Emerging Intuitionistic Fuzzy Classifiers for Intrusion Detection System

Kavitha B.
Research Scholar, Bharathiar University, Coimbatore, India
Email: kavitha_gana2006@yahoo.co.in

Karthikeyan S.
Assistant Professor, Department of Information Technology,
College of Applied Sciences, Sohar, Sultanate of Oman
Email: skaarthi@gmail.com

Sheeba Maybell P.
Lecturer, Department of Mathematics, Karpagam University, Coimbatore, India
Email:Sheeba.maybell@yahoo.com

Abstract— **One of the toughest challenges in Intrusion Detection System is uncertainty handling. The normal and the abnormal behaviors in networked computers are hard to predict as the boundaries cannot be well defined. The prediction of the normal or abnormal behaviors is done by the comparison with predefined classes to find the most similar one. This prediction process may generate false alarms in many anomaly based intrusion detection systems. Consequently, we observed uncertainty where there is a fair chance of the existence of a non-null hesitation part at each moment of evaluation of an unknown object. A new technique is implemented in this paper using Intuitionistic fuzzy logic which is a generalization of fuzzy logic. In this model the false alarm rate in determining intrusive activities can be reduced. A set of Intuitionistic fuzzy rules can be used to define the normal, abnormal and indeterministic behavior in a computer network. An Intuitionistic fuzzy inference algorithm can be applied over such rules to determine when an intrusion is in progress. The main problem with this approach is to generate good Intuitionistic fuzzy classifiers to detect intrusions. The rules generated by Intuitionistic fuzzy classifiers are fine tuned using improvised genetic algorithm that can detect anomalies and some specific intrusions. The main idea is to evolve three rules, one for the normal class, second for the abnormal class and third of indeterministic class using KDD Cup 99 Dataset. This paper exhibits the performance of Emergent Intuitionistic fuzzy classifiers in intrusion detection.**

Index Terms—**Intrusion Detection, Prediction, normal, abnormal, Intuitionistic Fuzzy Classifiers, indeterministic, and genetic algorithm.**

I. INTRODUCTION

An Intrusion Detection System (IDS) is a program that analyses what happened or what has happened during an execution and tries to find indications that the computer has been misused. An IDS does not eliminate the use of preventive mechanism but it works as the last defensive mechanism in securing the system. The demand for secured communication is increasing rapidly. Information security is about protecting the information against unauthorized disclosure, transfer or modification accidentally or intentionally when it is transmitted through the Network. A significant challenge in providing an effective defense mechanism is to a network perimeter is having the ability to detect intrusions and implement counter measures. Intrusion detection system is a system for detecting computer intrusions and misuse by monitoring system activity and classifying it as either normal or anomalous [1]. In Initial stages of intrusion detection on network focused on sensors how to get data how to display data and what kind of data to handle. Later when the analyst team grew to handle the load and training and team coordination were the issues of the day. When the data grew larger, system was inadequate for detecting the most dangerous attacks [2].

The concept of intrusion detection was first introduced by Anderson to complement conventional computer security approaches in 1980[3]. Anderson defined an intrusion attempt or a threat to be a deliberate unauthorized attempt to access information, manipulate information or render a system unreliable or unusable. IDS can be classified in two broad categories: Misuse Detection and Anomaly Detection.

Misuse Detection: It looks for previously described patterns of behavior that are likely to indicate an intrusion.

Anomaly Detection: It looks for deviations from stored patterns of normal behavior

The problem of intrusion detection has been studied extensively in computer security ([4], [5] and [6]) and recent research experiments show that the data mining approaches lead to new directions by creating models for intrusion detection. In complex classification domains of intrusion detection, attributes of datasets may contain false correlation, which hamper the process of detecting

doi:10.4304/jait.2.2.99-108

intrusions. Some attributes of datasets may be redundant because the information they add is contained in other attributes. Also some data in the dataset may not be useful for intrusion detection and thus can be eliminated before learning. In this paper Best First Search is adopted to reduce the problem of effective attributes selection [7].

Integration of learning algorithms provides a potential solution for the adaptation and accuracy issues. Many different solution based on machine learning techniques have been deployed for intrusion detection in both commercial systems and the state-of-the art. In standard set theory each element is either completely in or not in a set this leads to the problems in case of vague concepts. Recent works deal with the fuzzy system but they fail to solve the indeterminacy (unknown) problem.

In this paper, we show the certainty of improvised genetic algorithm [8] to evolve simplest set of Intuitionistic fuzzy rules that can solve some well–studied intrusion detection problems. In this approach, genetic algorithm can find good and simple Intuitionistic fuzzy rules to characterize normal, abnormal and indeterministic behavior of network systems. This Intuitionistic fuzzy logic can reduce the false signal rate in determining intrusive behaviors.

The paper is organized as follows, Section 2 describes the related work in the field of Intrusion Detection System, section 3 explains the basic Intuitionistic fuzzy logic and Intuitionistic fuzzy classifiers concepts used in this paper, section 4 presents the proposed approach to solve some intrusion detection problems. In section 5 experiments and results are presented. Finally in section 6 conclusion and future work are shown

II. RELATED WORK

In the recent past there has been a growing recognition of deploying intelligent techniques for the creation of efficient and reliable intrusion detection systems. A complete survey of these techniques is hard to be presented at this point, since there are more than hundred IDS based on machine learning techniques. MIT Lincoln Lab's DARPA intrusion detection evaluation dataset is employed to design and test intrusion detection systems. In 1999, recorded network traffic from the DARPA'98 Lincoln Lab dataset [9] was summarized into network connections with 41-features per connection. This formed the KDD'99 intrusion detection benchmark in international Knowledge Discovery and Data mining tools competition [10].

