Chapter 1

Introduction

1.1 Introduction

Fuzzy Logic was introduced by the researcher named Zadeh in year 1965 and since then it has been used in various classification fields and control applications.

The fuzzy system consists of following components and processes:

- **Crisp set** - Crisp sets are those sets that represent bivalent condition i.e. either 0(false) or 1(true) meaning, that an element either belongs to a set or not. It is taken as an input variable for fuzzification process.

- **Fuzzy Sets** - Fuzzy means vagueness or uncertainty. It is a multi-valued logic that allows values to exist between 0(false) and 1(true) which represents human reasoning. The crisp sets are converted to fuzzy sets after the fuzzification process.

- **Membership function** - It provides a measure of the degree of similarity of elements in the universe of discourse to the fuzzy set. It is not same as probability but represents membership in vaguely defined sets. It is represented by \( \mu A(x) \).

  For fuzzy sets, the membership function lies in the range of 0-1.

  \[ \mu A(x) \in [0, 1] \] for crisp sets, \( \mu A(x) = 0 \) if \( x \in X \)

- **Fuzzification**: The process of converting crisp sets into fuzzy sets with membership function lying between 0 and 1 and represented by linguistic labels.

- **Defuzzification**: The processes of converting fuzzy sets back to crisp values are called defuzzification.

- **Fuzzy Rule Based System**: A fuzzy rule based system is a collection of propositions containing linguistic variables and the rules are expressed in the form of: If X1 is A1 and X2 is A2 and .......and Xn is An, then Y is B. The crisp set is taken as input where the fuzzification process takes place and a fuzzy set is received as an output. This fuzzy set act as an input for the inference engine where a fuzzy reasoning process is conducted and a fuzzy
output is obtained. Now the fuzzy output undergoes a defuzzification (Crisp Logic Paradigm) process to obtain a crisp output as a resultant.

Fuzzy Rule Based systems consists of fuzzy if then rules and have been used in various pattern classification problems. It consists of following components:

**Knowledge Base:** The Knowledge Base is further divided into two types-

1. **Rule Base** - The rule base consist of set of rules. The rules can be represented in different ways. Here the rule is of the following structure-\( Rk \) – *If \( X1 \) is \( A1 \) \( k \) and \( \ldots \) \( \ldots \) \( Xn \) is \( k \), then class is \( Ck \) with \( CDk \).*

Where each input variable \( Xi \) has a value from a set of linguistic labels \( Ai \) \( k \) \( [Li \ 1 \ or \ Li \ 2 \ \ldots \ or \ Li \ i ] \), \( Ck \) represents the consequent class and \( CDk \) is the certainty degree or confidence \( (CDk \in 0,1) \). Confidence or certainty degree \( CDr = \beta_{Class \ hr \ Rj} - \beta_{Class \ h \ r \ Rj \ c \ h=1} \)

Where \( \beta_{Class \ hr \ Rj} \) represents the sum of membership function of training pattern in class \( h \) and \( \beta \) represents the sum of membership function of training pattern that do not belong to class \( h \).

2. **Database:** The fuzzy set related to linguistic terms which are used in Rule Base is defined in Database (DB).

**Fuzzy Inference Engine:** It is a reasoning procedure that derives inference from fuzzy if-then rules. Here the incoming pattern is matched with the antecedents of rules and the pattern is classified according to the rule consequent using the property of maximum matching. So to classify an unlabelled pattern, \( xp = \{xp1,xp2,\ldots,xpn\} \) from a given rule set \( C = \max h=1,2,..c (\tau h ) \) where \( \tau h = \max r=1,..N \ Cr \in h \ CDr \)

**Genetic Approach with Fuzzy:** Genetic Approach (paradigm) is inspired from Darwin’s theory of evolution-survival of fittest. It is an adaptive heuristic search Approach (paradigm) based on evolutionary ideas of natural selection and genetics and is very successful in search and optimization problems. Genetic Approach (paradigm) exploits historical information to direct the search into the region of better performance. Some of the terminologies which are used in genetic Approach (paradigm) are:-

- **Chromosome-** A chromosome is a set of genes and represents the solution.
Genes- Each gene represents a parameter of the whole problem space and has a specific meaning.

Population- It represents the number of individuals or chromosome participating to find the optimized solution.

Fitness- It is the value assigned to each individual to determine how close the individual is from the solution and is calculated with the help of fitness function.

Selection- A proportion of individuals is selected from existing population during each successive generation.

Crossover- It is a genetic operator in which two individuals are selected and they perform intermingling to exchange information and generate two new individuals.

