Survey on Intrusion Detection System using Machine Learning Techniques

Sharmila Kishor Wagh
Research Scholar
North Maharashtra University, Jalgaon

Vinod K. Pachghare, Ph.D
Department of Computer Engineering,
College of Engineering, Pune

Satish R. Kolhe, Ph.D
Professor, School of Computer Science
North Maharashtra University, Jalgaon

ABSTRACT

In today’s world, almost everybody is affluent with computers and network based technology is growing by leaps and bounds. So, network security has become very important, rather an inevitable part of computer system. An Intrusion Detection System (IDS) is designed to detect system attacks and classify system activities into normal and abnormal form. Machine learning techniques have been applied to intrusion detection systems which have an important role in detecting Intrusions. This paper reviews different machine approaches for Intrusion detection system. This paper also presents the system design of an Intrusion detection system to reduce false alarm rate and improve accuracy to detect intrusion.

Keywords

Intrusion Detection System (IDS) , Machine Learning Techniques, Anomaly Detection, False Alarm Rate (FAR).

1. INTRODUCTION

The security of computer networks has been in the focus of research for years. The organization has come to realize that information & network security technology has become very important in protecting its information. Any successful attempt or unsuccessful attempt to compromise the integrity, confidentiality, and availability of any information resource or the information itself is considered a security attack or an intrusion. Every day new kind of attacks is being faced by industries. One of the solutions to this problem is by using Intrusion Detection System (IDS).

Machine Learning is one of the technique used in the IDS to detect attacks. Machine learning is concerned with the design and development of algorithms and methods that allow computer systems to autonomously acquire and integrate knowledge to continuously improve them to finish their tasks efficiently and effectively. In recent years, Machine Learning Intrusion Detection system has been giving high accuracy and good detection of novel attacks. Intrusion detection system (IDS) is a security technique attempting to detect various attacks. They are the set of techniques that are used to detect suspicious activity both on host and network level.

2. TAXONOMY OF ANOMALY DETECTION

Several classifications of intrusion detection methods have been proposed in the earlier period, but there is still no universally accepted taxonomy. A taxonomy that is based on the synthesis of a number of existing ones is here presented, using six criteria to classify IDSs, as summarized in Fig. 1.

Currently the two basic methods of detection (analytical method) are signature-based and anomaly-based [1],[2]. The signature-based method, also known as misuse detection, seems for a specific signature to match, signaling an intrusion. They can detect many or all known attack patterns, but the weakness of signature based intrusion detection systems is the incapability of identifying new types of attacks or variations of known attacks.

Another useful method for intrusion detection is called anomaly detection. Anomaly detection applied to intrusion detection and computer security has been an active area of research since it was originally proposed in [3]. In anomaly based IDSs, the normal behavior of the system or network traffic are represented and, for any behavior that varies over a pre-defined threshold, an anomalous activity is identified. By the other side, in anomaly based IDSs, the number of false positives generated are higher than on those based on signatures. An important issue in anomaly based IDSs is how these systems should be trained, i.e., how to define what is a normal behavior of a system or network environment (which features are relevant) and how to represent this behavior computationally.