Srinivas et al [11] describes approaches to intrusion detection using neural networks and support vector machines. The key ideas of their research are to discover useful patterns or features that describe user behavior on a system, and use the set of relevant features to build classifiers that can recognize anomalies and known Intrusions. The temporal association rules technique generates fuzzy and classical rules [13]. Using short sequences of system calls that running programs perform as discriminators between normal and abnormal operating characteristics [14].The discriminator uses the Hamming distance as a distance function between short sequences of system calls. If the distance of a particular sequence to the normal sequences is higher than a threshold then the sequence is abnormal.

The idea of restarting the classic GA, so as to increase the performance, derives from the well known idea of restarting the Arnoldi's method for finding the eigenvalues [15]–[18]. Fuzzy Association rule are used to explore the possibility of integrating the fuzzy logic with Data Mining methods using Genetic Algorithms for intrusion detection [19]. A technique to generate fuzzy classifiers using genetic algorithms that can detect anomalies and some specific intrusion using two rules for normal class and other for the abnormal class are proposed using evolved fuzzy classifiers in intrusion detection.

In this paper to handle the uncertainty problem more precisely we will present Intuitionistic Fuzzy Logic (IFL) as a tool for reasoning in the presence of imperfect facts and imprecise knowledge. The rules generated using IFL is fine tuned by improvised genetic algorithm [8] which adapts the idea of restarting a Genetic Algorithm in order to obtain better knowledge of the solution space of the IDS problem.

III. INTUITIONISTIC FUZZY CLASSIFIERS FOR INTRUSION DETECTION

Handling uncertainty is one of the major concerns intrusion detection environments. The goal is to classify the patterns of the system in three categories (normal, abnormal and indeterministic), using patterns of known attacks, which belong to the abnormal class, patterns of the normal behavior and pattern of unknown class. Using intuitionistic fuzzy rules, the solution of this classification problem is based on the intuitionistic fuzzy logic concept.

A. Intuitionistic Fuzzy logic

A number of generalizations of fuzzy set theory are there in the literature developed by several authors with more general aims and objectives to deal with various types of problems of Computer Science, Information Technology, Social Science, Decision Science, Management Science, etc. to list a few only. Of these, the notion of intuitionistic fuzzy set theory (IFS theory) introduced by Atanassov ([20], [21]) is of interest to us. Fuzzy sets can be viewed as intuitionistic fuzzy sets, but the converse is not true. It has been asserted by many authors that there are a large and large number of real life problems for which IFS theory is a more suitable tool than fuzzy set theory for searching solutions. IFS can be useful in situations when description of a problem by a fuzzy variable, given in terms of a membership function only, seems too rough.

The key improvement of intuitionistic fuzzy set theory over fuzzy set theory is that in the latter, the membership value of an object also defines the non-membership value of it by means of a mathematical relation, whereas in the former the membership value and non-membership value

of an object are not, in general, related by a mathematical equation. Rather, the decision-maker (or the problem analyst or the intelligent agent) independently decides both, up to his best intellectual capability. This is because, when deciding the degree of membership of an object there may be some hesitation.

A fuzzy set could be viewed as a special case of intuitionistic fuzzy set, provided that at the processing stage for evaluation of membership value, there is no indeterministic situation with respect to any object of the universe of discourse.

An Intuitionistic fuzzy set(IFS) A on a universe X is defined as an object of the following form

$$A = \{< x, \mu_A(x), \nu_A(x) > | x \in X\}$$

Where the functions

$$\mu_A : X \rightarrow [0,1] \text{ and } \nu_A : X \rightarrow [0,1]$$

defines the degree of membership and the degree of non-membership of the element $x \in X$ in A, respectively and for every $x \in X$

$$0 \leq \mu_A(x) + \nu_A(x) \leq 1$$

Obviously, each ordinary fuzzy set may be written as

$$\{< x, \mu_A(x), 1-\nu_A(x) > | x \in X\}$$

Recently, the necessity has been stressed of taking into consideration a third parameter $\pi_A(x)$, known as the intuitionistic fuzzy index or hesitation degree, which arises due to the lack of knowledge or 'personal error' in calculating the distances between two fuzzy sets [22]. In fuzzy set, non-membership value is equal to 1 – membership value or the sum of membership degree and non-membership value is equal to 1. This is logically true. But in real world this may not be true as human being may not express the non-membership value as 1-membership value. This is due to the presence of uncertainty or hesitation or the lack of knowledge in defining the member ship function. This uncertainty is named as hesitation degree. Thus the summation of three degrees, i.e., membership, non-membership and hesitation degree is 1. It is obvious that $0 \leq \pi_A(x) \leq 1$, for each $x \in X$. So, with the introduction of hesitation degree, an intuitionistic fuzzy set A in X may be represented as

$$A = \{< x, \mu_A(x), \nu_A(x), \pi_A(x) > | x \in X\}$$

with the condition $\mu_A(x) + \nu_A(x) + \pi_A(x) = 1$.

In figure 1 the object x has denoted both by the degree of membership and non-membership set. The IFL allows an object to belong to different classes at the same time. This concept is helpful when the difference between classes is no well defined.

