Network Intrusion Detection System Using C4.5 Algorithm

Saurabh Fegade[#1], Amey Bhadkamkar[#1], Kamlesh Karekar[#1], Jaikishan Jeshnani[#1], Vinayak Kachare[#2]
BE Students, Department of Computer Engineering, Vivekananda Education Society's Institute of Technology[1]
Professor, Department of Computer Engineering, Vivekananda Education Society's Institute of Technology[2]
saurabh.fegade@ves.ac.in,amey.bhadkamkar@ves.ac.in,kamlesh.karekar@ves.ac.in,jaikishan.jeshnani@ves.ac.in,
vinayak.kachare@ves.ac.in

***Abstract*— There is a great concern about the security of computer these days. The number of attacks has increased in a great number in the last few years, intrusion detection is the main source of information assurance. While firewalls can provide some protection, they fail to provide protection fully and they even need to be complemented with an intrusion detection system (IDS). A newer approach for Intrusion detection is data mining techniques.IDS system can be developed using individual algorithms like neural networks, clustering, classification, etc. The result of these systems is good detection rate and low false alarm rate. According to a recent study, cascading of multiple algorithms gives a way better performance than single algorithm. Single algorithm systems have a high alarm rate. Therefore, to solve this problem, a combination of different algorithms are required. In this research paper, we use the hybrid algorithm for developing the intrusion detection system. C4.5 Support Vector Machine (SVM) and Decision Tree combined to achieve high accuracy and diminish the false alarm rate. Intrusions can be classified into types like Normal, DOS, R2L and U2R.Intrusion detection with Decision trees and SVM were tested with benchmark standard NSL- KDD, which is the extended version of KDD Cup 1999 for intrusion detection (ID).**

Keywords— **Support Vector Machine, NSL- KDD, Intrusion Detection System (IDS), Decision Tree.**

I. Introduction

The system which assists in detecting actions that attempt to compromise integrity, availability of a systems resources and its confidentiality can be defined as intrusion detection. Mahoney defined six types of attacks [1]. They are client attacks, network attacks and root attacks viruses, worms, server attacks. Efficient security policies like firewalls, anti-virus software, or other mechanism exist. To detect these attacks it is difficult, since every system has its own weakness and flaws. That's why the IDS are very much significant in today's world and can detect the new attacks. The Aim of an IDS is to detect unnecessary attempts at accessing, disabling, and manipulating of computer system. This paper is divided into four main areas. The first section describes various techniques for intrusion detection. The second part gives an overlook of the various different approaches for intrusion detection. The third section represents the proposed system. Finally the results and comparative analysis are presented.
The two most popular of the varied techniques of IDS are:

A] Anomaly detection

This technique constructs models for normal behavior of system by using machine learning. Any minor change from this built model is considered to be an intrusion.
Advantage: It can detect attacks based on past history. Disadvantage: Difficulty in training systems for highly changing environment, high false positives rates.

B] Misuse/signature detection:

These methods apply the knowledge gathered about particular attacks. The IDS contains information about these attacks. Any action is considered normal for an attack that is not explicitly recognized. A continuously updated data store is usually used to keep records of the signatures of known attacks. The technique of intrusion detection is similar to the way anti-virus software operates.
Advantages: low false positive rates.
Disadvantages: Need for updating the signatures, not able to detect novel attacks.
The intrusion detection system is classified into two main sections- Host-based and Network-based [2] on the basis of detecting a data target.

A] Host Based IDS:

The data is collected from records of various host activities, including system logs, application programs, audit record of operation system, information, and so on.
Advantages: It enables higher visibility of behavior of each application that is currently running on the host machine.
Disadvantage: IDS is needed for each and every machine but if attacker takes over machine can change its IDS binaries and modify audit logs.

B] Network Based IDS (NIDS):

Collection of audit data from the current network traffic, such as: Internet Packets.

Advantages: single NIDS is needed, which can protect many hosts and look for widespread matching patterns of activities in and around the network.
Disadvantages: All attacks do not arrive from the network. High working speed links are required to monitor, record and process huge amounts of traffic on the network.

II. INTRUSION DETECTION APPROACHES

Intrusion detection system uses a variety of approaches. A brief review of different approaches is analyzed below which are considered for the development of intrusion detection systems using C4.5 algorithm and SVM [3].

A. Statistical approach:

This approach involves theoretical comparison of specific actions based on a pre-initialized set of criterion. The statistical block model tests this collected data for attack analysis. This is very tedious and time consuming work with bulks of data.

