# Intrusion Detection Using Data Mining Along Fuzzy Logic and Genetic Algorithms

**Y.Dhanalakshmi [1] and Dr.I. Ramesh Babu [2]**

Research Scholar Professor
Dept of Computer Science & Engineering
Acharya Nagarjuna University, Guntur, A.P. India

**Summary**

Intrusion Detection is one of the important area of research. Our work has explored the possibility of integrating the fuzzy logic with Data Mining methods using Genetic Algorithms for intrusion detection. The reasons for introducing fuzzy logic is two fold, the first being the involvement of many quantitative features where there is no separation between normal operations and anomalies. Thus fuzzy association rules can be mined to find the abstract correlation among different security features. We have proposed architecture for Intrusion Detection methods by using Data Mining algorithms to mine fuzzy association rules by extracting the best possible rules using Genetic Algorithms.

.

***Key words:***

*: Data Mining algorithms, Apriori, Fuzzy logic, Genetic algorithms.*

## 1. Introduction:

A significant challenge in providing an effective defense mechanism to a network perimeter is having the ability to detect intrusions and implement counter measures. Components of the network perimeter defense capable of detecting intrusions are referred to as Intrusion Detection Systems (IDS). Intrusion Detection techniques have been investigated since the mid 80s and depending on the type and source of the information used to identify security breaches, they are classified as host-based or network based [7].

Host-based systems use local host information such as process behavior; file integrity and system logs to detect events. Network-based systems use network activity to perform the analysis. Combinations of these two types are possible. Depending on how the intrusion is detected an IDS is further classified as signature-based (also known as misuse system) or anomaly-based [7]. Signature-based systems attempt to match observed activities against well defined patterns which also called signatures. Anomaly-based systems look for any evidence of activities that deviate from what is considered normal system use. These systems are capable of detecting attacks for which a well-defined pattern does not exist (such as a new attack or a variation of an existing attack). A hybrid IDS is capable of using signatures and detecting anomalies [1].

More recently, techniques from the data mining area (mining of association rules and frequency episodes) have been used to mine the normal patterns from audit data. Typically, an IDS uses Boolean logic in determining whether or not an intrusion is detected and the use of fuzzy logic has been investigated as an alternative to Boolean logic in the design and implementation of these systems. Fuzzy logic addresses the formal principles of approximate reasoning [2]. It provides a sound foundation to handle imprecision and vagueness as well as mature inference mechanisms using varying degrees of truth. Since boundaries are not always clearly defined, fuzzy logic can be used to identify complex pattern or behavior variations [3]. This is accomplished by building an Intrusion Detection System that combines fuzzy logic rules with an expert system in charge of evaluating rule truthfulness.

Data Mining techniques have been commonly used to extract patterns from sets of data. Specifically two data mining approaches have been proposed and used for anomaly detection: association rules and frequency episodes. Association rule algorithm finds correlations between features or attributes used to describe a data set [6]. On the other hand, frequency episode techniques are effective at detecting sequential patterns of occurrences in a sequence of events. It is important to note that the association rule algorithms are used to produce rules suitable for anomaly-based and signature-based detection by mining normal (attack-free) and attack network traffic respectively [7]

Using genetic algorithms, 1) may result in the tune of the fuzzy membership functions to improve the performance and 2) select the set of features available from the audit data that provide the most information to the data mining component. These algorithms are often used for optimization problems[5]. When using fuzzy logic, it is often difficult for an expert to provide “good” definitions for the membership functions for the fuzzy variables. Those genetic algorithms can be successfully used to tune the membership functions of fuzzy sets used by the intrusion detection system [4]. Each fuzzy

Manuscript received February 5, 2008
Manuscript revised February 20, 2008

membership function can be defined using the standard function representation of fuzzy sets. Each chromosome for the Genetic Algorithms consists of a sequence of these parameters. An initial population of chromosomes is generated randomly where each chromosome represents a possible solution to the problem.

