A Comprehensive Survey on Machine learning based Intrusion Detection Technique

Roshan Lal[1], Pooja Ahlawat[2]

[1] M. tech Scholar, Department of Computer Science & Engineering, R N College of Engineering, Rohtak
[2] Head of department, Department of Computer Science & Engineering, R N College of Engineering, Rohtak

ABSTRACT

In the last few year , Intrusion Detection materializes the high network security. Thus tries to be the most perfect system to deal with the network security and the intrusions attacks. Monitoring activity of the network and that of threats is the feature of the ideal Intrusion Detection System. Intrusion Detection System is classified on the basis of the source of Data and Model of Intrusion. There are some challenges faced by the Intrusion Detection System. Neural Network and Machine Learning are the approaches through which the challenges can be overwhelmed. Anomaly in the Anomaly based Intrusion Detection System can be detected using various Anomaly detection techniques. This paper presents literature of the different approach developed by the author to prevent the network from serious threats such as: Denial of services, remote to local (R2L), user to root etc. together with its merits and demerits.

Keywords: Anomaly, IDS, K-means, KDD dataset, Misuse, PSO, Threats

I. INTRODUCTION

Today, political and commercial entities are increasingly engaging in sophisticated cyber-warfare to damage, disrupt, or censor information content in computer networks [6]. In designing network protocols, there is a need to ensure reliability against intrusions of powerful attackers that can even control a fraction of parties in the network. The controlled parties can launch both passive (e.g., eavesdropping, nonparticipation) and active attacks (e.g., jamming, message dropping, corruption, and forging). Intrusion detection is the process of dynamically monitoring events occurring in a computer system or network, analyzing them for signs of possible incidents and often interdicting the unauthorized access [4]. This is typically accomplished by automatically collecting information from a variety of systems and network sources, and then analyzing the information for possible security problems.

Motivation: Traditional intrusion detection and prevention techniques, like firewalls, access control mechanisms, and encryptions, have several limitations in fully protecting networks and systems from increasingly sophisticated attacks like denial of service. Moreover, most systems built based on such techniques suffer from high false positive and false negative detection rates and the lack of continuously adapting to changing malicious behaviors. In the past decade, however, several Machine Learning (ML) techniques have been applied to the problem of intrusion detection with the hope of improving detection rates and adaptability. These techniques are often used to keep the attack knowledge bases up-to-date and comprehensive.

Study Approach: In this paper, we study several papers that use ML methods for detecting malicious behavior in distributed computer systems. There is a huge body of work in this area thus, we decided to carefully select a few papers based on two factors: diversity and citations count. By diversity we mean most ML techniques for IDS are covered but only one paper is picked from the set of papers that use the same technique. Also, the papers are chosen based on their citations count as this factor greatly shows how much the corresponding work has influenced the community. All non-survey papers studied here are cited at least 100 times.

Paper Organization: In Section 2, we briefly state the main challenges in intrusion detection and describe two general approaches for solving these problems. In Section 3, we review several intrusion detection techniques based on traditional AI.

In section 4, we define various core methods of computational intelligence and describe several Ceased algorithms proposed in the literature.

RELATED WORK

A new technique of combining swarm intelligence (Simplified Swarm Optimization) and data mining algorithm (Random Forest) for feature selection and reduction. SSO is used to find more appropriate set of attributes for classifying network intrusions, and Random Forest is used as a classifier. In the preprocessing step, we optimize the dimension of the dataset by the proposed SSO-RF approach and find an optimal set of features. SSO is an optimization method that has a strong global search capability and is used here for dimension optimization [6].

Two techniques C5.0 and artificial neural network (ANN) are utilized with feature selection [7]. Feature selection techniques will discard some irrelevant features while C5.0 and ANN acts as a classifier to classify the data in either normal type or one of the five types of attack. KDD99 data set is used to train and test the models; C5.0 model with numbers of features is producing better results with all most 100% accuracy. Performances were also verified in terms of data partition size.

A technique which is divided into four steps: initial step, k-means clustering is used to generate different training subset then based on the obtained subset, various neuro-fuzzy data model are trained [8]. Consequently, a vector for SVM classification is obtained and in last, classification using radial SVM is applied to detect the intrusion occurred or not. To demonstrate the applicability and ability of the new method, the result of KDD dataset is confirmed in which it shows that the proposed methods produce better result than the BP, multi-class SVM and other approach such as decision tree etc.

Bharat et al., proposed a method for intrusion detection using Particle Swarm Optimization with Genetic Algorithm based feature selection and using Adaptive Mutation for sluggish convergence of optimization algorithm [9].

