Network Intrusion Detection by using Supervised and Unsupervised Machine Learning Techniques: A Survey

Abhinav Jain, Sanjay Sharma, Mahendra Singh Sisodia

Abstract—As network attacks have increased in number and severity over the past few years, intrusion detection system (IDS) is increasingly becoming a critical component to secure the network. Due to large volumes of security audit data as well as complex and dynamic properties of intrusion behaviors, optimizing performance of IDS becomes an important open problem that is receiving more and more attention from the research community. Intrusion poses a serious security risk in a network environment. The ever growing new intrusion types pose a serious problem for their detection. The human labeling of the available network audit data instances is usually tedious, time consuming and expensive. This paper compares the performance of Intrusion Detection System (IDS) Classifiers using various feature reduction techniques. To enhance the learning capabilities and reduce the computational intensity of competitive learning comparing the performance of the algorithms is performed respectively, different dimension reduction techniques have been proposed. These include: classifying and clustering Algorithms Naïve Bayes, Simple k mean, Decision tree and J48, Linear Discriminate Analysis, and Independent Component Analysis. Many Intrusion Detection Systems are based on neural networks. However, they are computationally very demanding. In order to mitigate this problem, dimension reduction techniques are applied to a given dataset to extract important features. This paper provides a review on current trends in intrusion detection together with a study on technologies implemented by some researchers in this research area. We try to build as system which create clusters from its input data by labeling clusters as normal or anomalous data instances and finally used these cluster to classify unseen network data instances as either normal or anomalous[1]. Both training and testing was done using different subset of KDD Cup 99[2] data which is very popular and widely used intrusion attack dataset.

Index Terms—Supervised Machine Learning, Unsupervised Machine Learning, Network Intrusion Detection, Network Security.

I. INTRODUCTION

The field of intrusion detection has received increasing attention in recent years. One reason for this is the explosive growth of the Internet and the large number of networked systems that exist in all types of organizations. The increase in the number of networked machines has lead to an increase in unauthorized activity, not only from external attackers, but also from internal attackers, such as disgruntled employees and people abusing their privileges for personal gain. Security is a big issue for all networks in today’s enterprise environment. Hackers and intruders have made many successful attempts to bring down high-profile company networks and web services. Many methods have been developed to secure the network infrastructure and communication over the Internet, among them the use of firewalls, encryption, and virtual private networks. Intrusion detection is a relatively new addition to such techniques. Intrusion detection methods started appearing in the last few years. Using intrusion detection methods, you can collect and use information from known types of attacks and find out if someone is trying to attack your network or particular hosts. The information collected this way can be used to harden your network security, as well as for legal purposes. Both commercial and open source products are now available for this purpose. Many vulnerability assessment tools are also available in the market that can be used to assess different types of security holes present in your network.

Along with the benefits, the Internet also created numerous ways to compromise the stability and security of the systems connected to it. Although static defense are describe the attack. Network intrusion detection system applied to detect and monitor network packets. The network intrusion detection system decides to load and operate detection rules according to a current load. The network intrusion detection system includes a network connection unit, a storage unit, and a processing unit. The processing unit operates an alert correlation program, a plurality of detection rules, and a plurality of operation policies according to the received network packets. The alert correlation program applied to detect whether contents of the network packets conform to the detection rules, assign a resource consumption level to each detection rule, and categorize the detection rules to the operation policies according to the resource consumption levels.
A loading level of the processing unit is decided according to a device load and an access load. The operation policies and the alert correlation program that the processing unit operates are decided according to the loading-level. Available in the market that can be used to assess different types of security holes present in your network.

II. NETWORK INTRUSION DETECTION

Intrusion Detection is a problem of identifying unauthorized users in a computer system. It is also defined as the problem of protecting computer network systems from being compromised. The first published renowned literature on computer network security is [3] where Denning discussed various security concerns, presented a definition of Intrusion Detection and discussed different types of Intrusion Detection.

An intrusion detection system is software and/or hardware designed to detect unauthorized attempts at accessing, manipulating, and/or disabling of computer system, mainly through a network, such as the internet. One of the main challenges in the security management of large-scale high-speed networks is the detection of anomalies in network traffic.

A secure network must provide the following:
- Confidentiality: Data that are being transferred through the network should be accessible only to those that have been properly authorized.
- Integrity: Data should maintain their integrity from the moment they are transmitted to the moment they are actually received. No corruption or data loss is accepted either from random events or malicious activity.
- Availability: The network should be resilient to Denial of Service attacks.