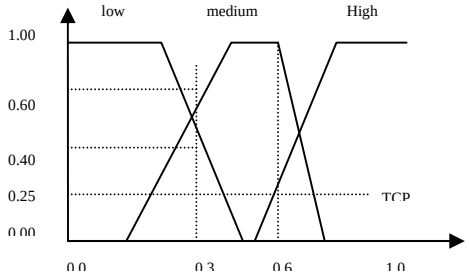


Fig 1 TCP linguistc variables.

The Fig 1 illustrates the use of fuzzy sets to describe a linguistic variable TCP with domain {0-1} using the terms Low, Medium and High as specified by their respective membership functions. Fuzzy membership functions determine degrees of membership for each category of term. In Fig 1, a TCP value of 0.3 belongs 40% to the low category and 60% to the Medium and a TCP value of 0.6 belongs 100% to Medium and 25% to High. Under this scheme, the TCP value 0.6 is "more important" than the value 0.3 since the sum of its degrees of membership (fuzzy support) is 125% as compared to 100% for a TCP value of 0.3. This shortcoming was eliminated by normalizing the fuzzy terms, ensuring that the fuzzy support for any value totals 100%.

## 2. Architecture:

The Hybrid Fuzzy logic IDS architecture has two modes of operations: rule-generation and detection. When operating in the rule-generation mode, the system processes network data and uses a fuzzy data mining algorithm to generate rules. A subset of the rules produced by the data mining algorithm is used as a model for the input data. The detection mode uses this rule subset for intrusion detection.

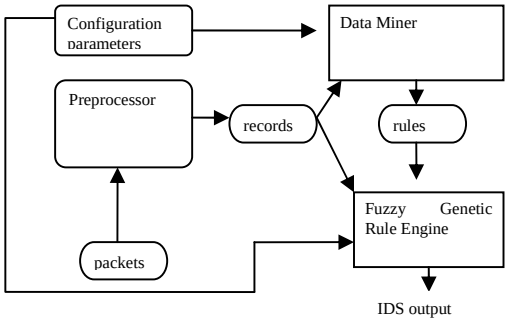


Fig 2 System Architecture

## 3. Preprocessor:

The preprocessor is responsible for accepting raw packet data and producing records. This component is used in both modes and is capable of reading packets from the wire or a tcpdump file. The output produced by this component consists of records. Records contain aggregate information for a group of packets. Using records and concentrating only on attributes of interest greatly helps in reducing the amount of information to be used by more computationally intensive components of the architecture. Most of the approaches in the literature [7] differ on how those attributes are selected. Here in this approach uses a light weight technique that employs positive and negative examples to identify the subset of attributes that provides the largest information gain. This is done by focusing on the subset attributes that contain the majority of the positive examples. The use of supervised learning in the preprocessor helps improve the effectiveness of the unsupervised learning algorithm used in the data miner by selecting relevant data subsets.

## 4. Configuration Parameters:

Parameter values stored in the configuration file regulate operation of the Data Miner and Fuzzy Inference Engine. The configuration file associates attributes with a term set and describes functions corresponding to the fuzzy membership functions associated with each term. The structured file identifies the number and names of attributes followed by a description of each attribute.

## 5. Data Miner:

The Data Miner integrates Apriori to produce fuzzy rules. With one pass through the records, the fast and efficient algorithm used by the Data Miner extracts rules with sufficient support and confidence.

## 6. Rules:

Rules are expressed as a logic implication $p \rightarrow q$ where p is antecedent and q is the consequence. Both p and q are assumed to be in conjunctive normal form, where $aa_i$, $ca_j$ and $cat_{attr}$ denote an antecedent attribute, a consequent attribute and an attribute category respectively.

