DATA MINING TECHNIQUES FOR EFFICIENT INTRUSION DETECTION SYSTEM: A SURVEY

J.S. SHANTHINI
PG Scholar
Jai Shri Ram Group of Institutions
Avinashipalayam, Tirupur

Dr. S. RAJALAKSHMI Ph.D
Head of the Department, CSE
Jai Shri Ram Group of Institutions
Avinashipalayam, Tirupur

ABSTRACT
In the present world huge amounts of data are stored and transferred from one location to another. The number of cyber attacks have increased and detecting the intrusion in networks become a very leathery work. Various Data Mining techniques are used to analyze data and determine the various kinds of attack. Various hybrid approaches have also been made in order to detect known and unknown attacks more accurately. An efficient detection method must provide proper diagnosis of any obstruction with greater accuracy and low false alarm rate. This paper reviews various data mining techniques for anomaly detection to provide better understanding among the existing techniques that provides accurate results and reduces false detection

Index Terms - Anomaly Detection, Data mining, Intrusion Detection,

I INTRODUCTION
Intrusions are a set of actions that threatens the integrity, availability or confidentiality of a network resource. Intrusion Detection is the process of monitoring and analyzing the events occurring in a computer and/or network system in order to detect signs of security problems. Due to the extensive growth of the internet and increasing availability of tools and tricks for intruding and attacking networks, new challenges arise in order to combat external attacks. Such attacks external to these bodies are deliberate in action against data, software or hardware and can destroy, degrade, disrupt or deny access to a network computer system. Intrusions are such deliberate attacks. IDS is the area, where Data mining is used extensively, this is due to limited scalability, adaptability and validity. In IDS data is collected from various sources like network log data, host data etc. Since the network traffic is large, the analysis of data is too hard. This gives rise to the need of using IDS along with different Data mining techniques for intrusion detection. Lee & Salvatore J. Stolfo, Columbia University were first to apply Data mining techniques in the IDS [1]

The remaining part of the paper is structured in this way.
Section I Introduction, Section II Data Mining in Intrusion Detection , Section III Intrusion Detection System , Section IV Literature Review, Section V Data mining based IDS Techniques are described and finally Section VI brings us to the conclusion.

II DATA MINING IN INTRUSION DETECTION
Data mining refers to a process of non-trivial extraction of implicit, previously unknown, and potentially useful information from data. It is a convenient way of extracting patterns, which represents mining implicitly stored in large data sets and focuses on issues relating to their feasibility, usefulness, effectiveness and scalability. It can be viewed as an essential step in the process of knowledge data discovery.

Fig 1 : Data Mining

Benefits to use data mining approaches in IDS
1. It is very hard to program an IDS using ordinary programming languages that require the exploitation and formalization of knowledge.

2. The adaptive and dynamic nature of machine-learning makes it a suitable solution for this situation.

3. The environment of an IDS and its classification task highly depend on personal preferences. What may seem to be an incident in one environment may be normal in other environments. This way, the ability of computers to learn enables them to know someone’s “personal” (or organizational) preferences, and improve the performance of the IDS, for this particular environment [2].

III INTRUSION DETECTION SYSTEM

The concept of IDS was proposed by Denning (1987), to identify, detect and trace the intrusion [3]. An IDS is a combination of software and hardware which are used for detecting intrusion [4]. It gathers and analyzes the network traffic & detect the malicious patterns and finally alert to the proper authority. The main function of IDS includes:

- Monitoring and analyzing the information gathered from both user and system activities.
- Analyzing configurations of system and evaluating the file integrity and system integrity.
- For static records, it finds out the abnormal pattern.
- To recognize abnormal pattern, it use static records
- and alert to system administrator.

A. Classification of IDS

According to techniques used for intrusion detection based on whether attack’s patterns are known or unknown, IDS classified into two category

1) Misuse detection
2) Anomaly detection

Misuse detection: It is Signature based IDS where detection of intrusion is based on the behaviors of known attacks like antivirus software. Antivirus software compares the data with known code of virus. In Misuse detection, pattern of known malicious activity is stored in the dataset and identify suspicious data by comparing new instances with the stored pattern of attacks.

