Analyze Different approaches for IDS using KDD 99 Data Set

Mr. Kamlesh Lahre
Asst. Professor CSE. Dept .
Dr. C.V. Raman University
Bilaspur(C.G.),India
lahrekamlesh@gmail.com

Mr. Tarun dhar Diwan
Asst. Professor CSE. Dept.
Dr. C.V. Raman University
Bilaspur(C.G.),India
taruncsit@gmail.com

Suresh Kumar Kashyap
M.Tech Research scholar
Dr. C.V. Raman University
Bilaspur(C.G.),India
s3.kashyap@gmail.com

Pooja Agrawal
Asst. Professor I.T. Dept
Dr. C.V. Raman University
Bilaspur(C.G.),India

Abstract: **the integrity, confidentiality, and availability of Network security is one of the challenging issue and so as Intrusion Detection system (IDS). IDS are an essential component of the network to be secured. Intrusion detection is the process of monitoring and analyzing the events occurring in a computer system in order to detect signs of security problems. Intrusion detection includes identifying a set of malicious actions that compromise information resources. Traditional methods for intrusion detection are based on extensive knowledge of signatures of known attacks. In the last three years, the networking revolution has finally come of age. More than ever before, we see that the Internet is changing computing, as we know it. The possibilities and opportunities are limitless; unfortunately, so too are the risks and chances of malicious intrusions There are two primary methods of monitoring these are signature-based and anomaly based. In this paper is to analyze different approaches of IDS. Some approach belongs to supervised method and some approach belongs to unsupervised method.**

Keywords: Firewall, IDS, AI, anomaly & misuse, DOS, R2L, NID.

*

I. INTRODUCTION

Computer security can be very complex and may be very confusing to many people. It can even be a controversial subject. Network administrators like to believe that their network is secure and those who break into networks may like to believe that they can break into any network. Intrusion detection is therefore needed as another wall to protect computer systems. The elements central to intrusion detection are: *resources* to be protected in a target system, i.e., user accounts, file systems, system kernels, etc; *models* that characterize the "normal" or "legitimate" behavior of these resources; *techniques* that compare the actual system activities with the established models, and identify those that are "abnormal" or "intrusive". It is very important that the security mechanisms of a system are designed so as to prevent unauthorized access to system resources and data. However, completely preventing breaches of security appear, at present, unrealistic. We can, however, try to detect these intrusion attempts so that action may be taken to repair the damage later. This field of research is called Intrusion Detection.

Intrusion detection techniques while often regarded as grossly experimental, the field of intrusion detection has matured a great deal to the point where it has secured a space in the network defense landscape alongside firewalls and virus protection systems. While the actual, the concept behind intrusion detection is a surprisingly implementations tend to be fairly complex, and often proprietary simple one: Inspect all network activity (both inbound and outbound) and identify suspicious patterns that could be evidence of a network or system attack.

Classification of Intrusion

Intrusions can be divided into 6 main types

- Attempted break-ins, which are detected by atypical behavior profiles or violations of security constraints.
- Masquerade attacks, which are detected by atypical behavior profiles or violations of security constraints.
- Penetration of the security control system, which are detected by monitoring for specific patterns of activity.
- Leakage, which is detected by atypical use of system resources.
- Denial of service, which is detected by atypical use of system resources.
- Malicious use, which is detected by atypical behavior profiles, violations of security constraints, or use of special privileges.

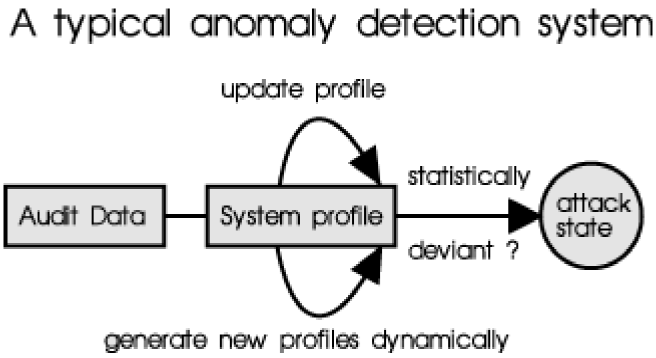
II.TYPES OF TECHNIQUES OF INTRUSION DETECTION SYSTEM

We can divide the techniques of intrusion detection into two main types.

