Integrated Bayes Network and Hidden Markov Model for Host Based IDS

NagarajuDevarakonda
Dept of Computer Science & Engineering
Acharya Nagarjuna University,
Nagarjuna Nagar, Guntur -522510, India

Srinivasulu Pamidi
Dept of Computer Science & Engineering
V R SiddhrathaEngg. College,
Kanuru, Vijayawada-520007, India

V Valli Kumari
Dept of Computer Science & SystemsEngineering
Andhra University
Visakhapatnam-530003, India

A Govardhan
Dept of Computer Science & Engineering
Jawaharlal Nehru Technological University, Hyderabad
Hyderabad-500085, India

ABSTRACT

Today Internet is more popular for many users and business applications such as banking, social networks, education, entertainment, scientific research, and recently cloud computing. The number of services provided by the internet service providers through Internet is rapidly increasing. For many applications security has become a serious issue for anyone connected to the Internet. Security should be provided by the ISPs to the Internet users in the form confidentiality, integrity, and authentication. These can be provided through IDS. In our paper we have proposed a simple, easy and efficient approach for building IDS using integrated model of Bayes Net with Hidden Markov Model. The first phase of the model is to build the Bayesian network using the dataset. Once the network is built the conditional probability or joint probability for each node can be determined. The Bayes network has been used as state transition diagram for HMM. The HMM parameters can be estimated using the Bayesian Network. We have used a standard kddcup99 dataset for building the model. This model can be able to differentiate the intruders from normal users with low false positive rate and high true positive rate. The model works for even high dimensional data streams with high performance detection rate and robust to noise.

Keywords

IDS, Bayes Network, HMM, training, and conditional probability tables

1. INTRODUCTION

To provide the highest level of secrecy for users of their information technology systems and to assure their intellectual freedom is protected.

The access to information at various levels of organization is dependent on the policy of the organization. In general users at lower level can’t have full access to the information, whereas at the top level has full access.

An Organization may have many types of information such as staff records, financial records, personnel records, and business records. Individual information includes academic, Scientific,engineering, research, personal and business correspondence, and other records created and managed by employees.

Network intrusion detection systems (NIDS)[11-14] is becoming a standard network security tool. In this section we give the basic infrastructure required to build IDS System [6]. Some of the IDS components on the network have been explained which are shown in the figure 1[7]. *IDS Sensors* gather information of the wire in the Internet. Now the way they gather information is by using a tap connected to the wire. The *tap* is like a splitter, it allows the traffic to flow across the wire normally. But at the same time a copy of everything which is going across the wire sent down to the sensor. These sensors send the information down to the *IDS collector*.

The collector does the analysis to determine whether data is just normal or is an attack is taking place. The other device, IDS Manager that would be a Personal Computer (PC) or a laptop that has client software on it that allows the use of the computer to manage the information of various IDS devices. Let us look at the various types of Intrusion Detection Systems. An active IDS is also known as Intrusion Detection and Prevention System (IDPS) [4].

A passive IDS is a system that is configured to monitor and analyze network traffic and alert the user to possible intrusions. A passiveIDS will not be capable of performing any protective or corrective functions on its own. NIDS[8-9] usually consists of a network appliance with a Network Interface Card operating in immoral mode and a separate interface. The IDS is placed along a boundary and monitors all traffic on that device.

A Host based IDS (HIDS) monitor the operating system and write data to log files and trigger alarms when it detects intrusions.
A HIDS can only monitor the individual workstations on which it is installed and it will not monitor the whole network. HIDS systems are used to monitor intrusion attempts on critical servers.

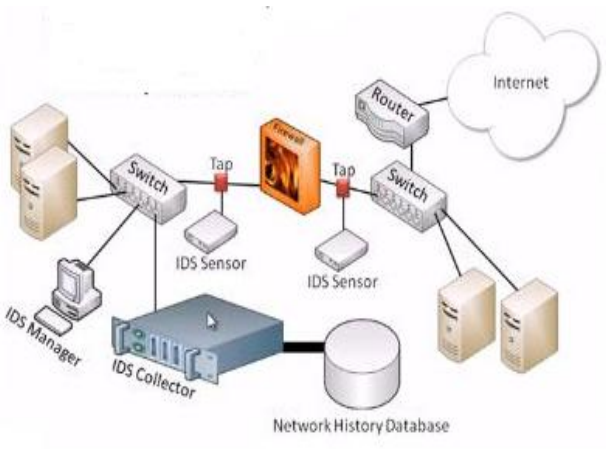

Fig 1. Network Environment Infrastructure for Intrusion Detection System

A Signature-based IDS references a database of previous attack signatures. The meaning of signature is recorded evidence of an intrusion or attack. Every intrusion leaves a mark behind. These path are called signatures and can be used to identify and prevent the same attacks in the future. Based on these signatures Knowledge-based (Signature-based) IDS identify intrusion attempts.

