An Optimized Feature Selection for Intrusion Detection using Layered Conditional Random Fields with MAFS

Mr.C.Saravanan[1] , Mr.M.V.Shivsankar[2] , Prof.P.Tamije Selvy[3],Mr.S.Anto[4]

[1,2,]UG Student , [3]Assistant Professor (SG),[4] Assistant Professor

[1,2,3,4]Department of Computer Science and Engineering, Sri Krishna College of Technology, Coimbatore, India

[1]saaraooty@gmail.com, [2]midhunvlbjcet@yahoo.com , [3]tamijeselvy@gmail.com [4] santocse@gmail.com

Abstract

Intrusion Detection systems are now an essential component in the overall network. With the rapid advancement in the network technologies including higher bandwidths and ease of connectivity of wireless and hand held devices, the main focus of intrusion detection has shifted from simple signature matching approaches to detecting attacks based on analyzing contextual information which may be specific to individual networks and applications. As a result, anomaly and hybrid intrusion detection approaches have gained significance. The Denial of Service Attacks (DoS), Probe, User to Root (U2R) and Remote to Local (R2L) are some of the common attacks that affect network resources. Intrusion detection faces a number of challenges; an intrusion detection system must reliably detect malicious activities in a network and cope up with large amount of network traffic. In this paper, we address these two issues of Accuracy and Efficiency using Conditional Random Fields and Layered Approach. Finally we demonstrate that high attack detection accuracy can be achieved by using Memetic algorithm for feature selection with Layered Conditional Random Fields.

Keywords

IDS, Layered Approach, Conditional Random Fields, Signature and Anomaly Based IDS, MAFS

1. INTRODUCTION

Intrusion detection is defined as art of detecting inappropriate, inaccurate, or anomalous activity inside and the network environment. It is the most high priority task for network administrator. The main purpose of intrusion detection is that finding the attack which are specified in the signatures and as well as finding the new or unseen attacks effectively and also cope up with large amount of network traffics. The main Key factor in any kind of IDS is having low FAR (False Alarm rate) value. That is the System must be accurate in nature for finding the attacks. Now a days different kinds of IDSs are introduced. Each having different features for finding the different kind of network attacks . **NIDS (Network Intrusion Detection Systems)** Network Intrusion Detection Systems are placed at a strategic point or points within the network to monitor traffic to and from all devices on the network. Ideally one would scan all inbound and outbound traffic

HIDS (Host Intrusion Detection Systems) Host Intrusion Detection Systems are run on individual hosts or devices on the network. A HIDS monitors the inbound and outbound packets from the device only and will alert the user or administrator if suspicious activity is detected. It will analyze network traffic and system-specific settings for effectively finding the attacks exist in the current network environment.

Signature Based: A signature based IDS will monitor packets on the network and compare them against a database of signatures or patterns of known malicious threats. This is similar to the way most antivirus software detects malware. The issue is that there will be a lag between a new threat being discovered in the wild and the signature for detecting that threat being applied to your IDS. During that lag time your IDS would be unable to detect the new threat.

Anomaly Based : An IDS which is anomaly based will monitor network traffic and compare it against an established baseline. The baseline will identify what is "normal" for that network, what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other and alert the administrator or user when traffic is detected which is anomalous or significantly different than the baseline.

Another approach for detecting intrusions is that integrating the Signatures and anomalies in the network environment. So that we can find the attacks which are specified in the database as well as newly found in the environment. This approach will give the high efficiency for finding the different kind of network attacks which results in better classification test on the observed data.The Generic representation of this system is shown in fig. 1

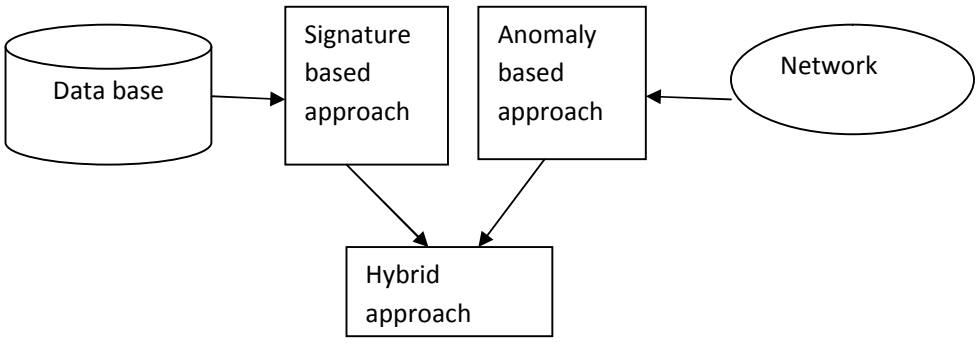

Figure 1. Generic representation of the system.