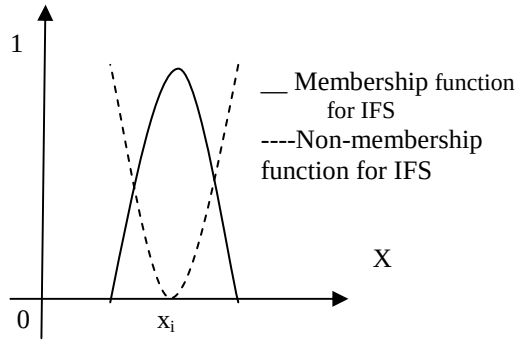

Figure 1. Triangular membership and non-membership function for a Intuitionistic fuzzy set

It is the case in the intrusion detection task, where the difference between the normal and abnormal class are not well defined.

Using these linguistic concepts, atomic and complex intuitionistic fuzzy logic expressions can be built. An atomic intuitionistic fuzzy expression is an expression:

parameter is [not] Intuitionisticset

Where, parameter is an object and Intuitionisticset is a defined intuitionistic fuzzy space for the parameter. The truth value(TV) of an atomic expression is the degree of membership and non-membership of the parameter to the intuitionistic fuzzy set. Because TVs are expressed by numbers between 0 and 1(0 means entirely false, 1 means entirely true and other values means partially true), the intuitionistic fuzzy expression evaluation process is reduced to arithmetic operations. Also for each classical logical operator and fuzzy logic arithmetic operator there is a common Intuitionistic fuzzy logic arithmetic operator which is shown in the table II

TABLE II

Intuitionistic Fuzzy logic operator

Logical operator	Fuzzy operator	Intuitionistic operator
p AND q	Min{p,q}	$<x, \min\{\mu_p(x), \mu_q(x)\}, \max\{\nu_p(x), \nu_q(x)\} \mid x \in X>$
p OR q	Max{p,q}	$<x, \max\{\mu_p(x), \mu_q(x)\}, \min\{\nu_p(x), \nu_q(x)\} \mid x \in X>$
NOT p	1.0 - p	$<x, 1.0 - \mu_p(x), 1.0 - \nu_p(x) \mid x \in X>$

IFS rules have the form:

R: IF condition THEN consequent [weight]

Where,

- *Condition* is a complex intuitionistic fuzzy expression, i.e. that uses Intuitionistic fuzzy logic operators and atomic intuitionistic fuzzy expressions
- *Consequent* is an atomic expression, and
- *Weight* is a real number that defines the confidence of the rule.

B. Intuitionistic Fuzzy Classifiers and three classes classification problem

In the three class classification problem, there are three classes where every object should be classified. These classes are called normal, abnormal and indeterministic. The data set used by the learning algorithms consists of a set of objects, each object with n+1 attributes. The first n attributes define the object characteristics (monitored parameters) and the last attribute defines the class that the object belongs to (the classification attribute).

A Intuitionistic fuzzy classifier for solving the three class classification problem is a set of three rules, one for the normal class next for the abnormal class and the last for indeterministic class, where the condition part is defined using only the monitored parameters and the conclusion part is an atomic expression for the classification attribute.

Example for both membership and non membership elements intuitionistic fuzzy rule as follows.

R_N: If x is high and y is low then
pattern is normal [0.3]

R_A: If x is medium and y is high then
pattern is abnormal [0.5]

R_I: If x is medium and y is medium then
pattern is indefinite [0.2]

R_N - is the rule for the normal class
R_A – is the rule for abnormal class and
R_I – is the rule for indeterministic class

The Intuitionistic fuzzy rule truth-value is calculated as the product of the condition truth-value by the weight, i.e.

$$TV(R) = TV(\text{Condition}) * \text{Weight}$$

For membership function:

$TV(R_N)$ = TV (If x is high and y is low) * [0.3]

$TV(R_A)$ = TV (If x is medium and y is high) * [0.5]

$TV(R_I)$ = TV (If x is medium and y is medium) * [0.2]

Example:

For membership, $\mu(A)$

$$TV_{\mu(A)}(R_N) = TV(x \text{ is high and } y \text{ is low}) * (0.3) = \min\{0.2, 0.7\} * 0.3 = 0.2 * 0.3 = 0.06$$

$$TV_{\mu(A)}(R_A) = TV(x \text{ is high and } y \text{ is low}) * (0.5) = \min\{0.2, 0.7\} * 0.5 = 0.2 * 0.5 = 0.10$$

$$TV_{\mu(A)}(R_I) = TV(x \text{ is high and } y \text{ is low}) * (0.2) = \min\{0.2, 0.7\} * 0.2 = 0.2 * 0.2 = 0.04$$

For non-membership, $\nu(A)$:

$$TV_{\nu(A)}(R_N) = TV(x \text{ is high and } y \text{ is low}) * (0.3) = \max\{0.5, 0.2\} * 0.3 = 0.5 * 0.3 = 0.15$$

$$TV_{\nu(A)}(R_A) = TV(x \text{ is high and } y \text{ is low}) * (0.5) = \max\{0.5, 0.2\} * 0.5 = 0.5 * 0.5 = 0.25$$

$$TV_{\nu(A)}(R_I) = TV(x \text{ is high and } y \text{ is low}) * (0.2) = \max\{0.5, 0.2\} * 0.2 = 0.5 * 0.2 = 0.10$$

In Intuitionistic fuzzy classifier for the three classes classification problem there are several techniques to determine the class that an object belongs to. One of these techniques is the maximum technique, which classifies the object as the class in the conclusion part of the rule that has the maximum truth-value, i.e.:

$$\text{Class} = \begin{cases} N \text{ - If } TV(R_N) > TV(R_A) > TV(R_I) \\ A \text{ - If } TV(R_A) > TV(R_N) > TV(R_I) \\ I \text{ - If } TV(R_N) = TV(R_A) \end{cases}$$

Where,
N - represents the Normal class,
A - represents the Abnormal class and
I- Indeterministic

If the two rules produces the same truth-value then it is considered as indeterministic class.

