Research Paper Presentation

Vanga Aravind Shounik

CS20BTECH11055

Title

Probabilistic Approach for Intrusion Detection System - FOMC Technique

Authors

1. A.S.Aneetha - Anna University, Chennai
2. S.Bose - Anna University, Chennai

What is Intrusion Detection System?

1. An Intrusion Detection System (IDS) is a device or software application that monitors a network for malicious activity or policy violations.
2. IDS can be either Hardware and/or Software which is capable of detecting any malicious or unusual activities in the network
3. There are 2 types of IDS
    1. Host based IDS
    2. Network based IDS

Host Based IDS

Host Based IDS consists of an agent on a host which identifies intrusions by analyzing file system modifications, system calls etc.

Network Based IDS

In a Network IDS environment, 2 types of detection system namely

1. Signature Based Detection:
   It compares ongoing observations with patterns of well known attacks.
   It cannot detect a new type of attack.
2. Anomaly Based Detection:
   It builds a model based on the normal profile and any deviation from this is signaled an anomaly.

Anomaly Based Detection

Here, new attacks can also be detected but the time taken to analyse it is high.
It has 2 stages:

1. It builds the model using training normal traffic profile.
2. It is evaluated using testing traffic profile.

Most anomaly based IDS detects the attacks after they cause serious damage to the system.
It is necessary to to design an Intrusion Detection Model which is capable of predicting the anomalous events in network before it causes serious damage to the system.

First Order Markov Chain

Definition

A Markov chain is a first order discrete time stochastic process that predicts the future by analyzing traditional characteristic from one state to another state. This model analyses the operation of the system with finite number of states and it's state transition property.

Properties

The property of First Order Markov Chain says that the state of a system at time $n+1$ depends only on the state at time n but not on $n-1$,$n-2$..

$$P(S_{n+1} = i_{n+1} | S_n = i_n, ...S_0 = i_0) = P(S_{n+1} = i_{n+1} | S_n = i_n) \quad (1)$$

1. n and $n+1$ represents the sampling time at which the states are defined and the time interval between n and $n+1$ can be either regular or irregular.

First Order Markov Chain continued...

Here, we take

$$P(S_{n+1} = i_{n+1}|S_n = i_n) = P(S_{n+1} = j|S_n = i) = p_{ij} \tag{2}$$

Where p_{ij} is the probability of state transition from state i to state j in the time interval of n and $n+1$

The Markov model can be defined for the system with k number of finite states using the probability transition matrix

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1k} \\ p_{21} & p_{22} & \cdots & p_{2k} \\ . & . & \cdots & . \\ . & . & \cdots & . \\ p_{k1} & p_{k2} & \cdots & p_{kk} \end{bmatrix} \tag{3}$$

with the constraint,

$$\sum_{j=1}^{k} p_{ij} = 1 \tag{4}$$

First Order Markov Chain continued...

There is another matrix Q which is initial probability distribution matrix.

$$Q = \begin{bmatrix} q_1 & q_2 & \dots & q_k \end{bmatrix}^T \tag{5}$$

These matrices are calculated from the past observed data events made on the system. The data observations $X_o, X_i, ..., X_{N-1}$ are taken from the system at time $n = 0, 1, ..., N-1$.

$$p_{ij} = N_{ij}/N_i \tag{6}$$

$$q_i = N_i/N, \text{ for } i, j = 1, ...k \tag{7}$$

N_{ij} is the number of observation pairs X_n and $X_{n+1}$with X_n in the state i and $X_{n+1}$is in state j,
N_i is the number of data items in the state i and
N is the total number of data item used to build the system.
The joint probability for a given sequence of T states $X_{n-T+1}..., X_n$ at time $n - T + 1, ..n$ is

$$P(S_{t-N}, S_{t-N+1}...S_t) = q_{S_{t-N}} \Pi_{i=N}^{1} P_{S_{t-i}S_{t-i+1}} \tag{8}$$

Proposed Framework

The proposed FOMC model for anomaly intrusion detection in the network based on first order Markov chain process is discussed.
Since the framework is designed to detect anomalous activities in the network, it has two phases

1. Training Phase
2. Testing Phase

There are 3 steps in Training Phase,

1. Pre-Processing
2. Defining the states
3. Building FOMC Model

Proposed Framework continued...

