Anomaly-Based – Intrusion Detection System using User Profile Generated from System Logs

Roshan Pokhrel*, Prabhat Pokharel, Arun Kumar Timalsina, PhD***

**Institute of Engineering, Pulchowk Campus, Tribhuvan University*
***Nepal College of Information Technology, Pokhara University*

DOI: 10.29322/IJSRP.9.02.2019.p8631
http://dx.doi.org/10.29322/IJSRP.9.02.2019.p8631

***Abstract*— Intrusion Detection System (IDS) is a form of defense that aims to detect suspicious activities and attack against information systems in general. With new types of attacks appearing continuously, developing adaptive and flexible security oriented approaches is a severe challenge. In this scenario, this thesis presents an anomaly-based intrusion detection technique as a valuable technology to protect the target system against malicious activities. This technique uses a semi-supervised learning model to identify and learn from past events as manifested in system logs and build a user behavior profile. The observed behavior of the user is analyzed to infer whether or not the normal profile supports the observed one. This is carried out using two-class classifiers. A new hybrid approach using Support Vector Machine (SVM) and Naïve Bayes (NB) is proposed to provide better accuracy and to reduce the problem of high false positive. The comparison of the proposed approach is made with other SVM and NB techniques. Hybrid approach is found to outperform SVM and NB. For the validation of the result, cross-validation is employed, and the result is presented using Receiver Operating Characteristics (ROC) curve. The experimentation is implemented in datasets from two different organizations.**

Keywords*— Intrusion Detection System, user behavior profile, hybrid approach, Support Vector Machine, Naïve Bayes

INTRODUCTION

Intrusion detection is a process of monitoring the events from the computer-based system and investigating them for possible signs of incidents which are violationss of security policies, guideliness, or standard security practices [1]. The most popular approaches for intrusion detection systems can be classified to signature based and anomaly based [2]. In this paper, anomaly-based intrusion detection has been considered. The anomaly-based intrusion detection system requires training datasets to learn general attacks and careful selection of the threshold level of detection [3]. For correct detection of anomalies, normal and abnormal network behavior profile needs to be defined. Once the normal and abnormal profile has been defined, the model can be tested on test dataset to check its accuracy.

RELATED WORK

M. Corney et al. research titled "Detection of Anomalies from User Profiles Generated from System Logs" to identify anomalous events and event patterns manifested in computer system logs. Prototype software was developed with a capability that identifies anomalous events based on usage patterns or user profile, and alerts administrators when such events are identified. More specifically the research attempted to detect unauthorized use of software applications by users from within an organization [4].

N. A. Durgin and P. Zhang researched on "Profile-Based Adaptive Anomaly Detection for Network Security." The research focused on enhancing current IDS capabilities by addressing some of these shortcomings. They identified and evaluated promising techniques for data mining and machine learning. The algorithms were "trained" by providing them with a series of data-points from "normal" network traffic. They also built a prototype anomaly detection tool that demonstrates how the techniques might be integrated into an operational intrusion detection framework [3].

J. P. Anderson introduced a term audit trail in a report titled “Computer security threat monitoring and surveillance” which includes information for tracking down the misuse and user behavior [5]. This paper introduced a misuse detection technique. This paper provides a base for IDS design and development.

M. Zhang et al. research on "An Anomaly Detection Model Based on One-Class SVM to Detect Network Intrusions" proposes an anomaly detection model based on One-class SVM to detect network intrusions. The one-class SVM adopts only normal network connection records as the training dataset. However, after being trained, it can recognize normal from various attacks [6].

N. B. Amor et al. research paper titled "Naïve Bayes v.s decision trees in intrusion detection systems" performed a comparison between two classifiers Naïve Bayes networks and decision tree using KDD Cup dataset 1999 [7]. Naïve Bayes and decision tree having their own decision capable of detecting the intrusion. Both performed equal, however, while detecting U2R and probe Naïve Bayes performed better and in normal, DOS and R2L decision tree performed better.

R. Jain et al. carried out a survey on "Network attacks, classification and models for anomaly-basedd network intrusion detection system.". This paper presents a selective survey of incremental approaches for detecting a annomaly in a normal system and network traffic [8].

