Intrusion Detection with Hidden Markov Model and WEKA Tool

Ashish T. Bhole
Associate Professor
Department of Computer Engineering
SSBT's College of Engineering & Technology,
Bambhori, Jalgaon, Maharashtra, India

Archana I. Patil
Research Scholar
Department of Computer Engineering
SSBT's College of Engineering & Technology,
Bambhori, Jalgaon, Maharashtra, India

ABSTARCT
The internet has become a convenient way for people exchanging information and doing business over the internet. Uptil now, intrusion detection technique has been using either anomaly based or signature based detection technique. The hybrid technique gives advantages of both the techniques. Anomaly detection strategy is to suspect of what is considered an unusual activity for the subject (users, processes, etc.) and carry on further investigation. This approach is particularly effective against novel (i.e. previously unknown) attacks. Signature based detection systems detect previously known attack in a timely and efficient way. The Hybrid technique gives better result than signature based or anomaly based technique alone.

General Terms
Layered Approach, Hidden Markov Model, WEKA Tool.

Keywords
Intrusion detection, Layered Approach, Hidden Markov Model, Decision Trees, Naive Bayes

1. INTRODUCTION
Intrusion detection is defined as the problem of identifying individuals who are using a computer system without authorization (i.e., `crackers') and those who have legitimate access to the system but are abusing their privileges. The identification of attempts to use a computer system according to Heady et al. where an intrusion is defined as "any set of action that attempts to compromise the integrity, confidentiality, or availability of a resource", disregarding the success or failure of those actions [1].

The definition of an intrusion detection system does not include preventing the intrusion from occurring, only detecting it and reporting it to an operator [3].

1.1 Types of Intrusion Detection
There are two types of intrusion detection depending on way their components are distributed.

A centralized intrusion detection system is one where the analysis of the data is performed in a fixed number of locations, independent of how many hosts are being monitored. The location of the data collection components is not considered, only the location of the analysis components such as IDES, IDIOT [5][10].

A distributed intrusion detection system is one where the analysis of the data is performed in a number of locations proportional to the number of hosts that are being monitored. Again, consider the locations and number of the data analysis components, not the data collection components such as DIDS, GrIDS.

Further IDS divided into two types as:

Host Based IDS

Typically monitors system, event, and security logs on Windows NT and syslog in UNIX environments Checks key system files and executables via checksums at regular intervals for unexpected changes Can use powerful regular-expressions to clearly define signatures.

Network Based IDS

Uses network packets as the data source. Typically utilizes a network adapter to analyze all traffic in real-time as it travels across the network [8].

Also Intrusion detection is divided into two categories.

Anomaly detection, where the strategy is to suspect of what is considered an unusual activity for the subject (users, processes, etc.) and carry on further investigation. This approach is particularly effective against novel (i.e. previously unknown) attacks. Its main drawback is the high rate of false positives, because any legitimate but new activity can rise an alert [2].

Signature detection, where the strategy is to look for some special activity (signature) of previously known attacks. Signature based detection a system detects previously known attack in a timely and efficient way. The main issue of this approach is that in order to detect an intrusion this must to be previously detected [7].

2. RELATED WORK
There are various techniques such as data mining, clustering, naive Bayes classifier, support vector machines, genetic algorithms, artificial neural networks, and others have been applied to detect intrusions.

Following are several different approaches for Intrusion Detection.

In 1998 by W. Lee and S. Stolfo, in this paper intrusion detection is through sequence call which takes too much of time to detect attack [12].

In 2001 by L. Portnoy, E. Eskin, and S. Stolfo, clustering methods such as the k-means and the fuzzy c-means have also been applied extensively for intrusion detection. One of the main drawbacks of the clustering technique is that it is based on calculating numeric distance between the observations, and hence, the observations must be numeric the clustering methods consider the features independently and are unable to capture the relationship between different features of a single record, which further degrades attack detection accuracy [10].

In 2003 by Y.-S. Wu, B. Foo, Y. Mei, and S. Bagchi, Support vector machines have also been used for detecting intrusions Support vector machines map real valued input feature vector to a higher dimensional feature space through nonlinear mapping and can provide real-time detection capability, deal

with large dimensionality of data, and can be used for binary-class as well as multiclass classification [11].