The results thus obtained are around 92% that proves the proposed method to be reasonably effective in intrusion detection. It developed a hybrid method of C5.0 and SVM and investigate and evaluate the performance of our proposed method with DARPA dataset. The motivation for using the hybrid approach is to improve the accuracy of the intrusion detection system when compared to using individual SVM and individual SVM [10].

It introduced a new host-based anomaly intrusion detection methodology using discontinuous system call patterns, in an Endeavour to increase detection rates whilst plummeting false alarm rates [8]. The main idea is to apply a semantic structure to kernel level system calls in order to replicate inherent activities hidden in high-level programming languages which can help comprehend program anomaly behaviour. Outstanding results were demonstrated using a multiplicity of decision engines evaluating the KDD98 and UNM data sets and a new, modern data set. The ADFA Linux data set was created as part of this research using a recent operating system and contemporary hacking methods and is now openly available. Additionally, the new semantic method possesses an inherent flexibility to mimicry attacks and demonstrated a high level of portability between dissimilar operating system versions.

The clustering method by using hybrid method based on Principal Component Analysis (PCA) and Fuzzy Adaptive Resonance Theory (FART) for classifying diverse attacks. The PCA is concerned to random selects the best provenance and reduction the feature space. The FART is implementing which is used to classifying dissimilarity in collection of data, regular and irregular. The proposed method can improves the high performance of the detection rate and to reduce the false alarm rate and this is computed approach on the benchmark data from KDD Cup 99 data set [11].

The method which is based on MAHALANOBIS Distance characteristic ranking and an improved comprehensive search to choose an improved combination of features. They evaluated the approach on the KDD CUP 1999 datasets using SVM classifier and KNN classifier. The results showed that classification is done with high classification rate and low misclassification rate with the reduced feature subsets [10].

AI-BASED TECHNIQUES

Laskov et al. [7] develop an experimental framework for comparative analysis of supervised (classification) and unsupervised learning (clustering) techniques for detecting malicious activities. The supervised methods evaluated in this work include decision trees, k-Nearest Neighbor (kNN), Multi-Layer Perceptron (MLP), and Support Vector Machines (SVM). The unsupervised algorithms include γ-algorithm, k-means clustering, and single linkage clustering. They define two scenarios for evaluating the aforementioned learning algorithms from both categories. In the first scenario, they assume that training and test data come from the same unknown distribution. In the second scenario, they consider the case where the test data comes from new (i.e., unseen) attack patterns. This scenario helps us understand how much an IDS can generalize its knowledge to new malicious patterns, which is often very essential for an IDS system since today's sophisticated adversaries tend to use several intrusion patterns to escape from modern IDS.

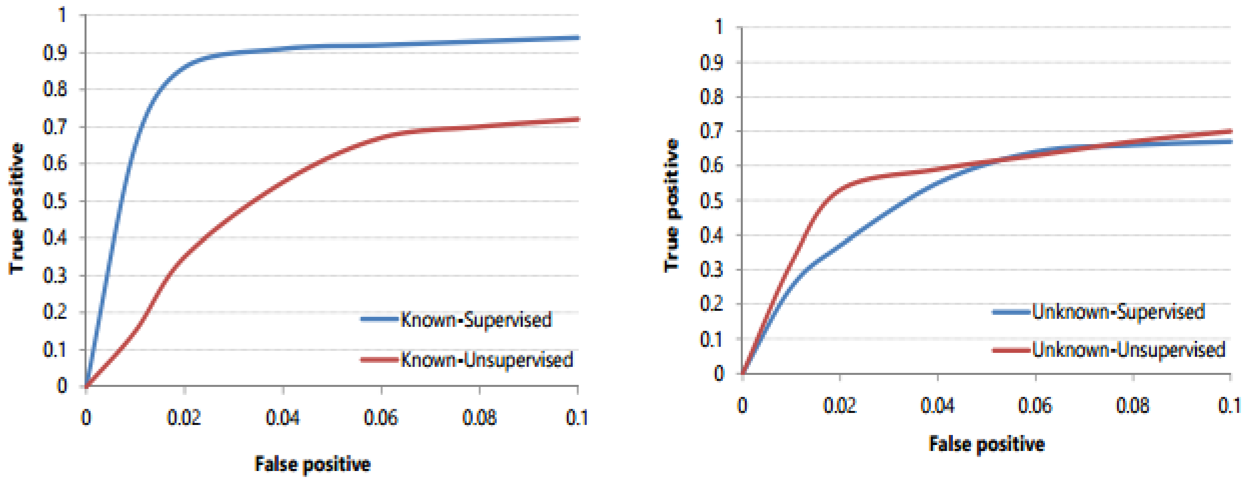

Figure 1: Average of detection rates for methods evaluated in [7] in two scenarios: test data contains only known attacks (left) and test data contains unknown attacks (right).