S. S. Murtaza et al. research on "A host-based anomaly detection approach by representing system calls as states of kernel modules" attempts to reduce the false alarm rate and processing time while increasing the detection rate. The

paper presents a novel anomaly detection technique based on semantic interactions of system calls which analyzes the state interaction, and identifies anomalies by comparing the probabilities of occurrences of states in normal and anomalous traces [9].

X. Yingchao et al. paper titled "Parameter Selection of Gaussian Kernel for One- Class SVM" proposes a novel method to solve the problem of kernel parameter selection in one class classifier, specifically, one-class SVM (OCSVM) [10].

A. J. Hoglund et al. paper titled "A computer host-based user anomaly detection system using the self-organizing map" aimed at designing a system that contains an automatic anomaly detection component [11]. A prototype UNIX anomaly detection system was constructed for anomaly detection attempts to recognize abnormal behavior to detect intrusions. The component for detection used a test based on the self-organizing map to test if user behavior is anomalous.

Shakya, S., & Sigdel, S. in their paper "An approach to develop a hybrid algorithm based on support vector machine and Naïve Bayes for anomaly detection" proposed an anomaly detection technique using an ensemble of Naïve Bayes and Support Vector Machine [12].

Zerpaet et al. proposed a weighting average model based onthe variance of prediction to combine the individual surrogates model to form a hybrid model [13].

METHODOLOGY

Anomaly detection assumes that intrusive behavior, by its nature, is anomalous. Under such scheme, a user can be categorized as outlier if it can be establised that a given user is acting in an unusual manner. And a behavior can be deemed to be unusual by means of comparison against a profile that represents a typical behavior.

The aim of this paper is to develop a prototype software to identify events of anomalous nature and a possible indication of an account misuse. For this work, the data from windows security audit log from computer running Windows Server 2008 is used. When various audits are enabled, the information about user logon sessions, failed logins, successful logins, account lockout, application or process start or stop by the user etc. are logged and these parameters are used to detect anomaly. The basic model contains two phases for anomaly detection: profile creation phase and detection phase.

A. Profile Creation

The user profile has been built from the computer security audit logs which records user's activities as events. The constant window user profile has been used in this paper. User profiles have been created from data recorded in the Windows Security log by identifying the following:

- Duration of Logon Session and
- Frequency of occurrences of the following:
  - Failed login
  - Account lockout
  - Process execution
  - Sessions during working hours

B. Anomaly Detection

For the detection of anomalous behavior, current user behavior profile is compared with the existing normal profile. A certain threshold is maintained to calculate the similarity between the current and normal profile. If current profile is similar to normal profile, i.e,. current profile does not cross the threshold, then the user is genuine otherwise current user behavior is anomalous.

C. Data Collection

Windows Security logs from computers running Windows Server 2008 were collected and examined during this research. The features of interest were extracted by analyzing the different types of windows events. All events contain common data such as date and timestamp, computer name, domain name, user name, event type, and further information specific to each type of event. Security event logging was enable, and all available auditing options were se, and data was collected for a period of nine and seven consecutive months from two different organizations: Organization A and Organization B.

Table I: Data Collected for Analysis

Organization	Data Collection	No of Users	Activity Session Recorded
A	7 months	50	56, 646
B	9 months	60	92, 920

Table II: Summary of Windows Security Audit Events

Event ID	Description
4624	An account was successfully logged on
4625	An account failed to log on
4647	User initiated logoff
4740	A user account was locked out
4688	A new process has been created
4689	A process has exited

D. Data Preparation and Correlation

The computer accounts (indicated by a $ sign at the end of user account) has been removed, as these do not contribute to the actual user behavior. The events related to log on, log off, process start, process exited, failed login, account lockout has been filtered from rest of the events as they provide details of user's log-in sessions, interaction with the application, account failed and locked.

A logon session is a session that begins with successful user authentication and ends with the user logoff. It is necessary to correlate logon events (event IDs 4624 and 4634) to determine the duration of a user's logon session and the number of sessions during working hours.