Anomaly detection: It is different from Misuse detection. Here baseline of normal data in network data in network eg load on network traffic, protocol and packet size etc is defined by system administrator and according to this baseline, Anomaly detector monitors new instances. The new instances are compared with the baseline, if there is any deviation from baseline, data is notified as intrusion. For this reason, it is also called behavior based Intrusion detection system.

B. Advantages and Disadvantages of Anomaly Detection and Misuse Detection

The main disadvantage of misuse detection approaches is that they will detect only the attacks for which they are trained to detect. Novel attacks or unknown attacks or even variants of common attacks often go undetected. The main advantage of anomaly detection approaches is the ability to detect novel attacks or unknown attacks against software systems, variants of known attacks, and deviations of normal usage of programs regardless of whether the source is a privileged internal user or an unauthorized external user. The disadvantage of the anomaly detection approach is that well-known attacks may not be detected, particularly if they fit the established profile of the user. Once detected, it is often difficult to characterize the nature of the attack for forensic purposes. Finally a high false positive rate may result for a narrowly trained detection algorithm, or conversely, a high false negative rate may result for a broadly trained anomaly detection approach.[5]

IV LITERATURE REVIEW

G. V. Nadianmai and M. Hemalatha in their paper —Effective approach toward Intrusion Detection System using data mining techniques, considered four issues namely Classification of Data, High Level of Human Interaction, Lack of Labeled Data, and Effectiveness of Distributed Denial of Service attack and solved them using the proposed algorithms EDADT algorithm, Hybrid IDS model, Semi-Supervised Approach and Varying HOPE RAA algorithm respectively. To solve the problem related to classification of data, an enhanced data adapted decision
A tree algorithm is implemented which effectively classifies the data into normal and attack without any classification. To minimize the workload of a network administrator, a high level of human interaction based on SNORT and anomaly based approaches are being used. This has a Hybrid IDS that automatically classifies the data based on the pre-defined rules within it. The issue related to belling the unlabeled data is solved using Semi-Supervised Approach where with the small amount of labeled data, the large amount of unlabeled data can be labeled. The last problem related to Distributed Denial of Service Attack is addressed by using varying clock drift. This varying clock drift in network based applications makes it difficult for the intruder to access the port that has been used by the legitimate client [6].

Basant Agarwal and Namita Mittal proposed in their paper—Hybrid Approach for Detection of Anomaly Network Traffic using Data Mining Techniques! a hybrid approach that exploits the benefits of both the techniques i.e. entropy based and support vector machine based respectively. Hybrid anomaly detection system learns the behavior of network traffic from the normalized entropy values of different network features. Entropy based techniques have the advantage of better representing the properties of the network traffic and support vector machine is good for classification. The normalized entropies are sent to SVM model for learning the behavior of the network. This trained SVM model can classify the network traffic in attack traffic or legitimate traffic. In entropy based anomaly detection system, firstly normalized entropy of network traffic features is calculated in every 60 seconds. Threshold value is fixed for each feature for identifying the anomalies based on experiments. Then voting system for each feature decides whether there is an attack or not. This method is able to produce good results in case of detecting attack traffic but it also produces high false alarms, because the entropy values can also deviate from the range or towards 0 or 1 in case of legitimate traffic.

Mr. Jaina Patel, Mr. Krunal Panchal proposed in their paper a hybrid model that integrates Anomaly based Intrusion detection technique with Signature based Intrusion detection technique which is divided into two stages. In first stage, the signature based IDS SNORT is used to generate alerts for anomaly data. In second stage, data mining techniques “k-means + CART” is used to cascade k-means clustering and CART (Classification and Regression Trees) for classifying normal and abnormal activities. The hybrid IDS model is evaluated using KDD Cup Dataset. The proposed assemblage is introduced to maximize the effectiveness in identifying attacks and achieve high accuracy rate as well as low false alarm rate.