The following three conditions hold for each rule:

$$cat_{aai} \in TERMS\ (aa_i) \qquad \forall i.0 \le i \le m$$
$$cat_{caj} \in TERMS(ca_j) \qquad \forall j.0 \le j \le n$$

$$\{aa_0, aa_1 .....aa_m\} \cap \{ca_0, ca_1 .....ca_n\} = \theta$$

A typical rule structure is shown in Fig.3

If $aa_0$ is cat $aa_0$
$\wedge$ $aa_1$ is cat $aa_1$
:
$\wedge$ $aa_m$ is cat $aa_m$
Then $ca_0$ is cat $ca_0$
$\wedge$ $ca_1$ is cat $ca_1$
:
$\wedge$ $ca_n$ is cat $ca_n$

Fig 3 Rule Structure

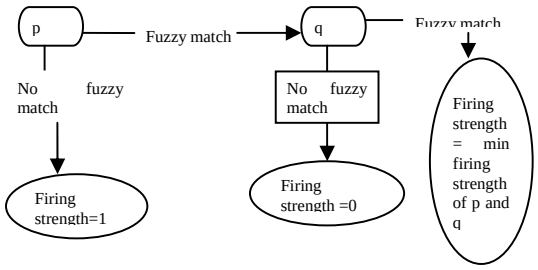


Fig 4 Analysis of fuzzy rules

The evaluation of rules of Fig 4 begins with the analysis of the antecedent, p. The following two cases are considered for the antecedent p.

- p does not have a fuzzy match so the rule does not apply to the record
- p does have a fuzzy match and the analysis of the consequent q begins

Note that a fuzzy match occurs when the truth value of the predicate is greater than zero. Similarly, the following two cases are considered for the consequence q:

- q does not have a fuzzy match and the firing strength of the rule is zero.
- q has a fuzzy match and the firing strength is determined using Mamdani inference mechanism

Fuzzy rules, as produced by the data mining algorithm, model a behavior represented by the data set employed to run the algorithm. The output of the Fuzzy Inference Engine is the firing strength of each rule for a given fact. This firing strength of each rule for a given fact determines whether or not the fact satisfies the modeled behavior. Firing strengths that have a value close to one indicate that observed behavior closely follows the model behavior, but when several facts register firing strengths at or close to zero for a given rule, then it is likely that a deviation from the model (an attack) has been detected [7].

## 7. Algorithm:

Input: Measurement from network traffic data
and
Threshold value for similarity

Output: Detected or null

Assumptions:

1) The parameters for network intrusion are assumed which form the bases for the input
2) The existence of trained normal data set (in the experiment conducted, we have assumed the data of one timing is chosen as the normal trained set)

Step1: Identify and collect relevant data from network traffic.

Step2:Convert the quantitative feature of the data in step 1 into fuzzy sets

Step 3: Define membership function for fuzzy variable

Step 4: Apply genetic algorithm to identify the best set of rules.

Step 5: For each of the rules identified in the step 4 do

a) Apply the fuzzy association rule algorithm to mine the correlation among them
b) Apply fuzzy frequency algorithm to mine sequential patterns

Step 6: For each test case generate new patterns using the fuzzy association algorithm for same parameters

Step 7: For each new pattern, compare it with normal patterns created by Training data for similarity

Step 8: IF the similarity > the threshold value
Then report “Detected” and the pattern
Else report “Null”.

## 8. Experimental Results:

The proposed work is implemented in a network environment having 60 nodes connected to a server at a local computer centre. As it is very hard to collect large amount of actual intrusion data some normal data with different behavior is treated as equivalent and used for training (many researchers in this area have proposed this method to be a sound method of expectation). One of the servers has been monitored and its real time network traffic data was collected by tcpdump. From the data four measurements, namely, SN, FN, RN and PN have been extracted.

SN indicates the number of SYN flags appearing in TCP header packets during last 10 seconds. And FN denotes the number of FIN flags appearing in TCP header packets during last 10 sec. While RN and PN respectively denote the number of RST flag appearing in TCP packet header and the number of destination ports during last 10 seconds. A statistical calculation is also done for overlapping the time periods of 10 seconds as shown in Fig. 5

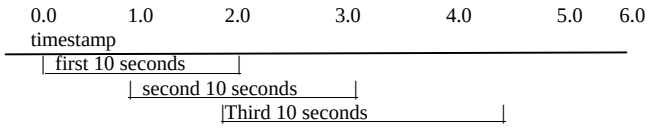


Fig 5 Specification of statistical measurements used in the experiments

Each of the above four features is viewed as fuzzy variable and they are divided into 5 fuzzy sets namely VERY_LOW, LOW, MEDIUM, HIGH, VERY_HIGH. Membership functions have been defined for fuzzy variables representing each of these features of the network being monitored.

