Probabilistic Approach for Intrusion Detection System - FOMC Technique

A.S. Aneetha, and S. Bose

Abstract—**Detection of unexpected and emerging new threats has become a necessity for secured internet communication with absolute data confidentiality, integrity, and availability. Design and development of such a detection system shall not only be new, accurate and fast but also effective in a dynamic environment encompassing the surrounding network. In this work, an attempt is made to design an intrusion detection model based on the probabilistic approach, first-order Markov chain process, to effectively detection and predict network intrusions. As a first step, the states are defined using clustering techniques for the network traffic profiles; secondly state transition probability matrix and initial probability distribution are determined based on the states defined. Based on the network states, the probability of event occurrence is stochastically measured if the value is lesser than the predefined probability then it event is predicted as anomaly. The proposed probabilistic model performance is evaluated through experiments using KDD Cup99 dataset. The proposed models achieve better detection rate while the attacks are detected in levels of stages.**

***Index** Terms— First - Order Markov Chain Process, Intrusion Detection System, Probabilistic Approach, State Transition Matrix.*

I. Introduction

Now-a-days, the prevention against cyber threats has assumed great importance in the field of internet communication owing to the increase in virtual attacks. Intelligent analysis needs to be performed on large volumes of data generated from the network devices, for secured communication. One of the popular ways of securing the communication system is an Intrusion Detection System (IDS) apart from firewall, authorization and authentication. IDS can either be hardware and/or software, which is capable of detecting unusual and malicious activities in the network. IDS can be classified into Network based and Host or Audit based IDS based on their traffic profile [1]. Host based IDS consists of an agent on a host which identifies intrusions by analyzing file system modifications, system calls etc., whereas network based IDS detects the intrusion in the network communication system. In a network IDS environment, two types of detection systems, namely, misuse or signature based detection and anomaly or behavioral based detection are very often employed [11]. The signature based IDS compares on-going observations (i.e. Network traffic profile) with the patterns of well known attacks and those matched are labeled as that attack type. But it has its own limitation of not being able to detect the continuously emerging new attacks. The anomaly based IDS builds the model based on the normal profile and any deviation from this is signaled as anomaly. Here new attacks can be detected but the time taken to analyze it is considerably high. Anomaly based IDS has two stages, first it builds the model using the training normal traffic profile and secondly, the model is evaluated using testing traffic profile, deviations are signaled as anomaly [1].

However, most of the anomaly based IDS detects the attacks after they cause serious damage to the system. There are different approaches such as machine learning algorithms, data mining approaches, artificial intelligence (AI), statistical techniques used to identify intrusions after they are caused [6][11][12]. As we know attacks occurs in stages in most of the scenarios. It is necessary to design an intrusion detection model which is capable of predicting the anomalous events in the network before it causes serious damages to the system.

In this work, an attempt is made to design a network IDS based on the probabilistic approach. The First – Order Markov Chain (FOMC) process is adapted to design the model, where the current state is determined with the knowledge of the previous state of the event. Based on the probability of event occurrences, the activities of the network traffic profile are decided as normal [9]. The higher order Markov model uses recent historical data or the last several states; the present state is determined by not only the previous state but all the previous states. Thus, the large number of computing resources is needed [2].

This paper is organized as follows: In section 2 the related works on different probabilistic models used in anomaly detection are discussed. Introduction to First – Order Markov chain process is clearly described in section 3. The proposed framework is explained in section 4 and results are discussed in section 5. The summary of the work is given in the last section.

II. Literature Review

Probabilistic model produces the result in terms of probability, which ranges from zero to one. From this probabilistic scale the administrator is able to understand the

Manuscript submitted on November 13, 2014

A.S. Aneetha, Reasearch scholar, Department of Computer Science and Engineering, Anna University, Chennai – 25. [corresponding author: e-mail: avsaneetha@yahoo.co.in; phone number: +91 9444545065]

Dr. S. Bose, Associate Professor, Department of Computer Science and Engineering, Anna University, Chennai – 25. [sbs@cs.annauniv.edu]

978-1-4799-8159-5/14/$31.00©2014 IEEE

Authorized licensed use limited to: Indian Institute of Technology Hyderabad. Downloaded on April 21,2021 at 18:01:30 UTC from IEEE Xplore. Restrictions apply.

degree of seriousness about the attack. This has a remarkable advantage when compared to all other detection methods which merely detects whether it is normal state or in attack state [2]. Anomaly detection in the network using probability methods has been carried out earlier using various models. Among those, Markov models plays a dominant role in the field of identifying intrusion since they decide the event based on previous encounters [15].

