IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

ANOMALY DETECTION USING MACHINE LEARNING

SAHIL DESWAL , SHIVANI THAKUR , SAMRITI BHAGAT

CGC JHANJERI, MOHALI

ABSTRACT

Intrusion detection system is still the subject of widespread interest among researchers. Even after years of research, the intrusion detection community still faces a difficult problem. Reducing the number of false positive during the detection process of unknown attack pattern remains an open problem. However, some recent research has shown that there is a potential solution to this problem. Anomaly detection is a key issue in intrusion detection. Disruption of normal operation indicates the presence of attacks, bugs, defects, etc. that may be intentionally or unintentionally induced. This white paper outlines research directions for applying supervised and unsupervised methods to address the problem of anomaly detection. Cited references cover major theoretical issues and guide researchers in interesting research directions.

Keywords

Supervised Machine Learning, Unsupervised Machine Learning, Network Intrusion Detection.

1. INTRODUCTION

Intrusion detection has been studied for approximately 20 years. Intrusions are the activities that violate the information system security policy, and intrusion detection is the identifying intrusions process. Intrusion detection is based on the assumption that the intruder behavior will be significantly diverse from the legitimate behaviors, which facilitates and enables the detection of a lot of non-authorized activities. Intrusion detection systems are usually used together with other protection systems such as access control and authentication as a second defense line to protect information systems. There are many reasons that make intrusion detection the important parts in the whole defense system. First, many of the traditional systems and applications have been built and developed without taking security seriously into account. Second, computer systems and applications may have flaws or bugs in their design that could be used by intruders to attack the systems or applications. Therefore, the preventive technique may not be as effective as anticipated. Despite their importance, IDSs are not replacement for preventive security mechanisms, but they complement the other protective mechanisms to enhance the security of the system. Actually, IDSs alone cannot offer sufficient protection for information systems. Therefore, IDSs should be used with other preventive security mechanisms as a part of a total protective system [59]. Intrusion detection systems are classified as a signature detection system and an anomaly detection system. A signature detection system identifies traffic or application data patterns assumed to be malicious,

while anomaly detection systems compare activities with ‘‘normal baseline. Both signature detection and anomaly detection systems have advantages and drawbacks. The primary advantage of signature detection is that it can detect known attacks fairly for all of the potential attacks against a network. Anomaly detection systems have two main advantages over signature based intrusion detection systems. The first advantage is their capability to detect unknown attacks because they can model the normal operation of a system and detect deviations from this model. The second advantage is the customization ability of the normal activity profiles for every system, application and network. This will increase the difficulty for an attacker to know what activities can be done without getting detected. However, the anomaly detection approach has its drawbacks such as the system complexity, high false alarms and the difficulty of detecting which event triggers those alarms. These are some of many technical challenges that have to be handled before the adoption of anomaly detection systems. This paper presents an overview of research directions for applying supervised and unsupervised methods for managing the problem of anomaly detection. The rest of this paper is organized as follows. In Section 2, the general architecture of anomaly intrusion detection systems and detailed discussions on the supervised and unsupervised techniques used in anomaly detection are described. Finally, the conclusion of this paper is presented in section 3

2. ANOMALY DETECTION TECHNIQUES

The general architecture of all anomaly based network intrusion detection systems (A-NIDS) methods is similar. According to [12] and [13], generally, all of them consist of the following basic modules or stages (Fig. 1). These stages are parameterization, training and detection. Parameterization includes collecting raw data from a monitored environment. The raw data should be representative of the system to be modeled, (e.g. Packet data from a network). The training stage seeks to model the system using manual or automatic methods. For the client-server architecture, the server is a host that keeps waiting for incoming connections. When a connection is established between client and server, the server would instantiate a socket, which will be used to instantiate a handler object that runs on a separate thread. These handlers will be kept in a collection object. The behaviors represented in the model will differ based on the technique used. Detection compares the system generated in the training stage with the selected parameterized data portion. Threshold criteria will be selected to determine anomalous data instance [13].

achine learning can build the required model automatically based on some given training data. A motivation for this approach is the availability of the necessary training data, or it can be at least obtained more easily compared to the effort needed to define the model manually. With the increase the complexity and the number of different attacks, machine learning techniques that allow constructing and maintaining anomaly detection system (ADS) with less human intervention look is the only practical approach to achieve the next generation of intrusion detection systems. Applying machine learning techniques for intrusion detection can automatically build the model based on the training data set, which contains data instances that can be described using a set of attributes (features) and associated labels. The attributes can be of different types such as categorical or continuous. The attributes nature determines the applicability of anomaly detection techniques.The labels associated with data instances are usually in form of binary values, i.e. normal and anomalous. On the other hand, some researchers have employed various attacks types such as DoS, U2R, R2L and Probe rather than the anomalous label. This learning technique is capable to provide more information about the anomalies types. Anomaly detection techniques include supervised techniques and unsupervised techniques (Fig.2) [20, 55].

