A New Intrusion detection system in data mining & fuzzy logic

Balaji.s.c.k [1], B.Kishore Kumar[2]
Assistant professor, SVIT, Hampapuram, Anantapuram (Dt), Andhra Pradesh

Abstract: *The computers abnormal activities are identified by system defense. Traditional Intrusion detection relays on extensive knowledge of traditional expertise, in particular, on the familiarity with the systems to be protected . To reduce this dependence, various data-mining and machine learning techniques have been used in the literature. In the proposed system, we have designed fuzzy logic-based system for effectively identifying the intrusion activities within a network. Currently available intrusion detection systems focus mainly on determining uncharacteristic system events in distributed networks using signature based approach. Due to its limitation of finding novel attacks, we propose a hybrid model based on improved fuzzy and data mining techniques, which can detect both misuse and anomaly attacks. Our aim is to reduce the amount of data retained for processing i.e., attribute selection process and also to improve the detection rate of the existing IDS using data mining technique. We then use improved Kuok fuzzy data mining algorithm, which in turn a modified version of APRIORI algorithm, for implementing fuzzy rules, which allows us to construct if-then rules that reflect common ways of describing security attacks. We applied fuzzy inference engine using mamdani inference mechanism with three variable inputs for faster decision making*

Keywords: *Fuzzy logic, apriori, intrusion detection, hybrid system.*

I. INTRODUCTION

In the current scenario, business does not have boundaries to carry out the task, as it is highly distributed across the globe. Securing these widely distributed data has become a hill-climbing task, as there are new types of threats arising from the people, especially in the current economic turmoil, who used to penetrate into the organization network, either for fun or for gaining some vital information which can fetch them monetary benefits. Network based defense systems normally combine network based IDS and packet filtering firewalls. The main drawbacks of these systems are its inability to identify and characterize new attacks and to respond them intelligently.

A significant challenge in providing an effective and efficient protective mechanism to a network is the ability to detect novel attacks at the initial phase and implement proper countermeasures. Information security process includes the risk assessment, which identifies the potential risk to the organizations networks, risk mitigation that suggests the methods to reduce the risks, and countermeasures. Intrusion Detection has become an integral part of the information security process. But, it is not technically feasible to build a system with no vulnerabilities; as such incursion detection continues to be an important area of research.

II. INTRUSION DETECTION SYSTEMS USING AI TECHNIQUES

Intrusion Detection Systems (IDS) is a powerful technology helping in protecting the precious data and networks from malicious and unauthorized access. An intruder is a hacker attempting to break into or misuse the resources of a computer system. Intruders come from outside an organizations network and may attempt to go around firewalls to attack machines on the internal network. Research and development of intrusion detection has been under way for nearly 27 years. The work that is most often cited is the technical report by James P.Anderson [1]. AI techniques like Fuzzy logic, Neural Networks, and recent development of Artificial Immune Systems (AIS) has played an important role in most misuse and anomaly detection. Few researchers [2] have applied the above mentioned techniques for the detection of malicious activities. Another direction of interest could be using ensemble approach [3], where by combining few approaches for detection individual classes.

A Intrusion detection using fuzzy logic and data mining

The method proposed by [4] extracts fuzzy classification rules from numerical data, applying a heuristic learning procedure. The learning procedure initially classifies the input space into non-overlapping activation rectangles corresponding to different output intervals. In this sense, our work is similar to that of [4,5].There is no overlapping and inhibition areas. However, the disadvantage listed is, the high false positive rates which is the primary scaling of all the IDS. Researcher [6] describes the approaches to address three types of issues: accuracy, efficiency, and usability. This works is an enhanced version of the work by Lee [4]. First issues of improving accuracy is achieved by using data mining programs to analyze audit data and extract features that can distinguish normal activities from intrusions. Second issue, efficiency is improved by analyzing the computational costs of features and a multiple-model cost-based approach is used to produce detection models with low cost and high accuracy. Third issue, improved usability, is solved by using adaptive learning algorithms to facilitate model construction and incremental updates; unsupervised anomaly detection algorithms are used to reduce the reliance on labeled data. Researchers [7] developed the Fuzzy Intrusion Recognition Engine (FIRE) using fuzzy sets and fuzzy rules. FIRE uses simple data mining techniques to process the network input data and generate fuzzy sets for every observed feature. The fuzzy sets are then used to define fuzzy rules to detect individual attacks. FIRE does not establish any sort of model representing the current state of the system, but instead relies on attack specific rules for detection. Instead, FIRE creates and applies fuzzy logic rules to the audit data to classify it as normal or anomalous. Dickerson et al. found that the approach is particularly effective against port scans and probes. The primary disadvantage to this approach is the labor intensive rule generation process. Our research work can be considered as an extension of the above work by automating the rule generation process.

