EVALUATION OF INTRUSION DETECTION TECHNIQUES AND ALGORITHMS IN TERMS OF PERFORMANCE AND EFFICIENCY THROUGH DATA MINING

PRABHDEEP KAUR & SHEVETA VASHISHT
Department of Computer Science and Engineering, Lovely Professional University, Phagwara, Punjab, India

ABSTRACT

Data mining has been one of the approaches used as a foundation then there is a erection of intrusion detection systems (IDS). Various algorithms and techniques have been developed to make out different types of vulnerabilities. These techniques help to recode the performance of the detection of vulnerabilities, calculates the overheads, and to make Dataset efficient without redundancy. However there is no heuristic to corroborate the accuracy of their results. In this paper we study the techniques and algorithms of the detection of known and unknown attacks through data mining. The methods, metrics and datasets are used to find the known and unknown attacks through data mining. The various techniques like Scoring Techniques, Attack generalization technique, Language model techniques, Novel hybrid technique, C4.5 decision algorithm, Nearest neighbour, Rough Set Theory(RST), Support Vector Machine(SVM), Statically Techniques, Ensemble boosted decision tree, apriori algorithm, Rule based techniques Association and Sequence techniques, are used to detect the vulnerabilities through Data mining. In this paper, we have described the outline of all the algorithms and techniques to identify their strengths and limitations.

KEYWORDS: KDD’99 Dataset, Attacks, IDS, Data Mining, Learning Techniques, Decision Techniques

INTRODUCTION

With the rapid development of electronics and information technology, computer network plays a significant role in our daily lives. However, more and more people use it as a tool for crime, and it has brought great losses to many companies and countries. [1] Data mining and machine learning technology has been extensively applied in network intrusion detection and prevention systems by discovering user behaviour patterns from the network traffic data. Intrusion Detection System (IDS) is necessary because traditional firewalls cannot provide the complete security against the intrusion. [2] Intrusion Detection (ID) is an active and important research area of network security. The functions for intrusion detection system.

- Monitoring and analyzing both consumer and structure activities.
- Analyzing structure configurations and vulnerabilities.
- Assessing structure and file reliability.
- Ability to be familiar with patterns typical of attacks.
- Analysis of uncharacteristic activity patterns.
- Tracking user strategy violations.

IDS are one of the key technologies to guarantee the systems security. IDS make a real time response to intrusion actions and intrusion processes. The goal of Intrusion Detection is to identify all the proper attacks and negatively identify all the non-attacks. The datasets (KDD’99CUP, NSL-KDD, and TCP Dump) are used. And the various techniques for
detection of vulnerabilities that improve the performance of the detection of known and unknown vulnerabilities, and use a dataset which is efficient means without redundancy.

**INTRUSION DETECTION TECHNIQUES**

No particular algorithm could detect all attack categories with a high possibility of detection and a low false alarm rate. It reinforce our conviction that different algorithms should be used to deal with different types of network attacks.

![Classification of Techniques for Detection of Malicious Activities](image)

**Figure 1: Classification of Techniques for Detection of Malicious Activities**

These techniques helps to detect the malicious activities from dataset efficiently. For choosing one of best technique is the objective of the user. And having one of best technique or algorithm for detecting intrusion is main task for a successful security.

**RELATED WORK**

This section discusses the related Works on IDS, Existing Pre-processing and Classifiers techniques applied on the IDS data. There are many research papers published regarding the pre-processing as well as classifiers in order to detect the intrusions in the dataset. The majority of them suffering with false positive type of errors while classifying the attacks and redundancy in the datasets. The following are the some of the related works suffering with low accuracy and false positive errors.

Mahbod Tavallaee *et.al* (2009) Anomaly detection has attracted the attention of many researchers to overcome the weakness of signature based IDSs in detecting novel attacks, and KDDCUP’99 is the mostly widely used dataset for the evolution of these systems. [3] Having conducted a statistical analysis on this data set, they found two important issues which highly affect the performance of evaluated system, and results in a very poor evaluation of anomaly detection approaches. To solve these issues, they proposed a new dataset, NSL-KDD, which consists of selected records of the complete KDD dataset and does not suffer from any of the mentioned shortcomings.

Lei, De-Zhan get *et.al* (2010) Network security is becoming an increasingly important issue, since the rapid Development of internet. Network intrusion detection systems, as the main security defending techniques, is widely used against such malicious attacks.[4] Data mining and machine learning technology has been extensively applied in network intrusion detection and prevention system by discovering user behaviour patterns from the network traffic data. Association rules and sequence rules are the main techniques of data mining for intrusion detection.
Radhika Goel et al. (2012) propose a novel hybrid model for misuse and anomaly detection. C4.5 based binary decision trees are used for misuse and CBA based classifier is used for anomaly detection. Firstly, the C4.5 based decision tree separates the network traffic into normal and attack categories. The normal traffic is sent to anomaly detector and parallel attacks are sent to a decision trees based classifier for labelling with specific attack type. The CBA based anomaly detection is a single level classifier where as the decision trees based misuse detector is a sequential multilevel classifier which labels one attack at a time in a step by step manner. Results show that 99.995% misuse detection rate with an anomaly detection rate of 99.298% is achievable. The overall attack detection rate is 99.911% and false alarm ratio of the integrated model is 3.229%. To overcome the deficiencies in KDD 99 dataset, a new improved dataset is also proposed. The overall accuracy of integrated model trained on new dataset is 97.495% compared to 97.24% of the old dataset. Mrudula Gudade et al. (2010) propose new ensemble boosted decision tree approach for intrusion detection system. Experimental results show better results for detecting intrusions as compared to others existing methods.