B. Rule based approach:

The rule based approach works on a set of “if-then” condition rules to specify the attacks. Every rule is mapped with a specific operation in the system. The intrusion detection mechanism continuously monitors, rules that are drafted in the audit record. In case the required conditions of a rule are satisfied by user activity the specified operation is executed. The flaws in this approach were that it was unable to detect new intrusion. This approach requires a continuous update of rules which is time taking process.

C. Expert System approach:

A system of software or hardware/software that is able to competently showcase a specific task performed by the human expert is an expert system. The major limitation of Expert Systems is it requires frequent and major updates by a System Admin. Lack of maintenance or update is the drawback of this approach.

D. Pattern recognition approach:

A series of penetration scenarios are encoded into the system. This approach is effective in minimizing the need to review a large amount of sample data. This approach cannot detect any new attacks

III.PROPOSED SYSTEM

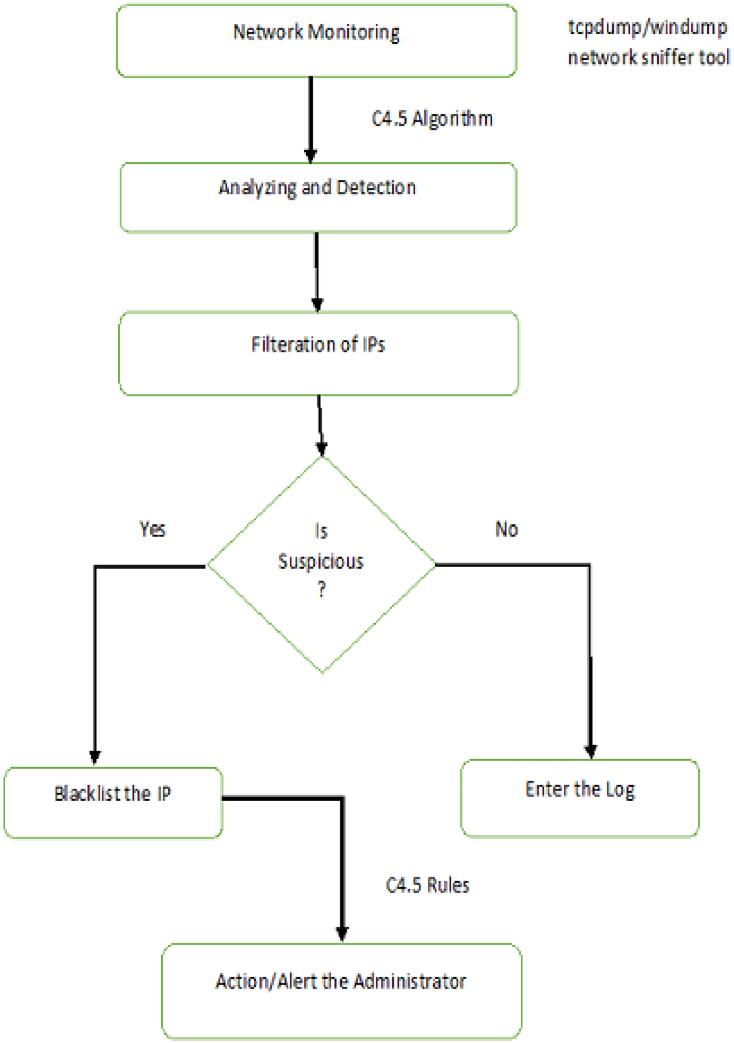


Fig.1 System Flowchart.

Explanation:
The system works in the following steps:-
1-The network has the IDS installed on the server. The server is linked with s database which will contain information about the various monitored and analyzed information.
2-The heart of the IDS consists of our packet sniffing tools like tcpdump or Wireshark followed by packet logger and then our algorithm which decides whether the IPs are suspicious or there exists an intruder somewhere.
3-The alarm manager notifies about intrusion detection by an intruder that is, if the algorithm finds any suspicious activity the alarm is triggered. Now there are two scenarios- 1. Detection alarm 2. False Alarm
4-As soon as the alarm is triggered the system, the intruder is traced that is, the suspicious IP is traced and then by selecting a suitable system protection strategy actions are taken accordingly.