2. RELATED WORK

The history of of intrusion detection has been start around since late 1980s. Since then, a number of methods and frameworks have been proposed and many systems have been built to detect intrusions

2.1. Data mining Approach

Wenke Lee Salvatore J. Stolfo and Kui W. Mok introduced data mining approaches for detecting intrusions[1]. Data mining approaches for intrusion detection include association rules , frequent episodes and outlier detection , which are based discovering relevant patterns of program and

user behavior for building up the Classifiers. These approaches(Association rules and Frequent episodes) are used to learn the patterns that describes behavior of the users in the network environment. These methods can deal with symbolic kind of data, and their features can be represented in the form of packet and connection details. Mining of features are very low at entry level but it is so large whenever the attributes of the particular feature is increased which results in more system complexity.

2.2 . Data clustering methods

Data clustering methods in intrusion detection includes the k-means and the fuzzy c-means clustering methods [2],[4]. The main drawbacks in these clustering technique is that it is based on determining the numeric distance between the observations resulting that, the observations must be numeric .Hence observations with symbolic features cannot be easily identified and used for clustering, resulting in inaccuracy for finding the attacks . In addition to that, these methods consider the features for intrusion must be independent and are not possible to capture the relationship between different features in a single record, which further degrades attack detection accuracy of the system .

2.3 . Naive Baye's classifiers

The next approach discussed here in intrusion detection is Naive Bayes classifiers [3]. Those approaches will make strict independent assumption between the features in an observation records which resulting in very low detection accuracy when the features in the observation are having correlation between them . Bayesian network can also be used for finding the anomalous activities in the environment . However, those networks may tend to be attack specific and construct a decision network based on special characteristics about individual attack groups . Thus, the size of a Bayesian network increases rapidly whenever the number of features and the type of attacks are increases.

2.4 . Decision trees

The intrusion detection also performed using Decision trees approach[3] this approach presents decision tree techniques that are used to automatically learn intrusion signatures and perform the classification activities in computer network systems as normal or intrusive. The decision trees will use some well defined criteria for selecting the best features during the construction of the tree. One such oftenly used criterion is usage of information gain ratio(which is used in C4.5 algorithm). This approach usually have very high speed of its operation and high accuracy of attack detection .

2.5 . Neural Networks

The neural network components also used for finding the intrusive events in the network [5].The neural network in intrusion will work well with correlated kind of data in the observation (noisy data) .However, the neural networks require large amount of data for training the observations and also it is often difficult to select the best possible architecture for a neural networks.

3. PROPOSED SCHEME

3.1. Applying conditional random fields

CRF was firstly proposed by Lafferty and his colleagues in 2001, whose model idea mainly came from MEMM (Maximum Entropy Markov Model). Conditional models are probabilistic systems that are used to model the conditional distribution over a set of random variables. Such models have been extensively used in the natural language processing tasks[8] and [9]. Conditional models offer a better framework as they do not make any unwarranted assumptions on the observations and can be used to model rich overlapping features among the visible observations.

Consider X is the random variable over data sequence in the observation to be labeled and Y is the corresponding label sequence for those observation . In addition, let G = (V, E) be a graph such that Y = (Yv) v(V)', so that Y is indexed by the vertices of G. Then, (X,Y) is a CRF, when conditioned on X, the random variables Yv obey the Markov property with respect to the graph:

p(Yv|X, Yw, w v) = p(Yv|X, Yw, w~ v)

where w ~ v means that *w* and *v* are neighbors in G, i.e., a CRF is a random field globally conditioned on X. For a simple sequence (or chain) modeling, as in our case, the joint distribution over the label sequence Y given X has the following form:

p(y|x) exp (eE,k kfk (e, y|e, x) + vV ,k kgk(v; y|v, x) , ……. ***(1)***

where x is the data sequence, y is a label sequence, and y|s is the set of components of y associated with the vertices or edges in subgraph S. In addition, the features fk and gk are assumed to be given and fixed

The prime difference between CRF and other graphical models such as the HMM is that the HMM, being generative. It models the joint distribution p(y,x), whereas the CRF are discriminative models and directly model the conditional distribution p(y|x), which is the distribution of interest for the task of classification and sequence labeling.