Mutation - It is also a genetic operator which alters a random gene or more number of genes in a chromosome from its initial state. The basic steps of a genetic Approach (paradigm) are as follows: - Before a genetic Approach (paradigm) is applied, a method to encode potential solution to that problem is required. Binary encoding is one of the most common techniques to represent the information. A random population of chromosomes is generated. Evaluate fitness f(x) of each chromosome on the basis of fitness function. Now create a new population by repeating the following steps until a new population is created.

Selection - A proportion of individuals is selected from the current population to procreate a new generation. The individuals are selected on the basis of different techniques on the basis of fitness value and the one with higher fitness will have higher probability of being selected. The different selection techniques are tournament selection, roulette wheel, rank selection and steady state selection.

Crossover - Two parent chromosomes are combined to produce two new off springs with user definable crossover probability which is kept as high as 0.9. The various types of crossover operators are one point crossover, two point, uniform, arithmetic and heuristic crossovers.
The following figure presents the view of fuzzifier and defuzzifier in the process of fuzzy logic based evaluations.

![Figure 1.1: Fuzzy Process](Image)

Following are the key components in the process:

- Crisp Input
- Base Set
- Fuzzy Scenario
- Fuzzifier
- Defuzzifier
- Intelligence Module
- Set of Rules
- Output Set
- Crisp Output
- Fuzzy based rules
- Rules
- Cumulative View
- Cumulative Fuzzy Module
- Cumulative Predictions
- Knowledge Extraction
Mutation - It is used to maintain genetic diversity by altering one or more gene. It occurs according to user definable mutation probability and is kept to low scale of 0.1. It helps to prevent the population to stagnate at local optima. Some of the techniques used in mutation are flip bit, boundary, non-uniform etc. The flip bit is the most famous mutation technique. Now a newly generated population is used for further run of Approach (Algorithm). If the termination condition is satisfied, then the best solution in the current solution is returned as the most optimized solution.

Genetic fuzzy rule based system is a fuzzy system augmented or hybridized with evolutionary computing to increase the learning capacity of the system, thus providing robust search capabilities in a complex environment. The main function is to adopt an evolutionary learning mechanism which can automatically generate /design knowledge base and thus help in search and optimization problems. The next section clearly explains how the genetic Approach (paradigm) helps in learning process. The generic code structure and domain independent features of genetic Approach (paradigm) make them desirable to incorporate apriori knowledge. This apriori information is in the form of linguistic labels, membership function or fuzzy rules in the fuzzy rule based system. Thus genetic Approach (paradigm) are used to acclimate or assimilate information from rule base or database. These systems are used in many applications such as classification, Fuzzy Data Analytics, control processes and modeling.

Genetic Approach (paradigm) performs very well in search and optimization problems. Thus they offer a domain independent search method in the process of fuzzy sets and integration associated machine learning. Here genetic Approach (paradigm) has been implemented on linguistic rule based classification systems.

They can be implemented in learning processes in three alternative approaches.

- Michigan Approach
- Pittsburgh Approach
- Iterative Rule Learning
In Michigan approach, each chromosome represents an individual rule and the set of rules represent the population size. Thus each rule is assigned a fitness value. The population of the classifier is maintained and rule discovery process and genetic operations are applied to each individual rule. With time, the rules are altered through interaction with the environment. Thus this approach is mainly used in online process and simulated environment.

The Approach (paradigm) of the Michigan approach is given as:-

1. Specify the parameters which are required i.e. number of linguistic rules or population size, crossover probability, mutation probability etc.
2. Generate an initial population of linguistic rules randomly.
3. Calculate fitness function corresponding to each linguistic rule. Select some linguistic rules from the current population through selection procedure and perform crossover and mutation on them.
4. Remove the worst linguistic rules and replace them with the newly generated rules in the current population.
5. If the stopping condition is not satisfied, then return to step 3 otherwise terminate the Approach (Algorithm).

The final solution is the population or rule set which has highest classification rate.

In Pittsburgh approach, a whole set of rules is encoded as an individual chromosome. Each substring of the chromosome is represented as an individual rule. The fitness function of a rule set is evaluated here instead of each single rule. A new set of rules is obtained by crossover while mutation leads to new rules. This approach is mainly used in batch processes. The basic steps of Pittsburgh approach is given below: -

1. Specify the population size or number of rule sets, number of linguistic rules in each rule set, crossover probability, and mutation probability and termination condition.
2. Randomly generate the rule sets containing given number of linguistic rules as an initial population.
3. Assign fitness function to each rule set and perform crossover and mutation on the rule sets selected in the current population.
4. Remove the rule sets which are worst and add the newly generated rule sets in the current population to perform next iteration.
5. Terminate the execution of Approach (paradigm) if the stopping condition is satisfied else go to step 3. The final solution is the best rule set selected from the population.