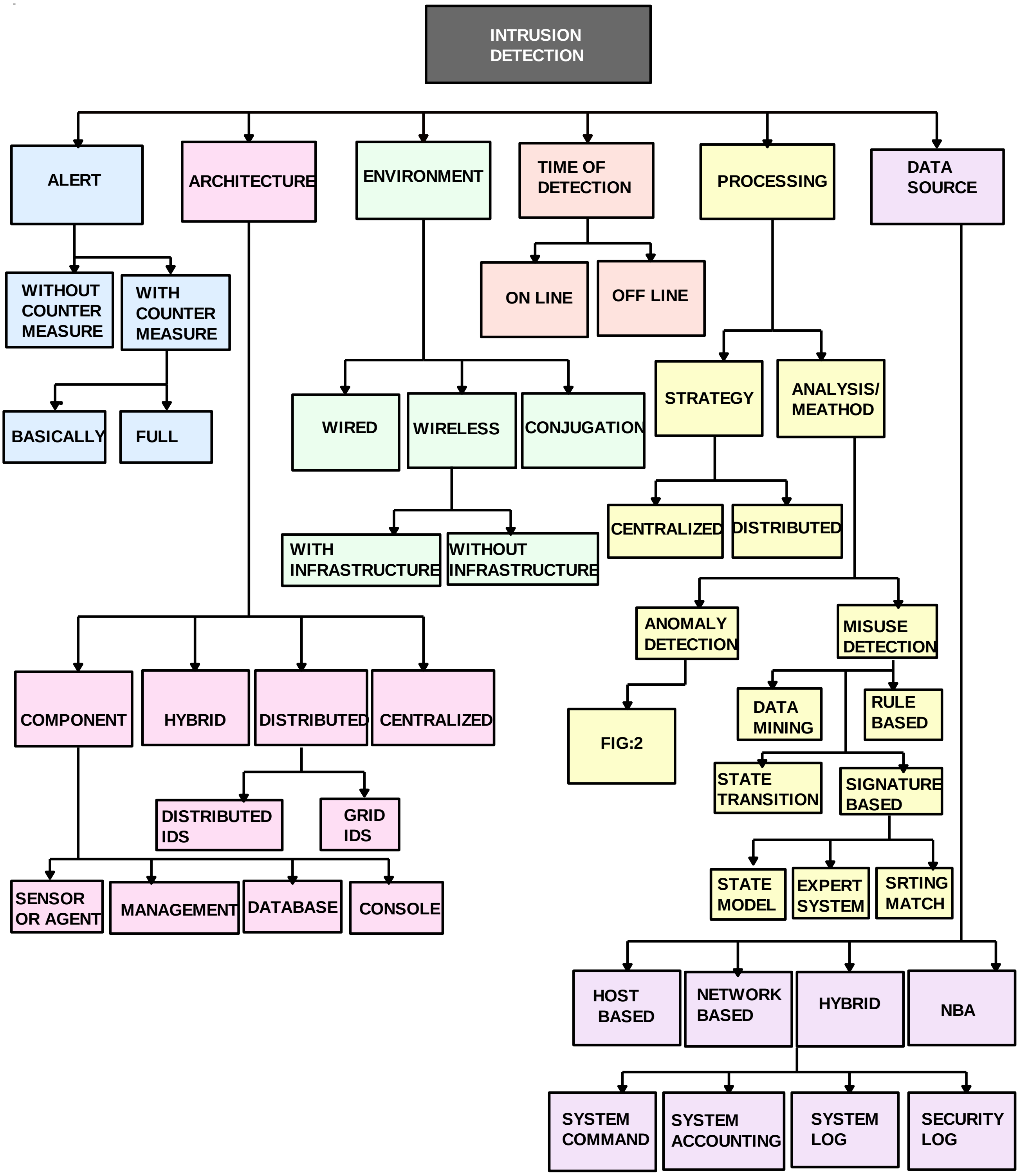

Fig 1: Taxonomy of IDS

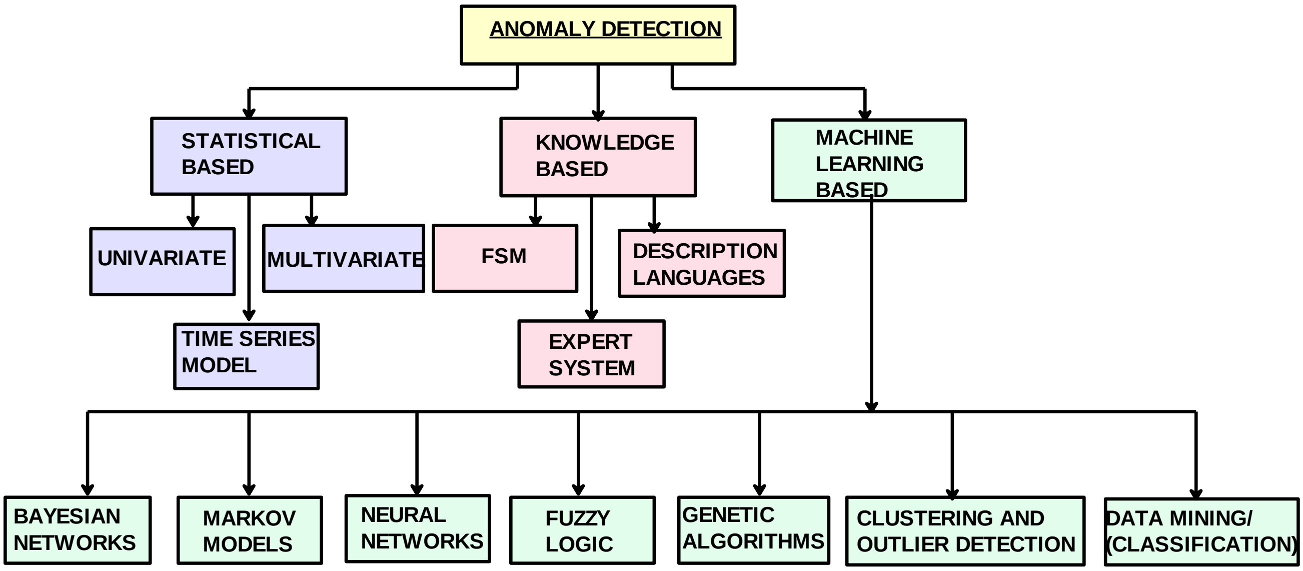

Fig 2: Classification of Anomaly detection

According to the type of processing related to the ‘‘behavioral’’ model of the target system, anomaly detection techniques can be classified into three main categories [4] statistical based, knowledge-based, and machine learning-based. In [18] the well-known intrusion detection approaches and Comparison of various approaches reviewed with the strength and weakness of those approaches.

2.1 Statistical anomaly-based IDS

A statistical anomaly-based IDS find out normal network activity like what sort of bandwidth is generally used, what protocols are used, what ports and devices generally connect to each other- and aware the administrator or user when traffic is detected which is anomalous (not normal) [5] [7]. It is again categorized into univariate, multivariate and time series model. Univariate model parameters are modeled as independent Gaussian random variables thus defining an acceptable range of values for every variable. The multivariate model considers the correlation between two or more variables. The time series model uses an interval timer, together with an event counter or resource measure and take into account the order and inter arrival times of observations and their values which are labelled as anomaly if its probability of occurrence is too low at a given time.

Pros: - 1) Prior knowledge about normal activity not required. 2) Accurate notification of malicious activities

Cons:- 1) Susceptible to be trained by attackers. 2)Difficult setting of parameters and metrics. 3) Unrealistic quasi-stationary process assumption

2.2 Knowledge-based techniques