The results of [7] show that the supervised algorithms in general show better classification accuracy on the data with known attacks (the first scenario). Among these algorithms, the decision tree algorithm has achieved the best results (95% true positive rate, 1% false-positive rate). The next two best algorithms are the MLP and the SVM, followed by the k-nearest neighbor algorithm. However, if there are unseen attacks in the test data, then the detection rate of supervised methods decreases significantly. This is where the unsupervised techniques perform better as they do not show significant difference in accuracy for seen and unseen attacks. Figure 1 shows the average true/false positive rates of all methods evaluated in [7]. As the plots show, the supervised techniques generally perform better although unsupervised methods give more robust results in both scenarios.

Zanero and Savaresi [17] introduce a two-tier anomaly-based architecture for IDS in TCP/IP networks based on unsupervised learning: the first tier is an unsupervised clustering algorithm, which build small-size patterns from the network packets payload. In other words, TCP or UDP packet are assigned to two clusters representing normal and abnormal traffic. The second tier is an optimized traditional anomaly detection algorithm improved by the availability of data on the packet payload content. The motivation behind the work is that unsupervised learning methods are usually more powerful in generalization of attack patterns than supervised methods thus, there is a hope that such an architecture can resist polymorphic attacks more efficiently.

Lee and Solfo [8] build a classifier to detect anomalies in networks using data mining techniques. They implement two general data mining algorithms that are essential in describing normal behavior of a program or user. They propose an agent-based architecture for intrusion detection systems, where the learning agents continuously compute and provide the updated detection models to the agents. They conduct experiments on Sendmail1 system call data and network tcpdump data to demonstrate the effectiveness of their classification models in detecting anomalies. They finally argue that the most important challenge of using data mining approaches in intrusion detection is that they require a large amount of audit data in order to compute the profile rule sets.

Sommer and Paxson [13] study the imbalance between the extensive amount of research on ML-based intrusion detection versus the lack of operational deployments of such systems. They identify challenges particular to network intrusion detection and provide a set of guidelines for fortifying future research on ML-based intrusion detection. More specifically, they argue that an anomaly-based IDS requires outlier detection2 while the classic application of ML is a classification problem that deals with finding similarities between activities. It is true that in some cases, an outlier detection problem can be modeled as a classification problem in which there are two classes: normal and abnormal. In machine learning, one needs to train a system with training patterns of all classes while in anomaly detection one can only train on normal patterns. This means that anomaly detection is better for finding variations of known attacks, rather than previously unknown malicious activity. This is why ML methods have been applied to spam detection more effectively than to intrusion detection.

OVERVIEW OF INTRUSION DETECTION TECHNIQUES

There are different data mining techniques used for Intrusion detection. In this section we are describing general idea of these techniques:

A. K-means Clustering

K-means clustering [12] is one of the simplest unsupervised clustering algorithms. The algorithm takes input parameter k cluster so that the intra-cluster similarity is high and inter-cluster similarity is low. "K number given in advance. K means clustering takes less time as compared to the hierarchical clustering and yields better results. With the help of clustering training dataset is clustered into 5 dataset wherein 4 dataset will be a type of intrusion called attack dataset and one with normal data type called normal dataset. Here are the four steps of the clustering algorithms:

1. Define the number of clusters K.
2. Initialize the K-cluster centroids. This can be done by randomly dividing all objects into K clusters, computing their centroids, and authenticating that all centroids are diverse from each other. Otherwise, the centroids can be initialized to K arbitrarily preferred, diverse objects.
3. Iterate over all objects and calculate the distances to the centroids of all clusters. Allot each object to the cluster with the adjoining centroid.
4. Recalculate the centroids of both customized clusters.
5. Reiterate step 3 until the centroids do not change any more.

A distance function is obligatory in order to calculate the distance (i.e. similarity) among two objects. The regularly used distance function is the Euclidean one which is defined as:

$$d(x,y)= \sqrt{}\sum(xi-yi)^2$$

Where x = (x1 . . . xm) and y = (y1…ym) are two input vectors with m quantitative features. In the Euclidean distance function, all features contribute equally to the function value. However, since different features are usually measured with different metrics or at different scales, they must be normalized before applying the distance function.