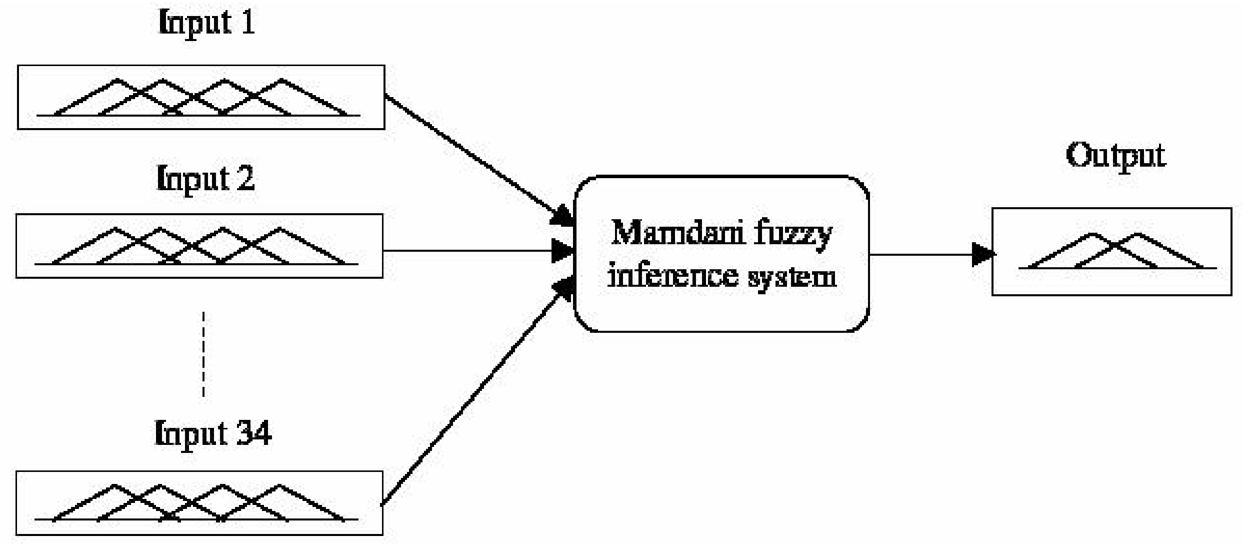

Figure 1. The designed Fuzzy system

The model proposed [8] combines neural networks and fuzzy logic. This system works by mapping a template graph and user action graph to determine patterns of misuse. The output of this mapping process will be used by the central strategic engine to determine whether an intrusion has taken place or not. The major drawback is that new type attacks rules need to be given by the external security officer i.e. it does not automate rule generation process and more number of components prevents it from working fast. Tombini used an approach wherein the anomaly detection technique is used to produce a list of suspicious items. The classifier module which uses a signature detection technique then classified the suspicious items into false alarms, attacks, and unknown attacks. This approach works on the premise that the anomaly detection component would have a high detection rate, since missed intrusions cannot be detected by the follow-up signature detection component. In addition, it also assumed that the signature detection component will be able to identify false alarms. While the hybrid system can still miss certain types of attacks, its reduced false alarm rate increases the likelihood of examining most of the alerts. Zhang and Zulkernine employed the method random forests algorithm in the signature detection module to detect known intrusions. Thereafter, the outlier detection provided by the random forests algorithm is utilized to detect unknown intrusions. Approaches that use signature detection and anomaly detection in parallel have also been proposed. In such systems, two sets of reports of possible intrusive activity are produced and a correlation component analyzes both sets to detect intrusions. Researchers use association based classification methods to classify normal and abnormal attacks based on the compatibility threshold.