Knowledge based stores information about subject domain. Information in knowledge based contains symbolic representations of expert’s rules of judgment in a format that allow the inference engine to perform deduction upon it. The expert system approach is one of the most widely used knowledge-based IDS schemes. Knowledge based techniques are divided into frame based model, rule based model and expert system. Rule based is modified form of the grammar based production rules. Frame based model localizes an entire body of expected knowledge and actions into a single structure. Expert systems are intended to classify the audit data according to a set of rules, involving three steps. First, different attributes and classes are identified from the training data. Second, a set of classification rules, parameters or procedures are deduced. Third, the audit data are classified accordingly [5], [6] .

Pros: - 1) Robustness. Flexibility and scalability

Cons: -1) Difficult and time-consuming availability of high-quality knowledge/data.

2.3 Machine learning-based IDS

Machine learning techniques are based on establishing an explicit or implicit model. A singular characteristic of these schemes is the need for labeled data to train the behavioral model, a procedure that places severe demands on resources. In many cases, the applicability of machine learning principles coincides with that for the statistical techniques, although the former is focused on building a model that improves its performance on the basis of previous results. Hence, machine learning for IDS has the ability to change its execution strategy as it acquires new information. This feature could make it desirable to use such schemes for all situations.

Pros:-1)Flexibility and adaptability capture of interdependencies.

Cons:-1) High depended on the assumption about the behavior accepted into the system.