It is similar to Michigan approach as each chromosome represents the rule but in contradiction to the Michigan approach, only the best rule is selected and all other remaining chromosomes are discarded. Thus it provides only the partial solution to the process of learning. Therefore GA has to be incorporated within an iterative approach in the following way:

1. A rule is obtained by applying genetic Approach (paradigm) in the system.
2. The rule is integrated with the new set of rules.
3. If the set of rules generated are sufficient to solve the problem, then they are considered as final set of rules otherwise go back to step 1.

In this approach, the fitness function of each chromosome is calculated individually without the requirement of cooperation of other chromosomes. Since only one rule is searched in iteration, so the search space is also reduced. These approaches are the most commonly used traditional approaches but some new heuristic approaches have also been proposed to give more optimized results and perform well on high dimensional dataset. An intrusion is defined as an encroachment to a person’s privacy or organization’s important information. Previously various conventional methods such as firewall, encryption techniques have been used to prevent the computer systems from unauthorized use. But these mechanisms were not sufficient enough to prevent the intrusion as hackers and attackers grew more proficient and were able to find vulnerabilities in the network, thus violating computer security policies. Hence, an additional mechanism i.e. intrusion detection system was established and it became a vital component in the field of security infrastructure. The first intrusion detection model was given by Denning in 1987. Since then many intrusion detection models have been constructed to determine the behavior of the network accurately and in an efficient manner.
Porras et al. proposed three standards to evaluate the functioning of intrusion-detection systems. These are:-

**Accuracy**: The attacks should be properly detected with no false alarms i.e. they should not be misclassified or misrepresented.

**Performance**: It is the rate of processing of audit data. For real time detection, the performance of IDS must be high.

**Completeness**: It is the ability to detect all the attacks. Thus the intrusion detection system must keep on updating itself and must have comprehensive information about the attacks.

Following is the brief description of each class of Fuzzy based IDS-

1. **Detection Method**: The features or the attributes of the detector is represented in this method. It consists of two schemes - (a) which extracts the information about the attacks from the database and triggers an alarm when vulnerability is evidently found out i.e. misuse detection.(b) when a baseline of the normal behavior is maintained and any abnormality from the baseline is detected as a malicious activity i.e. anomaly detection.

2. **Behavior on Detection**: It refers to the reaction or conduct of the intrusion detection system. When only alarms are triggered and corresponding to it no action is taken, then the IDS is considered to be passive but when some countermeasures to eradicate or control these attacks are taken, it is said to be active.

3. **Location of Audit Source**: The input information is examined and on the basis of which the intrusion detection system is classified. The information can be in the form of host log files, application log files or network packets.

4. **Usage Frequency**: It implies the time analysis of the intrusion detection system in a particular environment. Continuous monitoring performs a continuous time analysis to know about the activities which are happening immediately while the periodic analysis refers to analyzing the activities on a periodic basis.

5. **Detection Paradigm**: It describes the method used to detect the attacks in IDS with fuzzy integration. It can be either state based or on the basis of transition from one state to another. Subsequently various artificial intelligence, fuzzy sets and integration associated machine learning and
computational techniques have been applied on various intrusion detection models to obtain precise results while confronting various problems such as vast network traffic, noisy information, continuous adaptation to changing environment. To evaluate the performance of intrusion detection models, datasets are required which can clearly specify whether the Intrusion Detection System is able to comply with the standards or not. The data can be collected from various sources such as log files, data packets, command sequences etc. Two largely publically available and most used datasets are:- DARPA 98 Lincoln and KDD 99 dataset. In 1999, KDD99 dataset was derived from the DARPA98 network traffic dataset by ACM SIG-KDD International Conference on Knowledge Discovery and Fuzzy Data Analytics. It consists of TCP connections and consists of 9 weeks of training data and 2 weeks of testing data. Each connection consists of 41 attributes and the features of the dataset were defined by Stolfo et al.. Despite its high usage, the dataset has been criticized by McHugh on the basis of unrealistic data rates. Fuzzy systems with their potential to provide better accuracy and interpretability have the ability to build the models which resemble the real world systems. They are based on fuzzy logic and predicates which includes the knowledge of human experts and consolidates numerical and symbolic representation of data. Mamdani fuzzy rule based system consists of two parts-fuzzy knowledge base and fuzzy inference system. Ishibuchi et al. gave a fuzzy rule based system in which a fuzzy rule structure is matched against the patterns and corresponding to it a class and a discrete value is given as a consequent. The fuzzy rule based classification system can be differentiated on the basis of certainty factor or the confidence consorted to the class in the consequent in the following ways: -