III. BACKGROUND & RELATED RESEARCH

Statistical techniques have been widely used particularly for anomaly detection, and still form parts of many hybrid intrusion detection system based on machine learning techniques. Pattern matching, particularly string matching, and has also been successfully applied to this domain, with proposed algorithms such as ExB [6], E2xB [7].

IV. RULE BASED SYSTEMS

One of the most common forms of Rule Based Systems (RBSs) that have been applied to intrusion detection is expert systems [8].

A. Event correlation tools

There is a range of event correlation tools created with rule based systems, all of which operate similarly. The different tools have been somewhat specialized for different environments, allowing different types of rules. One tool that has been around for approximately two decades is the Production-Based Expert System Toolset (P-BEST), which has been integrated into several IDSs with a focus on handling SYN flooding and buffer overruns [9].

Lindqvist and Porras [9] briefly describe four IDSs that P-BEST have been used in: MIDAS, IDES, NIDES and EMERALD eXpert, and in [10], a fifth, eXpert-BSM; all systems being suitable for real-time misuse detection. The first three systems are host based, whilst the latter two have achieved support for distributed networks. Although eXpert-BSM was developed to analyze Sun Solaris audit trials on a host, it can be distributed by employing an alert collection application referred to as an eftunnel, which will produce a single event stream. Lindqvist and Porras [9] highlight some drawbacks of P-BEST, such as being poor at dealing with uncertainty and missing data due to being strictly forward chaining.

B. Fuzzy logic and rule induction

Fuzzy logic [11] is one approach to obtaining more flexible rules compared with traditional expert systems (which operate with crisp threshold boundaries to classify network traffic or user behavior). Fuzzy rules allow the IDS to quantify some degree of belonging to either intrusion or normal. Hence, this information may be incorporated in the alert reports, instead of only reporting intrusion alerts for traffic/behavior that exceed a crisp threshold. Bridges and Vaughn [12] contend that the very nature of intrusion detection is fuzzy; not always being clear whether events are a misuse or not. Furthermore, they also promote fuzzy logic because events may have membership in several categories, instead of just one or none.

C. Anomaly detection

Dickerson et al. [13] propose using a RBS for network based anomaly detection by creating a set of general fuzzy rules to deal with common intrusion scenarios. Their IDS architecture consists of three modules: first, network data is collected by a sniffer module, which is then organized and preprocessed by a second module to give fuzzy inputs to a fuzzy detector module. The second module includes data mining and feature selection to process the data before being input to the fuzzy detector. This was found to be very successful as the amount of input data is reduced and anomalies could be more easily detected.

D. Hybrid systems

Hybridizations that include RBSs appear in different forms in IDSs, which may be described as follows:
- Algorithmic: Techniques/algorithms are hybridized to perform a single task together. For example, using genetic algorithms to evolve rules for a RBS [15].
- Cooperative: Different techniques are employed to perform different, independent, tasks, which then are combined in some manner to form a holistic system. For example, one technique for misuse detection and another for anomaly detection [16].
- Hierarchical: There is a hierarchy in the architecture of the IDS, which includes different techniques performing different tasks at each level. For example, a RBS may be utilized at the top-level, correlating alerts from several detectors at a lower level [17].
V. INSTANCE BASED LEARNING

Many researchers employ Instance Based Learning (IBL) techniques in intrusion detection and event correlation/fault management as a means of obtaining a more flexible system compared with most expert systems, particularly for dynamic networks. The general drawbacks of expert systems highlighted in the literature [18] include:

- Knowledge engineering is time consuming and difficult (extracting expert knowledge of intrusions and coding this in rules).
- Difficult to manage and update the rule base.
- Many specific rules, which generally (unless fuzzy) cannot detect slight variations of know attacks.

VI. BAYESIAN REASONING

Bayesian reasoning is considered here a general phrase for a range of techniques that exploit Bayes theorem to deal with uncertainty. In short, Mitchell [19] provides the following definition:

"Bayesian reasoning provides a probabilistic approach to inference. It is based on the assumption that the quantities of interest are governed by probability distributions and that optimal decisions can be made by reasoning about these probabilities together with observed data."

A. Bayesian networks

In recent years, Bayesian networks have been utilized in the decision process of hybrid systems [20]. Bayesian networks offer a more sophisticated way of dealing with this compared with a RBS. Kruegel et al. [20] argue that most hybrid systems obtain high false alarm rates due to simplistic approaches to combining the outputs of the techniques in the decision phase. They propose a hybrid host based anomaly detection system consisting of four detection techniques: analyzing string length, character distribution, and structure, and identifying learned tokens, in which a Bayesian network is employed to decide the final output classification. The system was validated on the DARPA99 data set [21], compared with a simple threshold based approach. Both approaches (Bayesian and threshold) were given the same outputs from the detection techniques. With 100% true positives, the threshold based approach lead to twice as many false positives as the Bayesian network.