3. INTRUSION DETECTION AND MACHINE LEARNING

The idea of applying machine learning techniques for intrusion detection is to automatically build the model based on the training data set. This data set contains a collection of data instances each of which can be described using a set of attributes (features) and the associated labels. The attributes can be of different types such as categorical or continuous. The nature of attributes determines the applicability of anomaly detection techniques. For example, distance-based methods are initially built to work with continuous features and usually do not provide satisfactory results on categorical attributes. The labels associated with data instances are usually in the form of binary values i.e. normal and anomalous. In contrast, some researchers have employed different types of attacks such as DoS, U2R, R2L and Probe rather than the anomalous label. This way learning techniques is able to provide more information about the types of anomalies. However, experimental results show that current learning techniques are not precise enough to recognize the type of anomalies. Since labeling is often done manually by human experts, obtaining an accurate labeled data set which is representative of all types of behaviors is quite expensive. As a result, based on the availability of the labels, three operating modes are defined for anomaly detection techniques: as Supervised Learning, Unsupervised Learning, Semi supervised Learning.

4. CLASSIFICATION OF ANOMALY DETECTION

Several machine learning-based schemes have been applied to IDS. Some of the most important techniques are explained in following subsections.

4.1 Bayesian Network

A Bayesian network is a model that encodes probabilistic relationships among impotant variables. This technique is generally used for intrusion detection in combination with statistical schemes, a procedure that yields several advantages [9], including the capability of encoding interdependencies between variables and of predicting events, as well as the ability to incorporate both prior knowledge and data.

Conditional probability P (A|B) is used for calculating the probability of at once the condition B is present. However, in the real world applications. one needs to know about the conditional probability P (B|A) for B once its evidence A is present. In this Bayes theory, the goal is to calculate the probability of a given hypothesis H considering its sign or evidence E already exists. The H can be assumed to be a sampled column feature vector and noted as x = {x1 , x2 , . . .}. In the following text the E (Evidence) and the C (Class) sign can be replaced (where C = {c1 ,c2 , . . .}), if it makes it easier for the reader to understand the concept. The formula to calculate this probability is presented below

$$p\left(\frac{H}{E}\right) = \frac{\mathrm{P(\,H\,)*P(\,\frac{E}{H})}}{\mathrm{P(\,E\,)}} \qquad (1)$$

Where P (E |H) is the conditional probability of the evidence E once the hypothesis H is at hand. P(H) is the probability of the hypothesis H. P (E) is the probability of the evidence E. P (H |E) is the posterior probability of the hypothesis H once the evidence E is available.

A framework of NIDS based on a Naïve Bayes algorithm is proposed in [19]. The framework constructs the patterns of the network services over data sets labeled by the services. The framework detects attacks in the datasets using the naïve Bayes Classifier algorithm using the built patterns. Compared to the Neural network based approach, their approach achieves higher detection rate, less time consuming and has a low cost factor. However, it generates more false positives.Naïve Bayesian network is a restricted network that has only two layers and assumes complete independence between the information nodes. This poses a limitation of this research work. In order to minimize this problem so as to reduce the false positives, active platform or event based classification may be thought of using Bayesian network.

Researchers have designed several systems dealing with the problem of false alarms in recent years. In [12] author proposed to use Bayesian networks to perform reasoning on complementary security evidence, and thus to potentially reduce false alert rates.

4.2 Markov models

There are two subtypes of Markov models: Markov chains and hidden Markov models. A Markov chain is a set of states that are interconnected through certain transition probabilities, which determine the topology and the capabilities of the model. During a first training phase, the probabilities associated with the transitions are estimated from the normal behavior of the target system. The detection of anomalies is then carried out by comparing the anomaly score (associated probability) obtained for the observed sequences with a fixed threshold. In the case of a hidden Markov model, the system of interest is assumed to be a Markov process in which states and transitions are hidden. Only the so-called productions are observable. Markov-based techniques have been extensively used in the context of host IDS, normally applied to system calls.

A hybrid fuzzy-based anomaly IDS using hidden Markov model (HMM) detection engine and a normal database detection engine to reduce FAR is proposed in [13].

Development of host-based anomaly IDS has been studied with highlighting places on system call-based HMM training explained in [26].