The data set used in our experiments represents features of every session in relational form with only one label for the entire record. Here, using a conditional model would result in maximum entropy classifier. However, we represent the data in the form of a sequence and assign a label to every features in the observation record using the first-order Markov assumption instead of assigning a single label to the entire observation. This will improves the attack detection accuracy.

3.2. Layered approach for Intrusion detection

Let , We now describe the concept of Layer-based Intrusion Detection System (LIDS) in brief. There should be a number of security checks are performed one after the other in a sequence. So, the layers are deployed in a sequential manner. This approach has the advantages that we can increase or decrease the layered as we required, it will improves the high efficiency of the System. Each layer has the following units which are shown in Fig. 2

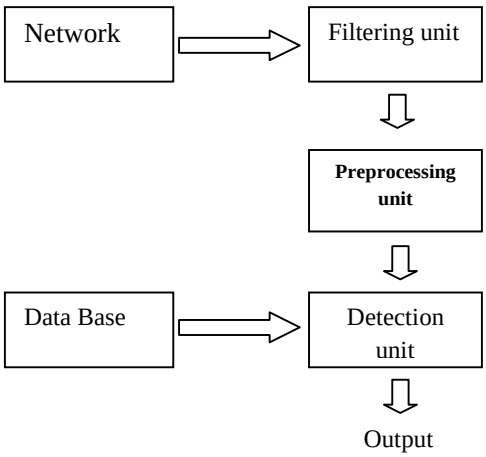

Figure 2. Layered representation

Filtering unit :

As shown in the figure 1, the WinPcap software provides facilities to capture raw packets and filter the packets according to user specified rules before dispatching them to the application.

Preprocessing unit :

The preprocessor defines one class called packet and this class will store all the packets that are generated by the Filtering unit.

Database:

It is a specially prepared pattern database. There is lot of signatures in the database for such analysis. The snort is used for this kind of purpose.

Detection unit :

It takes packets from preprocessor and compares them with special signatures from the database Result of the compassion is sent to the output module, where a report is prepared.

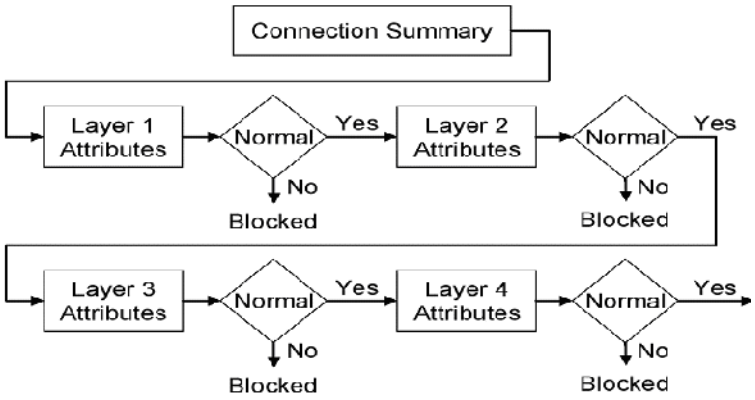

Figure. 3. Real-time representation of the system.

The figure 3 illustrates the entire representation of the system. The following two approaches are to be performed for getting better efficiency of the system.

3.2.1. Multi Layered Approach

A three layer system is proposed to ensure complete security viz. availability, confidentiality and integrity, each layer corresponding to one aspect of security. The layers are sequential and not overlapping i.e. layer one followed by layer two followed by layer three, where each layer has some unique features and some features from its previous layers. This ensures that each layer is stand-alone and is able to effectively block the type of intrusion which it is meant to block. Sharing of some features from previous layers is necessary to ensure that the layers are linked together. Various semantic features needs to be related to the non-semantic feature such as connection features to ensure better detection capabilities.