B. Naïve Bayes

Naïve Bayes (NB) is a simplified version of Bayesian networks, which offer machine learning capabilities. According to Mitchell [19], there are two particular drawbacks of Bayesian networks, namely the requirement of a priori knowledge about the problem to determine probabilities, and that the method is computationally expensive. For the former, it is possible to extract probabilities from training data, if available which is achieved with NB. However, NB does assume that all the features in the data are independent of each other [19], which is the reason for applying Bayesian networks to database intrusion detection instead of NB.

Nevertheless, NB (utilized as a classifier) has been successfully applied to network based intrusion detection by several researchers.

Ben Amor et al. [23] conducted an empirical investigation on the KDD Cup ’99 data set, comparing the performance of NB and a Decision Tree (DT). The DT obtains a higher accuracy (92.28% compared with 91.47%), but NB obtains better detection rates on the three minor classes, namely Probing, U2R and R2L intrusions. Most significantly, the DT detects merely 0.52% R2L intrusions whilst NB detects 7.11%.

Similar observations are made by Panda and Patra [24], as they compare NB with an ANN. ANNs and DTs are biased towards the major class(es) [25], and, therefore, are prone to perform worse on the minor class(es). Therefore, this can be seen as a benefit of the NB, provided that the FPR does not become too high.

VII. DECISION TREES

Decision trees (DTs) are popular in misuse detection systems, as they yield good performance and offers some benefits over other machine learning techniques. For example, they learn quickly compared with Artificial Neural Networks (ANNs), and DTs are not black boxes.

A. Classifier performance

DTs have been successfully applied to intrusion detection both as a stand alone misuse detector [26] [27] or as a part of hybrid systems [27], [28], [29]. A good example of the success of DTs is an application of a C5.0 DT by Pfahringer [29], which won the KDD Cup ’99 competition (although with bagging and boosting).

Sabhnani and Serpen [30] have examined the performance of several machine learning techniques on the KDD Cup ’99 data set, including a C4.5 DT. The DT obtained good accuracy, but does not perform as well as other techniques on some classes of intrusion, particularly U2R and R2L attacks, both of which are minor classes and include a large proportion of new attack types. An ANN and k-means clustering obtained higher detection rates on these classes, which are two techniques that are better able to generalize from learned data to new, unseen, data. Similar observations have been made by Gharibian and Ghorbani [31]. Furthermore, Gharibian and Ghorbani found that DTs and Random Forests (ensemble of DTs) are very sensitive to the data selected for training, i.e., the performance varied significantly on different folds (subsets) of the data.

B. Hybrids and classifier combination

Several researchers have benchmarked a range of machine learning algorithms, observing that different techniques appear better at detecting different classes of intrusion [32]. Nonetheless, creating classifier ensembles of different techniques has been shown to outperform the individual classifiers [32].
Although some researchers experience the instability of DTs as a drawback, that their performance is sensitive to the training data [31], others exploit this as a beneficial trait to construct successful ensembles of DTs (classifier combination) [34]. One commonly used ensemble approach; Random Forest (RF) [35] was first applied to intrusion detection by Zhang and Zulkernine [36] to perform network based misuse and anomaly detection. Their system makes use of a hierarchical hybridization, in which the misuse detection module operates at the first level, employing a RF to classify attacks in real time. Anomaly detection is then employed at a second level, utilizing RF to perform outlier detection on data that is not classified as intrusive by the misuse detection module. On a small subset of the KDD Cup ’99 data set, the hybrid systems obtain a 94.7% TPR and 2% FPR.

VIII. ARTIFICIAL NEURAL NETWORKS

The term Artificial Neural Network (ANN) encompasses a range of models, including Multi Layer Perceptrons (MLPs) and Self Organizing Maps (SOMs) [37], which are the main models applied to intrusion detection. The majority of the misuse detection applications of ANNs are implemented as feedforward MLPs [40]. Most of the misuse detection applications are network based; whilst host based applications are typically implemented as anomaly detection systems.

IX. SUPPORT VECTOR MACHINES

Similar to the MLP, Support Vector Machines (SVMs) [39] are supervised learning algorithms, which have been applied increasingly to misuse detection in the last decade.