A novel work on advanced probabilistic approach for effectively predicting and forecasting potential attacks in the network has been proposed by Seongjun et. al. in the 2013 [2]. The abnormality of the incoming data is stochastically measured using this model in real time. It has been claimed that the model is adaptive to changes in the normal profile along with robustness of the outlier factor. In 2001, Nong Ye et. al. presented a series of studies on probabilistic properties such as frequency, ordering and duration of an activity data for detecting intrusion. Different probabilistic techniques such as Hotelling T^2, chi – squared multivariate techniques and Markov chain have been used for detecting intrusion in the system [8]. Ye and Borror have proposed EWMA forecasting model in which Markov chain is used to learn and predict normal activities in that model [7].

Hence, building the Markov model is computationally simple since it only needs the state transition matrix and initial probability distribution for its construction. The Markov model is able to achieve a very low false positive rate. As the general task of intrusion detection process is to protect the system before serious damages are caused, the Markov model has been proved successful [5].

III. Introduction to First – Order Markov Model

A Markov chain is a first order discrete – time stochastic process that predicts the changes in the future by analyzing transitional characteristic from one state to another state. This model analyses the operation of the system with finite number of states and its state transition probability. The property of the First – Order Markov process says that the state of a system at time n+1 depends only on the state of the system at time n and does not depends on the previous states such as n-1, n-2 etc, [2][3]. It is represented by the equation as given in (1).

$$P(S_{n+1} = i_{n+1} \mid S_n = i_n, \ldots\ldots\ldots, S_0 = i_0) = P(S_{n+1} = i_{n+1} \mid S_n = i_n) \qquad (1)$$

Here 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. A new additional property says that, a state transition from time n to n+1 is independent of time, so Markov chain becomes stationary and is denoted by (2), which is derived from equation (1).

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

Where p_{ij} is the probability of state transition from state i to state j in a time interval of n and n+1. The Markov model can be defined for the system with k number of finite states using probability transition matrix and initial probability distribution [3][7]. The probability transition matrix P is calculated as given in (3) with the constrain as in (4).

$$P = \begin{vmatrix} P_{11} & P_{12} \cdots & P_{1k} \\ P_{21} & P_{22} \cdots & P_{2k} \\ \vdots & \vdots & \vdots \\ P_{k1} & P_{k2} & P_{kk} \end{vmatrix} \qquad (3)$$

$$\textstyle\sum_{j=1}^{k} p_{ij} \;\; =1 \qquad (4)$$

$$Q = [\, q_1, q_2, \quad \ldots , q_k]^{T} \qquad (5)$$

The probability transition matrix P and initial probability distribution matrix Q of the Markov model are calculated from the past observed data events made on the system. The data observations X_o, X_1, …., X_{N-1} are taken from the system at time n = 0,1, …., N-1. The value of P_{ij} and q_i are calculated using the equations given below.

$$P_{ij} = N_{ij} / N_i \qquad (6)$$

$$qi = N_i / N, \;\text{for } i, j = 1,\ldots., k \qquad (7)$$

where

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 q_i is the probability that the system is in state i at time zero. The joint probability for a given sequence of T states X_{n-T+1}, …., X_n at time [n-(T-1)] ….n is computed using the(8).

$$p(S_{t-N}\, ,S_{t-N+1,}\, \ldots .S_t) = q_{S_{t-N}} \textstyle\prod_{i=N}^{1} P_{S_{t-i} S_{t-i+1}} \qquad (8)$$

IV. Proposed Framework

In this section, 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 such as training and testing phase. The training phase consists of three steps such as data preprocessing, defining the states and development of FOMC model. 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 as explain in the previous section. 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

Authorized licensed use limited to: Indian Institute of Technology Hyderabad. Downloaded on April 21,2021 at 18:01:30 UTC from IEEE Xplore. Restrictions apply.

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.

A. Pre-Processing

Since network traffic profiles has redundant profiles which makes the model biased towards the repeated set of profiles so it needs to be removed. 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. Normalization has to be carried out for the features using Min Max technique [16] using the following formula given in (9).

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

Where

V_i - 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.