4.3 Neural networks

Artificial Neural Networks: - Inspired from known facts about how the brain works, researchers in the area of artificial intelligence (AI) have developed computational models which exhibit performance somewhat comparable to that of the brain [22]. Artificial neural networks (ANNs) are adaptive parallel distributed information processing models that consist of: (a) a set of simple processing units (nodes, neurons), (b) a set of synapses (connection weights), (c) the network architecture (pattern of connectivity), and (d) a learning process used to train the network. [20] Based on the advantages and disadvantages of the improved GA and LM algorithm, in this paper, the Hybrid Neural Network Algorithm (HNNA) is presented. Firstly, the algorithms uses the advantage of the improved GA with strong whole searching capacity to search global optimal point in the whole question domain. Then, it adopts the strong point of the LM algorithm with fast local searching to fine search near the global optimal point. The paper used respectively the three algorithms, namely the Improved GA, LM algorithm and HNNA, to adjust the input and output parameters of the ANN model, and adopt the theories of the fusion of the multi-classifiers to structure the Intrusion Detection System. By repeating an experiment, it is found that the HNNA is better in stability and convergence precision than LM algorithm and improved GA from the training result. The testing results are also proving that the detection rate of the multiple classifier intrusion detection system based on HNNA learning algorithm, including all attack categories that has a few or many training samples, is higher than the IDS that use LM and improved GA learning algorithm, and the false negative rate is less. So, the HNNA is proved to be feasible in theory and practice.

In [21] according to the difference between the attack categories, they adjust the 41-dimensional input features of the neural-network-based multiple classifier intrusion detection system. After repeated experiment, they find that the every adjusted sub-classifier is better in convergence precision, shorter in training time than the 41-features sub-classing, moreover, the whole intrusion detection system is higher in the detection rate, and less in the false negative rate than the 41-features multiple classifier intrusion detection system. So, the scheme of the adjusting input features is able to optimize the neural-network-based multiple classifier intrusion detection system, and proved to be feasible in practice

4.4 Fuzzy logic techniques

Fuzzy logic is derived from fuzzy set theory under which reasoning is approximate rather than precisely deduced from classical predicate logic. Fuzzy techniques are thus used in the field of anomaly detection mainly because the features to be considered can be seen as fuzzy variables [10]. The application of fuzzy logic for computer security was first proposed in [23]. Fuzzy Intrusion Recognition Engine (FIRE) for detecting intrusion activities is proposed in [24] and the anomaly based IDS is implemented using the data mining techniques and the fuzzy logic. The fuzzy logic part of the system is responsible for both handling the large number of input parameters and dealing with the inaccuracy of the input data. Three fuzzy characteristics used in this work are COUNT, UNIQUENESS and VARIANCE. The implemented fuzzy inference engine uses five fuzzy sets for each data element (HIGH, MEDIUM-HIGH, MEDIUM LOW and MEDIUM-LOW) and suitable fuzzy rules to detect the intrusion. In their report authors have not specified how they have derived their fuzzy set. The fuzzy set is a very important issue for the fuzzy inference engine\ and in some cases genetic approach can be implemented to select the best combination. The proposed system is tested using data collected from the local area network in the college of Engineering at Iowa State University and the results are reported in this paper. The reported results are descriptive and not numerical; therefore, it is difficult to evaluate the performance of the reported work.

4.5 Genetic algorithms

Genetic algorithms are classified as global search heuristics, and evolutionary computation that uses techniques inspired by evolutionary biology such as recombination, selection, inheritance and mutation. Thus, genetic algorithms represent another type of machine learning-based technique, capable of deriving classification rules [11] and/or selecting appropriate features or optimal parameters for the detection process [10] .

In [25] rule evolution approach based on Genetic Programming (GP) for detecting novel attacks on networks is proposed. In their framework, four genetic operators, namely reproduction, mutation, crossover and dropping condition operators, are used to evolve new rules. New rules are used to detect novel or known network attacks. Experimental results show that rules generated by GPs with part of KDD 1999 Cup data set has a low false positive rate (FPR), a low false negative rate (FNR) and a high rate of detecting unknown attacks. However, an evaluation with full KDD training and testing data is missing in the paper.