Duanyang Zhao et al. (2012) discuss the differences in host and network-based intrusion detection techniques to demonstrate how the two can work together to provide additionally effective intrusion detection and protection. They propose a hybrid IDS, which combines network and host IDS, with anomaly and misuse detection mode, utilizes auditing programs to extract an extensive set of features that describe each network connection or host session, and applies data mining programs to learn rules that accurately capture the behaviour of intrusions and normal.

K.Nageswara Rao et al. (2012) find it essential to evaluate machine learning techniques for web based intrusion detection on the KDD Cup 99 data set. This data set has served well to identify attacks using data mining. Furthermore, selecting the relevant set of attributes for data classification is one of the most significant problems in designing a reliable classifier. Existing C4.5 decision tree technology has a problem in their learning phase to detect automatic relevant attribute selection, while some statistical classification algorithms require the feature subset to be selected in a pre-processing phase. Also, C4.5 algorithm needs strong pre-processing algorithm for numerical attributes in order to improve classifier accuracy in terms of Mean root square error. In this paper, evaluated the influence of attribute pre-selection using statistical techniques on real-world kddcup99 data set. Experimental result shows that accuracy of the C4.5 classifier could be improved with the robust pre-selection approach when compare to traditional feature selection techniques.

COMPARISON OF DIFFERENT ALGORITHMS

Comparison between different algorithms and techniques can be done by the two. By going through the literature analysis of some of the important intrusion detection algorithms, it is concluded that each algorithm has some relative strengths and limitations. A tabular summary is given below in table 1, which summarizes the techniques, advantages and limitations of some of important intrusion detection algorithms and techniques. Some of the highlighted algorithms are given below just for knowledge.

Different Techniques

MACHINE LEARNING TECHNIQUES

SCORING TECHNIQUES
ATTACK GENERALIZATION TECHNIQUES

LANGUAGE MODEL TECHNIQUES

HYBRID MODEL TECHNIQUES

RULE BASED TECHNIQUES

DIFFERENT ALGORITHMS

DECISION TREE ALGORITHMS

ENSEMBLE BOOSTED ALGORITHMS

C4.5 DECISION ALGORITHMS

Table 1: Comparisons of Different Techniques and Algorithms for Performance and Efficiency

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Algorithm or Methods</th>
<th>Dataset</th>
<th>Working</th>
<th>Limitations</th>
<th>Overall Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA (principal component analysis)</td>
<td>Decision tree and, Nearest Neighbourhood</td>
<td>KDD’99 Dataset</td>
<td>1. Apply both algo’s on Dataset. 2. Projected data element into feature space.</td>
<td>Poor Prediction Rate of R2L class, Not Efficient technique for New attacks</td>
<td>Improve learning time Efficiency, Space representation of different dataset.</td>
</tr>
<tr>
<td>Machine Learning Techniques</td>
<td>Classifier Algorithms</td>
<td>KDD’99 Dataset</td>
<td>Classification of different algorithms and make comparisons which is best.</td>
<td>Not more efficient</td>
<td>Performance Improvement and real-time intrusion Detection</td>
</tr>
<tr>
<td>Machine Learning Techniques (RST &amp; SVM)</td>
<td>Classifier Algorithms</td>
<td>KDD’99 Dataset</td>
<td>Packet capture from network, then RST preprocess the data, Feature selected by RST sent to SVM to learn and test.</td>
<td>This method is Very effective to decrease the space density of data.</td>
<td>Compare the results with (PCA) and show RST and SVM schema could reduce the false positive rate and increase the accuracy.</td>
</tr>
<tr>
<td>Machine Learning Techniques (NN and SVM)</td>
<td>ANN and SVM</td>
<td>KDD’99Dataset</td>
<td>Applying this approach to the KDD CUP’99 data set demonstrates that the proposed approach performs high performance, especially to U2R and U2L type attacks.</td>
<td>Redundency in dataset, Not much performance</td>
<td>SVM is superior to NN in false alarm rate and in accuracy for Probe, DoS and U2R and R2L attacks, while NN is better than SVM only in accuracy.</td>
</tr>
<tr>
<td>Stastical Techniques</td>
<td>C4.5 Decision tree algorithm</td>
<td>KDD’99 Dataset</td>
<td>Evaluated the influence of attribute pre-selection using Statistical techniques on real-world kddcup99 data set</td>
<td>Problem in designing the reliable classifier, Problem in Classifier Accuracy</td>
<td>Well outperformed to detect the outliers.</td>
</tr>
<tr>
<td>Rule based approach on genetic programming.</td>
<td>Genetic Algorithm.</td>
<td>DARPHA Dataset</td>
<td>Detecting the novel attacks by using four genetic operators, namely reproduction, mutation, crossover, and dropping condition operators</td>
<td>Genetic operations are random, Result is not good for some runs.</td>
<td>Average, NOT more performance evaluated</td>
</tr>
<tr>
<td>Signature Based Techniques</td>
<td>Packet sniffing method And testing tools are used</td>
<td>Network communication between distributed environment</td>
<td>Detect and combat some common attacks on network systems</td>
<td>More network traffic.</td>
<td>Good.</td>
</tr>
<tr>
<td>Noval hybrid techniques or multilevel techniques used</td>
<td>Classification algorithms ,C4.5</td>
<td>KDD’99 Dataset</td>
<td>Detect known and unknown attacks based on decision tree algorithms, level wise detection</td>
<td>Still redundancy in dataset, not much performance with C4.5 algorithm</td>
<td>Good</td>
</tr>
<tr>
<td>Un-Supervised techniques</td>
<td>Language based models used</td>
<td>DARPHA’99 Dataset</td>
<td>Detect Unknown Attacks,</td>
<td>Not proper measure the unknown attacks</td>
<td>Detection Accuracy 80%, No false rate.</td>
</tr>
</tbody>
</table>
SELECTION OF AN EFFICIENT DATASET