B. Defining the states

After pre-processing the data to build the model, we need to define network states. As traffic records are large in number, the grouping has to be done for defining it as states. The grouping process has been carried out by the efficient k-means clustering technique for the traffic profile. Each cluster formed using k-means is defined as states in the FOMC model. Let a S = $\{S_1,S_2, . . . ,S_K\}$ be the set of k states of the system. A new state called as outlier state is introduced to define the abnormality in the test data. Since only normal traffic 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.

```
Algorithm for k-means Clustering

Choose k data points as the initial cluster center
For data points $X_i$ where i = 0, ..,N-1
    Compute the Euclidian distance from $X_i$ to each centroid;
    Assign $X_i$ to the closest centroid
    Re-compute the centroid using the current cluster
Repeat the steps for all data points until centroid gets stable
```

C. Building FOMC Model:

In this module the proposed FOMC model is build based on the state transition probability matrix, P, computed using (3) with the condition given in equation (4) along with the initial probability distribution, Q, using the equation(5). The transition of the data points is computed using (6) based on the data movement from one state to another in the specified time period t. 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 new matrix are computed based on the assumptions made [2]. The assumptions are explained here as

- The probability of state transition from a normal to outlier is considered as minimum, assumed as zero.
- The probability of state transition from outlier to outlier state is also less, assumed as zero.
- 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.

D. 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. If test observation is within the scope of a decision rule, the outlier state is assigned as the state for that observation. The degree of belongingness of the test observation with the normal states is need to be measured for that deviation factor (DF) is introduced. The deviation factor is used to measure the distances between the observation and the cluster centers. Let C = $\{C_1, C_2, . . . , C_K\}$ be cluster centers generated through in the clustering analysis based on the training dataset and DF(x), the deviation factor of test data x, which is computed using the equation given in (10).

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

where $d(C_i\text{-}x)$ is calculated as in (11),

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

where n is the number of attributes. In addition to DF the d_m value is calculated for data x. It is defined as the minimum distances value, calculated between x and the centroid of all the clusters. Calculate distance between data points and centroid of the clusters using equation in (11).

$$d_m = \min_{1 \le i \le k} \{d(X - C_i)\} \qquad (12)$$

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.

```
Algorithm for Decision Rule

Calculate the Euclidean distance between test data and cluster
        centers using (11)
Calculate deviation factor DF(x) using (10)
Identify the distance between data with closest cluster (i.e) $d_m$
        using (12)
If $d_m$ > deviation factor then data belongs to outlying state, else
```

Authorized licensed use limited to: Indian Institute of Technology Hyderabad. Downloaded on April 21,2021 at 18:01:30 UTC from IEEE Xplore. Restrictions apply.

normal state.

E. Anomaly Detection System

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. 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. 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. In the testing phase, the probability of event occurrences is calculated using the equation given in (8) for the specified T size.

The higher probability value for the sequence of events computed by T specified time sequence are considered as normal events, otherwise there is a possibility of the event sequence has abnormal event according to their probability value. 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. The higher and lower value decisions are based on the threshold, which can be varied to analysis the model and also makes the model more stable and robust.

V. Results and Discussions

In this section, we discuss the performance of the proposed FOMC anomaly detection model using the well known KDD Cup'99 bench mark dataset for network intrusion detection scenario [13][14]. As far as dataset is concern they have a profile for both normal and attacks, since we are interested in building the anomaly model the normal profiles alone are considered.

A. Dataset Description

The network traffic profile of KDD Cup '99 dataset consist of forty-one attributes along with one for representing the class. In that thirty-two are continuous attributes others are nominal, in which three are categorical and remaining six are binary valued attributes. For the experimental purpose, we assigned equivalent numeric values for the categorical attributes and class attribute are removed. In building phase the only 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. 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.

B. Results of FOMC Model

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. The proposed system uses N-fold cross validation method for evaluating the performance of the model and the best results are considered as outputs. In this work N value is set as ten and the entire training traffic profile are divided into ten blocks on which nine blocks are used to build the system and remaining one blocks is utilized for testing the model. As the same way the experimented is repeated ten times and the best results are discussed here. The number of states defined using clustering techniques has the impact on the system performance which is discussed and shown in the Fig 1. From this, it can be analyzed that when the number of states is two the system gives consistent performance whereas in the case of seven and three the system performance varies in much.