1. On the basis of class label only
2. On the basis of class label and certainty factor
3. On the basis of certainty factor only. But the major drawback of Mamdani Fuzzy Rule Based System is lack of accuracy in case of complex, high dimensional dataset due to lack of flexibility in linguistic variables. Therefore, various extensions were given in order to increase the accuracy of Mamdani fuzzy rule based system. Some of them include scatter fuzzy partitions,
weighted rules, disjunctive normal form etc. Genetic Approach (paradigm) given by Dejong are used in fuzzy rule based system to help in learning process and maintain parameter optimization. To find the best rules, the most important task is to find those parameters which can optimize the Knowledge Base. Karr gave a proposal on genetic fuzzy rule based system which utilized binary coded genetic Approach (paradigm) on fuzzy system based on triangular functions. Different coding schemes were applied on trapezoidal, triangular and Gaussian functions to obtain a more accurate linguistic fuzzy model. Real coded genetic tuning was given by Cordan et al. in Mamdani fuzzy rule based system. Binary coding was used to represent the chromosomes and it was first given by Ishibuchi as the foremost genetic rule selection process. MOGUL was given by Cordan et al. as a multi selection process in genetic Approach (paradigm) applying iterative rule learning approach and is used in both classification and control processes. The main GFS learning approaches i.e. Michigan, Pittsburgh, Iterative Rule Learning were used to obtain optimized results. In as proposed by Giordana and Neri, the Michigan and Pittsburgh approaches were collaborated to generate an efficient system consisting of rules which can be used in high dimensional problems as the unnecessary rules can be eliminated by the don’t care conditions. Another approach, GCCL (Genetic cooperative-competitive learning) approach was used where a chromosome is encoded by a single rule and each chromosome cooperates and competes to produce results which are more accurate and have good interpretability. Berjlanga et al. proposed DNF fuzzy rule based system while a hybrid genetic Approach (paradigm) has been proposed in which constitute a mathematical model to increase the accuracy and interpretability of each rule. The computational intelligence and he outlined it as: A computationally intelligent system is the one which is concerned with only low level data, incorporate constituents of pattern recognition, and does not apply knowledge in the sense of artificial intelligence; and manifest the following characteristics: - computational adaptability speed resembling turnaround like humans computational fault tolerance error rates that is roughly close to human performance. Computational Intelligence consists of four main paradigms as represented
1. **Artificial Neural Network** - It is an information processing model inspired from the biological neurons. It is an interconnected group of artificial neurons working in unison to solve complex problems. It is considered as the weighted directed graph in which the nodes depict the neurons while the edges represent the interconnections between the neurons. Artificial neural networks are generally used due to their ability to adapt, exhibit fault tolerance, self organization and real time operation.

2. **Evolutionary Computing** - The evolutionary computation is based on the mechanisms of genetics and natural selection is used in search and optimization problems. It uses an iterative approach where a population is selected from random space and different operations are applied to get a desired solution. It consists of genetic programming given by Koza, evolution strategies given by Rechenberg, evolutionary programming prescribed by Fogel and genetic Approach (paradigm) given by Holland.

3. **Swarm Intelligence** - Swarm intelligence is neighbor based system inspired from biological organisms where simple agents manifest collective intelligent neighbor to solve complex problems.