2.1 Supervised Anomaly

Detection Supervised methods (also known as classification methods) required a labeled training set containing both normal and anomalous samples to construct the predictive model. Theoretically, supervised methods provide better detection rate than semi-supervised and unsupervised methods, since they have access to more information. However, there exist some technical issues, which make these methods seem not accurate as they are supposed to be. The first issue is the shortage of a training data set that covers all areas. Moreover, obtaining accurate labels is a challenge and the training sets usually contain

some noises that result in higher false alarm rates. The most common supervised algorithms are, Supervised Neural Networks, Support Vector Machines SVM), k-Nearest Neighbors, Bayesian Networks and Decision Tree [60].

2.1.1 .K -Nearest Neighbor (k-NN)

K-nearest neighbor (k-NN) is one of the modest and conventional nonparametric techniques for classifying samples [4], [32]. It calculates the approximate distances between various points on the input vectors, and then assigns the unlabeled point to the class of its K-nearest neighbors. In the process of creating k-NN classifier, (k) is an important parameter and various (k) values can cause various performances. If k is very huge, the neighbors, which used for prediction, will consume large classification time and affect the prediction accuracy. Shailendra and Sanjay [51] introduced a hybrid approach for feature selection, which includes two phases filter and wrapper. The filter phase selects the features with highest information gain and feeds them to the wrapper phase that outputs the final feature subset. The final feature subsets are input to the K-nearest neighbor classifier to classify attacks. This algorithm effectiveness is demonstrated on DARPA KDDCUP99 cyber-attack dataset. Ming. Y [33] suggested a genetic algorithm combined with KNN (k-nearest-neighbor) for feature selection and weighting. All initial 35 features in the training phase were weighted, and the ones of highest weights were selected for testing. Many DoS attacks were applied to evaluate the systems.

2.1.2Bayesian Network (BN)

Heckerman [17] defined a Bayesian as "A Bayesian Network (BN) is a model that encodes probabilistic relationships among variables interest. This technique is generally used for intrusion detection in combination with statistical schemes. It has several advantages, including the capability of encoding interdependencies between variables and of predicting events, as well as the ability to incorporate both prior knowledge and data. " Johansen and Lee [22] stated that a BN system provides a proper mathematical foundation to make straightforward apparently a difficult problem. They have proposed that BN based IDS should distinguish attacks from normal network activity by comparing metrics of each network traffic sample. Moore and Zuev [35] used a supervised Naive Bayes classifier and 248 flow features to differentiate between different types of application such as packet length and inter arrival times, in addition to numerous TCP header derived features. Correlation-based feature selection was used to define stronger features, and it indicated that only a small subset of fewer than 20 features is needed for accurate classification.

2.1.2 Supervised Neural Network (NN)

The NNs learning predict different users and daemons behavior in systems. If they properly designed and implemented, NNs have the capability to address many problems encountered by rule-based approaches. The main NNs advantage is their tolerance to imprecise data and uncertain information, and their ability to conclude solutions from data without having previous knowledge of the regularities in the data. This, in combination with their ability to generalize from learning data, has made them a proper approach to ID. In order to apply this approach to ID, data representing attacks and non-attacks have to be introduced to the NN to adjust automatically network coefficients during the training phase [27]. Multilayer perceptron (MLP) and Radial basis function (RBF) are the most commonly supervised neural networks used. Multi Layered Perceptron (MLP). MLP can only classify linearly separable instances sets. If a straight line or plane can be drawn to separate the input instances into their correct categories, input instances are linearly separable and the perceptron will find the solution. If the instances are not linearly separable learning will never reach a point where all instances are classified properly. Multilayered perceptron (Artificial Neural Networks) have been created to try solving this problem [47]. There were researches implement an IDS using MLP, which has the capability of detecting normal and attacks connection as in [54] and [48]. They were implemented using MLP of three and four layers neural network. Moradi and Zulkernine [36], Mohammed et al. [34] used three layers MLP (two hidden layers) not only for detecting normal and attacks connection but also for identifying attack type. Yao et al. [57] proposed Hybrid MLP/CNN neural network, which is constructed in order to enhance the detection rate of timedelayed attacks. While obtaining a similarly detection rate of real-time attacks as the MLP does, the proposed approach