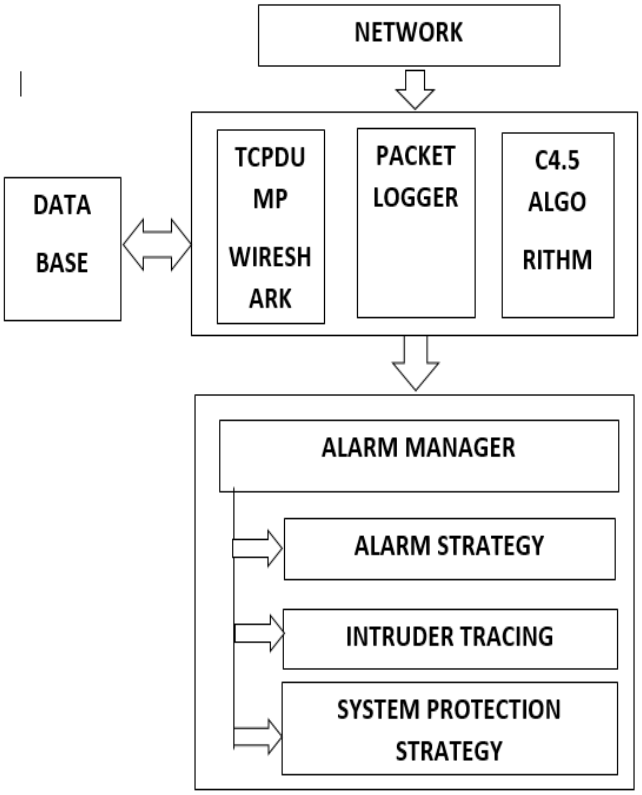


Fig.2 System Design.

The system is designed to analyze a network and detect intruders and malicious attacks in the network.

The inputs to the system are the various IP addresses from a network. The network analyzing tool, i.e Wireshark will capture the packets and analyze the IP addresses and will filter out IPs such as malicious, sniffing IPs etc .It will further forward the list of IP details to the C4.5 algorithm to process it further.

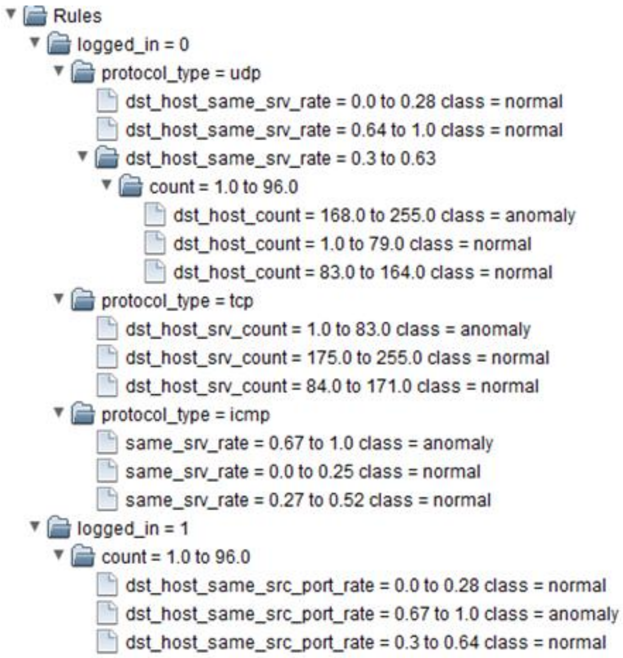
Packet Logger is the storage for keeping the logs of network traffic and IPs present over a network.C4.5 is a decision tree algorithm which will process the data along with a given a rule set which will give an output of whether the IPs are intruder IPs or not.

Based on the decision made by C4.5 algorithm, the alarm manager or the network administrator will notify of a network intrusion and take an appropriate action. The necessary action like warning the selected IP or blacklisting it.
The address can be either blocked temporarily or permanently.

V. EXPERIMAENTAL RESULTS

The results of the implementation of the algorithm are shown below. The NSL-KDD Intrusion detection contest data is used in our experiments. First; the algorithms are trained with the preprocessed dataset. Dataset was separated into two parts. Through the first part, the model was prepared and with the remaining of the dataset, the model was tested. CFS algorithm was used for feature selection. Out of 42 features only 12 features were selected that are count, dst_host_count, dst_host_same_srv_rate, same_srv_rate, dst_host_srv_count, dst_host_same_src_port_rate, protocol_type, serror_rate, dst_host_srv_serror_rate, dst_host_serror_rate, srv_ serror_rate and logged_in.