3.2.2. Feature Selection for Layers

Feature selection is performed automatically. The methods for automatic feature selection were not found to be effective. Every layer is separately trained to detect a single type of attack category. The attack groups are different in their impact, and hence, it becomes necessary to treat them differently. Features are selected for each layer based upon the type of attacks that the layer is trained to detect.

3.3. Integration of layered approach with Conditional Random fields

In previous sections we proposed two approaches called CRFs and Layered Approach for improving accuracy and efficiency respectively. Integration of these two approaches is performed to build the single system that is accurate in detecting attacks and efficient in operation.

Probe Layer.

Probe Layer is responsible for identifying the Probe attacks. These attacks are mainly focused on acquiring details about the target network from a source which is often external to the current network. So, the basic connection level features such as the “duration of connection” and “source bytes” are significant for finding those kind of attacks.

DoS Layer.

This layer is built for detecting the DoS (Denial of Service) kind of Attacks. These attacks are meant for forcing the target node to stop their service(s) that is (are) provided by flooding it with illegitimate requests. Hence, the traffic features such as the “percentage of connections having same destination host and same service” and packet level features such as the “source bytes” and “percentage of packets with errors” are enough for finding the DoS kind of attack.

R2L Layer.

This Layer will identify the R2L (Remote to Local) attacks. In a R2L kind of attack, attacker will exploits the vulnerability for local access. They are very difficult to detect which are present in the environment . They use the network level and the host level features for identification in the network. Network level features include the “duration of connection” and “service requested” and the host level features include the “number of failed login attempts” .

U2R Layer.

This layer is used for detecting the U2R(User to Root) attack .It is also very hard to identify. Such type of attacks is often content based thing. the features for U2R attacks include "number of file creations" and "number of shell prompts invoked.

We used domain knowledge as well as practical significance hence resulting in high efficiency for attack detection .Thus, from the total 41 features, we selected only 5 features for Probe layer, 9 features for DoS layer, 14 features for R2L layer, and 8 features for U2R layer. Since each layer is independent of every other layer, the feature set for the layers is not disjoint

3.4. Evaluation of performance analysis parameter

The system's performance is calculated by the three measures called precision, recall, f-value. The formula for calculating these measures are given as follows.

Precision : Precision refers to how the experimental values and observed values are close to each other. Precision = $\frac{TP}{TP+FP}$ where TP, FP are the number of True Positives and False Positives respectively.

Recall : Recall is defined as the fraction of relevant documents that are retrieved

Recall = $\frac{TP}{TP+FN}$ where TP ,FN are the number of True Positives and False Negatives respectively.

F-Values: F-measure is the harmonic-mean of *Precision* and *Recall* and takes account of both measures**.**

F-Value = $\frac{\left(1+\beta^2\right)*Recall*Precision}{\beta^2*(Recall+Precision)}$ where β correspondence to the relative importance of precision verses recall and it is usually set to 1.

3.5. Optimizing with MAFS

The optimization of the feature selection can be done with the help of memetic algorithm. Memetic frame work is the hybridization of wrapper and filter feature selection models. The goal of MAFS is to improve the classification performance accelerate to identify the important feature subset. It out performs recent existing methods in terms of classification accuracy. Optimization of feature selection can be performed for getting better precision, recall and F-Value for the given dataset

4. EXPERIMENTAL RESULT

For our experiments, we use the benchmark KDD '99 intrusion data set[4] . The data set contains about five million connection records as the training data and about two million connection records as the test data.

We use the WEKA tool to perform preprocessing operation. The preprocessing of dataset will reduce the features by dimensionality reduction operation. The preprocessing of data set is shown fig. 4.