IV. GENETIC ALGORITHM RESTARTINGS

To improve the performance of classic genetic algorithm we adopted the concept of restarting procedure for the classic GA which was introduced by Grigorios N. Beligiannis et.al, to achieve a better global exploration of the solution space while executing the minimum possible number of generations (function evaluations). In order to achieve this goal, they used the standard global exploration mechanism used by classic GAs (selection, crossover, mutation) but when the GA reaches the local refining phase, we restart the GA so as to preserve the global search procedure. This technique alleviates the enormous computational burden introduced by the local refining procedure, which is quite often useless in finding the optimal solution. The technique is described in Fig. 2. Of course, the new starting of the GA procedure should include all the valuable information gathered from the

previous global search. Thus, we propose a new operation called "insertion" to be included in the classic GAs' evolution procedure.

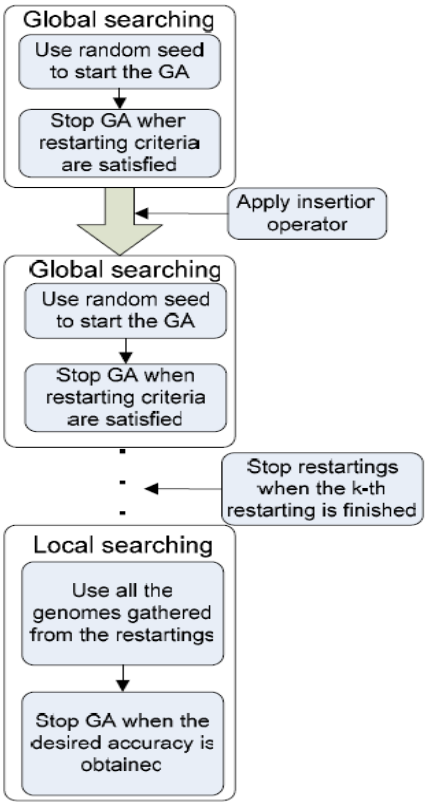

Figure 2. Procedure of the Genetic Algorithm with restartings

The insertion operator works as follows. It chooses randomly a constant percentage of the genomes of the population of the last generation (before the restarting procedure takes effect) and inserts them into the new initial population of the GA as shown in Fig.3.

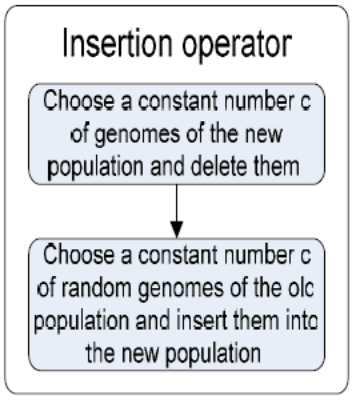

Figure 3. Insertion Operator

In this contribution, three different criteria for deciding when to apply restartings are proposed:

- Fitness function value
- Number of generations
- Mean fitness function value of population

Operator used in Genetic Algorithm Restartings

Crossover operator: Suppose if s1 and s2 are two chromosomes then they are represented as
$S_1 = \{S_{11}, S_{12}, S_{13}..... S_{1n}\}$,
$S_2 = \{S_{21}, S_{22}, S_{23}..... S_{2n}\}$ are

Two chromosomes, select a random integer number $0 \leq r \leq n$, S_3 and S_4 are offspring of crossover(S_1, S_2),

$S_3 = \{S_i \mid \text{if } i \leq r, S_i \in S_1, \text{ else } S_i \in S_2\}$,
$S_4 = \{S_i \mid \text{if } i \leq r, S_i \in S_2, \text{ else } S_i \in S_1\}$

Mutation Operator: Suppose a chromosome Si = $\{S_{11}, S_{12}, S_{13}..... S_{1n}\}$ Select a random integer number $0 \leq r \leq n$, S_3 is a mutation of S_1,

$S3 = \{Si \mid \text{if } i \neq r, S_i \in S_{1i}, \text{ else } S_i \in \text{random}(S_{1i})\}$

Selection Operator: Suppose there are m individuals, we select $[\frac{m}{2}]$ individuals but erase the others, the ones we select are having more fitness that means their profits are greater.

Insertion Operator: Suppose there are m individuals, choose a constant number C having genomes of the new population and delete them. At the same time, choose a constant number C of random genomes of the old population and insert them into the new population.

Improvised Genetic Algorithm:

1. Initialize the population: Producing a number of individuals randomly, each individual is a chromosome which is an n-length array, is the number of parameters.
2. Test if one of the stopping criteria (running time, fitness, generations etc) holds. If yes, stop the genetic procedure.
3. Selection: Select the better chromosomes. It means the profit under these parameters is greater.
4. Applying the genetic operators: such as crossover and mutation to the selected parents to generate an offspring.
5. Recombine the offspring and current population to form a new population with selection operator.
6. Insertion: Choose a 'C' constant number for new population and delete it. Add 'C' constant number of random population to form a new population.
7. Repeat step 2 to 6.

V EVOLVING INTUITIONISTIC FUZZY CLASSIFIERS

In order to use improvised genetic algorithm to evolve the intuitionistic fuzzy rules, each rule is represented by a complete tree. We opt to seek the best classification rule for each class separately because this leads to a much

faster and simpler search, has the potential to yield simpler rules, and also because this yields an approach that can be easily parallelized on several independent processors, especially in the presence of many classes. Because we run the genetic algorithm once for each class and we want compressible rules, the optimization problem is a three-goal objective function: maximize the sensitivity, maximize the specificity, and minimize the rule length.