In 2004 by N.B. Amor, S. Benferhat, and Z. Elouedi,Naive Bayes classifiers have also been used for intrusion detection. But in this the size of a Bayesian network increases rapidly as the number of features and the type of attacks modeled by a Bayesian network increases [3][6].

In Oct 2008 by Yusufovna, S.F "Integrating Intrusion Detection System and Data Mining" in this article Data mining[5] approaches for intrusion detection include association rules and frequent episodes. These methods can deal with symbolic data and the features can be defined in the form of packet and connection details .However, mining of features is limited to entry level of the packet and requires the number of records to be large and sparsely populated; otherwise, they tend to produce a large number of rules that increase the complexity of the system [5].

In 2009 by Oludele Awodele, Sunday Idowu, Omotola Anjorin, and Vincent J. Joshua this paper presents an Intelligent Intrusion Detection and Prevention System (IIDPS), which monitors a single host system from three different layers; files analyzer, system resource and connection layers. The approach introduced, a multi – layered approach, in which each layer harnesses both aspects of existing approach, signature and anomaly approaches, to achieve a better detection and prevention capabilities [14].

Aartificial neural networks are also used for network intrusion detection. Though the neural networks can work effectively with noisy data, they require large amount of data for training and it is often hard to select the best possible architecture for a neural network. Also genetic algorithm, autonomous and probabilistic agents are used for intrusion detection [11][12]. These methods are generally aimed at developing a distributed intrusion detection system.

3. HIDDEN MARKOV MODEL

A hidden Markov Model (HMM) is a statistical generative model in which the system being modelled is assumed to be a Markov process with unobserved state. An HMM can be considered as the simplest dynamic Bayesian network. An HMM is like a finite state machine in which not only transitions are probabilistic but also output [13].

An HMM is a doubly stochastic process with an underlying stochastic process that is not observable, and can only be observed through another set of stochastic processes that produce the sequence of observed symbols HMM is a useful tool to model sequence information. This model can be thought of as a graph with N nodes called 'state' and edges representing transitions between those states. Each state node contains initial state distribution and observation probabilities at which a given symbol is to be observed. An edge maintains a transition probability with which a state transition from one state to another state is made [9].

Fig. 1 shows the general architecture of an instantiated HMM. Each oval shape represents a random variable that can adopt any of a number of values. The random variable x(t) is the hidden state at time t (with the model from the above diagram(t) ∈ { x1, x2, x3 }). The random variable y(t) is the observation at time t (with y(t) ∈ { y1, y2, y3, y4 }). The arrows in the diagram (often called a trellis diagram) denote conditional dependencies.

Referring to Fig. 1, it is clear that the conditional probability distribution of the hidden variable x(t) at time t, given the values of the hidden variable x at all times, depends only on the value of the hidden variable x(t − 1): the values at time t − 2 and before have no influence. This is called the Markov property. Similarly, the value of the observed variable y(t) only depends on the value of the hidden variable x(t) (both at time t).

In the standard type of hidden Markov model considered here, the state space of the hidden variables is discrete, while the observations themselves can either be discrete (typically generated from a categorical distribution) or continuous (typically from a Gaussian distribution). The parameters of a hidden Markov model are of two types, transition probabilities and emission probabilities (also known as output probabilities). The transition probabilities control the way the hidden state at time t is chosen given the hidden state at time t-1.

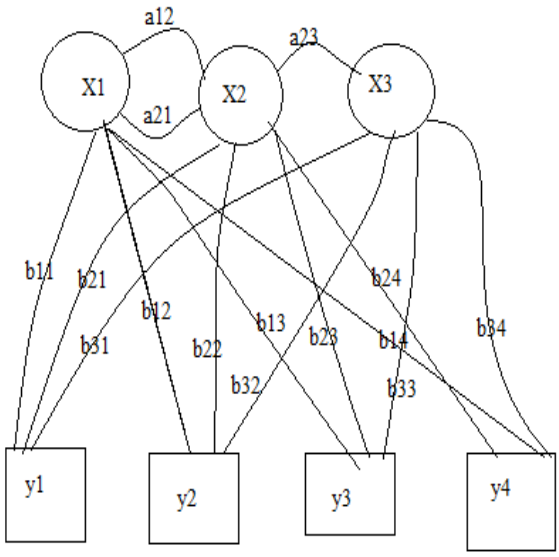

Figure 1: Probabilistic HMM

X - State
y - Possible Observation
a - State Transition Probabilities
b - Output Probabilities

4. DECISION TREE

Decision tree builds classification or regression models in the form of a tree structure. Dataset is a collection of data, usually presented in a tabular form. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed.