The fuzzy association algorithm has been applied to mine the correlation among the first four features after applying genetic algorithms technique on these association rules. As the data set is only a simulated one and not large enough the genetic algorithms approach has selected almost all rules. The fuzzy frequency algorithm is applied for mining the sequential patterns for the last feature is done by using fuzzy frequency episode algorithm. The network traffic data was partitioned into different parts according to different timings of their collection. The data was collected in 4 different timing, namely morning, afternoon, evening and late night. These slots are chosen because the patterns likely to exhibit different behavior. After choosing the morning data as the data anomaly detection was then conducted on traffic on four time slots. Patterns were established from training data by data mining. An example of fuzzy association rule from training data is

{SN=VERY_LOW,FN=LOW} ⟶ {RN= VERY_LOW}

0.863, 0.19 this means that

{SN=VERY_LOW,FN= LOW, RN=VERY_LOW} occurred in 19% of the training cases with confidence value of 86.3%. Further more when

SN = VERY_LOW, FN= LOW,

the probability of RN = VERY_LOW is 86.3%.

We have 24 test cases and for each test case new patterns were mined using same algorithms and same patterns. These new patterns were then compared to the normal patterns created from training data. If the similarity between them is less than a threshold value then no intrusion is detected otherwise a possible intrusion is detected. The similarity function for identifying the similarity between any two association rules between $R_1$ and $R_2$ as defined as follows. Here $R_1$ is normal association and $R_2$ is new association rule.

IF $R_1$: X ⟶ Y, c, s

And

$R_2$: X' ⟶ Y', c',s'

Then

$$similarity(R_1,R_2)=\begin{cases} & \text{if } ((X\neq Y)\vee(X'\neq Y')); \\ 0 & \\ \max\left(0,1-\max\left(\frac{|c-c'|}{c},\frac{|s-s'|}{s}\right)\right) & \text{if } ((X=Y)\wedge(X'=Y')) \end{cases}$$

Where $$s=\sum_{\substack{\forall R_1\in S_1\\ \forall R_2\in S_2}} similarity(R_1,R_2)$$

Here $S_1$ is the set of normal pattern and $S_2$ is the set of new patterns with similarity of $S_1$ and $S_2$ defined as

$$similarity(S_1,S_2)=\frac{s}{|S_1|}*\frac{s}{|S_2|}$$

here $|S_1|$ and $|S_2|$ denote the total number of rules of $S_1$ and $S_2$ respectively.

The training sets of different duration are from same time period is used to mine the fuzzy association rules. The similarity of each set of rules which were derived from these were compared to the test data for different time periods. The similarity evaluation for fuzzy episode rules is almost same as fuzzy association rule except one more parameter called window length for an episode rule. The window threshold should be identical when two episode rule is to be evaluated for their similarity. The results have demonstrated that the similarity of rules is likely for the normal and the new

association rules in the period of evening and late night. This brings to infer that the possibility of intrusion into the network occurs more during evening and late night time slots. Surprisingly the possibility of intrusion during the morning and afternoon time slots where, heavy network traffic is observed, is low. This may be attributed to minimal access to the network ports to enter into the network at that time. Another important observation is that the system is able to identify the patterns of the users who try to access the network in the time slots which they normally do not follow, though they are authorized users. For example, one observation is that a user who logs onto the system 5 times in the morning and repeats on to log on at occurred in different time slots later can also be detected.

## 8. Conclusion:

Intrusion Detection is one of the major concerns in any computer networks environment. Many techniques including that of Artificial Intelligence have been proposed and are in use presently. There are many researchers who developed intelligent Intrusion Detection Systems. The input to any Intrusion Detection System is some uncertain or fuzzy information that has to be processed. A part from being fuzzy in nature the information could be very large requiring data mining techniques for extracting the data. As the data for extracting has to follow certain rules, we need to have certain mechanism to pick up best possible rules. A genetic algorithm approach for identifying these rules is chosen. The present work has explored the possibility of integrating the fuzzy logic with Data Mining methods using Genetic Algorithms for intrusion detection. The reasons for introducing fuzzy logic is two fold, the first being the involvement of many quantitative features where there is no separation between normal operations and anomalies. Thus fuzzy association rules can be mined to find the abstract correlation among different security features.