For example a rule (((A or E) and C) or (B and D)) can be represented as free parenthesis logical expression {A or B and C and D or E}. The tree looks like Figure 4:

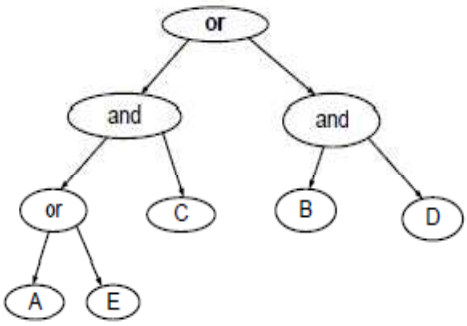

Figure 4. Complete Tree for free parenthesis expression.

To establish the process we used the following grammar (in Backus Normal Form) for a free parenthesis logical expression:
<EXP> → <EXP><OPER><ATOMIC> | <ATOMIC>
<ATOMIC> → variable is [not] set
<OPER> → or | and

Applying repeatedly the previous definition, the following logical expression can be obtained:

A or B and C and D or E

where *A*, *B*, *C*, *D*, and *E* are atomic expressions.

A Intuitionistic fuzzy classifier can be represented by a set of *m* rules, where *m* is the number of different classes (a rule per class):

R_1: IF $condition_1$ THEN data is $class_1$
....
R_m: IF $condition_m$ THEN data is $class_m$

Likewise the improvised genetic algorithm for the normal class tries to find a intuitionistic fuzzy rule with the following expression

(x1 is C OR y1 is not D) AND z1 is E

Can be represented without parenthesis using above mentioned complete tree expression as follows

x1 is C AND z1 is E OR y1 is not D

We codified a logic expression without parenthesis of n logic operators in a chromosome of n+1genes, where the i-th gene is composed by the atomic expression A_i and the logic operator O_i. In this way the last gene has an unused logic operator. Figure 5 shows the chromosome for an expression tree of n logic operators

Gen_1				...	Gen_n				Gen_{n+1}			
ac_1			op_1	...	ac_n			op_n	ac_{n+1}			*
var_1	ro_1	set_1		...	var_n	ro_n	set_n		var_{n+1}	ro_{n+1}	set_{n+1}	*

Figure 5. Chromosome for a complete rule condition

As an example expression encoded in the chromosome using operator priority is as follows:

Gen_1				Gen_2				Gen_3			
ac1			op1	ac2			op2	ac3			*
x1	yes	C	AND	z1	yes	E	OR	y1	NOT	D	*

Figure 6. Coding the expression x1 is C AND z1 is E OR y1 is not D

In our work the fitness of a chromosome for the normal class is evaluated according to the following set of equations:

$$TP = \sum_{i=1}^{p} \text{predicted}\ (\text{normal\_data}_i)$$

$$TN = \sum_{i=1}^{p} [1 - \text{predicted}\ (\text{abnormal\_data}_i) - \text{predicted}\ (\text{indeterministic\_data}_i)]$$

$$FP = \sum_{i=1}^{p} \text{predicted}\ (\text{abnormal\_data}_i)$$

$$FN = \sum_{i=1}^{p} [1 - \text{predicted}(\text{normal\_data}_i) - \text{predicted}(\text{indeterministic\_data}_i)]$$

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

$$\text{Specificity} = \frac{TN}{TN + FP}$$

$$\text{Length} = 1 - \frac{Chromelength\ (rules)}{10}$$

Fitness = w_1 * sensitivity + w_2 * specificity + w_3 * length

Here *p*, *q* and *r* are the number of normal, abnormal and indeterministic samples in the dataset used by each chromosome respectively.
predicted is the IFS value of the condition part of the codified rule.

In the above equations TP means true positive, TN means true negative, FP means false positive, FN means false negative for the codified rules respectively.
w_1, w_2, w_3 are the assigned weights for each rule characteristics respectively.
$normal\_data_i$ is the subset of normal training patterns
$abnormal\_data_i$ is the subset of abnormal training patterns and,

indeterministic_data$_i$ is the subset of indeterministic training patterns.

The same set of equations to calculate the fitness for the abnormal class can be obtained by replacing abnormal for normal in previous equations. The best chromosome in the population is chosen and the intuitionistic fuzzy rule:

If <condition> then pattern is <class>

is added to the intuitionistic fuzzy classifier. Here, <condition> is the condition represented by such chromosome, and <class> is the class pattern evolved by the improvised genetic algorithm.

VI DATASET DESCRIPTION

A. KDDcup'99 Dataset

In this paper, Kddcup'99 data set is used which is based on the 1998 DARPA ([6],[24]). Normal connections are created to profile that those expected in a military network and attacks fall into one of the following four categories namely Denial of Service (DoS), Remote to Local (R2L), User to Root (U2R) and Probe. The various types of attack in our experimental dataset which are classified into four categories are shown in the following table III

TABLE III
Various Attack Types

Categories	Attack Types
DOS	Apache2, Back, Land, Mail bomb, Neptune, Pod, process Table, Smurf, Tear drop, Udpstrom
PROBE	IPsweep, Mscan, nMap, Portsweep, Saint, Satan
U2R	Buffer Overflow, http tunnel, load module, perl, root kit, ps, sqlattack, xterm
R2L	Ftpwrite, guesspasswd, imap, multihop, named, phf, send mail, snmp getattack, snmpguess, warezmaster, worm, xlock, xsnoop

The KDDCup'99 Intrusion Detection benchmark is comprised of 3 components. In this work corrected KDD set is used because a dataset with different statistical distributions than either "10% KDD" or "Whole KDD" is provided by the "Corrected KDD" and is comprised of 14 additional attacks. Hence, the "Corrected KDD" dataset is being used for our experiment. The value of each connection is being predicted by this task

B. Exclusion of Dataset

As in our previous work [7] 65000 records have been selected as sample dataset out of 3, 11,029 Corrected KDD dataset connections for the work done by us. However, because the sample number of Probe, U2R, and R2L is being less, the number of records of above attack types will be constant in any sample rate. The remaining records out of 65,000 are 44,417 which are the outcome of excluding the Probe, U2R and R2L types of records. Out of 44417, 20% of Normal connection is selected, and remaining 80% of the dataset is accounted by the Dos. The data sampling number and ratio are shown in Table IV.