Anomaly Detection:

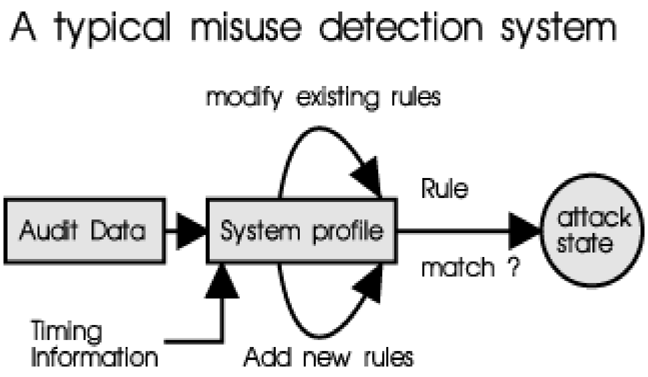
Anomaly detection techniques assume that all intrusive activities are necessarily anomalous. This means that if we could establish a "normal activity profile" for a system, we could, in theory, flag all system states varying from the established profile by statistically significant amounts as intrusion attempts. However, if we consider that the set of intrusive activities only intersects the set of anomalous activities instead of being exactly the same, we find a couple of interesting possibilities: (1) Anomalous activities that are not intrusive are flagged as intrusive. (2) Intrusive activities that are not anomalous result in false negatives (events are not flagged intrusive, though they actually are). This is a dangerous problem, and is far more serious than the problem of false positives.

The main issues in anomaly detection systems thus become the selection of threshold levels so that neither of the above 2 problems is unreasonably magnified, and the selection of features to monitor. Anomaly detection systems are also computationally expensive because of the overhead of keeping track of, and possibly updating several system profile metrics. Some systems based on this technique are discussed in Section 4 while a block diagram of a typical anomaly detection system is shown in Figure below.



Misuse Detection:

The concept behind misuse detection schemes is that there are ways to represent attacks in the form of a pattern or a signature so that even variations of the same attack can be detected. This means that these systems are not unlike virus detection systems -- they can detect many or all *known* attack patterns, but they are of little use for as yet unknown attack methods. An interesting point to note is that anomaly detection systems try to detect the complement of "bad" behavior. Misuse detection systems try to recognize known "bad" behavior. The main issues in misuse detection systems are how to write a signature that encompasses *all* possible variations of the pertinent attack, and how to write signatures that do not also match non-intrusive activity. A block diagram of a typical misuse detection system is shown in Figure below.



Advantages:

- Simplicity and nonintrusiveness (which translate into ease of deployment).

Disadvantages:

- Inspecting each packet on the wire is becoming increasingly more difficult with the recent advances in network and wireless technology in terms of complexity and speed.
- Most intrusion detection systems employ a combination of both techniques, and are often deployed on the network, on a specific host, or even on an application within a host.

Typically IDS has two types that are-

Network Based Intrusion Detection:

The most obvious location for an intrusion detection system is right on the segment being monitored. Network-based intrusion detectors insert themselves in the network just like any other device, except they promiscuously examine every packet they see on the wire.

Advantage:

- Network-based intrusion detection is straightforward to implement and deploy.

Disadvantage:

- Truly shared segments are rare nowadays, which means a single sniffer cannot be relied to monitor an entire subnet. Instead, detection systems must be integrated in the port of Ethernet switches (the ones that have visibility into all packets on the wire), which is not always feasible, even if such a port is available.

- The fact that a single intrusion detection system is servicing the entire segment makes it an easy target for a DoS attack. Such a system should not contain any user accounts other than the privileged (root/Administrator) user; host any unnecessary network services; offer any sort of interactive network access (console access only); or be hosted on an obscure, proprietary operating system.

