Adaptive Layered Approach using C5.0 Decision Tree for Intrusion Detection Systems (ALIDS)

Sherif M. Badr, Ph. D.
College of Computer science
Modern Academy, Cairo, Egypt

ABSTRACT

Intrusion Detection System (IDS) is one of a crucial issue and a major research problem in network security. This work, An Adaptive multi-Layer Intrusion Detection System (ALIDS) is designed and developed to achieve high efficiency, scalability, flexibility and improve the detection and classification rate accuracy. We apply C5 decision tree on our model. Our experimental results showed that the proposed ALIDS model with different order of training classes enhances the accuracy of U2R and R2L.

Keywords-component; network intrusion detection; Decision Tree.

1. INTRODUCTION

The rapid development and expansion of World Wide Web and local network systems have changed the computing world in the last decade. Therefore, network security needs to be carefully concerned to implement various systems to monitor data flow in computer networks. These systems are generally referred to as Intrusion Detection Systems (IDSs).

Intrusion Detection Systems (IDS) have become a critical technology to help protect systems from intruders by collecting and analyzing network data, to determine whether there is malicious network traffic.

Intrusion detection and prevention systems (IDPS) are primarily focused on identifying possible incidents, logging information about them, attempting to stop them, and reporting them to security administrators. In addition, organizations use IDPSs for other purposes, such as identifying problems with security policies, documenting existing threats, and deterring individuals from violating security policies. IDPSs have become a necessary addition to the security infrastructure of nearly every organization [1].

Intrusion detection systems are classified as network based, host based, or application based depending on their mode of deployment and data used for analysis. Additionally, intrusion detection systems can also be classified as signature based or anomaly based depending upon the attack detection method. The signature-based systems are trained by extracting specific patterns (or signatures) from previously known attacks while the anomaly-based systems learn from the normal data collected when there is no anomalous activity [2].

Another approach for detecting intrusions is to consider both the normal and the known anomalous patterns for training a system and then performing classification on the test data. Such a system incorporates the advantages of both the signature-based and the anomaly-based systems and is known as the Hybrid System. Hybrid systems can be very efficient, subject to the classification method used, and can also be used to label unseen or new instances as they assign one of the known classes to every test instance. This is possible because during training the system learns features from all the classes. The only concern with the hybrid method is the availability of labeled data. However, data requirement is also a concern for the signature-based and the anomaly-based systems as they require completely anomalous and attack free data, respectively, which are not easy to ensure [3].

This work aims to design and develop security architecture (intrusion detection and system) for computer networks that is capable of detecting known and unknown attacks and designing a model flexible to any situation desired to be implemented.

2. PREVIOUS WORK:

The purpose of IDS is to help computer systems with how to discover attacks and the IDS is collecting information from several different sources within the computer systems and networks and compares this information with preexisting patterns of discrimination as to whether there are attacks or weaknesses [4].

There are four major categories of networking attacks. Every attack on a network can be placed into one of these groupings [5].

1) Denial of Service Attack (DOS): is an attack in which the attacker makes some computing or memory resource too busy or too full to handle legitimate requests, or denies\ legitimate users access to a machine.

2) User to Root Attack (U2R): is a class of exploit in which the attacker starts out with access to a normal user account on the system (perhaps gained by sniffing passwords, a dictionary attack, or social engineering) and is able to exploit some vulnerability to gain root access to the system.

3) Remote to Local Attack (R2L): occurs when an attacker who has the ability to send packets to a machine over a network but who does not have an account on that machine exploits some vulnerability to gain local access as a user of that machine.

4) Probing Attack: is an attempt to gather information about a network of computers for the apparent purpose of circumventing its security controls

Decision Trees (DT) have also been used for intrusion detection [6]. Decision Tree is very powerful and popular machine learning algorithm for decision-making and classification problems. It has been used in many real life applications like medical diagnosis, radar signal classification, weather prediction, credit approval, and fraud detection etc [7]. The decision tree is a simple if then else rules but it is a very powerful classifier and proved to have a high detection rate. They are used to classify data with common attributes. Each decision tree represents a rule which categorizes data according to these attributes. A decision tree has three main components: nodes, leaves, and edges. Each decision tree represents a rule set, which categorizes data according to the attributes of dataset. The DT building algorithms may initially build the tree and then prune it for more effective classification. [8].

C5.0 is an extension of C4.5 which is one of the most popular inductive learning tools originally proposed by J.R. Quinlan [8]. C4.5 is a classic decision tree algorithm which is an extension of ID3 attribute-based machine learning system (Quinlan, 1993).