III. PROPOSED IIDS FRAMEWORK

The proposed IIDS framework (Figure 2) has several components, and the following sections describe briefly about each.

A. Attributes

Prior to any data analysis, attributes representing relevant features of the input data (packets) must be established. The amount of data being sniffed from the network will be heavy. If only vital attributes are chosen then the overhead of fuzzy inference engine will be reduced to the maximum extend. There were 42 variables from DARPA training data. The set of attributes provided to the Data Analyzer is a subset of all possible attributes pertaining to the information contained in packets headers, packet payloads, as well as aggregate information such as statistics in the number and type of packets or established TCP connections. Attributes are represented by names that will be used as linguistic variables by the Data Miner and the Fuzzy Inference Engine. We use traditional attribute selection algorithm that utilizes information gain for selection each attributes based on the information gain obtained for that 2193 particular class of variables available.

B. Data Analyzer

Once attributes of relevance have been defined and the data source identified, Data Analyzer is employed to compute configuration parameters that regulate the operation of IDS. This module analyzes packets and computes aggregate information by grouping packets. Packets can be placed in fixed size groups (s−group) or in groups of packets captured in a fixed amount of time (t−group). Each s−group contains the same number of packets covering a variable time range and each t−group contains a variable number of packets captured over a fixed period of time.

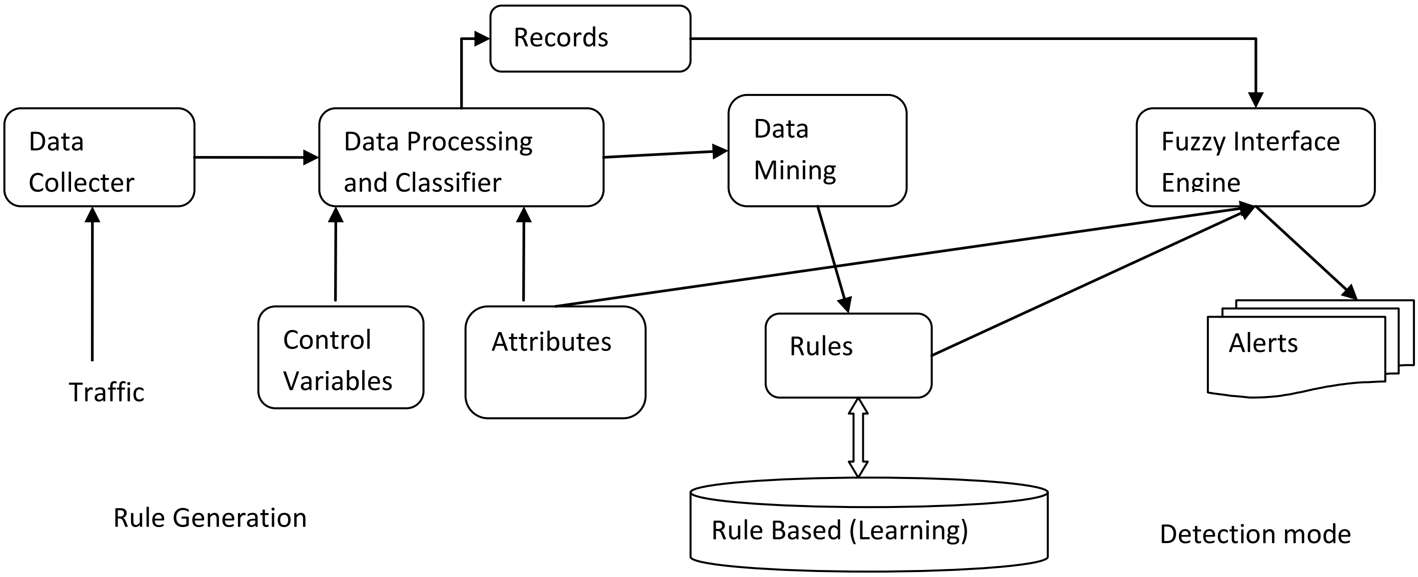

Figure 2. IIDS framework.