Host Based Intrusion Detection

While network-based intrusion detectors are straightforward to deploy and maintain, there is a whole class of attacks closely coupled to the target system and extremely hard to fingerprint. These are the ones that exploit vulnerabilities particular to specific operating systems and application suites. Only host-based intrusion detection systems (the ones running as an application on a network-connected host) can correlate the complex array of system-specific parameters that make up the signature of a well-orchestrated attack.

Advantage:

The host-based approach is ideal for those high-availability servers that enterprises rely on for everyday business. The most prevalent advantage of the host-based approach is its ability to detect an inside job-that is, an incident where a lawful user is using local host resources in a manner that violates the company's security policy. This type of offense would be virtually impossible to unveil with a network-based intrusion detection system; because the user could have console access to the system, his or her actions would not even traverse the wire.

Disadvantage:

Not all is well in the world of host-based intrusion detection, however: Since these systems are closely tied to the operating system, they become yet one more application to maintain and migrate. This is a critical point in an environment where operating system levels are upgraded often, as the intrusion detection system must be kept up to date for it to work efficiently. Also, deploying host-based detectors alone will not protect your enterprise against basic, Network-layer DoS attacks (SYN flooding, ping of death, land attack, and so on). These limitations withstanding, host-based detection should be an integral part of your overall intrusion defense.

In this paper, we will look at different intrusion detection approaches these are:-

1.Artificial Neural Network Intrusion Detection System:- Some IDS designers exploit ANN as a pattern recognition technique. Pattern recognition can be implemented by using a feed-forward neural network that has been trained accordingly. ANN is one of the oldest systems that have been used for Intrusion Detection System (IDS), Artificial neural networks are models designed to simulate specific organic brain functions such as pattern recognition. They consist of many similar building blocks – neurons. It is eligible to distinguish three types of units or layers [2]:

1. Input layer – receives an input data from external resources. Neuron's output is after processing passed to next layer.
2. Hidden layer(s) – receives an input from neuron at adjacent layer. Output signals are passed to output layer or remain within the ANN.
3. Output layer – receives an input from adjacent hidden layer. Output signals are sent out of ANN to post-processing.

ANN is one of the most used techniques and has been successfully applied to intrusion According to different types of ANN, these techniques can be classified into the following three categories: supervised ANN-based intrusion detection, unsupervised ANN-based intrusion detection, and hybrid ANN-based intrusion detection. Supervised ANN applied to IDS mainly includes multi-layer feed-forward (MLFF) neural networks and recurrent neural networks. However, the main drawbacks of ANN-based IDS exist in two aspects: (1) lower detection precision, especially for low-frequent attacks, e.g., Remote to Local (R2L), User to Root (U2R).

An approach for a neural network based intrusion detection system, intended to classify the normal and attack patterns and the type of the attack, has been presented in this paper.

2.Self Organizing Map Intrusion Detection System:-The Self-Organizing Map is one of the most popular neural network models. It belongs to the category of competitive learning networks. The Self-Organizing Map is based on unsupervised learning, which means that no human intervention is needed during the learning and that little need to be known about the characteristics of the input data. We could, for example, use the SOM for clustering data without knowing the class memberships of the input data. The SOM can be used to detect features inherent to the problem and thus has also been called SOFM, the Self-Organizing Feature Map. Purely based on a hierarchy of self-organizing feature maps (SOMs), an approach to network intrusion detection is investigated. Our principle interest is to establish just how far such an approach can be taken in practice. To do so, the KDD benchmark data set from the International Knowledge Discovery and Data Mining Tools Competition is employed. Extensive analysis is conducted in order to assess the significance of the features employed, the partitioning of training data and the complexity of the architecture. Contributions that follow from such a holistic evaluation of the SOM include recognizing that (1) best performance is achieved using a two-layer SOM hierarchy, based on all 41-features from the KDD data set. (2) Only 40% of the original training data is sufficient for training purposes. (3) The 'Protocol' feature provides the basis for a switching parameter, thus supporting modular solutions to the detection problem. The ensuing detector provides false positive and detection rates of 1.38% and 90.4% under test conditions; where this represents the best performance to date of a detector based on an unsupervised learning algorithm.

