

Experimental results show that each algorithm has a good detection performance for each table, and that the ensemble of Parzen and PCA shows generally excellent performance [13].

Another example of anomaly detection technology is Autoencode, which is implemented using CERT Dataset and Recurrent Neural Network (RNN).

Recurrent Neural Network (RNN) is different from other neural networks because it considers the present value and the previous input value together. Therefore, it is a model that is used widely for the problem of thinking about the input order like the natural language processing problem.

Autoencode is a machine learning algorithm that allows self-learning. In the experiment, we used Autoencode data after learning normal behavior and detected abnormal behavior by comparing pattern of normal pattern with pattern of day. The results showed more than 95% sensitivity, and abnormal behavior could be detected by comparing normal data with daily data [13].

In K. Veeramachaneni's "AI2: Protection of Large Data Systems Education" studies, analysts combined intelligent computer technology with the latest technology to detect new attacks and reduce their time lapse. For attack detection and successful prevention, five machine learning-based modelling techniques and features are shown in Table 6.

TABLE 6. MACHINE LEARNING-BASED MODELLING TECHNIQUES AND FEATURES

Name	Feature
Matrix Decomposition-based outlier analysis	Outlier detection methods based on matrix decomposition use Principal Component Analysis to find cases that violate the correlation structure of the main bulk of the data (Shyu et al. [2003]).
RNN (Replicator Neural Networks)	Network to compress and reconstruct the data in such away that the bulk of the data is reconstructed accurately, but outliers are not. This way,

	the reconstruction error can be directly translated into an outlier score.
Density-based outlier analysis	Test instances which fall within a low-density region of the distribution are considered outliers. The outlier score is simply the probability density of a point in the multidimensional space.
Outlier score interpretation	Combining scores from different methods is not straightforward. it is also difficult to exploit the robustness of multi-algorithm outlier detection ensembles
Transforming outlier scores into probabilities	matrix decomposition-based outlier scores with a Weibull distribution, which is flexible, and can model a wide variety of shapes
Outlier detection ensembles	Multi-algorithm ensembles are combinations of predictions from different machine learning models

An ensemble of method of detection of specificity was tested with actual datasets consisting of 3.6 billion log lines. The results of this thesis show that the system protects the system from invisible attacks. As feedback was collected over time, the detection speed and the detection speed increased. [14].

IV. CONCLUSIONS

The paper examines recent research trends of insider threat detection systems so as not to suffer much from insider threats. Based on rule-based abnormal behavior detection systems used in traditional internal threat detection techniques, many studies have been conducted on internal threat detection systems through recent mechanical studies. Machine learning based detection systems can detect intruder rules that can be found in complex rules and have not yet been discovered. You can also reduce workload by learning with artificial intelligence. Based on this research report, we will develop a system to detect abnormal behavior by insiders with up-to-date information.

ACKNOWLEDGMENT

This work was supported by the Technological Innovation R&D Program (Assignment number: S2457495) funded by the Small and Medium Business Administration (SMBA, Korea) and it was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (grant number: NRF-2016R1D1A1B01011908).

REFERENCES

- [1] S.M Park, "Fourth Industrial Revolution and Technology Protection of Small and Medium-Sized Enterprises", The Korean Operations Research and Management Science Society, pp. 1592-1597, Aug.2017.
- [2] "Guidelines for the separation of national Network", NIS, KOREA.
- [3] B.H.Sim, H.K.Kang, Y.S.Lee, J.H.Lee "A Study on Measures to Eliminate Internal Network Cyber Vulnerabilities", Defense and Technology, vol. 462, pp. 122-129, Aug.2017.

- [4] P.A.Legg, "Visualizing the insider threat: challenges and tools for identifying malicious user activity", Visualization for Cyber Security IEEE Symposium, pp. 1-7. Oct.2015.
- [5] "INSIDER THREAT", National Geospatial-Intelligence Agency, USA.
- [6] M.Theis, R.Trzeciak, J.Strozer, J.Clark, D.Costa, T.Cassidy, M.Albrethsen, and A.Moore, "Common Sense Guide to Mitigating Insider Threats, Fifth Edition", The CERT Insider Threat Center, TECHNICAL NOTE CMU/SEI-2015-TR-010, Dec.2016.
- [7] M.Theoharidou, S.Kokolakis, M.Karyda, and E.Kiountouzis, "The insider threat to information systems and the effectiveness of ISO17799", Computers & Security, vol.24(6), pp. 472-484, May.2005.
- [8] T.F.Lunt, R.Jagannathan, R.Lee, A.Whitehurst, S.Listgarten Knowledge-based intrusion detection. In AI Systems in Government Conference, Proceedings of the Annual, pp. 102-107, Mar.1989.
- [9] N. Liang, and D. Biros, "Validating Common Characteristics of Malicious Insiders: Proof of Concept Study", System Sciences (HICSS), 2016 49th Hawaii International Conference, pp. 3716- 3726, Mar.2016
- [10] H.Kaur , G.Singh, J.Minhas "A Review of Machine Learning based Anomaly Detection Techniques", International Journal of Computer Applications Technology and Research, vol.2- Issue 2, pp.185 - 187, Jul.2013.
- [11] J.Singh, M.J. Nene, "A Survey on Machine Learning Techniques for Intrusion Detection Systems", International Journal of Advanced Research in Computer and Communication Engineering Vol. 2.Issue 11, pp.40-46, Nov.2013.
- [12] H.D.Kim, J.H.Kim, M.S.Park, S.H.Cho, P.s.Kang, " Insider Threat Detection based on User behavior Model and Novelty Detection Algorithms", Journal of the Korean Institute of Industrial Engineers , vol 43(4), pp. 276-287, Aug. 2017.
- [13] D.W.Ha, K.T.Kang, Y.S.Ryu, "Detecting Insider Threat Based on Machine Learning : Anomaly Detection Using RNN Autoencoder", Journal of the Korea Institute of Information Security & Cryptology vol. 27(4), , 763-773, Aug. 2017.
- [14] K.Veeramachaneni, I.Arnaldo, A.Cuesta-Infante, V.Korrapati, C.Bassias, and K.Li, "AI2 : Training a big data machine to defend", 2016 IEEE 2nd International Conference on Big Data Security on Cloud (Big Data Security), Jul.2016.

Computer Society, IEEE Information Theory Society and etc.



Minkyu Kim

He received the B.S. degree in Computer Networking from Dongseo University, Republic of Korea in 2017. His current research interests are in Wireless Sensor Networks, Cryptography and Network Security, Side Channel Analysis



KiHwan Kim

He received the B.S. and M.S. degree in Computer Networking from Dongseo University, Republic of Korea in 2015. His current research interests are in Cryptography, Information Security and Network Security.



HoonJae Lee

He received the B.S., M.S. and Ph.D. degree in Electrical Engineering from Kyungpook national university in 1985, 1987 and 1998, respectively. He had been engaged in the research on cryptography and network security at Agency for Defense Development from 1987 to 1998. Since 2002 he has been working for Department of Computer Engineering of Dongseo University as an associate professor, and now he is a full professor. His current research interests are in security communication system, side-channel attack, USN & RFID security. He is a member of the Korea institute of Information security and cryptology, IEEE