A decision tree is composed of three basic elements:

1. A decision node specifying a test attributes.

2. An edge or a branch corresponding to the one of the possible attribute values which means one of the test attribute outcomes.

3. A leaf which is also named an answer node contains the class to which the object belongs.

The final result is a tree with decision nodes and leaf nodes. The topmost decision node in a tree which corresponds to the

best predictor called root node. Decision trees can handle both categorical and numerical data. Most of the decision trees algorithms use a top down strategy; i.e from the root to the leaves.

One of the greatest advantages of decision tree classification algorithm is that: It does not require user to know a lot of background knowledge in the learning process. To build a decision tree, two types of entropy using frequency tables should be calculated as follows:

1. Entropy using the frequency table of one attribute is shown in equation 1.

$$E(S) = \sum_{i=1}^{c} -p_i \log_2 p_i \quad (1)$$

2. Entropy using the frequency table of two attributes is shown in equation 2

$$E(T,X) = \sum_{c \in X} P(c)E(c) \quad (2)$$

Information Gain
The information gain is based on the decrease in entropy after a dataset is split on an attribute. Constructing a decision tree is all about finding attribute that returns the highest information gain (i.e., the most homogeneous branches)
Step 1: Calculate entropy of the target.

Step 2: The dataset is then split on the different attributes. The entropy for each branch is calculated. Then it is added proportionally, to get total entropy for the split. The resulting entropy is subtracted from the entropy before the split. The result is the Information Gain, or decrease in entropy
Step 3: Choose attribute with the largest information gain as the decision node.
Step 4a: A branch with entropy of 0 is a leaf node.
Step 4b: A branch with entropy more than 0 needs further splitting.
Step 5: The ID3 algorithm is run recursively on the non-leaf branches, until all data is classified.

5. NAIVE BAYESIAN

Bayes theorem provides a way of calculating the posterior probability, $P(c|x)$, from $P(c)$, $P(x)$, and $P(x|c)$. Naive Bayes classifier assumes that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors. This assumption is called class conditional independence[4].

$P(c \mid x) = P(x \mid c)P(c)/P(x)$

$P(c \mid X) = P(x1 \mid c) * P(x2 \mid c) *$……$*P(x \mid c) * P(c)$

$P(c|x)$ is the posterior probability of *class* (*target*) given *predictor* (*attribute*).
$P(c)$ is the prior probability of *class.*
$P(x|c)$ is the likelihood which is the probability of *predictor* given *class.*
$P(x)$ is the prior probability of *predictor*.
The probability density function for the normal distribution is defined by two parameters (mean and standard deviation).
Mean deviation is shown in equation 3.

$$\mu = 1/n \sum_{i=1}^{n} xi \quad (3)$$

Standard Deviation is shown in equation 4.

$$\sigma = [1/n\text{-}1 \sum_{i=1}^{n} (xi - \mu)2]0.5 \quad (4)$$

Predictors Contribution is given in equation 5.

$$\text{Log}_2 P(c \mid x) - \text{Log}2P(c) \quad (5)$$

Kononenko's information *gain* as a sum of information contributed by each attribute can offer an explanation on how values of the predictors influence the class probability.