The disadvantages of Signature-based IDS [10] are signature database must be continually updated and maintained, and Signature-based IDS may fail to identify unique attacks. A Behavior-based (Anomaly-based) IDS references a baseline or learned pattern of normal system activity to identify active intrusion attempts. Abnormalities from this model or pattern cause an alarm to be triggered. High false alarms are often associated with Behavior-based IDS.

2. DATASET DESCRIPTION

We conducted experiments on KDDCup99 dataset [3], [13]. The 1998 DARPA Intrusion Detection Evaluation Program was prepared and managed by MIT Lincoln Labs. The objective was to survey and evaluate research in intrusion detection. A standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment, was provided.

A connection is a sequence of TCP packets starting and ending at some well defined times, between which data flows to and from a source IP address to a target IP address under some well defined protocol. Each connection is labeled as either normal, or as an attack, with exactly one specific attack type. Each connection record consists of about 100 bytes. Attacks fall into four main categories:

- DOS: denial-of-service, e.g. syn flood;
- R2L: unauthorized access from a remote machine, e.g. guessing password;
- U2R: unauthorized access to local superuser (root) privileges, e.g., various ``buffer overflow" attacks;
- Probing: surveillance and other probing, e.g., port scanning.

The dataset contain a total of 41 attributes in which there are 6 symbolic (discrete) and remaining are continuous type. To use Bayes network the attributes must be discrete type and data should be complete (no missing values). So the data has been preprocessed which is explained as follows:

The rows have been removed which contains missing values for any attribute. We have chosen four attributes from 41 attributes of kddcup99 namely protocol_type, src_bytes, dst_bytes, and count. The continuous attributes src_bytes, dst_bytes, and count have been discretized. In the following way the variables have been discretized:

- src_bytes : ‘L’, ’M’, and ‘H’,
- dst_bytes : ‘l’, ‘m’, ‘h’,
- count: ‘low’, ‘mid’, ‘high’.

The dataset has 22 atack types such as smurf, Neptune, satan,etc. The 22 attacks have been replaced by four types namely DoS, R2L, U2R and Probe which are explained above.

3. BAYES NETWORK

Bayes Nets or Bayesian networks [15-17] are graphical representation for probabilistic relationships among a set of random variables. Given a finite set $X=\{X_1,...,X_n\}$ of discrete random variables where each variable X_i may take values from a finite set, denoted by $Val(X_i)$.

A Bayesian network is an annotated directed acyclic graph (DAG) G that encodes a joint probability distribution over X. The nodes of the graph correspond to the random variables $X_1,...,X_n$. The links of the graph correspond to the direct influence from one variable to the other. If there is a directed link from variable X_i to variable X_j, variable X_i will be a parent of variable X_j. Each node is annotated with a conditional probability distribution (CPD) that represents $p(X_i \mid Pa(X_i))$, where $Pa(X_i)$ denotes the parents of X_i in G. The pair (G, CPD) encodes the joint distribution $p(X_1,...,X_n)$. A unique joint probability distribution over X from G is factorized as:

$$p(X_1,...,X_n)=\prod_i (p(X_i \mid Pa(X_i))) \quad (1)$$

We have used tabu search algorithm for finding a well scoring Bayes network structure. Tabu search [1] is a higher level heuristic procedure for solving optimization problems, designed to guide other methods to escape the trap of local optimality.

It uses flexible structured memory to permit search information to be exploited more thoroughly than by rigid memory systems. It uses conditions for strategically constraining and freeing the search process embodied in*taburestrictions* and *aspiration criteria*. It uses memory functions of varying time spans for intensifying and diversifying the search. Tabu search is hill climbing till an optimum is reached. The best network found in this traversal is returned.

The conditional probability tables of a Bayes network can be estimated once the structure has been learned. The network structures has been created for both normal and attack record types and shown in figures 2 and 3 respectively. The Conditional Probability Tables (CPTs) for normal records dataset of figure 2 shown in tables 1 to 4. The CPTs for attack type records dataset of figure 3 shown in tables 5 to 8.