B. Genetic Algorithm

Genetic algorithms (GA) are search algorithms based on the principles of natural selection and genetics. The aim of development of GAs is developing a system as robust and as adaptable to the environment as the natural systems [5]. Genetic algorithms are search procedures often used for optimization problems. In this algorithm an initial population of chromosomes is generated randomly where each chromosome represents a possible solution to the problem (a set of parameters). From each chromosome different positions are encoded as, characters, bits or no.s. These positions could be known as genes. Goodness of each chromosome calculated by evaluation function, according to the desired solution; this function is known as "Fitness Function". It holds three phases after calculating fitness function i.e. selection, crossover, and mutation. In selection it selects most optimal solution of a problem calculated by using fitness function. The selected chromosomes are called parents. After selection phase crossover phase comes in which characteristics of different parent chromosomes exchange and they produce offspring, there are various methods for crossover, for example N point crossover, uniform crossover etc. Mutation involves flipping of one or more bits of chromosomes and then evaluated using some fitness criteria. After termination chromosomes having the highest fitness function called the best solution of the problem. Mutation maintains diversity in the population. Genetic algorithm from other algorithm because it implemented at machine code level, it is fast to detect in real time. Some of its good properties, e.g. robust to noise, no gradient information is needed to find global optimal or sub-optimal solution, self-learning capabilities made it best approach.

C. Support Vector Machine

The Support Vector Machine is one of the most successful classification algorithms in the data mining area.SVM uses a high dimension space to find a hyper-plane to perform binary classification.SVM approach is a classification technique based on Statistical Learning Theory (SLT). It is based on the idea of hyper plane classifier. The goal of SVM is to find a linear optimal hyper plane so that the margin of separation between the two classes is maximized. The SVM uses a portion of the data to train the system. It finds numerous support vectors that correspond to the training data. These support vectors will form a SVM model. According to this model, the SVM will categorize a given unknown dataset into target classes [11].

CONCLUSION

We reviewed several influential algorithms for intrusion detection based on various machine learning techniques. Characteristics of ML techniques makes it possible to design IDS that have high detection rates and low false positive rates while the system quickly adapts itself to changing malicious behaviors. We divided these algorithms into two types of ML-based schemes: Artificial Intelligence (AI) and Computational Intelligence (CI). Although these two categories of algorithms share many similarities, several features of CI-based techniques, such as adaptation, fault tolerance, high computational speed and error resilience in the face of noisy information, conform the requirement of building efficient intrusion detection systems.

REFERENCES

[1]. Steven A. Hofmeyr, Stephanie Forrest, and Anil Somayaji. Intrusion detection using sequences of system calls. Journal of Computer Security, 6(3):151–180, August 1998.

[2]. Peter Mell Karen Scarfone. Guide to intrusion detection and prevention systems (idps). National Institute of Standards and Technology, NIST SP - 800-94, 2007. Available at http://www.nist.gov/customcf/get_pdf.cfm?pub_id=50951.

[3]. Jungwon Kim, Peter J. Bentley, Uwe Aickelin, Julie Greensmith, Gianni Tedesco, and Jamie Twycross. Immune system approaches to intrusion detection – a review. Natural Computing, 6(4):413–466, December 2007.

[4]. Andrew F. Krepinevich. Cyber warfare: A nuclear option?, 2012. Center for Strategic and Budgetary Assessments, Washington, DC, USA, 2012.

[5]. Pavel Laskov, Patrick Dssel, Christin Schfer, and Konrad Rieck. Learning intrusion detection: Supervised or unsupervised? In Image Analysis and Processing ICIAP 2005, volume 3617 of Lecture Notes in Computer Science, pages 50–57. Springer Berlin Heidelberg, 2005.

[6]. Chih-Fong Tsai, Yu-Feng Hsu, Chia-Ying Lin, and Wei-Yang Lin. Review: Intrusion detection by machine learning: A review. Expert Syst. Appl., 36(10):11994– 12000, December 2009.

[7]. Mahdi Zamani, Mahnush Movahedi, Mohammad Ebadzadeh, and Hossein Pedram. A DDoS-aware IDS model based on danger theory and mobile agents. In Proceedings of the 2009 International Conference on Computational Intelligence and Security - Volume 01, CIS ’09, pages 516–520, Washington, DC, USA, 2009. IEEE Computer Society.

[8]. Mahdi Zamani, Mahnush Movahedi, Mohammad Ebadzadeh, and Hossein Pedram. A danger-based approach to intrusion detection. CoRR, abs/1401.0102, 2014.

[9]. Stefano Zanero and Sergio M. Savaresi. Unsupervised learning techniques for an intrusion detection system. In Proceedings of the 2004 ACM symposium on Applied computing, SAC ’04, pages 412–419, New York, NY, USA, 2004.

[10]. Feng Guorui, ZouXinguo, Wu Jian, (2012).“Intrusion detection based on the semi supervised Fuzzy C- Means clustering algorithm”, Department of Information Science Technology, Shandong University, China, pp. 2667-2670, 2012.