6. RESULTS AND DISCUSSION

Table 1 gives the detection rate to detect the attacks of Hidden Markov Model, Weka with decision Tree, and Weka with Naïve Bayes. Detection rate can be calculated as shown in equation 6.

$$\text{Detection rate} = \frac{\text{Number of attacks detected by Proposed Layered Approach}}{\text{Number of attacks present in Input File}} 100 \quad (6)$$

Table 1: Comparison with different techniques

Name of Technique	Hidden Markov Model	Weka with	
		Decision Tree	Naïve Bayes
Detection Rate in %	95.15	95.05	93.51

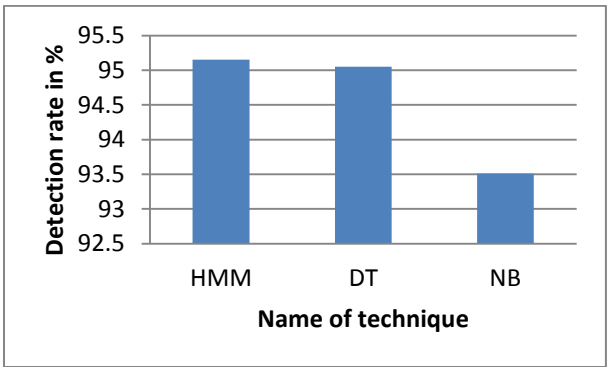

Figure 2: Comparison Detection Rate of HMM, DT and NB Techniques

Hidden Markov Model gives more accurate result while detecting the attacks. It uses anomaly based intrusion detection model. In signature based intrusion detection model, the attack can be detected on the basis of previously known attacks. Whereas in anomaly based intrusion detection, the attack can also be detected which is previously unknown. Using Hidden Markov Model, high attack detection rate can be achieved.

7. CONCLUSION AND FUTURE SCOPE

In this paper, an anomaly detection based IDS using privilege transition flows is analyzed. Hidden Markov Model is a generative model and it can give more accurate results than WEKA tool. Thus impact of attacks can be minimized by using Hidden Markov Model. As the technique uses anomaly based detection technique, it gives better result than signature based technique while detection of attacks. When Detection rate of HMM, WEKA with DT and WEKA with NB are compared, HMM gives better detection rate. In future, use of hybrid technique can be used to detect different types of attacks so that attack detection rate can be increased.

8. REFERENCES

[1] Kapil Kumar Gupta, Baikunth Nath, January –March 2010 Layered Approach Using Conditional Random Fields for Intrusion Detection.

[2] T. Abraham, 2008 IDDM: Intrusion Detection Using Data Mining Techniques.

[3] N.B. Amor, S. Benferhat, and Z. Elouedi, 2004 Naive Bayes vs. Decision Trees in Intrusion Detection Systems.

[4]https://blog.itu.dk/SPVCE2010/files/2010/11/wekatutorial.pdf , CMP: Data Mining and Statistics within the Health Services.

[5] Yusufovna, S.F,Oct 2008 Integrating Intrusion Detection System and Data Mining.

[6] Christopher Kruegel ,Darren Mutz William ,Robertson Fredrik Valeu , Reliable Software Group University of California , Bayesian Event Classification for Intrusion Detection.

[7] SANS Institute InfoSec Reading Room , Understanding Intrusion Detection Systems.

[8] Ozalp Babaoglu , 2006 IDS:Intrusion Detection Systems.

[9] Y. Du, H. Wang, and Y. Pang , 2004 A Hidden Markov Models-Based Anomaly Intrusion Detection Method.

[10] L. Portnoy, E. Eskin, and S. Stolfo, 2001 Intrusion Detection with Unlabeled Data Using Clustering.

[11] Y.-S. Wu, B. Foo, Y. Mei, and S. Bagchi, 2003 Collaborative Intrusion Detection System (CIDS): A Framework for Accurate and Efficient IDS .

[12] W. Lee and S. Stolfo, "Data Mining Approaches for Intrusion Detection," Proc. Seventh USENIX Security Symp. (Security '98),pp. 79-94, 1998.

[13]http://www.cse.sc.edu/research/isl/agentIDS.shtml,2010 Probabilistic Agent Based Intrusion Detection.

[14] Awodele, Oludele; Idowu, Sunday; Anjorin, Omotola; Joshua, Vincent J., "A Multi-Layered Approach to the Design of Intelligent Intrusion Detection and Prevention System (IIDPS)", Academic journal article from *Issues in Informing Science & Information Technology*, Vol. 6.2009