can detect time-delayed attacks efficiently with chaotic neuron. Radial Basis Function Neural Networks (RBF) is another common type of feed forward neural networks. Since they perform classification by measuring distances between inputs and the centers of the RBF hidden neurons, RBF networks are much faster than the time consuming back propagation, and most suitable for problems with large sample size [6]. Research, such as Hofmann et al. [18], Liu et al. [31], Rapaka [45] employed RBFs to learn multiple local clusters for wellknown attacks and for normal events. Other than being a classifier, the RBF network is also used to fuse results from multiple classifiers [6]. It outperformed five different decision fusion functions, such as a Dempster–Shafer combination and weighted majority vote. Jiang et al. [21] introduced a new approach, which combines both misuse and anomaly detections in a hierarchical RBF network. In the first layer, an RBF anomaly detector defines the event nature if it is normal or anomaly. Anomaly events then pass through a RBF misuse detector chain, where each detector detects a specific type of attack. Un classified anomaly events by any misuse detectors were saved into a database. If enough anomaly events were collected, they were clustered by a C-means clustering algorithm into different groups, which used to train a misuse RBF detector, and added to the misuse detector chain. This mannar leads to detect and lable and all intrusion events automatically.

2.1.3 Decision Tree (DT)

Quinlan [43] defined Decision Trees as "powerful and common tools for classification and prediction. A decision tree is a tree that has three main components: nodes, arcs and leaves. Each node is labeled with a feature attribute, which is most informative among the attributes not yet considered in the path from the root. Each arc out of a node is labeled with a feature value for the node's feature, and each leaf is labeled with a category or class. A decision tree can then be used to classify a data point by starting at the root of the tree and moving through it until a leaf node is reached. The leaf node provides the classification of the data point. ID3 and C4.5 developed by Quinlan are the most common implementations of the Decision Tree. " Peddabachigari et al. [41], proposed decision trees (DT) and support vector machines (SVM) as intrusion-detection model.

2.1.4 Support Vector Machine (SVM)

Support vector machines (SVM) are proposed by Vapnik [56]. SVM first maps the input vector into a higher-dimensional feature space and then obtains the optimal separating hyperplane in the high dimensional feature space. Moreover, a decision boundary, i.e. the separating hyper-plane, is determined by support vectors rather than the whole training samples and thus is extremely robust to outliers. In particular, an SVM classifier is designed for binary classification. That is, to separate a set of training vectors, which belong to two different class's notes that the support vectors are the training samples close to a decision boundary. The SVM also provides a user-specified parameter called a penalty factor. It allows users to make a tradeoff between the number of misclassification samples and the width of a decision boundary. Mukkamala et al. [37] designed model to network anomaly detection problems by "applied kernel classifiers and classifier design methods to network anomaly detection problems. They evaluated the impact of kernel type and parameter values on the accuracy with which a support vector machine (SVM) performs intrusion classification. Jun et al. [25] proposed PSO–SVM model is applied to an intrusion detection problem, the standard PSO is used to determine free parameters of support vector machine and the binary PSO is to obtain the optimum feature subset at the building intrusion detection system. Paulo et al. [40] proposed an intrusion detection system model based on the behavior of network traffic through the analysis and classification of messages. Two artificial intelligence techniques named Kohonen neural network (KNN) and support vector machine (SVM) are applied to detect anomalies.

2.2 Unsupervised Anomaly Detection Techniques

These techniques do not need training data. As alternative, they based on two basic assumptions. First, they presume that most of the network connections are normal traffic and only a very small traffic percentage is abnormal. Second, they anticipate that malicious traffic is statistically various from normal traffic. According to these two assumptions, data groups of similar instances which appear frequently are assumed to be normal traffic, while infrequently instances which considerably various from the majority of the instances are regarded to be malicious [7]. The most common unsupervised algorithms are, K-Means, Self-organizing maps (SOM), C-means, Expectation-Maximization Meta algorithm (EM), Adaptive resonance theory (ART), Unsupervised Niche Clustering (UNC) and One-Class Support Vector Machine

2.2.1 Clustering Techniques

Rawat [45] and many more found that Clustering techniques work by grouping the observed data into clusters, according to a given similarity or distance measure. There exist at least two approaches to clustering based anomaly detection. In the first approach, the anomaly detection model is trained using unlabeled data that consist of both normal as well as attack traffic. In the second approach, the model is trained using only normal data and a profile of normal activity is created. The idea behind the first approach is that anomalous or attack data forms a small percentage of the total data. If this assumption holds, anomalies and attacks can be detected based on cluster sizes large clusters correspond to normal data, and the rest of the data points, which are outliers, correspond to attacks.