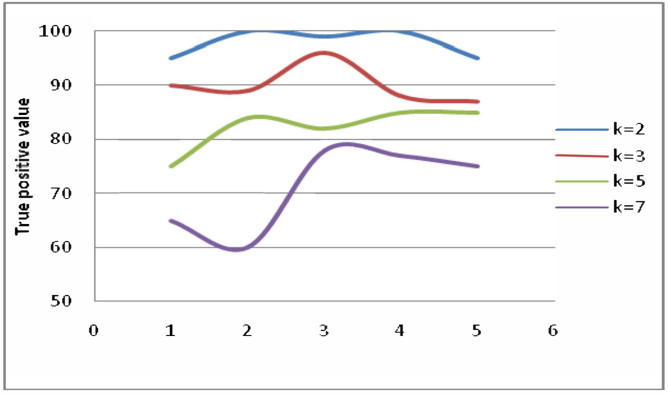

Fig. 1 True positive rate Vs States

The intrusion detection performance of the FOMC system is evaluated with two different sets of testing data as explained above. The test set 1 uses only the normal profiles but not used in the model development while test set 2 consist of fifty percent of unused new normal profile and remain fifty percent of DoS attack profiles are utilized. The detection rate of the system is same as the Fig 1 since the true positive values places a major component. The error rate of the FOMC model for the test data 1 is given in Fig 2, when the graph is drawn for different number of states ranging from two to seven.

It explains clearly that the number of states is an important factor which should also be taken into consideration. The

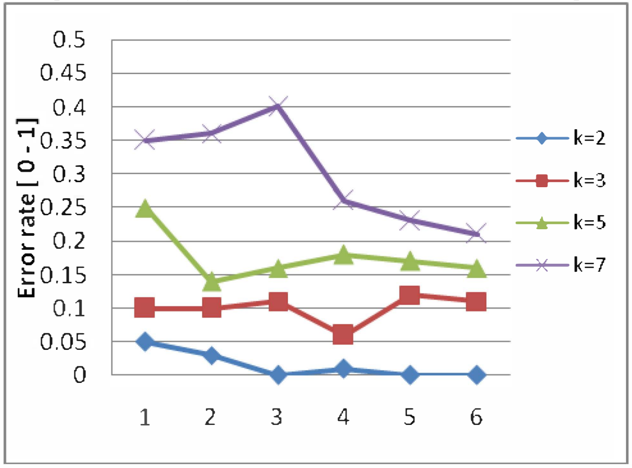

Fig. 2 Performance of the system Vs States

Authorized licensed use limited to: Indian Institute of Technology Hyderabad. Downloaded on April 21,2021 at 18:01:30 UTC from IEEE Xplore. Restrictions apply.

performance of the FOMC model with test data 2 achieves very less false positive rate as 0.01 or even zero in most of the times. It assures that the system is capable of identifying all type known and unknown attacks perfectly, which makes the FOMC model more advantage and also this is our claim towards the system performance. However, we need to compromise in the detection rate where normal profiles are considered as attacks. The detection rates of the system with both testing datasets are given in the Table1.

Table 1 shows the results of proposed FOMC model based on tenfold cross validation technique. It is clear that ten different set of training data are used to build the model and it also evaluated with both test data sets. From the table 1 we can say test data 2 performance is better than test data 1 because all types of attacks in test data 2 are identified correctly. As results shows that test set 2 gives better performance than the previous one even though fifty percent of DOS attacks are included.

TABLE I

DETECTION RATES FOR NUMBER OF SATES

S.No	K=2		K=3		K=7	
	Test set 1	Test set 2	Test set 1	Test set 2	Test set 1	Test set 2
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%

VI. CONCLUSION

In this work, the probabilistic approach of analyzing intrusion in the network traffic has been proposed. The probabilistic approach deploys first order Markov chain process to predict the anomalous activities in the network. This has been carried out in three steps such as defining the states, building the state transition matrix and probability distribution of event occurrence. 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. The FOMC anomaly detection model is built on the normal traffic profile since the anomaly detection model character says that. The FOMC model performance is evaluated through experiments with network intrusion detection bench mark dataset, KDD Cup99. The proposed models achieve better detection rate while the attacks are detected in levels of stages and also gives very low false positive rate. In the future the higher order Markov chain process can be applied for predict and forecasting the intrusions. The detection rate can be improved since it is based on previous history for a longer duration.

REFERENCES

[1] Chen, Tieming, Xu Zhang, Shichao Jin, and Okhee Kim. "Efficient Classification using Parallel and Scalable Compressed Model and Its Application on Intrusion Detection." *Expert Systems with Applications,* vol. 41, no. 13, pp : 5972-5983, 2014.

[2] Shin, Seongjun, Seungmin Lee, Hyunwoo Kim, and Sehun Kim. "Advanced Probabilistic Approach for Network Intrusion Forecasting and Detection." *Expert Systems with Applications,* vol. 40, no. 1, pp: 315-322, 2013.

[3] Ye, Nong, Timothy Ehiabor, and Yebin Zhang. " First-Order Versus High-Order Stochastic Models For Computer Intrusion Detection." *Quality and Reliability Engineering International,* vol. 18, no. 3, pp: 243-250, 2002.