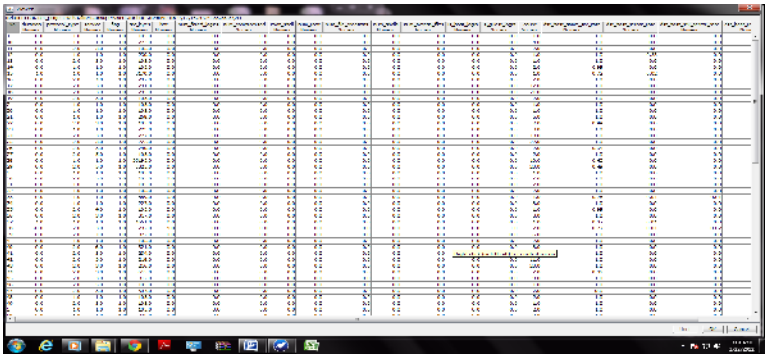

Figure 4. Preprocessing

We develop the java source for implementing conditional random fields. This module is separated out into two phases called Training and Testing. Training uses the Signature based approach. So, the input for the Training phase is randomly selected packets from the dataset. Testing used the Anomaly based approach. So, the Input for the testing phase is data collected from the current environment.

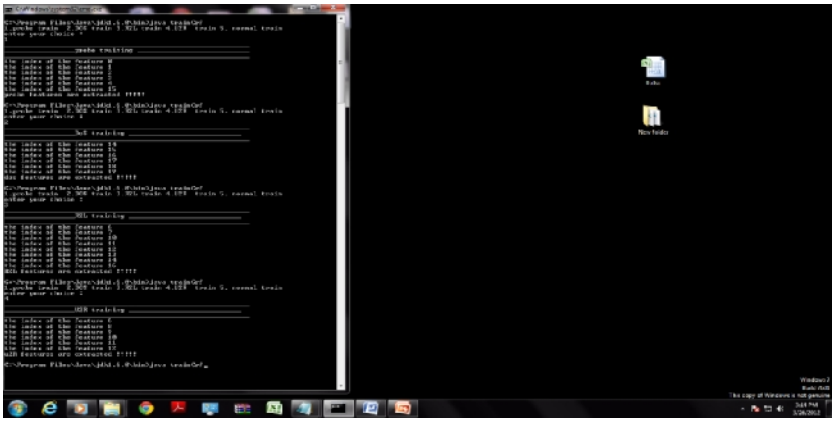

Figure 5 . Training

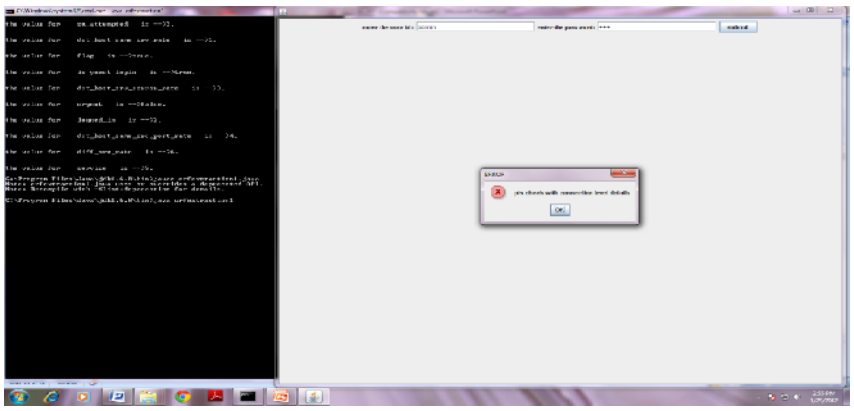

Figure 6. Testing

The Figure 5 and 6 show the training and testing phase of the system. The Result of this module is feature to be extracted from the dataset. Those extracted features are belongs to the each attack groups and they are stored in separate files.

The java source is written for implementing layered approach. The fig 7, 8,9,10 are shown the layering of each attack group. Each layer is responsible for each attack group. Fig 11 shows the optimization with MAFS algorithm.