2.2.1.1 Unsupervised Neural Network

The two typical unsupervised neural networks are selforganizing maps and adaptive resonance theory. They used similarity to group objects. They are adequate for intrusion detection tasks where normal behavior is densely concentrated around one or two centers, while anomaly behavior and intrusions spread in space outside of normal clusters. The Self-organizing map (SOM) is trained by an unsupervised competitive learning algorithm [26]. The aim of the SOM is to reduce the dimension of data visualization. That is, SOM outputs are clustered in a low dimensional (usually 2D or 3D) grid. It usually consists of an input layer and the Kohonen layer, which is designed as the two-dimensional arrangement of neurons that maps n dimensional input to two dimensions. Kohonen's SOM associates each of the input vectors to a representative output. The network finds the node nearest to each training case and moves the winning node, which is the closest neuron (i.e. the neuron with minimum distance) in the training course. That is, SOM maps similar input vectors onto the same or similar output units on such a two-dimensional map, which leads to self-organize the output units into an ordered map and the output units of similar weights are also placed nearby after training. SOMs are the most popular neural networks to be trained for anomaly detection tasks. For example Kayacik et al. [28], they have created three layers of employment: First, individual SOM is associated with each basic TCP feature. Second layer integrates the views provided by the first-level SOM into a single view of the problem. The final layer is built for those neurons, which win for both attack and normal behaviors. Oh and Chae [39] proposed an approach a real-time intrusiondetection system based on SOM that groups similar data and visualizes their clusters. The system labels the map produced by SOM using correlations between features. Jun et al. [24] introduced a novel methodology to analysis the feature attributes of network traffic flow with some new techniques, including a novel quantization model of TCP states. Integrating with data preprocessing, the authors construct an anomaly detection algorithm with SOFM and applied the detection frame to DARPA Intrusion Detection Evaluation Data. Adaptive Resonance Theory (ART). The adaptive resonance theory embraces a series of neural network models that perform unsupervised or supervised learning, pattern recognition, and prediction. Unsupervised learning models Include ART-1, ART- 2, ART-3, and Fuzzy ART. Various supervised networks are named with the suffix ''MAP'', such as ARTMAP, Fuzzy ARTMAP, and Gaussian ARTMAP. Amini et al. [1] Compared the performance of ART-1 (accepting binary inputs) and ART-2 (accepting continuous inputs) on KDD99 data. Liao et al. [29] deployed Fuzzy ART in an adaptive learning framework which is suitable for dynamic changing environments. Normal behavior changes are efficiently accommodated while anomalous activities can still be identified.

2.2.1.2 K-Means

K-means algorithm is a traditional clustering algorithm. It divides the data into k clusters, and guarantee that the data within the same cluster are similar, while the data in a various clusters have low similarities. K-means algorithm is first selected K data at random as the initial cluster center, for the rest data add it to the cluster with the highest similarity according to its distance to the cluster center, and then recalculate the cluster center of each cluster. Repeat this process until each cluster center doesn't change. Thus data are divided into K clusters. Unfortunately, K-means clustering is sensitive to the outliers and a set of objects closer to a centroid may be empty, in which case centroids cannot be updated [16]. [30] proposed K-means algorithms for anomaly detection. Firstly, a method to reduce the noise and isolated points in the data set was advanced. By dividing and merging clusters and using the density radius of a super sphere, an algorithm to calculate the number of the cluster centroid was given. By more accurate method of finding k clustering center, an anomaly detection model was presented to get better detection effect. Cuixiao et al. [7] proposed a mixed intrusion detection system (IDS) model. Data are examined by the misuse detection module and then the detection of abnormal data is performed by anomaly detection module. In this model, unsupervised clustering method is used to build the anomaly detection module. The algorithm used is an improved algorithm of K-means clustering algorithm and it is demonstrate to have a high detection rate in the anomaly detection module.