More efforts using GA for intrusion detection are made in [14, 4, 8] proposes a linear representation scheme for evolving fuzzy rules using the concept of complete binary tree structures. GA is used to generate genetic operators For producing useful and minimal structural modifications to the fuzzy expression tree represented by chromosomes. However, the training process in this approach is computationally very expensive and time consuming. Bridges and Vaughn employ GA to tune the fuzzy membership functions and select an appropriate set of features in their intelligent intrusion detection system. GA as evolutionary algorithms was successfully used in different types of IDS. Using GA returned impressive results; the best fitness value was very close to the ideal fitness value. GA is a randomization search method often used for optimization problem. GA was successfully able to generate a model with the desired characteristics of high correct detection rate and low false positive rate for IDS [15].

4.6 Clustering and outlier detection

Clustering techniques work by grouping the observed data into clusters, according to a given similarity or distance measure. The procedure most commonly used for this consists in selecting a representative point for each cluster. Clustering techniques to determine the occurrence of intrusion events only from the raw audit data, and so the effort required to tune the IDS is reduced. One of the most popular and most widely used clustering algorithms is K-Means [19], which is a non-hierarchical Centroid-based approach.

4.7 Data Mining

Data mining is an information activity to discover hidden facts contained in the database. These techniques are used to find patterns and intelligent relationships in data and infer rules that allow the prediction of future result.

Association rule learning is one of many data mining techniques that describe events that tend to occur together. Association rule discovery is to define normal activity by which discovery of anomalies is easily enabled. Classification is to classify each audit record into one of the possible categories normal and anomaly.

In [17] authors discussed the uses of data mining approach in Intrusion Detection. This data mining technique works by learning the training data know to be free of attacks (normal) and then uses an algorithm group an attack from the data. It uses associates rules to store knowledge data about the nature of pattern about individual records that can improve the classification efficiency.

5. SYSTEM DESIGN FOR INTRUSION DETECTION SYSTEM

Figure 3 shows system design for intrusion detection system.It has following main modules.

5.1 Preprocessing Phase

On-line (real-time) IDS: - In this phase packet capturing and extraction of packet features is done with the help of packet sniffing tools (for e.g Wireshark ,Capsa) which are used to capture the packet information like, IP/TCP/ICMP headers, from each of the packets. After that partition the packet header with source addresses, destination address etc. In this phase need of some techniques for selection of essential feature. And finding whether the packet is normal or intrusion.

Off-line IDS: - In this phase packet capturing is done from dataset (for e.g KDD dataset/NLS KDD) to serve for the data source of the IDS.

5.2 Classification

In classification phase utilize the data received from the previous phase for detecting whether the normal packet or attack packet. Depending on feature values the corresponding algorithms will classify the packet into similar groups. It consists of two processes:

(a) Training data (b) Testing data

In training phase answer class is provided along with the packet features which will help to formulate rules deciding mapping domains. These rules may get changed replaced depending on further training. Every algorithm has its own strategy of classification.

In the Testing Phase, untrained data are given to the system for sampling whether true answers are obtained or not. The system process is performed providing input as packets without specifying answer class.