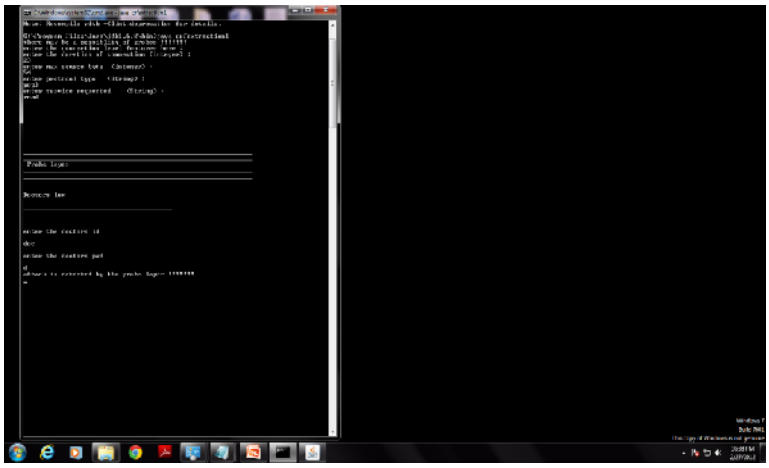

Figure 7. Probe layer

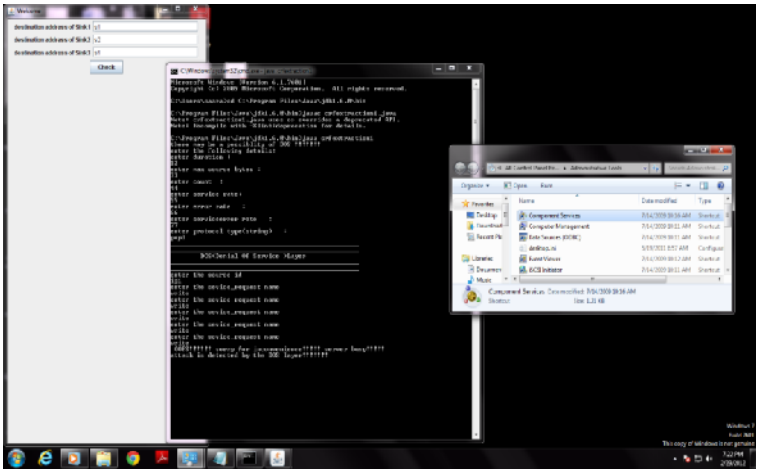

Figure 8. DoS layer

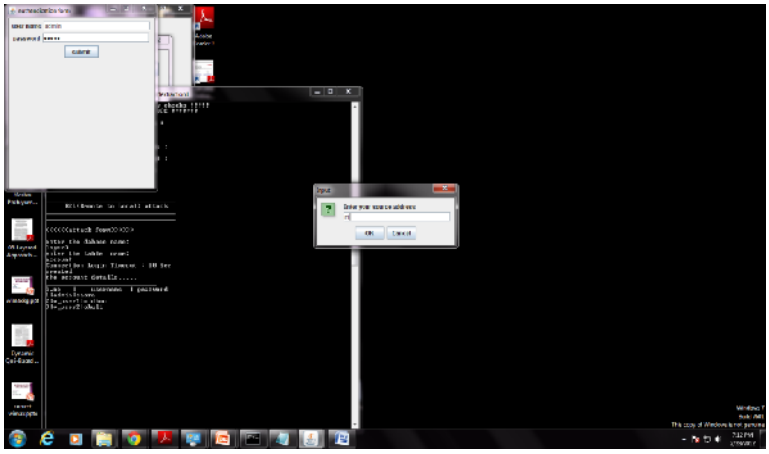

Figure 9. R2L layer

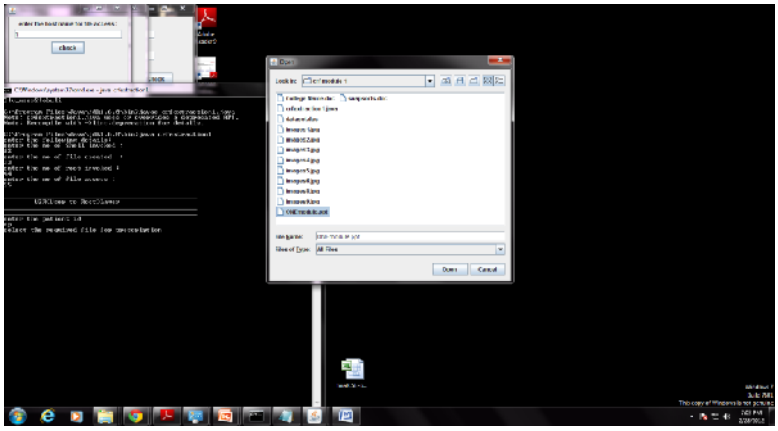

Figure 10. U2R layer

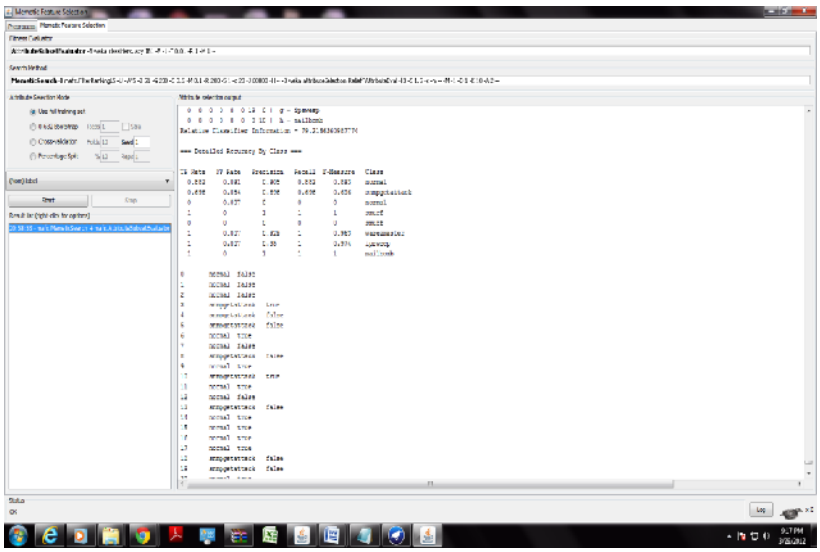

Figure 11. Optimization using MAFS

Table 1 represents the performance of system using LCRF approach based on precision, recall, f-value and probability of attack detection measures.

Table 1. Results based on LCRFs

Attack group	Precision	Recall	F-value	Prob. Of Attack detection
Probe	1	0.8	0.888889	0.8
DoS	0.8	0.8	0.800002	0.16
R2L	0.8	0.66667	0.72727272	0.032
U2R	0.5714285	0.5714285	0.5714285	0.0064

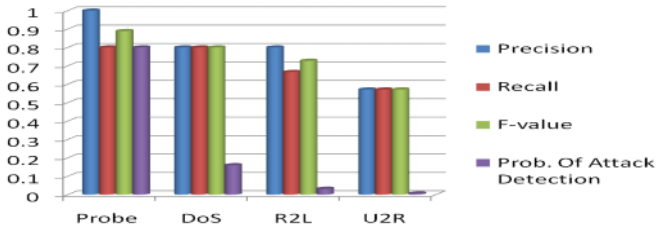

figure 12. Attack detection by various measures using LCRFs

