Feature Subset Selection using Multi-Objective Genetic Algorithm

A thesis submitted to

Department of Computer Science &Software Engineering
For the Partial Fulfillment of the Requirement of
MS (CS) Degree
By

Umm-e-Zahoora
595-FBAS/MSCS/F09

Supervisor

Dr. Ayyaz Hussain
Assistant Professor

Department of Computer Science & Software Engineering
International Islamic University Islamabad Campus
(2012)

1. Algorithms; Computer programming

2. Computer programming

Department of Computer Science & Software Engineering

International Islamic University Islamabad

Final Approval

Dated: 30th Jan '2012

This is to certify that the Department of Computer Science and Software Engineering, International Islamic University Islamabad, accepts this dissertation submitted, in its present form as satisfying the dissertation requirements for the degree of MS (CS).

Committee

Supervisor:

Dr.Ayyaz Hussain

Assistant Professor,

Department of Computer Science & Software Engineering,

International Islamic University Islamabad

(Dr. Ayyaz Hussain)

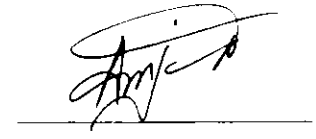
External Examiner:

Dr.Amjad Iqbal

Assistant Professor,

Department of Information Technology,

University of Centeral Punjab, Lahore



(Dr. Amjad Iqbal)

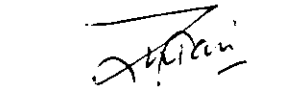
Internal Examiner:

Dr.Zunera Jalil

Assistant Professer,

Department of Computer Science & Software Engineering,

International Islamic University Islamabad



(Dr.Zunera Jalil)

Declaration

I, Ms. Umm-e-Zahoora solemnly declare that this thesis, neither as a whole nor as a part thereof has been copied out from any source. It is further declared that this dissertation is made entirely on the basis of my personal efforts under the sincere guidance and help of my supervisor. No portion of the work presented in this report has been submitted in support of any application for any other degree or qualification of this or any other university or institute of learning.

Umm_e_Zahoora

595-FBAS/MSCS/F09

Abstract

The main theme of the report is the application of the genetic algorithm in biomedical sciences for diagnosing of the diseases. Feature Subset selection is an approach that promises the better or same results as with using all the features. This approach can be helpful for diagnosing disease with rather less clinical tests and can be proved in future an effective tool for diagnosing disease. There are basic three approaches of feature subset selection filter, wrapper and embedded. The main idea of this thesis is feature subset selection using opposition based multi-objective genetic algorithms which is a type of wrapper approach for supervised learning. In past several techniques have been proposed on this topic but most of them are based on simple genetic algorithm. This thesis is emphasizing on selection of feature subset using multi-objective genetic algorithms. In this regard we have used non-dominated sorting Genetic Algorithm (NSGAII). The classifier used is SVM (RBF). This technique is being tested on several biomedical datasets taken from the UCI machine repository. Due to the better spread of candidate solution at initial step It has also been observed that our technique converge faster towards best optimal solution as compare to other techniques in literature not only this our technique also shows improved accuracy and more reduce feature as compare to the previous work .This research not ends here there are many ways to extend this work.

Glossary

This section lists symbols and acronyms that frequently appear in this thesis.

Acronyms and Abbreviations

ALife	Artificial life
CMMO	Coello's Min-Max Optimisation
FSSMC	featureselection via supervised model construction
GA	Genetic Algorithm
Gbest	Global best
HLGA	Hajela & Lin's Weighting-based Genetic Algorithm
Lbest	Local best
MOEA's	Multi objective evolutionary algorithms
MDD	Major Depressive Disorder
MOGA	Multi-Objective Genetic Algorithm
NPGA	Nafpliotis & Goldberg's Niched Pareto Genetic Algorithm
NSGA	Nondominated Sorting Genetic Algorithm
PAES	Pareto Archived Evolution Strategy
PSO	Particle Swarm Optimization
PO	Pareto Optimal
SGA	Simple Genetic Algorithm
SA	Simulated Annealing
SPEA	Strength Pareto Evolutionary Algorithm
VEGA	Vector evaluated Genetic Algorithm
CFS	Correlation-based Feature Selection
FCBF	Fast Correlation-Based Filter
ROC	Receiver Operating Characteristic

UCI	University of California at Irvine (UCI) Machine Learning Repository
MFFN	Multi Layer Feed Forward Neural Network
MO	Multi-objective
SO	Single Objective
RIIM	Rapid Image Information mining
IIM	Image Information mining

Acknowledgement

First of all I am very much thankful to Al Mighty Allah for making me able to understand the things around me, to interpret them, provided me with His divine help and made every hurdle in my way as wall of sand.