Mr. Sahu and Mr. Mehtre, Sch. of Comput. & Inf. Sci., Univ. of Hyderabad, in their paper used a new labelled network dataset, called Kyoto 2006+ dataset. In Kyoto 2006+ data set, every instant is labelled as normal (no attack), attack (known attack) and unknown attack. They used Decision Tree (J48) algorithm to classify the network packet that can be used for NIDS. For training and testing they used 134665 network instances. The generated rules works with 97.2% correctness for detecting the connection i.e., no attack, known attack or unknown attack.

The proposed paper by Yogita B. Bhavsar et al. uses SVM method for classification, which can reduce the time required to build model for classification and increase the intrusion detection accuracy when Gaussian RBF kernel is used. The experimental results show that, when data sets are properly processed and proper SVM kernel is selected i.e. Radial Basis Function (RBF), it can overcome the drawback of SVM i.e. extensive time required to build model. They conducted experiment with 10 fold cross validation and Gaussian RBF kernel of SVM, the time required to build model was 77.07 seconds and attack detection accuracy achieved was 94.1857 %. This attack detection accuracy was increased to 98.5749 %, when they changed classification to 10 fold cross validation and re-evaluation using supplied test set with same RBF SVM kernel function.
Praveen P. Naik and Mr. Prashantha S.J proposed a new approach by utilizing data mining techniques such as neuro-fuzzy and radial basis support vector machine (SVM) for helping IDS to attain higher detection rate. The proposed technique has four major steps: primarily, k-means clustering is used to generate different training subsets. Then, based on the obtained training subsets, different neuro-fuzzy models are trained. Subsequently, a vector for SVM classification is formed and in the end, classification using radial SVM is performed to detect intrusion has happened or not. The results of experiments on KDD CUP 1999 dataset is demonstrated. Experimental results shows that our proposed new approach do better than Conditional random fields (CRF) with respect to specificity and detection accuracy.

Srilatha Chebrolu, Ajith Abraham and Johnson P. Thomas have investigated new techniques for intrusion detection and performed data reduction and evaluated their performance on the DARPA benchmark intrusion data. The feature selection method using Markov blanket model and decision tree analysis have been used. Following this, they explored the general Bayesian network (BN) classifier and Classification and Regression Trees (CART) as intrusion detection models and have also demonstrated performance comparisons using different reduced data sets. The proposed ensemble of BN and CART combines the complementary features of the base classifiers. Finally, a hybrid architecture involving ensemble and base classifiers for intrusion detection has been proposed. From the empirical results, it is seen that by using the hybrid model Normal, Probe and DOS could be detected with 100% accuracy and U2R and R2L with 84% and 99.47% accuracies, respectively.

V. DATA MINING BASED IDS TECHNIQUES

A. Classification [1][13]

Classification is the task of taking each and every instances of dataset under consideration and assigning it to a particular class normal and abnormal means known structure is used for new instances. It can be effective for both misuse detection and anomaly detection, but more frequently used for misuse detection. Classification categorized the datasets into predetermined sets. It is less efficient in intrusion detection as compared to clustering. Different classification techniques such as decision tree, naive bayes classifier, K-nearest neighbour classifier, Support vector machine etc are used in IDS.

1) Decision Tree [7]

Decision tree is a recursive and tree like structure for expressing classification rules. It uses divide and conquer method for splitting according to attribute values. Classification of the data proceeds from root node to leaf node, where each node represents the attribute and its value & each leaf node represent class label of data. Tree based classifier have highest performance in case of large dataset. Different decision tree algorithms are described below

ID3 algorithm

It is famous decision tree algorithm developed by Quinlan. ID3 algorithm basically attribute based algorithm that constructs decision tree according to training dataset. The attribute which has highest information gain is used as a root of the tree.

J48 algorithm

It is based on ID3 algorithm and developed by Ross Quinlan. In WEKA, C4.5 decision tree algorithm is known as J48 algorithm. It construct decision tree using information gain, attribute which have highest information gain is selected to make decision. The main disadvantage of this algorithm is that it takes more CPU time and memory in execution. Another different tree based classifier [8]:

AD Tree

Alternating decision tree is used for classification. AD Tree have prediction node as both leaf node and root node.

NB Tree

NB Tree algorithm uses both decision tree and naive bayes classifier. Root node uses decision tree classifier and leaf nodes uses naive bayes classifier.