4. **Fuzzy Systems** - Fuzzy systems as discussed earlier are based on fuzzy logic and deal with vague and imprecise data to get optimized results. They are manly used in control and classification problems. Different techniques have been proposed to evaluate KDD 99 dataset on the basis of computational intelligence and provide an efficient and fault tolerant intrusion detection system. The Approach (paradigm) was examined using SVM classifier for the evaluation of performance. A multilayered Self Organizing Map was built by Rhodes and Sarasamma et al. drawing conclusion that different subsets of features were efficient enough to detect different attacks. Decision trees, association rules and fuzzy implication segment were used to generate fuzzy rules. The collaboration of Hidden Markov Model with fuzzy inference engine was done by Cho to detect normal connections. Fuzzy C-Means and Fuzzy Mining Approach (paradigm) are two clustering approaches used to detect abnormal behavior through the concept of outliers. REGAL, a distributed genetic Approach (paradigm) given by Mischiatti and Neri used combination of Pittsburgh and Michigan learning approach to model network
traffic. Linear Genetic Programming outperforms Support Vector Machines and Artificial Neural Network in terms of detecting intrusion detection with accuracy as proposed by Abraham et al. and Song. Artificial Immune System can be used to model intrusion detection system and the first Artificial Immune System model based on anomaly detection, was given in to detect file alterations and call sequences. Swarm Intelligence technique was also used in intrusion detection due to its self organizing and distributed properties to get optimized results but suffered from the problems of clustering high dimensional network data. An ant based clustering and sorting Approach (paradigm) given by the researchers was used by Abraham to detect intrusion in KDD-99 dataset. A standard Nature Inspired Fuzzy Approach technique was engrafted in genetic fuzzy system by Abadeh et al. Soft Computing techniques soon started to be hybridized to built fault tolerant, precise and robust intrusion detection systems. Since there are five classes in KDD-99 dataset, hence a five neuro-fuzzy classifier was developed by Toosi et al.. Fuzzy cognitive map was introduced by Kosko to provide a graphical representation of the work and was used by Xin to detect complicated or intricate attacks. Tsang highlighted the illustration of fuzzy if then rules in a genetic fuzzy system. Different approaches such as Michigan, Pittsburgh and Iterative Rule Learning were used by Abadeh et al. to detect attacks in the network infrastructure. Thus, soft computing models were more effective in building accurate, robust system with high performance. The KDD-99 dataset constitutes of five major classes. But two of them i.e. U2R and R2L are very less in number in the training dataset. So these classes are not properly trained to get accurate results and perform poorly in the testing datasets which consists of 11 different types of attacks. These attacks are very difficult to be detected and therefore a major drawback in existing IDS. Wu et al. analyzed different approaches and proved that soft computing techniques perform better than other techniques. The computational intelligence approach performed better than the decision trees. Evolved classification rules did not perform well because when overlapping occurs, the data cannot be separated into two classes. Also it is easier to apprehend the fuzzy rules alone. Self organizing maps suffered from problems such as high
dimensionality, higher detection rates with false positives and computational overhead. Evolutionary computing techniques do not have fair termination criteria and does not give accurate results when the data distribution is unbalanced. Swarm Intelligence techniques are mainly used to learn clusters and classification rules but prove to be a constraint in high dimensional network and cannot differentiate dissimilar objects. Therefore collaboration of various soft computing techniques is required which has the ability to learn in an uncertain and imprecise network. It encloses all the complementary features of different techniques and builds a robust and fault tolerant system.

Computational intelligence systems have the ability to adapt, exhibit fault tolerance, high computational speed and error resilience against noisy information. The fuzzy rule based system performs well in an uncertain and imprecise environment and establishes more concise and pliant patterns which enhances the adaptation capability and robustness of the intrusion detection system and classifies normal and abnormal connections correctly. Evolutionary computing has the capability to learn with the changing environment and is used in designing optimized fuzzy rules. These fuzzy rules are constructed from the training dataset. Genetic Approach (paradigm) are applicable in tuning membership functions of the fuzzy sets. The crossover operator interchanges the chromosomes between two parents to get more prominent rule child while the mutation operator generates new rules. Thus new suspected attacks can also be detected with the adaptive capability. The genetic Approach (paradigm) continues for specified number of generations and the best rules are extracted. These rules undergo a compatibility model which will yield more precise rules and therefore invigorate the performance of intrusion detection system. The main objective behind the proposed work is to build a model which has high detection rate and low or minimal false alarm rate. It should be accurate and complete to classify all the attacks in their true classes and should exhibit the property of high adaptability i.e. the ability to adjust according to changing behavior of the users and networks and modifying itself for proper functioning. Thus the proposed model should be fault tolerant in nature. The malicious activities may create faults in the system but the IDS should have the potential to maintain reliability and accuracy in
the system so as to prevent the systems from abuse. It should remain intact and update its database with the contemporary information about the network connections and therefore able to detect new anomalous activities. The proposed approach is amalgamation of the fuzzy systems with that of genetic Approach (paradigm) to bring out a hybridized genetic fuzzy rule based system which provides robust platform to detect intrusions existing in the network distinctly and classifies them into normal and different types of attacks according to their signatures. The work has been performed on KDD-99 data set which is a standard dataset used to detect intrusions in the network. The KDD-1999 intrusion detection dataset uses a version of database which was prepared in 1998 DARPA Intrusion Detection Evaluation Program (MIT Lincoln Labs) to evaluate their research in intrusion detection. It consisted of 9 weeks of raw TCP dump data as training dataset and 2 weeks of testing dataset. The KDD-99 dataset was used in Third International Knowledge Discovery and Fuzzy Data Analytics Tools Competition to prepare an intrusion detector which can identify good or bad connections. The dataset is containing 41 attributes and labeled with either normal or specific attack type. Here 10% of the KDD-1999 dataset has been used to evaluate the whole process. The 10% KDD dataset consists of 4,94,021 records /connections in the training data. The training dataset consists of 24 training attack types which are under 4 major classes of attacks while the testing dataset consists of 14 additional attacks to differentiate some signatures and check whether these variants are captured or not to increase efficiency.