Session duration and number of session in a working hour is given by the formula below:

$$Session\ duration = Logoff\ time - Logon\ time \quad (1)$$

$$Number\ of\ sessions = Count\ of\ Logon \quad (2)$$

Logon failure events are combined into one event ID 4625 with the proper status codes to identify the different reason for logon failure.

$$Number\ of\ failed\ logons = Count\ of\ event\ id\ 4625 \quad (3)$$

Account lockouts events are combined into one event ID 4740.

$$Number\ of\ Account\ Lockouts = Count\ of\ event\ id\ 4740 \quad (4)$$

Process executed and terminated are combined into event ID 4688 and 4689 respectively.

$$Number\ of\ processes = Count\ of\ event\ id\ 4688\ or\ 4689 \quad (5)$$

E. One Class Support Vector Machine (OCSVM)

OCSVM is a form of SVM which maps input data into a high dimensional feature space and iteratively finds the maximal margin hyperplane that best separates the training data from the origin. The hyperplane corresponds to the classification rule:

$$f(x) = < \boldsymbol{w}, \boldsymbol{x} > + b \quad (6)$$

where,
w is the normal vector and *b* is a bias term. If $f(\boldsymbol{x}) < 0$ we label **x** as an anomaly, otherwise it is labeled normal.

F. Naïve Bayes

Naïve Bayes classifier is statistical classifier and classification is based on Bayes' theorem. Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. Given a series of n attributes, the Naïve Bayes classifier makes 2n! independent assumptions. By Bayes theorem

$$P(C_i|\boldsymbol{X}) = \frac{P(\boldsymbol{X}|C_i)P(C_i)}{P(\boldsymbol{X})} \quad (7)$$

where,
$P(C_i) = Prior\ probability\ of\ class\ C$
$P(\boldsymbol{X}) = Prior\ probability\ of\ training\ set\ \boldsymbol{X}$
$P(\boldsymbol{X}|C_i) = Probability\ of\ \boldsymbol{X}\ given\ C$
$P(C_i|\boldsymbol{X}) = Probability\ of\ C\ given\ \boldsymbol{X}$

G. Proposed Hybrid Model

The weighted ensemble of SVM and Naïve Bayes is proposed, to detect the anomalous behavior of a user. The weights are calculated based on the variance of prediction. The prediction of the hybrid model can be modeled by [13]:

$$y_h = \sum_{i=1}^{N} w_i y_i(x) \quad (8)$$

The weights for each algorithm is computed as follow [13]:

$$w_i = 1/v_i \sum_{j=1}^{m} \frac{1}{v_j} \quad (9)$$

where,
$y_h = Hybrid\ model\ prediction$
$N = Number\ of\ classifiers$
$w_i = Weight\ factor\ for\ i^{th}\ classifier$
$y_i = Response\ estimated\ by\ i^{th}\ classifier$
$x = Vector\ of\ input\ variables$
$v_i = Prediction\ Variance\ of\ i^{th}\ classifier$

Experiment and Results

The experiment was carried out in the dataset prepared for two organizations, i.e., Organization-A and Organization-B. The various quantifiers like Accuracy, Precision, Recall, F1-Score and Area Under Curves were measured.