Secondly thanks to my parents and family whose utmost efforts and dedication made me to face the world with dignity.

I am really thankful to **Dr Ayyaz Hussain**, he provided me with an opportunity to learn and understand things from an entirely new prospective.

Last but not least thanks to my friends and class fellows for their unending support throughout the course of project.

Table of Contents

List of Figures

List of Tables

Chapter 1

1. Introduction

Features high dimensionality is a major problem in many fields since long especially in biomedical sciences e.g. Gene Array based cancer classification, tumor detection etc [1,3].

High dimensionality is a great curse in other fields as well such as Dos attacks classification in networking , image processing. data mining etc .Many efforts has been made using machine learning and pattern recognition techniques to reduce the features dimensions.

Irrelevant features include extra overhead in problem domain by increasing the complexity of the model that may cause distraction from actual problem. Thus there is need of mechanism which can intelligently allow only relevant features to be select. Consequently feature subset selection is a great emerging field in this aspect.

In Feature subset selection, features are reduced intelligently so that this reduction may not affect the classification accuracy. Hence the feature selection can be define as the search problem for the best optimal subset features 'm' out of original feature set 'M' with the aim of maximizing classification accuracy minimizing learning time saving database resources and reducing measurements cost.

In medical diagnosis used of the classifier system is in trend now. There is no doubt experts decisions are the most important factor in identifying disease. But the expert systems and different artificial intelligence techniques can serve as the tool for the experts by making their job easier and free of error in case of fatigue or inexperience experts. Moreover these artificial intelligence classification techniques facilitates in analyzing medical data in more detail and less time.

In inductive learning, machine learning algorithm learn the model by different examples or instances. each example increases the experience of the algorithm to generate the rules for classification. For example, algorithm can be provided by the information (case history of patients or clinical testes) of set of patients in a hospital. The algorithm would then be

informed which patient is suffering from particular cancer and which not. This information about the patient will then be used by machine learning algorithm to contrast a model to diagnose particular disease. Major problem for this artificial diagnose system is to identify which information is more important in correctly predicting the particular disease and which one information is extra overhead on the classifier .For the identification of useful features that best describe the patients optimal feature selection problem arises.

Different approaches to optimal feature selection are filter approach ,wrapper approach and embedded approach. In filter technique feature selection is a step before applying inductive algorithm. Later this subset is used as input to the inductive learning algorithm during the training step. In wrapper approach search start from empty set later on features are added or removed under the evaluation of the selected features that is based on learning algorithm. Hence the inductive learning algorithm is the part of the search engine. In embedded approach optimal features selection is part of the inductive algorithm.

1.1 Motivation

Following factors motivate us to check the application of MOGA for optimal feature selection for biomedical domain

- Limitations of the simple genetic algorithm[9,10]
- Easy formulation of Objective factor calculation using MOGA
- Fast Convergence of the MOGA towards true optimal
- Conflicting Multi-objective nature of biomedical data
- Better results of MOGA in other domains

1.2 Goals and Challenges

Mostly in real-life problems, concern objectives have conflicts. Thus, optimizing features with respect to a single objective often results in undesirable results with respect to the other objectives. As it's almost impossible to simultaneously solve multi-objectives functions faithfully Hence it's a great challenge to find out the way out of multi-objective problems. As there is always tradeoff between multi-objectives. No single solution can be

consider as best in all aspects. So one solution to a problem is to. don't rely on one solution. But considering many best alternative solutions, each best solution will represent satisfaction of demand at specific level.