The features of KDD-99 dataset with fuzzy blend as defined by Stolfo et al. have been classified into following categories-

1. Basic Features- This category enclose all the attributes that can be retrieved from an individual TCP connection and comprise 9 attributes.
2. Time Based Traffic Features- It includes features which are calculated on the basis of time interval and is subdivided into two types:-
   a) Same Host Features-It examines only the connections in the past two seconds that have the same destination host as the current connection.
   b) Same Service Features-It examines only the connections in the past two seconds that have the same service as the current connection.
3. Host Based Traffic Features- Here features were constructed using a window of 100 connections to the same host instead of a time window of 2 seconds.

4. Content Based Features- It consists of 13 features that are extracted from domain knowledge and are used to indicate suspicious behavior in the network or unstructured data portions in the packet. The data describes the 41 attributes of the KDD- Cup 99 dataset and shows whether the feature is of symbolic type or continuous.

Fuzzy Rule Based System
Step1- Representation of data
Step2- Normalization of data
\[ y = \frac{x - \text{min}}{\text{max} - \text{min}} \]
Where \( x \) represents the original attribute value, \( \text{min} \) represents the minimum boundary value i.e. 0, \( \text{max} \) represents the maximum boundary value i.e. 1 and \( y \) represents the normalized value i.e. \( 0 \leq y \leq 1 \).

Step3- Calculating membership function
After normalizing the data, the membership function of every feature of each training pattern is determined by the following formula:
\[ \mu_j x = \max 0,1 - x - x_j \nu \]
Where, \( x \) represents the normalized value of each feature,
\[ x_j = j - 1 \ L - 1 \]
where \( j = 1,2,...,L \), and \( \nu = 1 \ L - 1 \)
where \( L \) represents the number of linguistic labels. Here \( L \) varies from 1-5 and help in calculating the membership function of each attribute.

Step4- Determining compatibility of each training pattern
The compatibility of each training pattern \( x_p \) is calculated with the fuzzy if then rule \( R_j \) by using the following formula:
\[ \mu_j x_p = \mu_j (x_{pi}) n_i =1 \]
Where \( p \) refers to 1,2,........,m training patterns and \( \mu_j (x_{pi}) \) is the membership function of the \( i \)th attribute of \( p \)th pattern.

Step5- Finding compatibility grade for each class
After obtaining the compatibility of each training pattern, relative sum of compatibilities grades of the training patterns with rule \( R_j \) for each class is calculated.
This is given by: \( \beta \text{Class } h \ R_j = \epsilon \text{Class } h \ \mu_j (xp) \ N \text{Class } h \)

Step 6 - Selecting fuzzy if-then rule for a particular class

The consequent class \( C_j \) for a given rule \( R_j \) is calculated as follows:

\[
\beta \text{Class } h(R_j) = \max_{h=1,\ldots,c} (\beta \text{Class } h(R_j)) \tag{10}
\]

The maximum of \( \beta \text{Class } h(R_j) \) is evaluated and the one with the maximum value is considered to be the class of that fuzzy if then rule \( R_j \). If the maximum value comes out to be true for more than one class, then the consequent class \( C_j \) cannot be determined uniquely and is taken as \( \varphi \) and the corresponding rule is rejected.

Step 7 - Assessing the certainty degree \( CD_j \)

\[
CD_j = \beta \text{Class } h_j (R_j) - \beta \text{Class } h_{r c} (R_j) + 1 \tag{11}
\]

Where \( \beta = h \neq h_r \beta \text{Class } h (R_j) - c - 1 \tag{12} \)

The value of \( CD_j = 1 \) represents very high confidence which denotes that the rule belongs to that specific class.