C5.0 can deal with missing attributes by giving the missing attribute the value that is most common for other instances at the same node. Or, the algorithm could make probabilistic calculations based on other instances to assign the value. C5.0 handles continuous-valued functions by dividing them into a set of discrete valued functions. This can be repeated at each step of the algorithm to make the divisions that yield the largest information gain.

C5.0 supports boosting of decision trees. Boosting is a technique for generating and combining multiple classifiers to give improved predictive accuracy. By this process error rate is reduced on some datasets. C5.0 incorporates variable misclassification costs. Algorithm allows a separate cost to be defined for each predicted/actual class pair; if this option is used, C5.0 then constructs classifiers to minimize expected misclassification costs rather than error rates. [9]

This Paper is the enhancement of reference [10].

In [10], they applied for each layer a fixed attack category but ALIDS is flexible to any combination of classes desired to be implemented to get higher classification rate, the proposed model as shown in figure 1.

3. THE NEW PROPOSED MODEL ALIDS:

The proposed ALIDS has the capability of classifying network intruders into two stages. The first stage classifies the network records to either normal or attack. The second stage consists of four sequential Layers which can identify four categories / classes and their attack type. The data is entered in the first stage which identifies if this record is a normal record or attack. If the input record is identified as an attack then the module would raise a flag to the administrator that the coming record is an attack then the module inputs this record to the second stage which consists of four sequential Layers, one for each class type (DOS, Probe, U2R, and R2L). Each Layer is responsible for identifying the attack type of coming record according to its class type. Else the attack passes through the next layer, the new proposed model [13] as shown in figure 2.

If attack record couldn't be classified in the four layers, it will be labeled as unknown attacks.

If the attack type or category of the second stage is misclassified then at least the admin was notified that this record is malicious after the first stage network and the records which identified as unknown attack can be relabeled by the admin

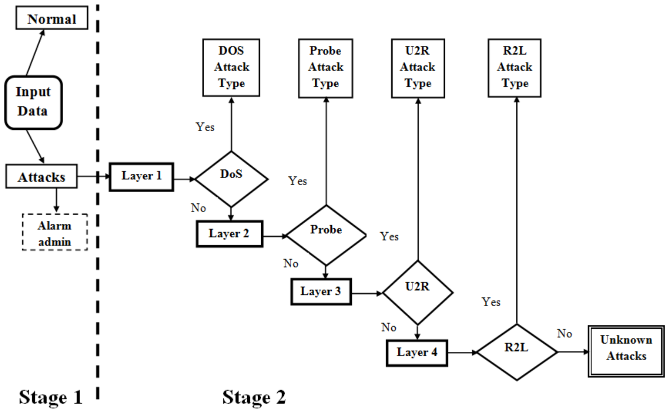

Figure 1 the Layered Intrusion Detection System

Algorithm:

Step 1: Each input record will be detected if either normal or attack in the first stage.

Step 2: If the input record is identified as an attack, it will be pass through the second stage, and the administrator will be alarmed with the suspicious record.

Step 3: Separately perform Gain Ratio Feature selection for each layer (DOS, Probe, U2R, and R2L).

Step 4: Train each layer with three machine learning techniques (C5, MLP, and Naïve Bayes).

Step 5: In each Layer, the attack records are tested; if it is categorized correctly then its attack class/type will be identified.

Step 6: if the attack record pass to the next layer then it couldn't be classified in previous layer.

Step 7: Records that couldn't be classified in any layer then it will be classified as unknown attack and will be relabeled by the administrator.

4. EXPERIMENTS AND RESULTS:

4.1 Input Data Description:

KDDCUP'99 is the mostly widely used data set for the evaluation of these systems. The KDD Cup 1999 uses a version of the data on which the 1998 DARPA Intrusion Detection Evaluation Program was performed. They set up environment to acquire raw TCP/IP dump data for a local area network (LAN)