Proposed Framework

In pre-processing step, data cleaning and transformation are carried out to the form needed for model development, finite number of states are defined with processed data using clustering approach and FOMC model is build with state transition matrix and initial probability distribution.

Proposed Framewok continued...

In the testing phase the test data undergoes the same pre-processing as in training phase, the deviation factor is calculated to classify test data belongs to normal or abnormal state. The probability of event occurrence for the specified time period T for the test data is calculated based on FOMC model. If the probability of event occurrence value is lesser than the predefined threshold, then the test data is considered as anomaly event otherwise it is considered as the normal event.

Pre-Processing Step

1. The data contains many unnecessary profiles which are needed to be removed so that the results will not be biased towards the repeated set of profiles.
2. The data need to be transformed into the form suitable for developing the model, by removing the labels and assigning numerical equivalent for categorical and symbolic attributes to perform the mathematical operations.
3. Normalization has to be carried out for the features using Min Max technique so that all values can be represented in a given range which in this case is [0,1]

$$V_{i(new)} = \frac{V_{i(old)} - V_{min}}{V_{max} - V_{min}} \tag{9}$$

Where $V_{i(new)}$ - new normalized value for i^{th} record of that attribute,
V_{max} - maximum value of that attribute and
V_{min} - minimum value of that attribute.

Defining the States

1. After pre-processing the data, we will define states using k-means clustering.
2. Each cluster formed by k-means clustering is defined as a state in FOMC model.A new state called outlier state is defined to represent any anomaly in the test data.
3. Since only normal profiles are used in the Markov model building, the outlier state is needed for representing unusualness in the test data along with the normal states formed by clustering.

Building FOMC Model

FOMC model is build based on the state transition probability matrix, P, computed using (3) using the constraint in (4) along with initial probability distribution using (5).
Here, In the FOMC model apart from defined states, new outlier state is also added. The new probability transition matrix is computed by considering the outlier state along the normal states defined by clustering process.
The outlier state row and column of the matrix are computed based on the assumptions made.

1. The probability of state transition from a normal to outlier is considered as minimum, assumed as zero.
2. The probability of state transition from outlier to outlier state is also less, assumed as zero.
3. The probability of the state transition from the outlier to normal state is proportional to the number of data points belong to each state. It is same as initial probability distribution Q.

Deviation Factor Analysis

Since normal records are only used in the Markov model, the outlier state is needed to represent unusualness in the test data in addition to the normal states. So outlier state is introduced in the probability transition matrix to define the abnormality in the test traffic profile.
We introduce a new variable called Deviation Factor(DF) which is computed as

$$DF(x) = \frac{\sqrt{\sum_{i=1}^{k} d(C_i - x)^2}}{K} \tag{10}$$

Where $d(C_i - x)$ is Euclidean distance which is calculated as

$$d(C_i - x) = \sqrt{\sum_{k=1}^{n} (C_{ik} - x_k)^2} \tag{11}$$

Where n is the number of attributes.

Deviation Factor Analysis continued

We write another variable d_m which is given by

$$d_m = min_{0 \leq i \leq k}\{d(C_i - x)\} \tag{12}$$

The Decision rule for the deviation factor is that if the d_m value of the data point is greater than the $DF(x)$, then the data point belongs to outlying state, otherwise, it belongs the normal state which is closest, i.e, if $d_m > DF(x)$, it lies in the outlier state.

Anomaly Detection System

1. The objective of the FOMC model is to test the observed traffic profile for any anomaly, based on probability of event occurrence of that observation.
2. As the intrusions are made in steps the changes in the behavioral pattern are identified from sequences of transition of the data from state to state, which may also help in suspecting the traffic.
3. The probability of event occurrences for the specific sequence of states is calculated and it is compared with the sequence of states has one or more abnormal activities.
4. In the testing phase, the probability of event occurrences is calculated using the equation given in (8) for the specified T size.
5. The probability of event occurrence value of the test data is considered as normal if the value is greater than the threshold otherwise anomaly to the degree of that probability.