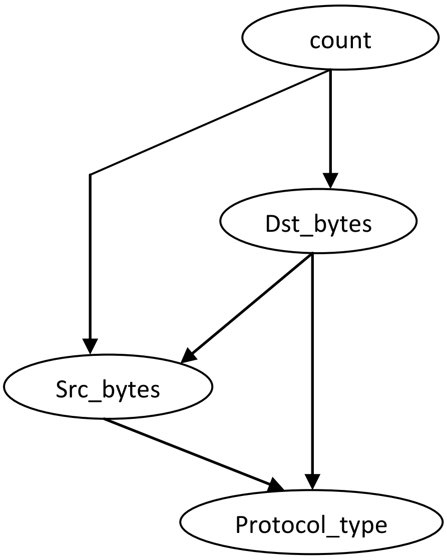

Figure 2: BayesNetwork for Normal Records

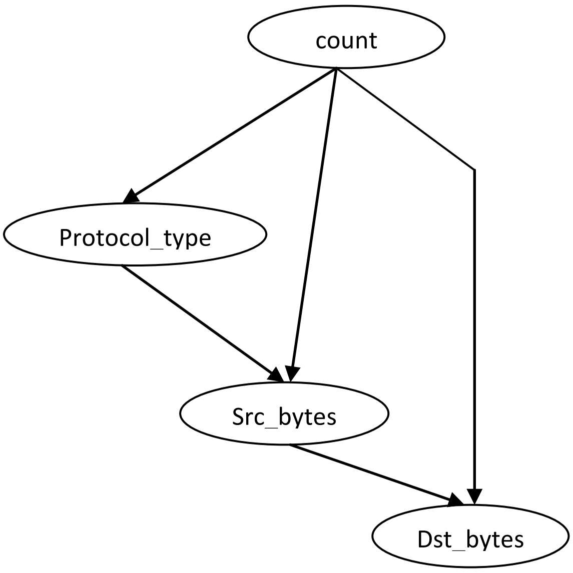

Figure 3: BayesNetwork for Intrusion Record

Table 1. CPT for count variable

Count		
Low	Mid	high
0.997028652	0.002759109	2.12E-04

Table2. CPT for src_bytes variable

Src_bytes				
dst_Bytes	count	M	H	L
H	low	0.9263	0.045779	0.027851
H	mid	0.3333	0.333333	0.333333
H	high	0.3333	0.333333	0.333333
M	low	0.2932	0.233365	0.473416
M	mid	0.0731	0.024390	0.902439
M	high	0.6	0.2	0.2
L	low	0.1101	0.183134	0.706733
L	mid	0.3333	0.333333	0.333333
L	high	0.3333	0.333333	0.333333

Table3. CPT for dst_bytes variable

dst_bytes			
count	h	m	L
low	0.6647	0.22270	0.112593118
mid	0.0243	0.95121	0.024390244
high	0.2	0.6	0.2

Table4. CPT for protocol_type variable

protocol_type				
src_bytes	dst_bytes	tcp	udp	icmp
M	h	0.9997	1.15E-04	1.15E-04
M	m	0.9935	0.005393	0.001078
M	l	0.7740	0.005649	0.220338
H	h	0.9953	0.002320	0.002320
H	m	0.9972	0.001360	0.001360
H	l	0.9931	0.003412	0.003412
L	h	0.9923	0.003802	0.003802
L	m	0.0583	0.940983	6.56E-04
L	l	0.6035	0.301333	0.095111

Table5. CPT for count variable

count		
high	low	mid
0.623876	0.2276258	0.148498

Table6. CPT for protocol_type variable

	protocol_type		
Count	Icmp	tcp	udp
High	0.9365916	0.063347	6.13E-05
Low	0.1231726	0.831961	0.04487
Mid	0.003863	0.995364	7.73E-04

Table 7.CPTfor src_bytes Variable

count	Protocol_type	H	L	M
high	icmp	0.910835	6.55E-05	0.0891
high	tcp	9.66E-04	0.99807	9.66E-04
high	udp	0.333333	0.333333	0.333333
low	icmp	0.755102	0.24354	0.001361
low	tcp	0.108015	0.86836	0.023622
low	udp	0.003717	0.97777	0.018587
mid	icmp	0.882353	0.05882	0.058824
mid	tcp	2.59E-04	0.99948	2.59E-04
mid	udp	0.2	0.6	0.2

Table 8. CPT for dst_bytes variable

		dst_bytes		
Count	src_bytes	l	h	M
High	H	0.998068	9.66E-04	9.66E-04
High	L	0.998533	7.34E-04	7.34E-04
High	M	0.523373	0.47571	9.17E-04
Low	H	0.952351	0.01708	0.030571
Low	L	0.853659	0.13821	0.00813
Low	M	0.882353	0.05882	0.058824
Mid	H	0.999483	2.58E-04	2.58E-04
Mid	L	0.333333	0.333333	0.333333
Mid	M	0.333333	0.333333	0.333333

4. HIDDEN MARKOV MODEL WITH BAYES NETWORK

Hidden Markov Model represents a very useful tool to model data-sequences, and to capture the underlying structure of a set of strings of symbols. HMM [5] is a state full model, where the states are hidden. Two probability density functions are associated to each hidden state: one provides the probability of transition to another state, and the other provides the probability that a given symbol is emitted from that state.

According to [2], an HMM" is characterized by the following expression.

$\lambda = \{A, B, \pi\}$ (2)

Where

N is the number of states in the model.