[4] Ye, Nong, Yebin Zhang, and Connie M. Borror. "Robustness Of The Markov-Chain Model For Cyber-Attack Detection." *Reliability, IEEE Trans.,* vol. 53, no. 1, pp: 116-123, 2004.

[5] Goonatilake, Rohitha, Susantha Herath, and Ajantha Herath. "Probabilistic Models for Anomaly Detection Based on Usage of Network Traffic." *Journal of Information Engineering and Applications* 3, no. 9, pp: 28-40, 2013.

[6] Govindarajan, M., and R. M. Chandrasekaran. "Intrusion Detection Using Neural Based Hybrid Classification Methods." *Computer networks,* vol. 55, no. 8, pp:1662-1671, 2011.

[7] Ye, Nong, Qiang Chen, and Connie M. Borror. "EWMA Forecast Of Normal System Activity For Computer Intrusion Detection." *Reliability, IEEE Trans.,* vol. 53, no. 4, pp: 557-566, 2004.

[8] Ye, Nong, Xiangyang Li, Qiang Chen, Syed Masum Emran, and Mingming Xu. "Probabilistic Techniques for Intrusion Detection based on Computer Audit Data."*Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Trans.,* vol.31, no. 4 pp: 266-274, 2001.

[9] Sha, Wenyao, Yongxin Zhu, Tian Huang, Meikang Qiu, Yan Zhu, and Qiannan Zhang. "A Multi-order Markov Chain Based Scheme for Anomaly Detection." *Computer Software and Applications Conference Workshops (COMPSACW), 2013 IEEE 37th Annual*, pp. 83-88. IEEE, 2013.

[10] Ye, Nong, Syed Masum Emran, Qiang Chen, and Sean Vilbert. "Multivariate Statistical Analysis Of Audit Trails For Host-Based Intrusion Detection."*Computers, IEEE Transactions,* vol. 51, no. 7 (2002): 810-820.

[11] Yasami, Yasser, and Saadat Pour Mozaffari. "A Novel Unsupervised Classification Approach For Network Anomaly Detection By K-Means Clustering And ID3 Decision Tree Learning Methods." *The Journal of Supercomputing,* vol. 53, no. 1 pp: 231-245, 2010.

[12] Lin, Shih-Wei, Kuo-Ching Ying, Chou-Yuan Lee, and Zne-Jung Lee. "An Intelligent Algorithm With Feature Selection And Decision Rules Applied To Anomaly Intrusion Detection." *Applied Soft Computing,* vol. 12, no. 10 pp: 3285-3290, 2012.

[13] Tavallaee, Mahbod, Ebrahim Bagheri, Wei Lu, and Ali-A. Ghorbani. "A Detailed Analysis Of The KDD CUP 99 Data Set." In *Proceedings of the Second IEEE Symposium Computational Intelligence for Security and Defence Applications 2009*. 2009.

Authorized licensed use limited to: Indian Institute of Technology Hyderabad. Downloaded on April 21,2021 at 18:01:30 UTC from IEEE Xplore. Restrictions apply.

[14] The UCI KDD archive, KDD Cup'99 Data Set, http://kdd.ics.uci.edu/databases/ kddcup99/ kddcup99.html

[15] Khanna, R., and Liu, H., "System Approach To Intrusion Detection Using Hidden Markov Model " , *Proc. of the ACM international conference on Wireless communications and mobile computing,* pp. 349–354,July,2006.

[16] Han, Jiawei, Micheline Kamber, and Jian Pei, *Data mining: concepts and techniques*, Morgan Kaufmann,2009.

A.S.Aneetha Received MCA from Bharathiyiar University, Coimbatore, Tamil Nadu and M.Phil from Alagappa University, Karaikudi, Tamil Nadu. She is currently a Research Scholar, pursuing her Ph.D programme in the area of intrusion detection system, in the Department of Computer Science and Engineering, Anna University Chennai, Tamil Nadu. Her area of interest is Data Mining and Network Security.

Dr. S. Bose Received M.E from Madurai Kamaraj University and Ph.D from Anna University, Chennai. He is currently working as an Associate Professor in the Department of Computer Science and Engineering, Anna University Chennai. His area of interest is Networks, Network Security, Web Technology, Multimedia Streaming and Artificial Intelligence.

Authorized licensed use limited to: Indian Institute of Technology Hyderabad. Downloaded on April 21,2021 at 18:01:30 UTC from IEEE Xplore. Restrictions apply.