Table 2 represents the performance of the system using optimized LCRF i.e LCRF with MAFS will result in better precision, recall, f-value and probability of attack detection values. Hence our

approach has the better method for attack detection when compare to all other well known methods

Table 2. Results based on after optimizing with MAFS

Attack group	Precision	Recall	F-value	Prob. Of Attack detection
Probe	0.95	1.0	0.974	1.0
DoS	1.0	1.0	1.0	1.0
R2L	0.929	1.0	0.963	1.0
U2R	0.696	0.696	0.696	0.696

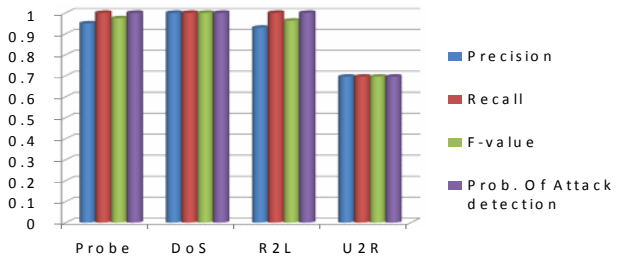

figure 13. Attack detection by various measures using LCRF with MAFS

The attack detection of the each attack each groups is calculated by various measures using both normal and optimized methods . Based on these measures we can say the probability of attack existence in the environment. These measures are shown in the fig. 12,13