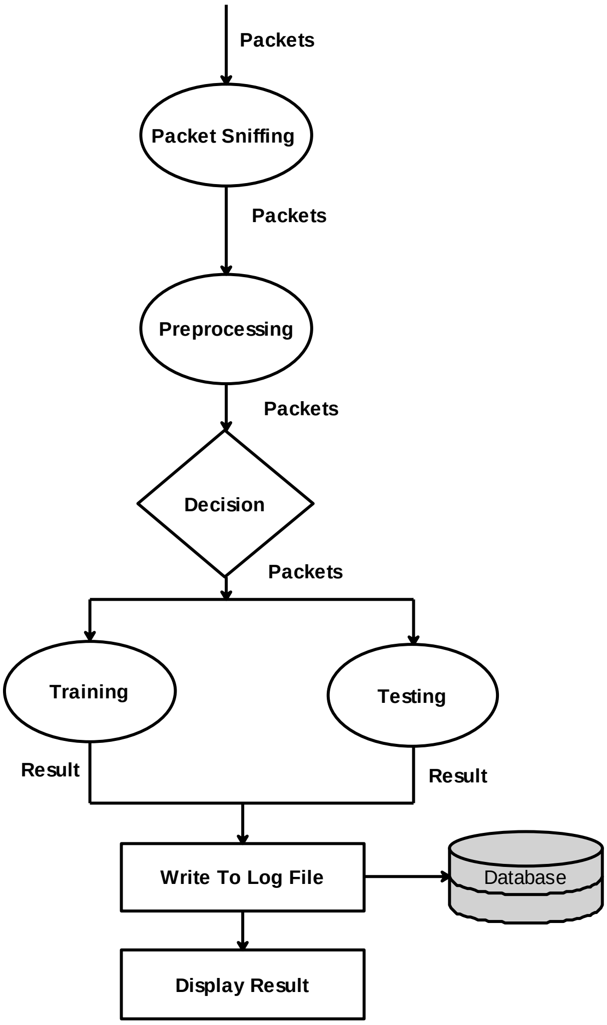

Fig 3: System Design for IDS

5.3 Post Processing

The result got in preprocessing phase is evaluated against answer class and system performance is measured in combinations of correctness and false alarms. i.e. True Positive, True Negative, False Positive and False Negative.

5.4 Reducing False Alarms

If the system is still giving some false alarms for all the algorithms some more training is needed to be given. This is the machine learning mechanism i.e. the system will keep on learning on its own without human interference. And hence there is no updating required.

6. CONCLUSIONS

In this paper authors have presented an overview of machine learning technologies which are being utilized for the detection of attacks in IDS and system design of effective IDS. The security of information in computer based systems is a major concern to researchers. The work of IDS and methodologies which has been a major focus of information security related research. Machine learning is a vast and advanced field still relatively immature and definitely not optimized for IDS.

7. FUTURE DIRECTION

In recent years, the challenges that lie ahead of us in intrusion detection system are huge, which are listed as follows

1. Inability to lessen the number of false positives which reduce efficiency of IDS. Good IDS should perform with a high precision and a high recall, as well as a lower false positive rate and a lower false negative rate. How one can have confidence in the result is a major issue.

2. Time taken to process the huge amount of data for training is very large.

3. To improve classification accuracy is a major task in IDS. Impose to focus on multi classifier system.

4. Because of inadequate computing resources and tremendous increase of targeted attacks necessity of real time Intrusion detection system. However, its implementation in real life environment is challenging.

5. Need of the standard evaluation dataset which simulate for real time IDS .

6. Feature reduction work -Many studies use feature selection for data reduction, to decrease the computational complexity. Need to concentrate more to perform the task of data deduction

7. Need to implement a combination technique for misuse detection and anomaly detection

The machine learning technique could turn out very good field for IDS by resolving these challenges.

8. RREFERENCES

[1] S. Chebrolu, A. Abraham, and J. P. Thomas, “Feature deduction and ensemble design of intrusion detection systems,” Comput. Secure., vol. 24, no. 4, pp. 295–307, Jun. 2005

[2] W. Lee and S. J. Stolfo, “A framework for constructing features and models for intrusion detection systems,” ACM Trans. Inf. Syst. Secur. vol. 3, no. 4, pp. 227–261, Nov. 2000.

[3] Denning D, “An Intrusion-Detection Model,” IEEE Transactions on Software Engineering, Vol. SE-13, No 2, Feb 1987.

[4] Lazarevic A, Kumar V, Srivastava J. Intrusion detection: “A survey, Managing cyber threats: issues, approaches, and challenges,” Springer Verlag; 2005. pp. 330.

[5] Denning DE, Neumann PG. “Requirements and model for IDES – a real-time intrusion detection system,” Computer Science Laboratory, SRI International; 1985. Technical Report #83F83- 01-00

[6] Anderson D, Lunt TF, Javitz H, Tamaru A, Valdes A. “Detecting unusual program behavior using the statistical component of the next-generation intrusion detection expert system (NIDES),” Menlo Park, CA, USA: Computer Science Laboratory, SRI International; 1995. SRIO-CSL-95-06.

[7] Ye N, Emran SM, Chen Q, Vilbert S. “ Multivariate statistical analysis of audit trails for host-based intrusion detection,” IEEE Transactions on Computers 2002;51(7).

[8] Wenke Lee and Salvatore J. Stolfo, “A framework for constructing features and models for intrusion detection systems,” 2000, ACM Trans. Inf. Syst. Secur., 3(4):227–261.