TABLE IV
Amount and ratio of data sampling

Category	Corrected Dataset		Randomly Selected Sampled Records	
Normal	60593	19.48%	8883	13.67%
Probe	4166	1.34%	4166	6.4%
DOS	229853	73.9%	35534	54.67%
U2R	70	.02%	70	.11%
R2L	16347	5.26%	16347	25.15%
Total	3,11,029	100%	65000	100%

VI EXPERIMENTAL RESULT

In order to demonstrate the efficiency and performance of the Intuitionistic fuzzy classifier technique, we constructed our Emerging Intrusion Detection System (EIFS) and tested their performance on the KDD-99 intrusion detection contest dataset. The feature selection and detection rules generation are two key steps in any intrusion detection system based learning algorithm. Feature extraction done using Best First Search and the category of attacks is identified using intuitionistic fuzzy classifier. Our proposed model generates the detection rule based on the Intuitionistic Fuzzy classifier. The main aim of this work is to generate good intuitionistic fuzzy classifiers to detect intrusions. This paper adopted improvised classical genetic algorithm to fine tune the rules generated by IFS. All the experiments were carried out on a Intel(R) Core(TM) i3 2.13GHz PC with 2 GB RAM. The implementation is done using MATLAB Software.

A. Dataset Preprocessing

Using uniform distribution algorithm we created a dataset from the original data set with the following property: If the sample number of k patterns is m and the original data set has n samples. Probability to find a sample of class y = $\frac{n}{m}$ samples of the final dataset in 1.0 Each dataset is normalized between 0.0 & 1.0 using the equation.

$$X = \frac{x - \min}{\max - \min}$$

Where,

x - numerical value

Min- minimum value for the attribute that x belongs to

Max- maximum value for the attribute that x belongs.

Non Numerical data: The degree of membership and Non-membership value is 0 for false and for the value of True it is 1.

B. Dimensionality Reduction

The original dataset is comprised of 41 attributes and one class label. Using Best First Search method [7] we obtained set of reduced dimensionality to 7 potential

attributes which is listed as follows : Protocol Type, Service, Srcbytes, Dstbytes, count, diff_srv_rate, dest_host_srv_count.

In this paper we used these 7 attributes to frame the atomic expressions and developed the complete expression tree which eliminates most of the insignificant rules.

C. 10 Fold Cross Validation

Ten-fold cross-validation is a Standard method for evaluation. Extensive experiments have shown that this is the best choice to get an accurate estimate. It is calculated by following ways which is represented in the figure 7

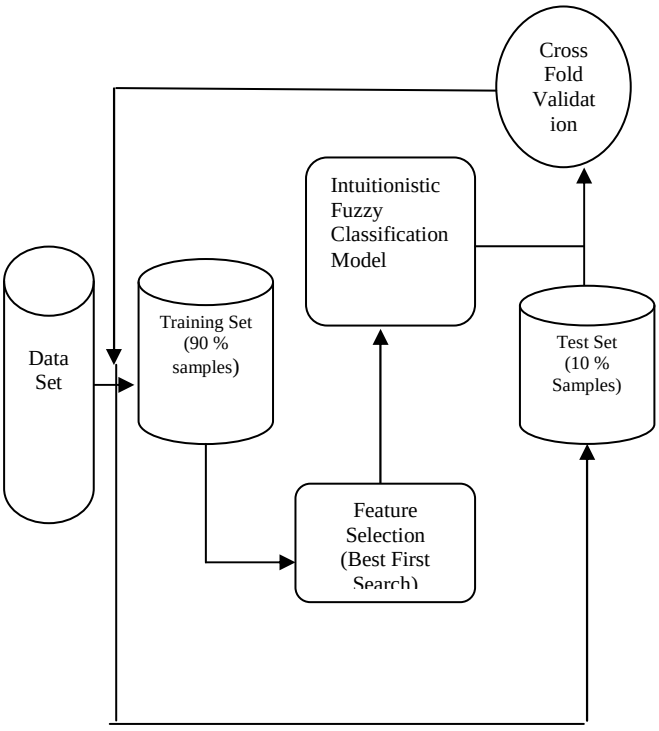

Figure 7. Repeat 10 x for 10 – Fold Cross Validation

First, the data section is divided in 2 sections randomly one section is included 90% of all dataset and called as learning dataset. Another section is included 10% of all dataset and called as validation datasets. Second, the learning dataset is used for acquiring rules and the validation dataset is used for validating the rules externally. Third, the process is repeated 10 times.

D. Implementation of Intuitionistic Fuzzy classifer

In this proposed work each detection model is trained with normal and intrusive data set. The training test data set is divided randomly in ten group. Each group was taken as testing set for the intuitionistic fuzzy classifier trained by the improvised genetic algorithm with the other nine groups. We repeated the process ten times and the score of the trained classifier was calculated as the average of the hundred tests applied.