The present work is the extension of in the areas of fuzzy association rules based on mining and genetic algorithms. We have proposed architecture for Intrusion Detection methods by using Data Mining algorithms to mine fuzzy association rules by extracting the best possible rules using Genetic Algorithms.

## 9. Future Work:

The work can be extended further by using Dempster-Shafer theory. This Dempster-Shafer theory approach considers sets of propositions and assigns to each of them an interval.

[Belief, Plausibility]

In which the degree of belief lies. Belief is a measure that gives the strength of evidence in favor of set propositions. It ranges from 0 to 1, where 0 indicating no evidence and 1 indicating certainty. The plausibility measures the extent to which evidence in favor of negation of S leaves no room for belief in the attribute S. Plausibility also ranges from 0 to 1. The belief – Plausibility interval measures not only the level of belief in some propositions but also the amount of information it has.

The work on be fuzzy logic based reasoning can be replaced by Dempster -Shafer theory, the fuzzy association rules in data mining. However, sufficient care needs to be taken for assigning belief values for proposition. A statistical analysis has to be done on past data for assign belief values which in turn can be used further.

## References:

[1] Susan M.Bridges, Rayford B. Vaughan “Fuzzy Data mining and Genetic Algorithms Applied to Intrusion Detection”, Conference onNational Information Systems Security, Oct. 2000.

[2] K.C.C.Chan and W.H.Au, “Mining Fuzzy Association Rules” Proc.Of ACM CIKM, 1997,pp.209- 215.

[3] M.Delgada, Nicolas Marin, Daniel Sanchez and Maria Amparo Vila“Fuzzy Association Rules:General Model and Applications”, IEEE Transactions on Fuzzy Systems, Vol. 11, No. 2, April 2003, pp.214-225.

[4] K.M. Faraoun, and A. Boukelif ”Genetic Programming Approach for Multi-Category Pattern Classification Applied to Network Intrusions Detection”, International Journal of Computational Intelligence Vol.3, No.1 2006,pp.79-90.

[5] Francisco Herrera “Genetic Fuzzy Systems: Status,Critical Considerations and Future Directions” International Journal of Computational Intelligence Research, Vol 1. No.1, 2005, pp.59-67.

[6] Ramakrishna Srikant and Rakesh Agrwal “Mining quantative Association rules in large relational tables” Proc. Of ACM IGMOD,1996, pp. 1-12.

[7] Aly EI Semary, Jamica Edmonds, Jesus Gonzalez-Pino, Mauricio Papa “Applying Data mining of Fuzzy Association Rules to Network Intrusion Detection”, IEEE Proc. On Information Assurance, West Point, New York, 2006,pp.100-107.

**MrsY.Dhanalakshmi** received her MCA and M. Phil degrees in 2000 and 2003 respectively. At present she is pursuing her Ph.D from Acharya Nagarjuna University A.P, India. She has published papers in journals. Her Area of interest is Data Mining, Network Security.

**Prof. I.Ramesh Babu** received his B.E. and M.E. degrees in 1981and1984 respectively. And Ph.D from Nagarjuna University in 1994. He joined as an Assistant Professor in the Department of Computer Science and Engineering in Acharya Nagarjuna University in 1988,and became a Professor in 2004. He held many positions in Acharya Nagarjuna University as Executive council member, Chairman Board of Studies, Head, Director Computer Centre, Member of academic senate, member of the standing committee of academic senate. He was a special officer, convenor of ICET. He is also a member of Board of Studies for other universities. He has published many research papers in International, national journals and presented papers in international conferences also. His research areas of interest include Image Processing, Computer Graphics, Cryptography, Network Security and Data Mining. He is member of IEEE, CSI, ISTE, IETE, IGISS, Amateur Ham Radio (VU2UZ