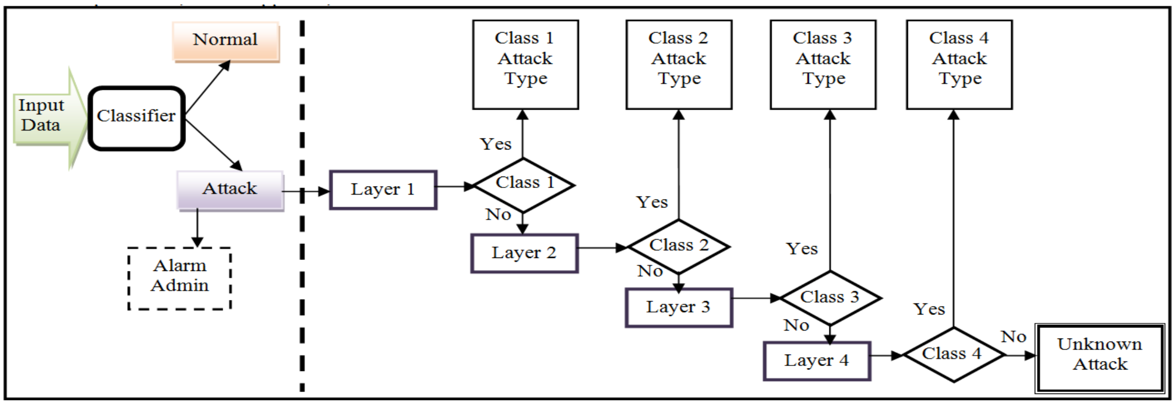

Figure 2 The new Proposed ALIDS

simulating a typical U.S. Air Force LAN.

There are some inherent problems in the KDDCUP'99 data set [11], which is widely used as one of the few publicly available data sets for network-based anomaly detection systems

The data in the experiment is acquired from the NSLKDD dataset which consists of selected records of the complete KDD data set and does not suffer from mentioned shortcomings by removing all the repeated records in the entire KDD train and test set, and kept only one copy of each record [5]. Although, the proposed data set still suffers from some of the problems and may not be a perfect representative of existing real networks, because of the lack of public data sets for network-based IDSs, but still it can be applied as an effective benchmark data set to help researchers compare different intrusion detection methods. The NSL-KDD dataset is available at [12].

We used attacks from the four classes to check the ability of the intrusion detection system to identify attacks from different categories.

The ALIDS is examined by applying new attacks on the testing dataset. The sample dataset contains 83644 record for training (40000 normal and 43644 for attacks) and 19784 for testing (9647 normal, 6935 for known attacks and 3202 for unknown attacks).

Table 1 The Possible Sequence Combination of Layers

Experiment Number	Attack Type Classification Sequence
1	DOS-Probe-R2L-U2R
2	DOS-Probe-U2R-R2L
3	DOS- R2L- Probe -U2R
4	DOS- R2L-U2R- Probe
5	DOS-U2R- Probe- R2L
6	DOS-U2R- R2L- Probe
7	Probe- DOS -R2L-U2R
8	Probe- DOS -U2R- R2L
9	Probe- R2L- DOS -U2R
10	Probe- R2L-U2R- DOS
11	Probe-U2R- DOS- R2L
12	Probe-U2R- R2L- DOS
13	R2L -DOS-Probe -U2R
14	R2L -DOS -U2R- Probe
15	R2L - Probe -DOS -U2R
16	R2L - Probe -U2R-DOS
17	R2L -U2R-DOS- Probe
18	R2L -U2R -Probe- DOS
19	U2R -DOS-Probe-R2L
20	U2R -DOS -R2L- Probe
21	U2R - Probe -DOS -R2L
22	U2R - Probe -R2L- DOS
23	U2R-R2L- DOS- Probe
24	U2R-R2L- Probe- DOS

4.2. Data Sets:

The sequences of layers were swapped to see how it will affect in the accuracy of each layer.

The experimental works use 24 possible sequence combinations of Layers as shown in table 1.

4.3. Experimental Results:

Actually each combination was performed by 4 experiments. For example in the experiment number 1, after 1st stage classifier, the attacks connection start with DOS attack type classifier as class 1 attack type, then Probe attack type classifier as class 2 attack type, then U2R attack type classifier as class 3 attack type, finally R2L attack type classifier as class 4 attack type, finally the connection classified as unknown attack.

This means 94 experiments for the 24 model that were performed, table 2 summeraze all the resultsof the 96 experiments.

All possible combinations of attacks classes can be categorized in one of the following groups:

A) First Group:

First six combinations are:

1) *DOS-Probe-U2R-R2L*
2) *DOS-Probe-R2L-U2R*
3) *DOS- R2L- Probe -U2R*
4) *DOS- R2L-U2R- Probe*
5) DOS-U2R- Probe- R2L
6) DOS-U2R- R2L- Probe

Chart 1 shows the results of the first group.