In practical life problems, as in case of industry product, there consider two objectives, reliability and cost, if we decrease cost reliability degrades, increase in reliability influence cost. So, optimizing multiple objectives at same time is impossible. So there is certain level of compromise we should have to accept in one objective in order to gain in other objective.

Another example of such a problem is diagnosing disease through different clinical tests,each clinical test associate with the clinical cost and have certain level of importance in diagnosing disease, So selecting clinical testes out of original set of clinical testing with aim of increase in accuracy of diagnosing disease and less testes so less cost at the patients end, is another multi-objective problem. If we try to increase accuracy with less cost we need to find some important clinical tests that are best with respect to both aims. In such a problems single solution is not enough, we should find all alternate best solution in multiple scenarios. Solution to such a problems is to identify the set of best solution out of many possibilities.

Solution to this problem is to find out all possible Pareto optimal solution sets or a representative subsets. A Pareto optimal contains set of solutions that are non dominated as compare to each other. Each Pareto solution always has a certain amount of surrender in one objective(s) to achieve a certain amount of gain in the other objective. Pareto optimal solution sets are preferred because they can be practical when taking into consideration real-life problems [16].

Another challenge in population based optimizations method is to maintain the best solution of previous generation in next generations in a way that diversity of the population not affects. Consistent spread of solution is a great curse in population based techniques in efficient way with fewer computations.

This thesis summarizes the research in Feature selection using Multi objective Genetic Algorithm and its application in biomedical domain. Many attempts have been made using Simple Genetic Algorithm and some advanced optimization techniques .Here is the layout of the whole thesis.

1.3 Contribution

Following are the main contribution of this report

- Deals with the conflicting multi objectives in a better way unlike simple GA fitness functions that tries multiple objective to fit in one function that causes biased results in many cases.
- Improved population initialization technique by incorporating opposition population in population initialization step . That provides better spread of population then ordinary random population initialization. Hence provides better candidates solutions.
- Promises to survive the best solutions in every generation. Random selection not ensures the existence of the best solutions in the next generations. A simplest approach to elitism in a multi-objective genetic algorithm is to copy all the best solutions of parent population to the newly generated child population. To fill the rest of offspring population, reproduction operator will be applied on other fronts except first front of parent population. This technique will not be successful in case , total count of offspring plus non dominated parent exceeds than the size of the population.
- Improves the diversity so that weaker solution (having some good features) may have chance to select for mutation, without increasing additional complexity. Crowding distance approaches intend to find a consistent spread of solutions beside the best-known Pareto front devoid of using fitness sharing parameter.

1.4 Thesis Layout

Chapter2 gives brief overview of, difference between simple and multi objective genetic algorithm, optimization techniques (SGA, MOGA, and PSO), and Simple Genetic Algorithm application in feature selection.

Chapter3 presents the literature review of feature selection using GA, non GA and Multi objective Genetic Algorithm.

Chapter4 introduces the proposed technique that is opposition based multi objective genetic algorithm.

Chapter 5 reports performance measures, experimentation and results and comparison with other techniques.

Chapter 6 Concludes this work, summarize the work its application and future work.

Chapter 7 Contains references.

Chapter 2

2.Optimization Techniques

Optimization is comparing and finding feasible solutions from the search space until no better solution is found .The resulting found solution can be consider as good or bad depending on objective. Objective can be single or multiple. In real world problems more than one objective has to be optimized simultaneously.

This chapter presents overview about optimization, difference between single and multi objective optimization, feature selection optimization with genetic algorithm ,detail working of the genetic algorithm and overview of particle swarm optimization.

2.1 Function Optimization

Objective optimization means that there is/are some goal(s) that needs to be met by using the given resources in optimal manner. Goal can be maxima or minima in either case we have to consider not to stick in local maxima or minima. But sometimes the search space is so large that it becomes so difficult to search through whole space. There are two types of function optimization.

- Single Objective Optimization
- Multi Objective Optimization

2.1.1 Single Objective Optimization

Optimizing single objective requires minimizing one objective at a time. So it is called single objective optimization. [21]

Ex ample 2.1:

Minimize F(x, y) =z= 10- x + x y + x2

Where x ,y> -80

In Example 2.1 there is only one objective function, and the goal is to minimize the function. In Figure 2.2 each point 'z' against the corresponding x and y variables forms the three dimensional decision spaces.