The free parameters improvised genetic algorithm proposed in [23] is used to identify three different classes such as normal, abnormal and indeterministic. The degree of membership and the non-member ship function is applied to condition defined using only the monitored parameters and based on the truth value each record in the dataset is classified.

The improvised genetic algorithm was run for one hundred generations with one hundred individuals. With the length between one and seven genes one for each of the seven connection record attributes. It provided the necessary population breeding, randomizing, and statistics gathering functions, from which this genetic algorithm was written in MATLAB 7.0. In the initial generation, an individual with a fitness value of 0.70997 was created in run 1 and an individual with a fitness value of 0.710028 was created in run 2. The fitness value of the best individual for each generation had an approximate steady increase until generation hundred, at which point it is apparent that the best possible individual possible by the current methods had been created. This demonstrated the ability of the genetic algorithm to successfully evolve an individual's model. Information collected on each generation consisted of the mean fitness of all of the individuals within the generation, the fitness of the best performing individual, the correct detection rate and the false positive rate.

This kind of approach showed good performance in the evolution of intuitionistic fuzzy classifiers for the intrusion detection problem. Using the selected Features the conditions are framed and performed several test with different values for the fitness function weight. This allowed us to increase the speed for the genetic algorithm with huge datasets.

E. Results and Analysis

The cost function of an intrusion detection system is defined using the false alarm rate and the undetected attacks rate. The false alarm rate means the system which produces an alarm in a normal condition. The undetected attacks rate means the system considers an abnormal behavior as normal. The performance of the Emerging Intuitionistic fuzzy classifier over the hundred test performed is shown in table V.

TABLE V
Amount and ratio of data sampling

Algorithm	False Alarm %	Detection Rate %	o(n)
CTree	15.7%	**93.33**	658.29n
FRIDS	12.63%	95.98	453.07n
EIFRID	5.03%	98.89%	305.02n

CTree – Completion Tree
FIDS- Fuzzy Rule based Intruison Detection system

The performance of the Emerging Intuitionistic Fuzzy Rule for Intrusion Detection (EIFRID) outperforms remaining to algorithms CTree and Fuzzy rules of Intrusion Dectection System.

The research activities involved a process to establish if classification could be found in the data. These processes involved the statistical manipulation of the dataset in Excel. The aim of the research was to determine the maximum percentage of correctly classified instances. The process involved the creation of analysis tools and charting the data so that the percentage of correctly classified instances by each model is displayed and experts can interpret the findings

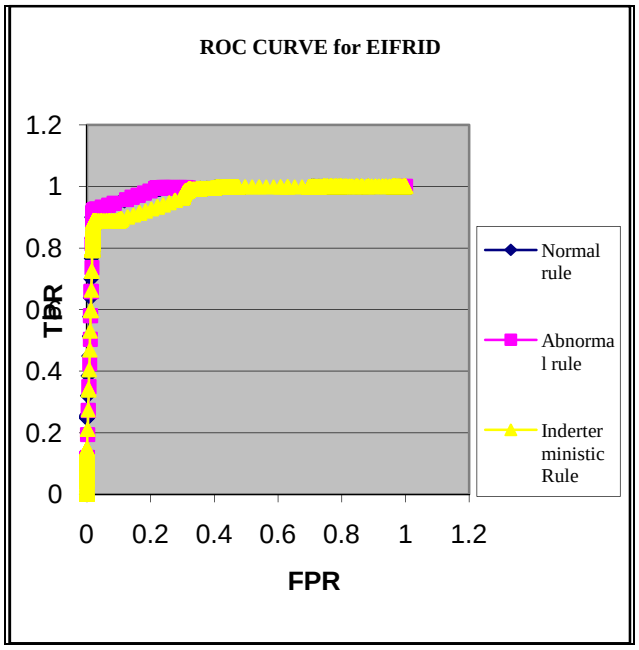

Figure 8. Repeat 10 x for 10 – Fold Cross Validation

The plotted points define the ROC curve for the given classifier. This ROC curve can be used to identify the better classifier based on domination of classifiers. The figure 7 shows that the Normal and Abnormal classifiers are similar and both the rules contribute best rules. The use of member ship and non-membership function it is possible to identify the indeterministic rule which needs more importance when expressing the hesitation examined objects. From the results obtained, it is evident that the improvised genetic algorithm adapted along with the Intuitionistic Fuzzy logic for this experiment was successfully able to generate a model with the desired characteristics of a high correct detection rate and a low false positive rate from learning over training data set.

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 improvised genetic algorithm approach for identifying these rules is chosen. The present work has explored the possibility of integrating the Intuitionistic fuzzy logic with Data Mining methods using Genetic Algorithms Restartings for intrusion detection. The present work is the extension of in the areas Dimensiionality reduction using Best First Search. We have proposed architecture for Intrusion Detection methods by using Intuitionisitc fuzzy rule classifier by tuning he best possible rules using Genetic Algorithms Restartings The main contribution of this work is employing the intuitionistic fuzzy classifiers for intrusion detection problem and handling the dataset in the presence of imperfect facts and imprecise knowledge. The application of intuitionistic fuzzy sets instead of fuzzy set means the introduction of another degree of freedom into a set description(i.e. in addition $\mu(A)$ to we also have $\nu(A)$).

REFERENCES

[1] Wenke Lee and Sal Stolfo, “Data Mining Approaches for Intrusion Detection” Proceedings of the seventh USENIX Security Symposium (SECURITY ‘ 98), San Antonio, TX, January 1998.

[2] V. Dhanalakshmi and Dr. I. Ramesh Babu, “Intrusion Detection using data Mining along Fuzzy logic and genetic algorithms”,IJCSNS, International Journal of Computer Science and Network Security, Vol.8 No.2,February 2008.