[9] Heckerman D.“A tutorial on learning with Bayesian networks,” Microsoft Research; 1995. Technical Report MSRTR-95-06.

[10] Bridges, Vaughn, “Fuzzy Data mining and genetic algorithms applied to intrusion detection,” In: Proceedings of the National Information Systems Security Conference; 2000. pp. 13–31.

[11] Li W. “Using genetic algorithm for network intrusion detection,” C.S.G. Department of Energy; 2004. pp. 1–8.

[12] Y. Zhai, P. Ning, P. Iyer, D.S. Reeves, “Reasoning about complementary intrusion evidence,” in: Proceedings of the 20th Annual Computer Security Applications Conference (ACSAC 04), December 2004.

[13] X.D. Hoang, J. Hu, P. Bertok, “A program-based anomaly intrusion detection scheme using multiple detection engines and fuzzy inference,” Journal of Network and Computer Applications 32 (2009) 1219–1228.

[14] Kamra, Bertino, “Design and Implementation of an Intrusion Response System for Relational Databases,” IEEE Transaction on Knowledge and Data Engineering, Volume: 23, Issue: 6 doi 10.1109 /TKDE.2010.151 ,2011, pp: 875 – 888

[15] Suhail Owais ,Václav Snášel, Pavel Krömer,Ajith Abraham ,”Survey: Using Genetic Algorithm Approach in Intrusion Detection Systems Techniques “,978-0-7695-3184-7/08 DOI 10.1109/CISIM7th Computer Information Systems and Industrial Management Applications./2008 IEEE

[16] C. Xiang and S. M. Lim, “Design of multiple-level hybrid classifier for intrusion detection system,” in Workshop on Machine Learning for Signal Processing, 2005, pp. 117–122.

[17] B. Daniel, C. Julia, J. Sushil, P. Leonard, N. N. Wu, “ADAM: Detecting intrusions by data mining”, Proceedings of the 2001 IEEE, workshop on Information Assurance and Security, West Point, NY, 2001.

[18] Murali A, Rao M, “A Survey on Intrusion Detection Approaches,” Information and Communication Technologies, 2005. ICICT 2005. First International Conference on DOI: 10.1109/ICICT.2005.1598592, Year: 2005, pp: 233 – 240

[19] Mrutyunjaya Panda, and Manas Ranjan Patra " NETWORK INTRUSION DETECTION USING NAÏVE BAYES ", IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.12, December 2007

[20] Li Xiangmei Qin Zhi "The Application of Hybrid Neural Network Algorithms in Intrusion Detection System "978-1-4244-8694-6/11 ©2011 IEEE

[21] Xiangmei Li ,"Optimization of the Neural-Network-Based Multiple Classifiers Intrusion Detection System ",978-1-4244-5143-2/10 ©2010 IEEE

[22] Naeem Seliya Taghi M. Khoshgoftaar, "Active Learning with Neural Networks for Intrusion Detection", IEEE IRI 2010, August 4-6, 2010, Las Vegas, Nevada, USA 978-1-4244-8099-9/10

[23] H.H. Hosmer, Security is fuzzy!: applying the fuzzy logic paradigm to the multipolicy paradigm, Proceedings of the 1992-1993 workshop on New security paradigms, ACM New York, NY, USA, 1993, pp. 175-184.

[24] John E. Dickerson and Julie A. Dickerson, Fuzzy network profiling for intrusion detection, Proceedings of NAFIPS 19th International Conference of the North American Fuzzy Infor mation Processing Society (Atlanta, USA), July 2000, pp. 301-306.

[25] T.Lunt and I.Traore, Unsupervised Anomaly Detection Using an Evolutionary Extension of K-means Algorithm,International Journal on Information and computer Science, Inderscience Pulisher 2 (May, 2008), 107-139.

[26] Jiankun Hu, Xinghuo Yu, Qiu D, Hsiao-Hwa Chen; "A simple and efficient hidden Markov model scheme for host-based anomaly intrusion detection," IEEE Transaction on Network, Volume: 23, Issue: 1 DOI: 10.1109/MNET.2009.4804323, Year: 2009, Page(s): 42 – 47.