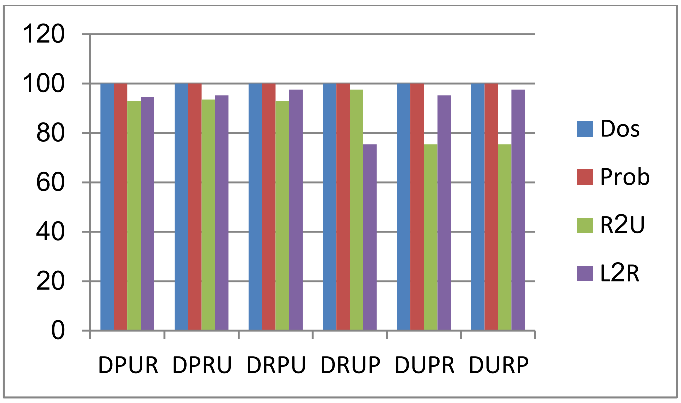

Chart 1: the results of the first group.

B) Second Group:

Second six combinations are:

7) Probe- DOS -R2L-U2R

8) Probe- DOS -U2R- R2L

9) Probe- R2L- DOS -U2R

10) Probe- R2L-U2R- DOS

11) Probe-U2R- DOS- R2L

12) Probe-U2R- R2L- DOS

Chart 2 shows the results of the second group.

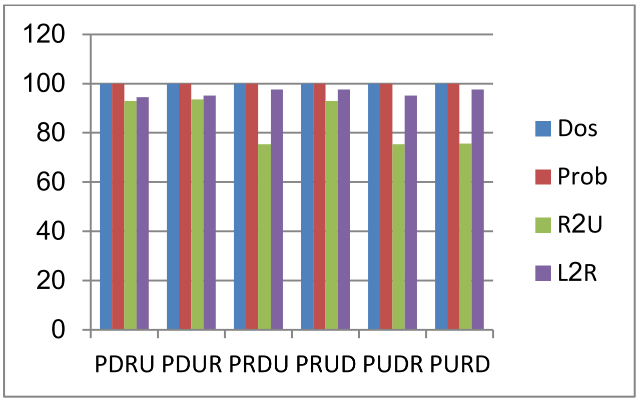

Chart 2: the results of the second group.

C) Third Group:

Third possible combinations are:

13) R2L -DOS-Probe -U2R

14) R2L -DOS -U2R- Probe

15) R2L - Probe -DOS -U2R

16) R2L - Probe -U2R-DOS

17) R2L -U2R-DOS- Probe

18) R2L -U2R -Probe- DOS

Chart 3 shows the results of the third group.

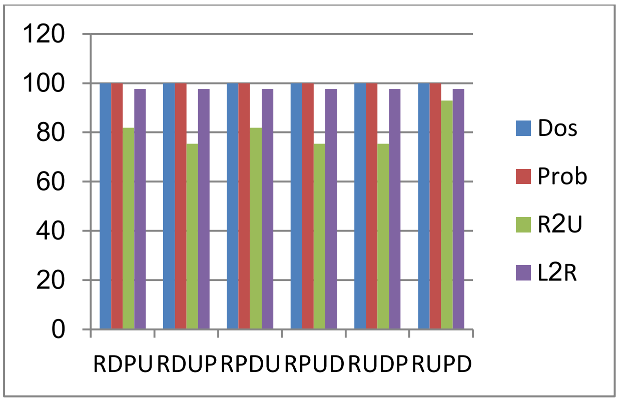

Chart 3: the results of the third group.

D) Fourth Group:

Fourth and last possible combinations are:

19) U2R -DOS-Probe-R2L

20) U2R -DOS -R2L- Probe

21) U2R - Probe -DOS -R2L

22) U2R - Probe -R2L- DOS

23) U2R-R2L- DOS- Probe

24) U2R-R2L- Probe- DOS

Chart 4 shows the results of the fourth group.

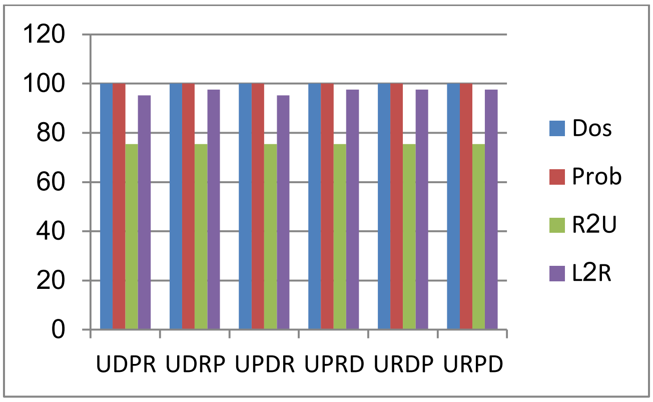

Chart 4: the results of the fourth group.