One of the primary benefits of SVMs is that they learn very effectively from high dimensional data [39]. Furthermore, they are trained very quickly compared with MLPs. For example, Mukkamala et al. [40] conducted a comparative study of feed forward MLPs and SVMs for misuse detection. Almost identical detection rates were obtained, and the SVM was trained in 17.77 seconds compared with 18 minutes for the MLP [40].

A. Misuse detection

Most SVM algorithms are binary classifiers, which is sufficient when only distinguishing between normal and intrusive data, as in [40]. Although more recent SVM algorithms have been proposed that directly support multi-class learning, e.g., [41], a common approach is to combine several two-class SVMs [42].

Peddabachigari [32] conducted an empirical investigation of SVMs and DTs, in which they analyzed their performance as stand alone detectors and as hybrids. Two hybrid models were examined, a hierarchical model (DT-SVM), with the DT as the first layer to produce node information for the SVM in the second layer, and an ensemble model comprising the standalone techniques and the hierarchical hybrid. For the ensemble approach, each technique is given a weight according to detection rate of each particular attack type during training. Thereafter, when the system is tested, only the technique with the largest weight for the respective attack prediction is chosen to output the classification.

The approaches were tested on the KDD Cup ’99 data set. Their results indicate that the ensemble performs better on two attack classes, Probing and R2L, and equally as good as the other techniques on the other attack types. The hybrid DT-SVM performs better or equally as good as the SVM alone.

B. Anomaly detection

Kim and Cha [43] and Seo and Cha [44] applied SVMs to host based anomaly detection of masquerades. Both studies analyze sequences of UNIX commands executed by users on a host. Kim and Cha applied a SVM with a Radial Basis Function (RBF) kernel, analyzing commands over a sliding window. They adopt the data set used in [45], which gave a detection rate of 80.1%. This was over 10% higher than other techniques applied to this data [44], however, with the highest FPR (9.7%). Seo and Chan examine two different kernels, K-gram and String kernel, which yield higher detection rates; 89.61% and 97.40%, respectively. The drawback is the same as with the RBF kernel employed by Kim and Cha, that the FPR is even higher; above 20% for the String kernel. Seo and Chan also examine a hybrid of the two kernel methods, which gave nearly identical results as Kim and Cha [43] with a RBF kernel.

X. ARTIFICIAL IMMUNE SYSTEMS

Artificial Immune Systems (AISs) have been extensively researched in the last decade, mainly for anomaly detection. Much research has been conducted on using negative selection, as that model lends itself conveniently to anomaly detection. However, within a decade of the proposition of negative selection, several researchers came to the conclusion that the model has problems with scalability, limiting its application to real problems. Consequently, some researchers considered alternative models, whilst others have, in recent years, proposed enhancements to negative selection to address scalability. For a complimentary review of AISs applied to intrusion detection, refer to Kim et al. [42], and reviews by Dasgupta et al. [46] and Timmis [47] for a general treatment of AIS.

XI. MOBILE AGENTS

Mobile Agents (MAs) are distributed by nature. They are typically written in a scripting language and roam around a system to perform designated tasks. Several agents may be present in the same system but perform different tasks, which are generally correlated by a higher level monitor.

XII. HIDDEN MARKOV MODELS

The Hidden Markov Model (HMM) is the only machine learning technique that explicitly learns state based classification (sequential modelling), and, thus, is not limited to conducting stateless intrusion detection. This enables HMMs to perform more comprehensive intrusion detection and detect multi-stage intrusions [48] [49]. Despite this advantage, HMMs have not been as extensively applied to intrusion detection as the other machine learning techniques reviewed here.
XIII. CLUSTERING

One of the main benefits of clustering is unsupervised learning. Labelling of data is not necessary and natural patterns in the data are extracted. Most clustering approaches are unsupervised and are commonly applied to anomaly detection. It is possible, however, to perform supervised clustering if labeled data is available. This generally obtains better results on benchmark data.

XIV. POPULATION BASED SEARCH AND OPTIMIZATION TECHNIQUES

The techniques referred to here are methods typically used to solve search and optimization problems, such as Genetic Algorithms (GAs) [50] [51], Genetic Programming (GP) [52], Particle Swarm Optimization (PSO) [53] and Ant Colony Optimization (ACO) [54]. One of the strengths of these techniques is their parallel nature, and that their application is very diverse, provided that the problem can be quantified into some form of fitness measure.