3.Fuzzy logic Intrusion Detection System:-The fuzzy based network intrusion detection system. Intrusion detection system is increasingly a key part of system defence is used to identify abnormal activities in a computer system.Fuzzy systems have demonstrated their ability to solve different kinds of problems in various applications domains. In general, the traditional intrusion detection relies on the extensive knowledge of security experts, in particular, on their familiarity with the computer system to be protected.

Fuzzy systems based on fuzzy if-rules have been successfully used in many applications areas. Fuzzy if-then rules were traditionally gained from human experts. It is possible to develop an anomaly based intrusion detection system which detects the intrusion behaviour within a network.Recently, various methods have been suggested for automatically generating and adjusting fuzzy if-then rules without using the aid of human experts. The fraction of IDS over the total number of them that predicts a given event will determine whether such event is predicted or not. The performance obtained from the application of fuzzy thresholds over such fraction is compared with the corresponding crisp thresholds.

4.Support Vector Machine Intrusion Detection System:- **SVMs**, also **support vector networks** are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier.Support vector machines(SVM) is a learning technique which has been successfully applied in many application areas. Support Vector Machines (SVM) are the classifiers which were originally designed for binary classification. The classification applications can solve multi-class problems. Intrusion detection can be considered as two-class classification problem or multi-class classification problem. We used dataset from 1999 KDD intrusion detection contest. SVM are learning systems that use a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory. SVM IDS was learned with triaing set and tested with test sets to evaluate the performance of SVM IDS to the novel attacks. And we also evaluate the importance of each feature to improve the overall performance of IDS. The SVM is one of the most successful classification algorithms in the data mining area, but its long training time limits its use. Many applications, such as Data Mining and Bio-Informatics, require the processing of huge data sets. The results of experiments demonstrate that applying SVM in Intrusion Detection System can be an effective and efficient way for detecting intrusions. Self-organizing maps (SOM) and support vector machine have also been used as anomaly intrusion detectors.

III. IDS REQUIREMENTS

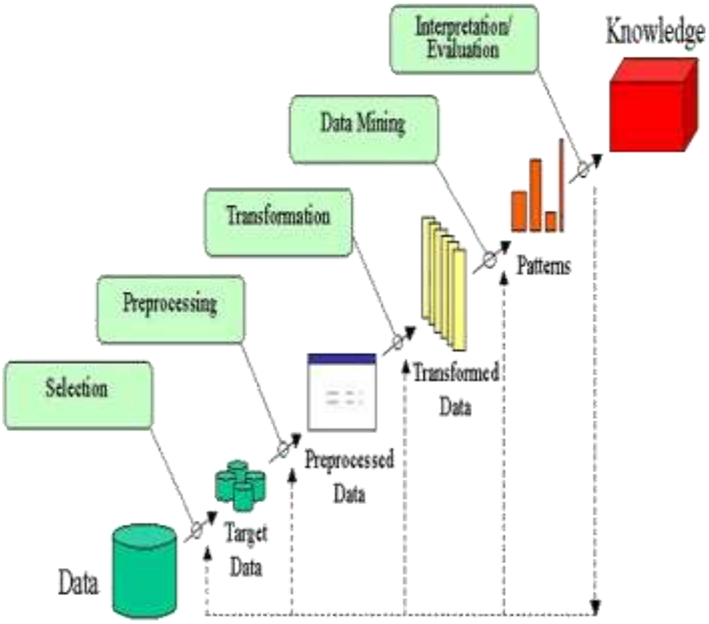
At least one past effort has identified desirable characteristics for an IDS. Regardless on what mechanisms an IDS is based, it must do the following:

- Run continuously without human supervision,
- Be fault tolerant and survivable,
- Resist subversion,
- Impose minimal overhead,
- Observe deviations from normal behavior
- Be easily tailored to a specific network
- Adapt to changes over time, and
- Be difficult to fool.