C. Data Miner

A variation of Kuok's algorithm is used to implement the Data Miner, which allows for efficient, single–pass, record processing by partitioning data into hierarchical files. From the data sample the algorithm read the records and creates a file for each term of attribute. The algorithm carries over the same process for all the record in the data file. Once the entire data set is processed, the support value for every item set is calculated. The algorithm extracts all large item sets, and constructs rules from those large item sets. The rules with confidence value greater than or equal to min confidence moved to the rule base.

D. Inference Engine

An inference determines which rules are relevant to the given data. Our prototype utilizes FuzzyJess to check the rules. The engine uses three inputs as explained in the figure. We have configured the inference engine to use Mamdani inference mechanism. The shooting strength concludes whether the modeled behavior satisfies the fact. If the shooting strength is somewhere near to zero indicates a possibility of attack. F. Data set used for testing the prototype We used 1999 DARPA data which is an extension of 1998 data set available. The latest dataset was added with few more attacks that reflect more real-time data. A quick glance at the 1999 DARPA dataset shows it contains 3 weeks of training data and two weeks of testing data. First and third week data do not contain any attack and this was used as training data for anomaly intrusion detection. Second week contains all types of attacks so this was used for creating attack profiles. We used two weeks of test data for our testing. The record files were broken down into smaller files as there were few limitations like processing power of the hardware and the size of the main memory.

IV. CONCLUSIONS AND FUTURE WORK

A few contributions have been made by this research namely, improved apriori algorithm for faster rule generation and reduced querying frequency, reduced features for faster detection of attacks and reduced false positives. Even though the system was tested using off- 22156 line data and in a small live environment, our next goal is to turn this system into a light weight system, by overcoming the existing limitations like bottle neck in packet processing, searching for known rules and very high false positive rates and improve the performance by means of faster detection and alert correlation and acceptable rate of false positives. As a token of appreciation to open-source community, our plan is to make this system an open source project once the system is complete and ready by all means to take the high level of real world challenge.

REFERENCES

[1] Anderson, P. Computer security threat monitoring and surveillance. Fort Washington.1980.

[2] Peyman Kabiri and Ali A. Ghorbani. “Research on Intrusion Detection and Response: A Survey”. International Journal of Network Security, Vol.1, No.2, PP.84–102, Sep. 2005

[3] Anazida Zainal, Mohd Aizaini Maarof and Siti Mariyam Shamsuddin “Ensemble Classifiers for Network Intrusion Detection System”, Journal of Information Assurance and Security 4 (2009) 217-225

[4] Lee, W. A “Data Mining framework for construction features and models for intrusion detection systems”, Columbia University.2001.

[5] Manganaris, S.. "A data mining analysis of RTID alarms." Computer Networks2000. 34(4): 571-577.

[6] Stolfo, S. J,. "Data mining-based intrusion detectors: an overview of the columbia IDS project." ACM SIGMOD2001. 30(4): 5-14.

[7] Dickerson, J. E. and J. A. Dickerson Fuzzy Network Profiling for Intrusion Detection. 19th International Conference of the North American Fuzzy Information Processing Society,2000,301-306.

[8] Botha, M., Solms VR, Perry K, Loubser E, “The utilization of Artificial Intelligence in a Hybrid Intrusion Detection System”. Annual research conference of the South African institute of computer scientists and information technologists on Enablement through technology,2002, 149-155.

AUTHORS BIOGRAPHY

Balaji .s.c.k M.tech, Assistant Professor Dept of CSE, SVIT, Hampapuram, Anantapur. He has 3 years of teaching experience and 10 years of I.T experience .His areas of interest are Data mining and Data Warehousing .

B. Kishore Kumar M.tech, Assistant Professor Dept of CSE, SVIT, Hampapuram, Anantapur. He has 5 years of teaching experience .His areas of interest are Data mining and Data Warehousing