Results and Discussions

Dataset Description

1. We use the network traffic profile of KDD Cup'99 benchmark Dataset.
2. It consists of 41 attributes.
3. In building phase only the normal profiles are utilized, whereas testing has done in two ways, one is with the normal profile alone but not used in model development and other one is test data with fifty percent of attack profile along with fifty percent of normal profile.
4. Both training and testing data are pre-processed by assigning numeric values for the categorical attributes and attribute normalization process which equalizes all the attributes irrespective for their greater or lesser original attribute values.

Results and Discussions continued...

1. The first step of FOMC model design is defining the network states using the k-means clustering technique based on which the actual detection system is developed.
2. The proposed system uses N-fold cross validation method for evaluating the performance of the model and the best results are considered as outputs.
3. Here N is set to 10 for the entire training traffic profile.

Graphs of the Observations

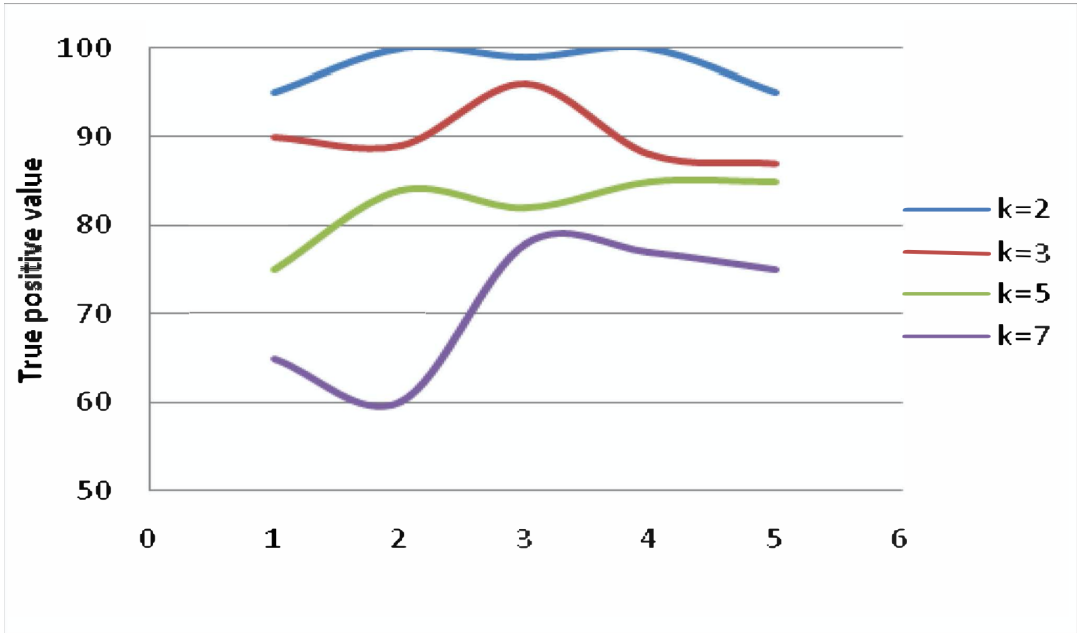

Figure: True positive rate Vs States

From this, we can see that when the states are 2, the system gives consistent performance. Whereas in the case of seven, five and three the system performance varies much.

Results and Conclusions continued...

S.No	K=2		K=3		K=7	
	Testset1	TestSet2	Testset1	TestSet2	Testset1	TestSet2
1.	95%	100%	90%	100%	65%	80%
2.	100%	90%	89%	82%	60%	—
3.	97%	96%	78%	75%	85%	78%
4.	99%	100%	96%	88%	82%	80%
5.	94%	100%	78%	96%	60%	78%
6.	90%	80%	60%	72%	72%	80%
7.	100%	96%	88%	74%	80%	84%
8.	97%	100%	79%	82%	63%	66%
9.	90%	100%	67%	83%	55%	—
10.	95%	100%	79%	90%	75%	75%

Table: Detection Rates for Number of States

Here, we can see that testdata 2 performance is better than testdata 1 even though there are 50% Dos attacks included.

Conclusion

1. The probabilistic approach of analyzing intrusion in the network traffic has been proposed.
2. The probabilistic approach deploys first order Markov chain process to predict the anomalous activities in the network.
3. The traffic is suspected as abnormal if the computed probability of event occurrences, based on state transition matrix and initial probability distribution, of the test data is lesser than the predefined threshold value, otherwise it is considered as normal traffic.

Thank You