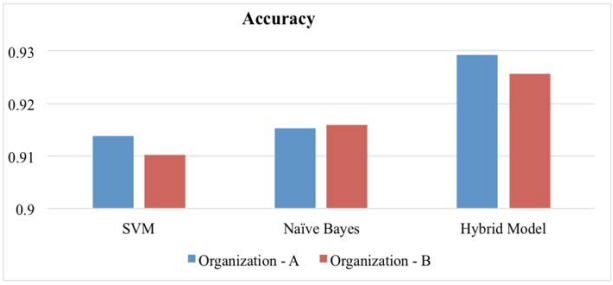

Fig. 1. Comparison of Accuracy of Different Classifier with Hybrid Model

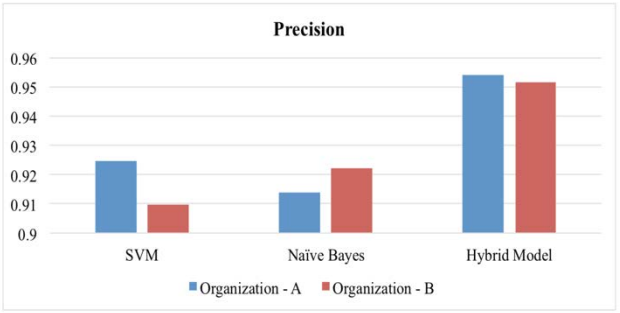

Fig. 2. Comparison of Precision of Different Classifier with Hybrid Model

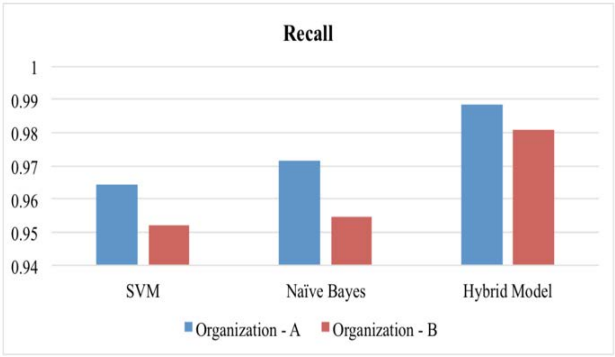

Fig. 3. Comparison of Recall of Different Classifier with Hybrid Model

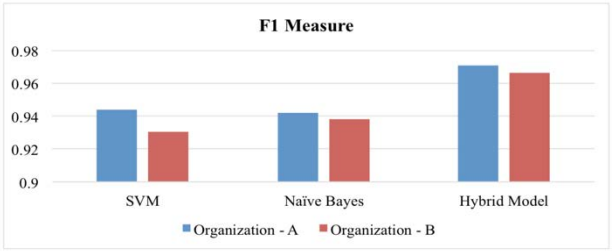

Fig. 4. Comparison of F1-Score of Different Classifier with Hybrid Model

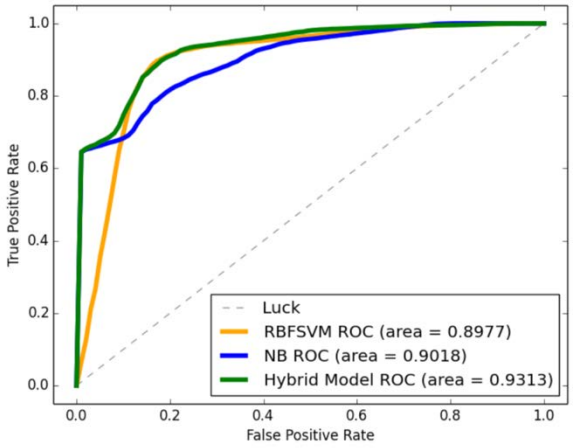

Fig. 5. ROC Curve Comparison for Organization-A

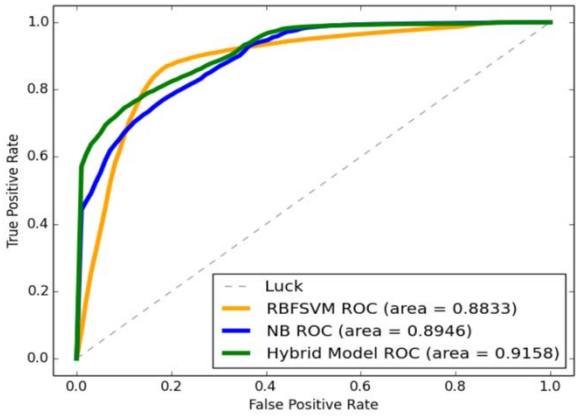

Fig. 6. ROC Curve Comparison for Organization-B

CONCLUSIONS

With the proliferation of new technologies prevention of security breaches by the use of existing security technologies is simply unrealistic. As a result of this anomaly detection is an important component of security. To improve the accuracy rate of intrusion detection in anomaly-based detection different data mining and machine learning technique is used. In this paper hybrid approach is implemented which is an amalgam of two different techniques namely support vector machine and Naïve Bayes. The results of the hybrid approach are compared with the results of other techniques and it is seen that the hybrid approach out performs them. It can be concluded that this hybrid approach is simple and efficient in terms of reducing the false alarm ratio.