IV. INTRODUCTION OF KDD 99 DATA SET

The term *Knowledge Discovery in Databases*, or KDD for short, refers to the broad process of finding knowledge in data, and emphasizes the "high-level" application of particular data mining methods. It is of interest to researchers in machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition for expert systems, and data visualization. **KDD** refers to the overall process of discovering useful knowledge from data. It involves the evaluation and possibly interpretation of the patterns to make the decision of what qualifies as knowledge. It also includes the choice of encoding schemes, preprocessing, sampling, and projections of the data prior to the data mining step.

An Outline of the Steps of the KDD Process



The overall process of finding and interpreting patterns from data involves the repeated application of the following steps:

1. Developing an understanding of
    - the application domain
    - the relevant prior knowledge
    - the goals of the end-user
2. Creating a target data set: selecting a data set, or focusing on a subset of variables, or data samples, on which discovery is to be performed.
3. Data cleaning and preprocessing.
    - Removal of noise or outliers.
    - Collecting necessary information to model or account for noise.
    - Strategies for handling missing data fields.
    - Accounting for time sequence information and known changes.
4. Data reduction and projection.
    - Finding useful features to represent the data depending on the goal of the task.
    - Using dimensionality reduction or transformation methods to reduce the effective number of variables under consideration or to find invariant representations for the data.
5. Choosing the data mining task.
    - Deciding whether the goal of the KDD process is classification, regression, clustering, etc.
6. Choosing the data mining algorithm(s).
    - Selecting method(s) to be used for searching for patterns in the data.
    - Deciding which models and parameters may be appropriate.
    - Matching a particular data mining method with the overall criteria of the KDD process.
7. Data mining.
    - Searching for patterns of interest in a particular representational form or a set of such representations as classification rules or trees, regression, clustering, and so forth.
8. Interpreting mined patterns.
9. Consolidating discovered knowledge.

The unifying goal of the KDD process is to extract knowledge from data in the context of large databases.

Knowledge Discovery in Databases (KDD) is the automated discovery of patterns and relationships in large databases.

Characteristics and nature of KDD applications:-

• KDD operates on large data sets

• KDD data sets are large in terms of number of attributes and number of records

• KDD attempts to deal with real world problems and real world data;

• Usually accesses input data several times;

• Builds dynamic and recursive data structures Hash tables, Linked lists, and Trees;

• Size and access of the data structure is data dependent;

• Complex core routines;

• KDD process consists of a number of interacting, iterative stages involving various data manipulation;

• KDD process is explorative;

V. STRUCTURE OF DATA SET

The KDD 99 intrusion detection datasets are based on the 1998 DARPA initiative, which provides designers of intrusion detection systems (IDS) with a benchmark on which to evaluate different methodologies. To do so, a simulation is made of a factitious military network consisting of three ‘target’ machines running various operating systems and services. Additional three machines are then used to spoof different IP addresses to generate traffic. Finally, there is a sniffer that records all network traffic using the TCP dump format. The total simulated period is seven weeks. Normal connections are created to profile that expected in a military network and attacks fall into one of four categories: User to Root; Remote to Local; Denial of Service; and Probe.

• **Denial of Service (dos):** Attacker tries to prevent legitimate users from using a service.

• **Remote to Local (r2l):** Attacker does not have an account on the victim machine, hence tries to gain access.

• **User to Root (u2r):** Attacker has local access to the victim machine and tries to gain Super user privileges.

• **Probe:** Attacker tries to gain information about the target host.

In 1999, the original TCP dump files were preprocessed for utilization in the Intrusion Detection System benchmark of the International Knowledge Discovery and Data Mining Tools Competition. To do so, packet information in the TCP dump file is summarized into connections. Specifically, “a connection is a sequence of TCP packets starting and ending at some well defined times, between which data flows from a source IP address to a target IP address under some well defined protocol”. his process is completed using the Bro IDS, resulting in 41 features for each connection, Features are grouped into four categories:

- Basic Features: Basic features can be derived from packet headers without inspecting the payload.
- Content Features: Domain knowledge is used to assess the payload of the original TCP packets.
- Time-based Traffic Features: These features are designed to capture properties that mature over a 2 second temporal window.
- Host-based Traffic Features: Utilize a historical window estimated over the number of connections – in this case 100 – instead of time. Host based features are therefore designed to assess attacks, which span intervals longer than 2 seconds.