2.2.1.3 Fuzzy C-Means (FCM)

Fuzzy C-means is a clustering method, which grants one piece of data to belong to two or more clusters. It was developed by Dunn [9] and improved later by Bezdek [3], it is used in applications for which hard classification of data is not meaningful or difficult to achieve (e.g, pattern recognition). C-means algorithm is similar to K-Means except that membership of each point is defined based on a fuzzy function and all the points contribute to the relocation of a cluster centroid based on their fuzzy membership to that cluster. Shingo et al. [52] proposed a new approach called FC-ANN, based on ANN and fuzzy clustering to solve the problem and help IDS achieving higher detection rate, less false positive rate and stronger stability. Yu and Jian [58] proposed an approach integrating several soft computing techniques to build a hierarchical neuro-fuzzy inference intrusion detection system. In this approach, principal component analysis neural network is used to reduce feature space dimensions. The preprocessed data were clustered by applying an enhanced fuzzy C-means clustering algorithm to extract and manage fuzzy rules. Another approach that uses a fuzzy approach for unsupervised clustering is presented by Shah et al. [50]. They employed the Fuzzy C-Medoids (FCMdd) in order to index cluster streams of system call, low level Kernel data and network data.

2.2.1.4 Unsupervised Niche Clustering (UNC)

(UNC) is a robust clustering algorithm, which uses an evolutionary algorithm with a niching strategy (Nasraoui et al. [38]. The evolutionary algorithm helps to find clusters using a robust density fitness function, while the niching technique allows it to create and maintain the niches (candidate clusters). Since UNC is based on genetic optimization, it is much less susceptible to suboptimal solutions than traditional techniques. The algorithm main advantage is the ability to handle noise and to determine clusters number automatically Elizabeth et al. [10] combined the UNC with fuzzy set theory for anomaly detection and applied it to network intrusion detection. They associated to each cluster generated by the UNC a member function that follows a Gaussian shape using evolved cluster center and radius. Such cluster membership functions will define the normalcy level of a data sample.

2.2.1.5 Expectation-Maximization Meta Algorithm (EM)

EM is another soft clustering method based on ExpectationMaximization Meta algorithm Dempster et al. [8]. Expectation-Maximization is an algorithm for finding maximum probability estimates of parameters in probabilistic models. EM clustering algorithm alternates between performing expectation (E) step, by computing an estimation of likelihood using current model parameters (as if they are known), and a maximization (M) step, by computing the maximum probability estimates of model parameters. The model parameters new estimations contribute to an expectation step of next iteration. Hajji [15] used

Gaussian mixture models to characterize utilization measurements. Model parameters are estimated using Expectation-Maximization (EM) algorithm and anomalies are detected corresponding to network failure events. Animesh and Jung [2] proposed an anomaly detection scheme, called SCAN to address the threats posed by network-based denial of service attacks in high speed networks. The noteworthy features of SCAN include: (a) it rationally samples the incoming network traffic to reduce the amount of audit data being sampled while retaining the intrinsic characteristics of the network traffic itself; (b) it computes the missing elements of the sampled audit data by using an enhanced Expectation-Maximization (EM) algorithm-based clustering algorithm; and (c) it enhances the convergence speed of the clustering process by employing Bloom filters and data summaries.

2.2.2 One -Class Support Vector Machine (OCSVM)

The one-class support vector machine is a very specified sample of a support vector machine which is geared for anomaly detection. The one-class SVM varies from the SVM generic version in that the resulting problem of quadratic optimization includes an allowance for a specific small predefined outliers percentage, making it proper for anomaly detection. These outliers lie between the origin and the optimal separating hyper plane. All the remaining data fall on the opposite side of the optimal separating hyper plane, belonging to a single nominal class, hence the terminology “one-class” SVM. The SVM outputs a score that represents the distance from the data point being tested to the optimal hyper plane. Positive values for the one-class SVM output represent normal behavior (with higher values representing greater normality) and negative values represent abnormal behavior (with lower values representing greater abnormality) [42]. Eskin et al. [11] and Honig et al. [19] used an SVM in addition to their clustering methods for unsupervised learning. The SVM algorithm had to be modified a little to work in unsupervised learning domain. Once it was, it performs better than both of their clustering methods.Shon and Moon [53] suggested a new SVM approach, named Enhanced SVM, which merges (soft-margin SVM method and one-class SVM) in order to provide unsupervised learning and low false alarm capability, similar to that of a supervised SVM approach. Rui et al. [46] proposed a method for network anomaly detection based on one class support vector machine (OCSVM). The method contains two main steps: first is the detector training, the training data set is used to generate the OCSVM detector, which is capable to learn the data nominal profile, and the second step is to detect the anomalies in the performance data with the trained detector.