John McHugh was the first to criticize the KDDCUP data sets, pointing out that the synthetic data generated are far from what one collects from real networks.[10] But, due to the lack of publicly available data sets, they are still used. Sabhani and Serpen showed that no pattern classification or machine learning algorithm can be trained successfully with the KDD data set to perform misuse detection for two of the four attack categories included in the KDD data set user-to-root or remote-to-local attack categories. Tavallaee conducted a statistical analysis of the KDDCUP’99 data set and found two important issues which affect the performance of machine learners: the huge number of redundant records in both train and test sets.[11] To correct this, they proposed a new data set – which they called NSL. The main task for the KDD 99 classifier learning contest was to provide a predictive model able to distinguish between legitimate (normal) and illegitimate (called intrusion or attacks) connections in a computer network.

Each attack type falls exactly into one of the following four categories:

- Probing: observation and other penetrating, e.g., port scanning;
- DOS: denial-of-service, e.g., syn flooding;
- U2R: unofficial access to local marvellous user (root) privileges, e.g., various “buffer overflow” attacks;
- R2L: unofficial access from a remote machine, e.g. password guessing.[9]

IMPROVING EFFICIENCY

- By taking competent properties for a dataset upon where the attacks must be noted down.

- The KDD-dataset is run by the most realistic environment for the evaluation of efficiency and redundancy.
• Above environment shows the result without redundancy.
• So in future we can use the KDD dataset for evaluation of detection of attacks.

Here we have so many datasets available but choosing an efficient dataset is one of the prevalent objectives for providing the efficiency. The performance enhancement by using algorithms or techniques also is a objective of by means of an efficient dataset. So the KDD dataset is one of dataset which proves the efficiency by removing the redundancy. The 39 dissimilar attack types present in the 10% datasets are available here.

CONCLUSIONS

In this paper we see most of the techniques that are based on machine learning techniques, Decision based techniques, classification and hybrid techniques for detection of known and unknown vulnerabilities and datasets (KDD’99CUP, NSL-KDD) is used. But still problems are there these all techniques cannot detect the both type of attacks known or unknown. Because some have the accuracy problem false alarm rate problems Some techniques only efficient for identification of known attacks so, it is found that decision models are mostly used for detection of attacks that increase the usage of public datasets for attack detection in future. And we can conclude that by using the KDD’99 Dataset we improve the efficiency and redundancy.

REFERENCES

1. Lei Li, De-Zhang Yang, Fang-Cheng Shen,” A Novel Rule-based Intrusion detection System Using Data Mining “international conference 2010


6. Mrudula Gudadhe, Prakash Prasad” A New Data Mining Based Network Intrusion Detection Model”” International conference on computer and communication technology, [ICCCT’10]

7. Duanyang Zhao,” Analysis of an intrusion detection system” in second international conference on security and management 2012(3):127-13ISSN.


9. Yacine Bouzida, Fr’ed’eric Cuppens, Nora Cuppens-Boulahia and Sylvain Gombault,” Efficient Intrusion Detection Using Principal Component Analysis”, D’epartement RSM GET/ ENST Bretagne 2, rue de la Ch’ataigneraie, CS 17607 35576 Cesson S’evoign’e CEDEX, FRANCE


AUTHOR’S DETAILS

Prabhdeep Kaur, Department of CSE, Lovely Professional University, Phagwara, Punjab, India, 9878282180, (prabhdeep_aulakh@yahoo.com).

Sheveta Vashi, Department Of CSE, Lovely Professional University, Phagwara, Punjab, India, 7508280568, (sheveta.16856@lpu.co.in)