Number of leaves produced by C4.5 algorithm is shown in fig 3.



```
▼ Rules
  ▼ logged_in = 0
    ▼ protocol_type = udp
        dst_host_same_srv_rate = 0.0 to 0.28 class = normal
        dst_host_same_srv_rate = 0.64 to 1.0 class = normal
      ▼ dst_host_same_srv_rate = 0.3 to 0.63
        ▼ count = 1.0 to 96.0
            dst_host_count = 168.0 to 255.0 class = anomaly
            dst_host_count = 1.0 to 79.0 class = normal
            dst_host_count = 83.0 to 164.0 class = normal
    ▼ protocol_type = tcp
        dst_host_srv_count = 1.0 to 83.0 class = anomaly
        dst_host_srv_count = 175.0 to 255.0 class = normal
        dst_host_srv_count = 84.0 to 171.0 class = normal
    ▼ protocol_type = icmp
        same_srv_rate = 0.67 to 1.0 class = anomaly
        same_srv_rate = 0.0 to 0.25 class = normal
        same_srv_rate = 0.27 to 0.52 class = normal
  ▼ logged_in = 1
    ▼ count = 1.0 to 96.0
        dst_host_same_src_port_rate = 0.0 to 0.28 class = normal
        dst_host_same_src_port_rate = 0.67 to 1.0 class = anomaly
        dst_host_same_src_port_rate = 0.3 to 0.64 class = normal
```

Fig 3.Number of leaves produced by C4.5 algorithm

Detection of attack can be measured by following metrics:
False positive (FP): Or false alarm, Corresponds to the number of detected attacks but it is in fact normal.
False negative (FN): Corresponds to the number of detected normal instances but it is actually attack, in other words these attacks are the target of intrusion detection systems.
True positive (TP): Corresponds to the number of detected attacks and it is in fact attack.
True negative (TN): Corresponds to the number of detected normal instances and it is actually normal.
The accuracy of an intrusion detection system is measured regarding to detection rate and false alarm rate.
The efficiency of IDS System is calculated by using Following two terms:

Detection rate:
Detection rate refers to the percentage of detected attack Among all attack data, and is defined as follows:
Detection rate = TP/ (TP+TN) × 100

False alarm rate:
False alarm rate refers to the percentage of normal data which is wrongly recognized as attack, and is defined as Follows:
False alarm rate = FP/ (FP+TN) × 100

The results that we got for C4.5, SVM Algorithm for 1000 records are shown below in Table 1 with their corresponding values

Parameters	**C4.5**	**SVM**
Accuracy	95.7%	90.4%
False Alarm Rate	8.6	7.26
Detection Rate	97.68	88.26

Table 1.

The results that we got for C4.5, SVM Algorithm for 20% records are shown below in Table 2 with their corresponding values.

Parameters	**C4.5**	**SVM**
Accuracy	96.71%	93.47%
False Alarm Rate	8.12	5.61
Detection Rate	99.86	91.45

Table 2.

IV. CONCLUSION & FUTURE WORK

After the implementation we will be able to actually increase the detection rate and lessen false alarm rate of IDS by combination of two algorithms C4.5 Decision Tree and Support Vector Machine. Comparison is done using various parameters like False Alarm Rate, Accuracy, and Detection rate. Building an efficient and proactive intrusion detection model with good rate of accuracy and real-time performance are vital. However, other kinds of preprocessing techniques and data mining approaches like AI-artificial intelligence, neural network models may be implemented for a better detection rate in the future research in IDS System. A serious effort will be made in the future to differentiate types of attack into different categories like DOS, Probe, U2R and R2L. A more efficient feature selection algorithm can be used in future.

ACKNOWLEDGEMENT

We are working on this project under the guidance of our Institution Mentor Mr. Vinayak Kachare (Assistant Professor, VESIT). nd also thank the Computer Department of VESIT for extending their support.

REFERENCES

[1] M. Mahoney, Computer security: A survey of Attacks and Defenses, 2000. http://www.cs.fit.edu/~mmahoney/ids.html

[2] S. Wu, E. Yen. “Data mining-based intrusion detectors,” Elsevier computer Network, 2009.

[3] Prevention Systems (IDPS)”. Computer Security Resource Center (National Institute of Standards and Technology). February 2007.

[4] R. Graham, “FAQ: Network Intrusion Detection Systems”. March 21, 2000.

[5] Badr HSSINA, Abdel Karim MERBOUHA, A comparative study of decision tree ID3 and C4.5, (IJACSA) International Journal of Advanced Computer Science and Applications,

[6]Ms.Lata Jadhav1, Prof.C.M.Gaikwad2,,“Implementationof Intrusion Detection System Using GA”International Journal Of Innovative Research In Electrical, Electronics, Instrumentation And Control EngineeringVol. 2, Issue 7, July 2014