FUTURE WORK

There are many possibilities to exploit and extend the learning approach used in this thesis. For example, this thesis includes five parameters to build a normal user profile. This could be extended to include more parameters that explains user profile tracking type of applications (system tuning, programming tool, file sharing, office package etc.) can help determine specific duties each user is assigned. Also, different approach to user profile can be carried out. This thesis has included static approach to user profiling. With this approach detection of alerts is carried out on an after the training period and the user profile remains the same for each week of testing. Alerts are always generated based on that initial user profile. This approach is likely to generate many alerts as the person's usage changes due to their role changing within their organization or when software updates are applied. Sliding window approach can be used where width of profile training time period window remains constant. After training the user profile and anomaly detection based on the data from the testing period, the user profile is recalculated by removing the oldest time profile data and adding the new time period data that has had any events causing false alerts.

REFERENCES

[1] M. Peter S. Karen, "Intrusion Detection and Prevention Systems," in *Handbook of Information and Communication Security*, Peter Stavroulakis and Mark Stamp, Ed. New York, USA: Springer, 2010, pp. 177-192.

[2] A. Banerjee, V. Kumar V. Chandola, "Anomaly Detection: A Survey," *ACM Computing Surveys*, vol. 41, no. 3, p. 58, July 2009.

[3] N. A. Durgin and P. Zhang, "Profile-Based Adaptive Anomaly Detection for Network Security," Sandia National Laboratories, Livermore, California, SANDIA REPORT SAND2005-7293, 2005.

[4] G. Mohay and A. Clark M. Corney, "Detection of Anomalies from User Profiles Generated from System Logs," in *9th Australasian Information Security Conference (AISC 2011)*, vol. 116, Perth, Australia, January 2011, pp. 23-32.

[5] J. P. Anderson, "Computer security threat monitoring and surveillance," James P. Anderson Company, Pennsylvania, Technical report, -, 1980.

[6] M. Zhang, "An Anomaly Detection Model Based on One-Class SVM to Detect Network Intrusions," *2015 11th International Conference on Mobile Ad-hoc and Sensor Networks (MSN)*, vol. -, no. -, pp. 102 - 107, December 2015.

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

[8] T. Singh, and A. Sinhai R. Jain, "A Survey on Network Attacks, Classification and models for Anomaly-based network intrusion detection systems," *Internationa Journel of Engineering Research and Science & Technology*, vol. 2, no. 4, pp. 63-74, November 2013.

[9] S.S. Murtaza et al., "A host-based anomaly detection approach by representing system calls as states of kernel modules," *Software Reliability Engineering (ISSRE), 2013 IEEE 24th International Symposium on*, vol. -, no. -, pp. 431-440, November 2013.

[10] X. Yingchao et al., "Parameter Selection of Gaussian Kernel for One-Class SVM," *IEEE Transactions on Cybernetics*, vol. 45, no. 5, pp. 941 - 953, April 2015.

[11] K. Hatonen, A. S. Sorvari A. J. Hoglund, "A computer host-based user anomaly detection system using the self-organizing map," in *Neural Networks*, 2000. IJCNN 2000, Proceedings of the IEEE-INNS-ENNS International Joint Conference, vol. 5, Como, 2000, pp. 411 - 416.

[12] Shakya, S., & Sigdel, S. (2017, May). "An approach to develop a hybrid algorithm based on support vector machine and Naïve Bayes for anomaly detection." *In Computing, Communication and Automation (ICCCA)*, 2017 International Conference on (pp. 323-327). IEEE.

[13] Zerpa, Luis E., et al. "An optimization methodology of alkaline–surfactant–polymer flooding processes using field scale numerical simulation and multiple surrogates." Journal of Petroleum Science and Engineering 47.3 (2005): 197-208.

AUTHORS

First Author – Roshan Pokhrel, MSc. Computer System and Knowledge Engineering, Tribhuvan University, roshanpokhrel@gmail.com

Second Author – Prabhat Pokharel, MSc. Computer Science, Nepal College of Information Technology, pokharelprabhat@gmail.com.

Third Author – Arun Kumar Timalsina, PhD, Deputy Director at Center for Applied Research & Development, Institute of Engineering, Pulchowk Campus, Tribhuvan University, t.arun@ioe.edu.np

Correspondence Author – Roshan Pokhrel, roshanpokhrel@gmail.com, 069mscs664@ioe.edu.np, +977-9841624444