5. CONCLUSION

In this paper, we have focused the two main problem of Accuracy and Efficiency for constructing robust and efficient intrusion detection systems. CRFs are proven to be successful framework for improving the attack detection accuracy rate and decreasing the FAR layered approach is used for high efficiency . We compared our integrated approach with some well-known methods and it is to be found that most of the present methods for intrusion detection will fail to reliably detect some of the network attacks (R2L and U2R attacks). Our system can help in identifying an attack once it is detected at a particular layer, which expedites the intrusion response mechanism, thus minimizing the impact of an attack. We showed that our system is robust to noise and performs better than any other compared system even when the training data is noisy .This can further be extended to implement pipelining of layers in multicore processors, which is likely to result in very high performance

REFERENCES

[1] Wenke Lee Salvatore J. Stolfo and Kui W. Mok " A Data Mining Framework for Building Intrusion Detection Models" submitted to the 1999 IEEE Symposium on Security and Privacy.

[2] Witcha Chimphlee and .Dr.Abdul Hanan Abdullah" Unsupervised Anomaly Detection with Unlabeled Data Using Clustering" Proceedings of the Postgraduate Annual Research Seminar 2005

[3] N.B. Amor, S. Benferhat, and Z. Elouedi, "Naive Bayes vs. Decision Trees in Intrusion Detection Systems," Proc. ACM Symp. Applied Computing (SAC '04), pp. 420-424, 2004

[4] Andrew Honig, Andrew Howard, Eleazar Eskin, Salvatore J. Stolfo; "Adaptive Model Generation: An Architecture for the Deployment of Data Mining-based Intrusion Detection Systems;" Data Mining for Security Applications; Kluwer; 2002

[5] H. Debar, M. Becke, and D. Siboni, "A Neural Network Component for an Intrusion Detection System," Proc. IEEE Symp.

[6] Kapil Kumar Gupta, Baikunth Nath, Kotagiri Ramamohanarao." Network Security Framework" submitted to International Journal of Computer Science and Network Security (IJCSNS),vol 6(7B), pages 151 - 157, 2006

[7] KDD'99 cup dataset

[8] J. Lafferty, A. McCallum, and F. Pereira, "Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data," Proc. 18th Int'l Conf. Machine Learning (ICML '01), pp. 282-289, 2001.

[9] C. Sutton and A. McCallum, "An Introduction to Conditional Random Fields for Relational Learning," Introduction to Statistical Relational Learning, 2006.

[11] Kapil Kumar Gupta, Baikunth Nath, and Kotagiri Ramamohanarao "Robust Application Intrusion Detection using User Session Modeling "Submitted to the ACM Transactions on Information and Systems Security (TISSEC).

[12] E. Tombini, H. Debar, L. Me, and M. Ducasse, "A Serial Combination of Anomaly and Misuse IDSes Applied to HTTP Traffic," Proc. 20th Ann. Computer Security Applications Conf. (ACSAC '04), pp. 428-437, 2004.

[13] Kapil Kumar Gupta, Baikunth Nath, and Ramamohanarao Kotagiri "Layered Approach Using Conditional Random Fields for Intrusion Detection" IEEE Transactions on dependable and secure Computing, vol. 5, no. 4, october-december 2008.

[14] Autonomous Agents for Intrusion Detection, http://www.cerias. purdue.edu/research/aafid/, 2010.

Authors

Mr. C.Saravanan is currently pursuing Bachelor of Engineering degree in Computer Science and engineering under Anna University, Coimbatore, India. His areas of Interests are Network security and Data mining

Mr. M.V Shivsankar is currently pursuing Bachelor of Engineering degree in Computer Science and engineering under Anna University, Coimbatore, India. His areas of Interest is Network security

Prof. P.Tamije Selvy received B.Tech (CSE), M.Tech (CSE) in 1996 and 1998 respectively from Pondicherry university. Since 1999, she has been working as faculty in reputed Engineering Colleges. At Present, she is working as Assistant Professor(SG) in the department of Computer Science & Engineering, Sri Krishna College of Technology, Coimbatore. Presently She is doing Ph.D under Anna University, Chennai. Her Research interests include Image Processing, Data Mining, Pattern Recognition and Artificial Intelligence.