5. CONCLUSION AND FUTURE WORK:

In this paper, ALIDS model is enhanced by not specifying the layer class type so it will be more flexible and adaptive in any environment. Then we applied different possible combination sequence of attack categories on the proposed ALIDS.

The proposed model ALIDS with fixed sequence enhances the accuracy of all categories (DOS-Probe-U2R-R2L).

Our experimental results showed that the proposed ALIDS model with different order of training classes enhances the accuracy specially of U2R and R2L.

The experimental results showed that ALIDS takes less training computations because each layer act as a filters that classifies the attacks of each layer category which eliminate the need of further processing at subsequent layers.

ALIDS is flexible to any combination of classes desired to be implemented.

Table 2 Classification Rate of (ALIDS) Possible Combination

Experiment Number	Correct Classification Rate			
	Layer 1	Layer 2	Layer 3	Layer 4
1	100	100	94.53	92.86
2	100	100	93.58	95.18
3	100	97.6	100	92.86

4	100	97.6	75.41	100
5	100	75.41	100	95.18
6	100	75.41	97.6	100
7	100	100	94.53	92.86
8	100	100	93.58	95.18
9	100	97.6	100	75.41
10	100	97.6	92.86	100
11	100	75.41	100	95.18
12	100	75.41	97.6	100
13	97.6	100	100	81.9
14	97.6	100	75.41	100
15	97.6	100	100	81.9
16	97.6	100	75.41	100
17	97.6	75.41	100	100
18	97.6	92.86	100	100
19	75.41	100	100	95.18
20	75.41	100	97.6	100
21	75.41	100	100	95.18
22	75.41	100	97.6	100
23	75.41	97.6	100	100
24	75.41	97.6	100	100

The experimental results show that DOS & Probe attacks has 100% classification rate at any layer sequence, while R2L has high classification rate at first layer (97.6 %), also U2R has high classification rate at third layer (93.58 %).

So the best combination sequence R2L, DOS, U2R, and Probe which get the highest classification rate for all attacks categories.

The training module can be retrained at any point of time which makes its implementation adaptive to any new environment and/or any new attacks in the network by notifying the network administrator. If the attack record was not detected at any layer, then it will be detected as unknown until it relabeled by the admin.

The Future work will be directed towards training each layer on separate computer in parallel which provides less training time. Also other Machine learning techniques can be used in our experiments for detecting more types of intrusions.

6. REFERENCES

[1] Naelah okasha, Abd El Fatah Hegazy, Sherif M. Badr, 2010. “Towards Ontology-Based Adaptive Multilevel Model for Intrusion Detection and Prevention System (AMIDPS)”, Egyptian science journal (ESC), Vol. 34, No. 5, September 2010.

[2] R. Bace and P. Mell, Intrusion Detection Systems, Computer Security Division, Information Technology Laboratory, Nat’l Inst. of Standards and Technology, 2001.

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

[4] Asmaa Shaker Ashoor, Prof. Sharad Gore,"Importance of Intrusion Detection System (IDS)", International Journal of Scientific & Engineering Research (IJSER), Volume 2, Issue 1, January-2011.

[5] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set,” Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.

[6] 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, and 2004.

[7] T. M. Mitchell. Machine Learning. McGraw Hill, 1997.

[8] Quinlan JR. "C4.5: programs for machine learning," Log Altos, CA: Morgan Kaufmann; 1993. SPSS. Clementine 12.0 modeling nodes. Chicago: SPSS; 2007.

[9] SPSS. Clementine 12.0 modeling nodes. Chicago: SPSS; 2007.

[10] Heba Ezzat Ibrahim, Sherif M. Badr and Mohamed A. Shaheen," Adaptive Layered Approach using Machine Learning Techniques with Gain ratio for Intrusion Detection Systems," International Journal of Computer Applications(IJCA), pp. 10-16 ,Volume 56, No.7, October 2012.

[11] KDD Cup 1999. Available on: http://kdd.ics.uci.edu/databases/kddcup 99/kddcup99.html, October 7002

[12] "NSL-KDD data set for network-based intrusion detection systems, "Available on: http://nsl.cs.unb.ca/NSL-KDD/, March 2009.

[13] Heba Ezzat Ibrahim, "Adaptive Layered Approach using Machine Learning Techniques for Intrusion Detection Systems", master thesis, Arab Academy for Science and Technology & Maritime Transport, Cairo, Jan 2013.