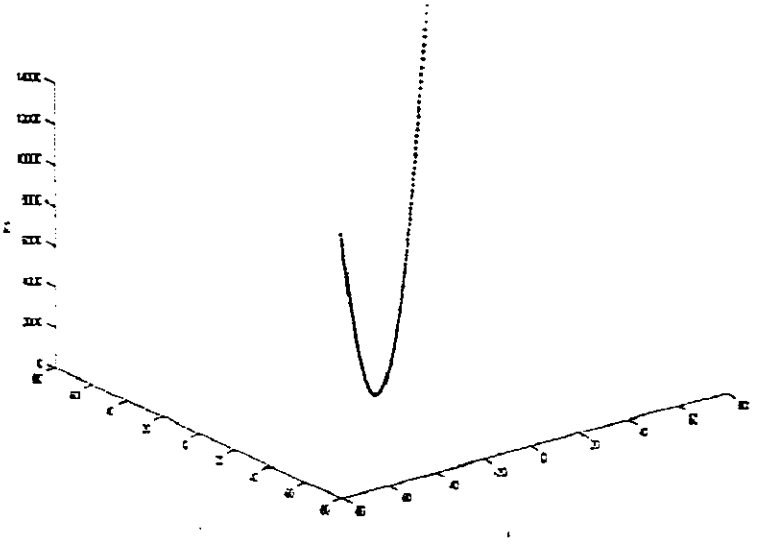


Figure 2.1: decision space and its function value [11]

In real world there are several scenarios where multiple conflicting goals have to be achieved simultaneously. For example In case of industries to maximize profit there exist many objectives that are conflicting to each other (minimizing cost, maximizing productivity), but have to adjust them simultaneously in a way so that overall firm profit can be maximize .In real world problems all objectives may not be independent to each other they normally depends on other objective. It's a great challenge to deal such problems.

2.1.2 Multi objective Function Optimizations

When two or more objective need to be optimizes at a time is called multi objective optimization [21].

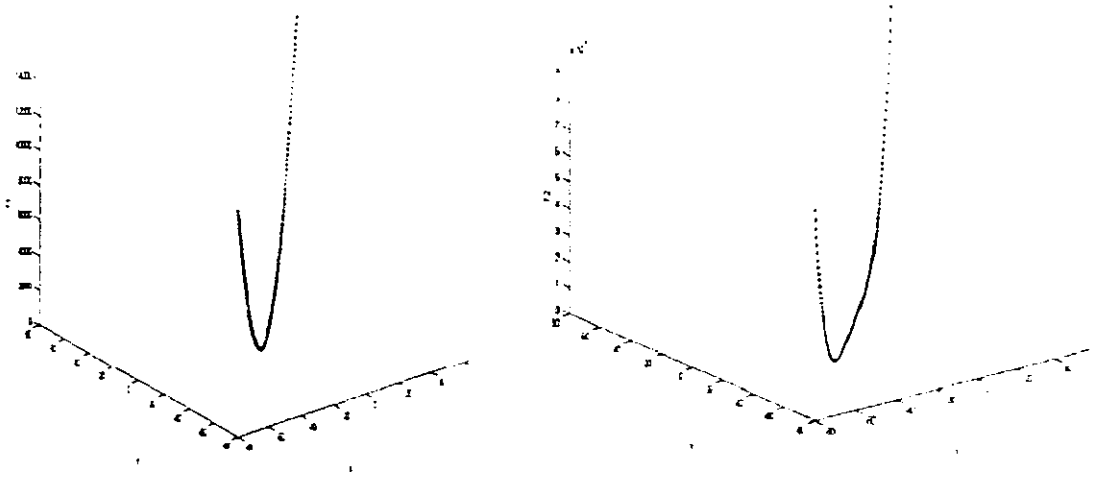
Example 2.2:

A. Minimize F1(x , y) = 10- x + xy – x2 Where x and y both are greater than -80

B. Minimize F2(x , y) = (x2 + y – 11) 2 + (x + y2 – 7) 2

Where x and y both are greater than -80

In Example 2.2 there are two objectives functions and require minimizing both of them simultaneously. In multi objective optimization each function will have its own separate decision space. Number of decision spaces will be equal to the number of the objectives or function need to be optimized. Decision spaces for example 2.2 are given in Figure 2.2 , it shows there exist two independent decisions spaces .