[3] James P. Anderson, “Computer Security Threat Monitoring and Surveillance”, Technical report, James P. Anderson Co., Fort Washington, Pennsylvania. April 1980.

[4] Edward Amoroso, "Intrusion detection", Intrusion.net Books, January 1999.

[5] Julia Allen at all, "State of the practice of intrusion detection technologies", Technical Report CMU/SEI99 - TR-028, ESC-99-028, Carnegie Mellon, Software Engineering Institute, Pittsburgh, Pennsylvania, 1999.

[6] Stefan Axelsson, "Intrusion detection systems: A survey and taxonomy", Technical Report No 99-15, Dept.of Computer Engineering, Chalmers University of Technology, Sweden, March 2000.

[7] B. Kavitha, S. Karthikeyan, and B. Chitra, “Efficient Intrusion Detection with Reduced Dimension Using Data Mining Classification Methods and Their Performance Comparison”, V.V Das et al. (Eds.): BAIP 2010, CCIS 70, pp. 96–101, 2010. Springer-Verlag Berlin Heidelberg 2010

[8] Grigorios N. Beligiannis, Georgios A. Tsirogiannis and Panayotis E. Pintelas, “Restartings: A Technique to Improve Classic Genetic Algorithms’ Performance”, World Academy of Science, Engineering and Technology 1 2005.

[9] The 1998 intrusion detection off-line evaluation plan. MIT Lincoln Lab., Information Systems Technology Group. http://www.11.mit.edu/IST/ideval/docs/1998/id98-eval-11.txt, 25 March 1998.

[10] Knowledge discovery in databases DARPA archive. Task description. http:\\www.kdd.ics.edu/databases/kddcup99/Task.html, as visited on 20 January 2009.
[11] Srinivas Mukkamala, Guadalupe Janoski, Andrew Sung (2002) “Intrusion Detection Using Neural Networks and Support Vector Machines”, Proceedings of IEEE Intrnational Joint Conference on Neural Networks, pp. 1702-1701.
[12] Yingjiu Li at all, "Enhancing profiles for anomaly detection using time granularities", Center for secure information systems. To appear in Journal of Computer Security, 2002.
[13] Susan Bridges and Rayford Vaughn, “Fuzzy data mining and genetic algorithms applied to intrusion detection”, Proceedings twenty third National Information Security Conference, October 1-19, 2000.
[14] Steve Hofmeyr at all, "Intrusion detection using sequences of systems call", Journal of Computer Security, 6:151-180, 1998.
[15] H. A.. van der Vorst, “Computational Methods for large Eigenvalue Problems”, in Handbook of Numerical Analysis, vol. 8, P. G. Ciarlet and J. L. Lions, Eds. Amsterdam: North-Holland (Elsevier), pp. 3-179, 2002.
[16] Y. Saad, Numerical methods for large eigenvalue problems, Manchester, UK: Manchester University Press, 1992.
[17] S. G. Petition, “Parallel subspace method for non-Hrmitian eigenproblems on the connection machine (CM2)”, Applied Numerical Mathematics, vol.10, pp. 19-36, 1992.
[18] D. C. Sorensen, “Implicit application of polynomial filters in a k-step Arnoldi method”, SIAM J. Matrix Anal. Applic., vol. 13(1), pp. 357-385, 1992
[19] J. Gmez and D. Dasgupta Evolving Fuzzy Rules for Intrusion Detection In Proceedings of the Third Annual IEEE Information Assurance Workshop 2002, New Jersey, June 2002.
[20] Atanassov, K., Intuitionistic fuzzy sets, Fuzzy Sets and Systems. 20(1986)87- 96.
[21] Atanassov, K., Intuitionistic Fuzzy Sets: Theory and Applications, Physica- Verlag, Heidelberg(1999).
[22] C.Tamalika and A.K Raya, A new measure using intuitionistic fuzzy set theory and its application to edge detection, Applied Soft Computing.,2007
[23] J. Gmez, D. Dasgupta, O. Nasraoui, and F. Gonzlez Complete ExpressionTrees for Evolving Fuzzy Classi¯ers Systems with Genetic Algorithms andApplication to Network intrusion Detection, In Proceedings of NAFIPS-FLINT joint conference, pages 469-474, New Orleans, LA, June 2002
[24] KDDCup99datasets,The UCI KDD Archive : http:/ kdd.ics.ucs.edu / databases /kddcup99 /kddcup99.html.

Kavitha B. completed her M.Phil in Computer Science from Bharathiar University in 2007. She is working as a Lecturer in School of Computer Science, Karpagam University, Coimbatore. Her experience is 6 yrs. Currently she is pursuing Ph.D in Bharathiar University. She has published 2 papers in International Journals and presented paper in 1 International Conference. Her research interests are Data mining, Network Security and Cryptography.

Karthikeyan S. received the Ph.D. Degree in Computer Science and Engineering from Alagappa University, Karaikudi in 2008. He is working as a Professor and Director in School of Computer Science and Applications, Karpagam University, Coimbatore. At present he is in deputation and working as Assistant Professor in Information Technology, College of Applied Sciences, Sohar, Sulatanate of Oman. He has published more than 14 papers in Natrional/International Journals. His research interests include Cryptography and Network Security

Sheeba Maybell P. received her M.Phil in Mathematics from Bharathiar University in 2007. She is working as Lecturer in Department of Mathematics, Karpagam University, Coimbatore. Her teaching experience is 3 years. Her research interest are Operator Theory and Fuzzy operators.