Step 8 - Generation of fuzzy if then rules

Thus the fuzzy if then rules are generated in the following manner:- Rule \( R_j = \) If \( x_1 \) is \( A_{j1} \) and \( x_2 \) is \( A_{j2} \) and \( \ldots \ldots \), \( x_n \) is \( A_{jn} \), then the class is \( C_j \) with \( CD_j \), \( j = 1,2,\ldots,N \) where \( R_j \) is the label of the \( j \)th fuzzy if then rule, \( A_{j1}, A_{j2}, A_{jn} \) denotes the antecedent fuzzy sets

Evaluation Parameters of IDS with Fuzzy

The IDS is basically evaluated through the confusion matrix or contingency segment which was given by Provost and Kohavi. The matrix contains information about the actual and predicted classifications done by the system.

It consists of the following records:-

True Negative (TN) – It refers to the number of correct events which predicted them as authentic connections.

False Positive (FN) – It implies the number of erroneous predictions which analyzed genuine events as fake events.

False Negative (FN) – The number of incorrect predictions which got evaluated false connections as correct connections.

True Positive (TP) – It refers to the number of correct predictions which anticipate that the instance is fake or anomalous.

Different methods such as RIPPER, EFRID and other genetic fuzzy systems have been used to compare with the current approach and perceive which
approach has overall better detection rate and lower false alarm rate. The Michigan Approach (paradigm) and IRL approach perform moderately in detecting class with 88.13% and 93.2% but very low false alarm rate of 0.11% and 0.18% which depicts a very less possibility of misclassification. The Pittsburgh approach has a very high recall which shows its efficiency in detecting the attacks but the false positive rate is comparatively higher rounding about to 2%. The EFRID approach has a good detection rate but exorbitant false alarm rate of 7%. Similarly the RIPPER approach has high false positive rate of 2.02%. The proposed approach gives more adept result with high recall value rounding to about 99% which is nearly equal to Pittsburgh approach and false alarm to about 1.27% which is comparatively very less than Pittsburgh approach. Hence the genetic fuzzy systems perform more skillfully than other existing approaches and are more reliable. The proposed methodology gives high performance than all other genetic fuzzy systems with suitable detection rate and meager false alarm rate, therefore stabilizing all the parameters.

This section contains the outcome which provides clear view about the anomaly detection scheme implemented. The genetic operators guarantee substantial individuals, as the class of generated rules is matched with their parent class. If they are same, then only the newly generated rule is accepted, otherwise it is rejected and the whole process is repeated. This approach reduces misclassification of rules and thereby increases accuracy. The genetic Approach (paradigm) is continued for some specific generations which provide more validity and accuracy to the rules. The fuzzy if then else rules perform exceedingly well on imprecise and uncertain data, therefore the system is fault - tolerant. As the rules keep on updating themselves in the database with the changing connections therefore the intrusion detection system has the property of high adaptability. Thus, the classification rules are able to classify the normal and abnormal behavior in the network with good accuracy, thus leaving fewer loopholes for the misjudgment. The proposed approach intermingles the features of fuzzy rule based system and genetic approach to build a robust intrusion detection system. The fuzzy if then rules extract features even from such small and imprecise dataset to extract
relevant features which will help in further classification. The genetic Approach (paradigm) helps in obtaining optimized rules due to their adeptness in exploiting historical information and yielding better results. The genetic fuzzy rule based inference engine has the ability to work in high dimensional network and is able to handle huge amount of audit data. KDD-99 data set is a high dimensional dataset with 41 attributes containing over 4,00,000 connections. So the IDS have the capability to handle large amount of data with ease. Genetic Approach (paradigm) are used in combination with fuzzy if then rules to prevent overlapping of rules and therefore each variable is distinguishable. For each connection, compatibility factor is calculated at the stage of fitness evaluation and then is classified into respective class which has higher compatibility factor. Genetic Approach (paradigm) are also used to maintain diversity. Also, the mathematical model seeks to cover each sample of the training dataset by classifying them through rule set, thus strengthening precision. Reducing the false detection rate and substantially increasing the recall is one of the major concerns of any successful intrusion detection system. The proposed approach has a very high detection rate but a little higher false alarm rate of about 1.27%, so the primary aim would be decrementing the false alarm rate so as to avoid misclassification of connections.

A fuzzy decision making model is characterized by a set of goals $G_i$ $(i = 1,2,...,m)$, along with a set of constraints $(1,2,...,n)$ $j C_j = n$, each of which is expressed by a fuzzy set on $X$. For such a model of decision making, Bellman and Zadeh in their pioneering approaches proposed that a fuzzy decision is determined by an appropriate aggregation of fuzzy sets. The main feature in this approach is the symmetry between goals and constraints.