Random Forest [9]
Random Forest is first introduced by Lepetit et.al. and it is ensemble classification technique which consists of two or more decision trees. In Random Forest, every tree is prepared by randomly select the data from dataset. By using Random Forest improve the accuracy and prediction power because it is less sensitive to outlier data. It can easily deals with high dimensional data.

2) K-Nearest Neighbor [8]
It is one of the simplest classification technique. It calculates the distance between different data points on the input vectors and assigns the unlabeled data point to its nearest neighbor class. K is an important parameter. If k=1, then the object is assigned to the class of its nearest neighbor. When value of K is large, then it takes large time for prediction and influence the accuracy by reduces the effect of noise.

3) Naive Bayes classifier [13]
Naive Bayes classifier is probabilistic classifier. It predicts the class according to membership probability. To derive conditional probability, it analyzes the relation between independent and dependent variable.

Bayes Theorem:

\[ P(H/X) = \frac{P(X|H) \cdot P(H)}{P(X)} \]

Where, X is the data record and H is hypothesis which represents data X and belongs to class C. P(H) is the prior probability, P(H/X) is the posterior probability of H conditioned on X and P(X/H) is the posterior probability of X conditioned on H. Construction of Naive Bayes is easy without any complicated iterative parameter. It may be applied to large number of data points but time complexity increases.

4) Support Vector Machine [8]
Support Vector Machine is supervised learning method used for prediction and classification. It separate data points into two classes +1 and -1 using hyperplane because it is binary classification classifier. +1 represents normal data and -1 for suspicious data. Hyperplane can be expressed as: W. X +b=0 Where W=[w1,w2,......,wn] are weight vector for n attributes A={A1,A2,.........,An}, X={x1,x2,......,xn} are attribute values and b is a scalar. The main 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.

B. Clustering [1]
Since the network data is too huge, labelling of each and every instances or data points in classification is expensive and time consuming. Clustering is the technique of labelling data and assign into groups of similar objects without using known structure of data points. Members of same cluster are similar and instances of different clusters are different from each other. Clustering technique can be classified into four groups: Hierarchical algorithm, Partitioning algorithm, Grid based algorithm and Density based algorithm. Some clustering algorithms are explained here.

1) K-Means Clustering algorithm [12][11]
K-Means clustering algorithm is simplest and widely used clustering technique proposed by James Macqueen. In this algorithm, number of clusters K is specified by user means classifies instances into predefined number of cluster. The first step of K-Means clustering is to choose k instances as a center of clusters. Next assign each instances of dataset to nearest cluster. For instance assignment, measure the distance between centroid and each instances using Euclidean distance and according to minimum distance assign each and every data points into cluster. K –Means algorithm takes less execution time, when it applied on small dataset. When the data point increases to maximum then it takes maximum execution time. It is fast iterative algorithm but it is sensitive to outlier and noise.

K-Medoids is clustering by partitioning algorithm as like as K-means algorithm. The most centrally situated instance in a cluster is considered as centroid in place of taking mean value of the objects in K-Means clustering. This centrally located object is called reference point and medoid. It minimizes the distance between centroid and data spoints means minimize the squared error. KMedoids algorithm performs better than K-Means algorithm when
the number of data points increases to maximum. It is robust in presence of noise and outlier because medoid is less influenced by outliers, but processing is more expensive.

![Usage of Data Mining Techniques in Intrusion System](image1)

**Fig 2: Usage of Data Mining Techniques in Intrusion System**

![Usage of Number of Hybrid Models in intrusion Detection](image2)

**Fig3: Usage of Number of Hybrid Models in intrusion Detection**

**V CONCLUSION**

On the basis of detection rate, accuracy, execution time and false alarm rate, the paper has analyzed different classification and clustering data mining techniques for Intrusion detection. The results also suggest that classification is widely in use for solving intruder-based problems and Support Vector Machine (SVM) remains popular with the researchers.

**REFERENCES**


[12] Iwan Syarif, Adam Pruge Bennett and Gary Wills, “Unsupervised clustering approach for network anomaly detection”, IEE.