A. Background and problem representation

GAs and PSO are commonly associated with the optimization of continuous numerical functions, and ACO with combinatorial optimization. Some of the benefits of adopting such techniques are flexibility in retraining, online/continuous learning and the potential for parallelism in the algorithms, which can be exploited both in the training and detection process. However, the challenge is to represent the intrusion detection problem in a form that can be processed and evaluated by these algorithms. For example, for evolving the weights of ANNs, each weight becomes a gene value of a chromosome or individual (solution) in a population that is evolved (optimized) by a GA. A set of weights can be applied to an ANN and evaluated on a training or estimation data set, from which the mean of squared errors (MSE) is typically used as a fitness measure. The lower this value, the fitter the solution, which is what drives the evolutionary process.

B. Rule induction

It is over a decade ago using GAs and DTs to generate rules for an expert system. Although they did not validate their approach empirically, many researchers have since proposed several successful rule induction approaches. For example, Banković [55] use a GA to evolve rules for detecting network based intrusions, demonstrating success on Probing and DoS intrusions from the KDD Cup ’99 data set. Lu and Traore [56] adopt GP to evolve existing rules to be able to detect novel intrusions. They performed experiments with the DARPA99 data set, using data from the first day (10,000 instances) for evolving the rules, then testing with data from the second day. The test data consisted of 10,000 instances, containing two new types of intrusion, which they were able to detect.

C. Clustering and stand alone detection applications

In an early study applying GAs to intrusion detection, Balajinath and Raghavan [57] emphasize the benefit of being able to continuously learn user behavior, to keep track of user drift. Similarly to Balajinath and Raghavan, Neri [57] adopts a distributed GA, REGAL [58], to determine patterns of normal network behavior.

D. Evolutionary neural networks

Michailidis et al. [59] [60][61] employ PSO to optimize the weights of MLPs. The principle is similar to that of using GAs to evolve the weights. Michailidis evaluate the approach on a small (under sampled) subset of the 10% KDD Cup ’99 data set. Their approach detects more Probing and U2R attacks, but at a higher FPR (approximately 3%). However, it is unclear whether this can be attributed to the use of PSO to train the MLPs, or due to the under sampling.

XV. CONCLUSION

Rule Based Systems (RBSs) are commonly used in commercial Intrusion Detection Systems (IDSs), and are better established than the other Artificial Intelligence (AI) techniques reviewed here. RBSs are well suited for event correlation to perform misuse detection. However, other techniques are better suited for anomaly detection, such as statistical methods and clustering.

The ability to facilitate anomaly detection is one of the benefits that have motivated much research on machine learning for intrusion detection. In the last decade, an increasing amount of research on machine learning for misuse detection can also be observed in this review. The application of techniques such as Artificial Neural Networks (ANNs) to misuse detection offers some desirable flexibility in the detection process compared with conventional RBSs, i.e., variations of learned attacks can be detected. The inflexibility of RBSs, due to operating with ‘crisp’ rules, has been considered one of their main drawbacks. However, this observation is no longer entirely accurate, since researchers have proposed several applications of fuzzy RBSs, which have also been shown to be capable of performing anomaly detection. A general benefit of machine learning is that knowledge engineering/extraction/acquisition is avoidable. Furthermore, for techniques that are not ‘black boxes’, such as Decision Trees (DTs) and rule mining/induction, knowledge of novel intrusions may be extracted. These techniques can be employed as detectors in IDS, or the extracted rules may be incorporated in a RBS. Rule induction approaches are typically used for misuse detection, mining rules of intrusions. Fuzzy association rule mining has been found to successfully determine patterns for anomaly detection. Optimization techniques such as Genetic Algorithms (GAs), Genetic Programming (GP) and Particle Swarm Optimization (PSO) have also been applied successfully to rule induction.
REFERENCES


Abhinav Jain is pursuing MTech(CSE) from Oriental Institute of Science & Technology, Bhopal. He had done BE(CSE) from MIT, Mandsaur. His research interest include Data Mining, Machine Learning, Network Security and Algorithm Optimization. e-mail: jainabhinav2008@gmail.com

Mr. Sanjay Kumar Sharma received M. Tech in Computer Science & Engineering and currently working as Assistant Professor in Oriental Institute of Science & Technology, Bhopal (M.P.) India. He has 8 years teaching experience and he has published two (International Journal), two (International Conference) and three (National Conference). His research area in Ad hoc Network, Network Security. Email id: sanjaysharmaemail@gmail.com

Mr. Mahendra Singh Sisodia received M. Tech in Computer Science & Engineering and currently working as Assistant Professor in Oriental Institute of Science & Technology, Bhopal (M.P.) India. He has 6 years teaching experience and he has published two (International Journal), five (International Conference) and two (National Conference). His research area in Network Security, Data Mining, Web Mining, Algorithm Optimization, Machine Learning etc. Email id: mahendrasi14783@gmail.com