1.2 Problem in Hand

This research work is having the key focus on the use of nature inspired approach with the integration of fuzzy genetic optimization to achieve the dynamic clustering for the datasets in assorted domains.

There is no reasonably "conform" clustering Approach (Algorithm), however as it was noted, "clustering is in the eye of the beholder." The most sensible clustering Approach (paradigm) for a particular issue continually ought to be picked likely, unless there is a numerical inspiration to slant toward one get-together model over another. It should be seen that an Approach (paradigm) that is set up for one kind of model has no believability to get on a data set that contains a basically enchanting kind of model. For example, k-supports can't find non-raised clusters.

It is one of the highly effectual approach that is widely used and implemented for the optimization of issues in the mathematics as well as other segments of research domains. Fuzzy logic is extremely useful for many people involved in research and development including engineers (electrical, mechanical, civil, chemical, aerospace, agricultural, biomedical, computer, environmental, geological, industrial, and mechatronics), mathematicians, computer software developers and researchers, natural scientists (biology, chemistry, earth science, and physics), medical researchers, social scientists (economics, management, political science, and psychology), public policy analysts, business analysts, and jurists.

Indeed, the applications of fuzzy logic, once thought to be an obscure mathematical curiosity, can be found in many engineering and scientific works. Fuzzy logic has been used in numerous applications such as facial pattern recognition, air conditioners, washing machines, vacuum cleaners, antiskid braking systems, transmission systems, control of subway systems and unmanned helicopters, knowledge-based systems for multi-objective optimization of power systems, weather forecasting systems, models for new product pricing or project risk assessment, medical diagnosis and treatment plans, and stock trading. Fuzzy logic has been successfully used in numerous fields such as control systems engineering, image processing, power engineering, industrial automation, robotics, consumer electronics, and
optimization. This branch of mathematics has instilled new life into scientific fields that have been dormant for a long time.
The fuzzy based systems can be developed for the following applications with the higher degree of accuracy and optimization level

- Solving Combinatorial Optimization Problems
- Engineering Problems with Higher Complexity Issues
- Medical Problems
- Biometric
- Fuzzy Integrations
- Diagnosis Domain
- Corporate Segment
- Stock Market Predictions
- Fuzzy based Optimization
- Engineering Processes
- Meta-Heuristic Approaches
- Deep Learning Approaches
- Machine Learning based Fuzzy Integrations
- Real Time Analytics
- Market Predictions
- Satellite Data Evaluations
- Maps and Cloud based Fuzzy Integrations
- Manufacturing and Optimization Issues
- Predictive Mining
- Knowledge Discovery
- Finding the hidden patterns
- Dynamic extraction and mining
- Cavernous evaluations and analytics of records
- Bio-Informatics with Fuzzy Integrations
1.3 Research Objectives

1. To devise a new approach for clustering using advance genetic Approach (Algorithm)
2. To evaluate the performance of novel approach on assorted parameters
3. To evaluate and devise the factors affecting cluster formation and outlier analysis
4. To investigate the performance of approach based on outliers detection aspects.
5. Testing and Evaluation of approach on assorted datasets
1.4 Scope of the Research


2. Evaluation of the work on assorted factors and parameters so that the consistency of the approach can be analyzed effectually.

3. Evaluation of the cluster formation strategy with the related approaches in the literature presented by the assorted researchers, academicians and practitioners.

4. Investigation of the issues and perspectives related to outlier detection after fuzzy genetic optimization in specified domains.

5. Analytics of the dataset using fuzzy based clustering

6. Evaluation of the outliers from assorted datasets using fuzzy evaluations and analytics patterns

7. Integration of fuzzy based patterns mining

8. Presentation and visualization of the hidden patterns

9. Evaluation of the dataset in association with cavernous integration of fuzzy sets

10. Association of dynamic clustering with fuzzy and unbiased sets.

11. Implementation of fuzzy or nebulous based clustering with dynamic perspectives

12. Evaluation of the aspects integrated with the fuzzy mathematics

13. Analytics on the fuzzy based dimensions in clustering

14. Integration of dynamic clustering in fuzzy associated modules

15. Integration of the datasets for clustering using fuzzy approach

16. Association of the nebulous processing for the dynamic datasets and real time clustering

17. Presentation of the clustered outcome using nebulous analytics

18. Cavernous analytics on the fuzzy based mining and extraction

19. Deep mining and fuzzy based clustering on the real time datasets for assorted domains

20. Integration of the fuzzy based plotting with the dynamic clustering