M is the number of distinct observation symbols per state, i.e. the discrete alphabet size.

A is the state transition probability distribution matrix[18]. In our case it is NxN matrix.

$\theta_{i=1..N}$ is the emission parameter for an observation associated with state i

B is the observation symbol probability distribution matrix[18]. In our case it is NxM matrix.

π is the initial state distribution[19]. Each element π_i is the probability that the initial state is the i^{th} state.

Mapping the Bayes Network (BN) parameters to HMM parameters as follows:

- The parameters which are same for both BN and HMM: N, M, A, and B.
- The values of these parameters have been initialized for the KDDCUP99 dataset as N = 4, M = 12.
- The parameters A and B have been initialized to some random values as follows:

Table 9. State Transition Probability Matrix A

	protocol_type	src_bytes	dst_bytes	count
protocol_type	0.50	0	0.30	0.20
src_bytes	0.50	0.3	0.1	0.1
dst_bytes	0.20	0.1	0.5	0.2
count	0.60	0.2	0.1	0.1

Table 10.Emission Transition Probability Matrix B

	udp	icmp	tcp	L	M	H	l	m	h	low	mid	high
protocol_type	0.10	0.21	0.177	0.52	0.0477	0.02785	0	0	0	0	0	0
src_bytes	0.133	0.13	0	0.1	0.1333	0.133	0	0.1333	0	0.1333	0.1	0
dst_bytes	0	0	0	0.13	0.1333	0.1333	0.27	0.01	0.02	0.1	0.1	0.1
count	0.1	0.1	0.1	0.1	0.133	0.173	0	0	0	0.128	0.1	0.624

5. CONCLUSION

Our IDS model is designed with Bayes Hidden Markov Model for Intrusion detection. The attributes have been selected with high entropy. This will reduce the data size with the same performance. It is trained and tested with kddcup99 dataset. The joint probabilities have estimated with full data which is also called model learning has done as the first phase. This part of the work is Bayes Network and BN parameters have been initialized. Some of these BN parameters are same as the HMM parameters. In the second phase the model has been tested which shows high performance in detecting intrusions for large and high dimensional datasets. The same is to be tested on other real time data streams. Our approach is simple and easy to implement for real time applications to the network security applications.

REFERENCES

[1] Staff.kfupm.edu.sa/DAD/sanaulla/work/**TABU**1.doc

[2] L. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE 77 (2) (1989) 257–286.

[3]http://www.kdd.org/kddcup/index.php?section=1999&method =info

[4] http://www.omnisecu.com/security/infrastructure-and-email-security/types-of-intrusion-detection-systems.htm

[5] Hidden Markov Model: http://en.wikipedia.org/wiki/Hidden_Markov_model

[6] Intrusion Detection Systems - INTRODUCTION, DETECTION METHODOLOGIES: http://encyclopedia.jrank.org/articles/pages/6646/Intrusion-Detection-Systems.html

[7] Dorothy E. Denning. An intrusion detection model. IEEE Transactions on Software Engineering, SE-13(2):222–232, 1987.

[8] James P. Anderson. Computer security threat monitoring and surveillance.Technical report, James P. Anderson Co., 1980.

[9] Richard Heady, George Luger, Arthur Maccabe, and Mark Servilla. The architecture of a network level intrusion detection system.Technical report, University of New Mexico, 1990.

[10] Frederick K. K.: *Network Intrusion Detection Signatures*.December 19, 2001, http://online.securityfocus.com/infocus/1524

[11] Elson D. :*Intrusion Detection, Theory and Practice*.March 27, 2000, http://online.securityfocus.com/infocus/ 1203

[12]Frederick K. K.: *Network Intrusion Detection Signatures*.December 19, 2001, http://online.securityfocus.com/infocus/1524.

[13] P Srinivasulu, D Nagaraju, P Ramesh Kumar, and K NageswaraRao, "Classifying the Network Intrusion Attacks using Data Mining Classification Methods and their Performance Comparison" JCSNS International Journal of Computer Science and Network Security, VOL.9 No.6, June 2009.

[14] Nagaraju Devarakonda, Srinivasulu Pamidi, V Valli Kumari, A Govardhan "Outliers Detection as Network Intrusion Detection System Using Multi Layered Framework" Advances in Computer Science and Information Technology: First International Conference on Computer Science and Information Technology, Springer, CCSIT 2011,Jan 2011.

[15] http://www.bayesnets.com/#BayesNetsStructureLearning

[16] http://www.cs.ubc.ca/ ~murphyk/ Bayes/bnintro.html

[17] http://www.autonlab.org/tutorials/bayesnet.html

[18] figment.cse.usf.edu/~sfefilat/data/papers/WeBCT8.23.pdf

[19]www.cs.brown.edu/research/ai/dynamics/tutorial/Documents/ HiddenMarkovModels.html