The KDD cup 99 data set we have given number to different types attack including normal attack as shown in table

Table Classification of dataset

Attack type	**Class**	**Group**	**Sub attacks types**
Normal	1	A	normal
DoS	3	B	smurf, teardrop, pod, back, land, apache2, udpstrom, mailbomb, processtable, neptune
Probe	4	C	ipsweep, portsweep, nmap, satan, saint, mscan
R2L	2	D	dictionary, ftp_write, guess_password, imap, named, sendmail, spy, xlock, xsnoop, snmpgetattack, httptunnel, worm, snmpguess, multihop, phf, wraezclient, wrazemaster
U2R	5	E	perl, ps, xterm, loadmodule, eject, buffer_overflow, sqlattack

VI. CONCLUSION

IDS come in a variety of “flavors” and approach the goal of detecting suspicious traffic in different ways. There are network based (NIDS) and host based (HIDS) intrusion detection systems.

This paper present to different approaches of IDS with KDD 99. Some approach belongs to supervised method and some approach belongs to unsupervised method. Many hybrid methods using different types of IDS Method. All these methods can be simulated using the Matlab and KDD99 dataset. Fuzzy rules will be identified by fuzzifying the definite rules .These rules will be fed to fuzzy system, which will classify the test data. It is decided to use KDD cup 99 dataset for evaluating the performance of the proposed system and the proposed method is effective in detecting various intrusions in computer networks. We use Support Vector Machines (SVM)for classification. The SVM is one of the most successful classification algorithms in the data mining area, but its long training time limits its use.

REFERENCES

1. Kumar.S “Classification and Detection of Computer Intrusion ” .
2. Sundaram A., "An Introduction to Intrusion Detection",http://www.acm.org/crossroads/xrds2-4/intrus.html
3. O. Depren, M. Topallar, E. Anarim and M. Kemal, An Intelligent Intrusion DetectionSystem (IDS) for Anomaly and Misuse Detection in Computer Networks, Expert Systemswith Applications, 29:713-722, 2005.
4. S. Peddabachigari, Ajith Abraham, C. Grosan, J. Thomas, “Modeling intrusion detection system using hybrid intelligent systems”, Journal of Network and Computer Applications, Volume 30, Issue 1, January 2007, Pages 114–132
5. M. Saniee Abadeh, J. Habibi, C. Lucas, “Intrusion detection using a fuzzy genetics-based learning algorithm”, Journal of Network and Computer Applications, Volume 30, Issue 1, January 2007,Pages 414-428.
6. Paul Innella Tetrad, “The Evolution of Intrusion Detection Systems”, Digital Integrity,LLC on November 16, 2001.
7. Harley Kozushko, “Intrusion Detection: Host-Based and Network-Based Intrusion Detection Systems”, on September 11, 2003.
8. http://kdd.ics.uci.edu/databases/kddcup99/kddcup.testdata.unlabeled.gz (11.2M; 430M Uncompressed)
9. http://kdd.ics.uci.edu/databases/kddcup99/kddcup.testdata.unlabeled_10_percent.gz (1.4M;45M Uncompressed)
10. http://kdd.ics.uci.edu/databases/kddcup99/corrected.gz Test data with corrected labels.
11. training_attack_types A list of intrusion types.
12. typo-correction.txt A brief note on a typo in the data set that has been corrected.
13. http://www.sigkdd.org/kddcup/index.php?section=1999&method=data
14. http://matauranga.wordpress.com/rana/kdd-cup-1999-data-evaluation/
15. http://www.scribd.com/doc/2346440/Neural-Networks-and-Fuzzy-logic-Control-IC-1403