(a): Decision space for F1 (b): Decision space for F2

Figure2.2: Decision spaces

To optimize the multiple objectives simultaneously have to map these all decision spaces to one objective space and this can be map when we have one single calculated value for each decision variable against each objective.

The main difference between the single and multi objective optimization is this additional objective space. Objective space for the above example 2.2 is given in figure 2.3. The dimensions of the objective space is equal to the total objectives need to optimized in given example there are two dimensions as two objectives are given for optimizing simultaneously.

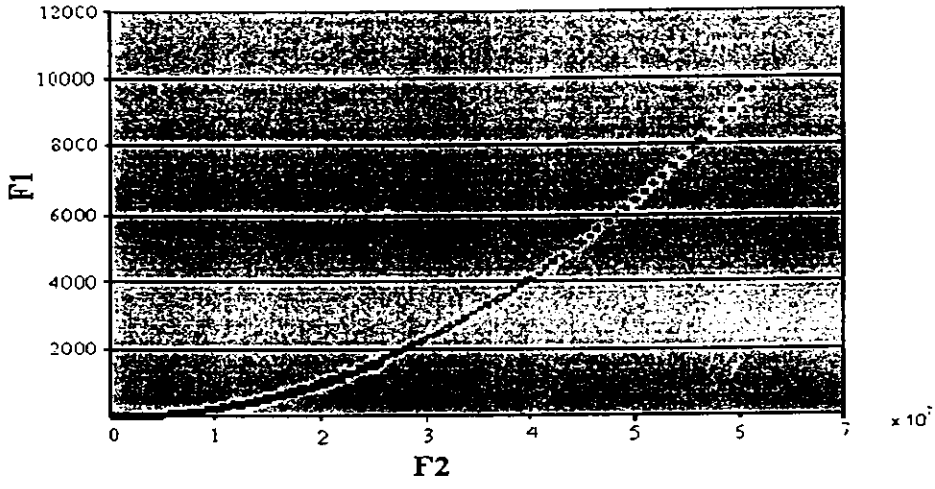


Figure 2.3 : Objective space of F1 and F2 [11]

2.2 Search Techniques

There are three types of basic search methods, calculus based methods ,guided random search techniques and enumerative techniques.

Enumerative techniques, visit every possible point while searching, and visit one point at a time. These techniques are simple to implement, but in the case when search space is too large it will be a time consuming procedure to visit each and every point. In some cases, we can restrict our search to not to visit specific region that cannot contains the solution e.g. game playing. Some examples of enumerative techniques are dynamic programming, depth first search and bread first search.

Calculus based techniques considers the feature space like a multi-dimensional continuous function and find out minima or maxima by finding out the derivative of function.

Indirect methods employ the information, that the extreme(a) lies where the function derivative is equal to zero. Where there derivative of the function is zero the density of the search points is also very low as compare to the whole search space.

Direct calculus techniques try to locate the nearby extrema by using the values of the gradient of the function, such as Newton technique. These types of techniques are recognized as **Hill Climbing** techniques, as they approximate for extrmea and move to that

point and again estimate and again move until it reaches the top of the hill. Such techniques are helpful in the case of well behaved problem or the problems that can be altered to turn out to be well behaved.

The above mentioned techniques are direction less like random they have probability of getting stuck in local maxima or minima so they are not consider as effective. Whereas, enumerative search techniques are not feasible for large search spaces. Hence we go for heuristics for large search spaces.

Stochastic refers to the system whose behavior intrinsically non-deterministic. These search techniques estimate about the next point through the some predictable actions and random elements. They are universal in their capacity, being able to resolve some very difficult problems that are away from the capabilities of either calculus or enumerative techniques.

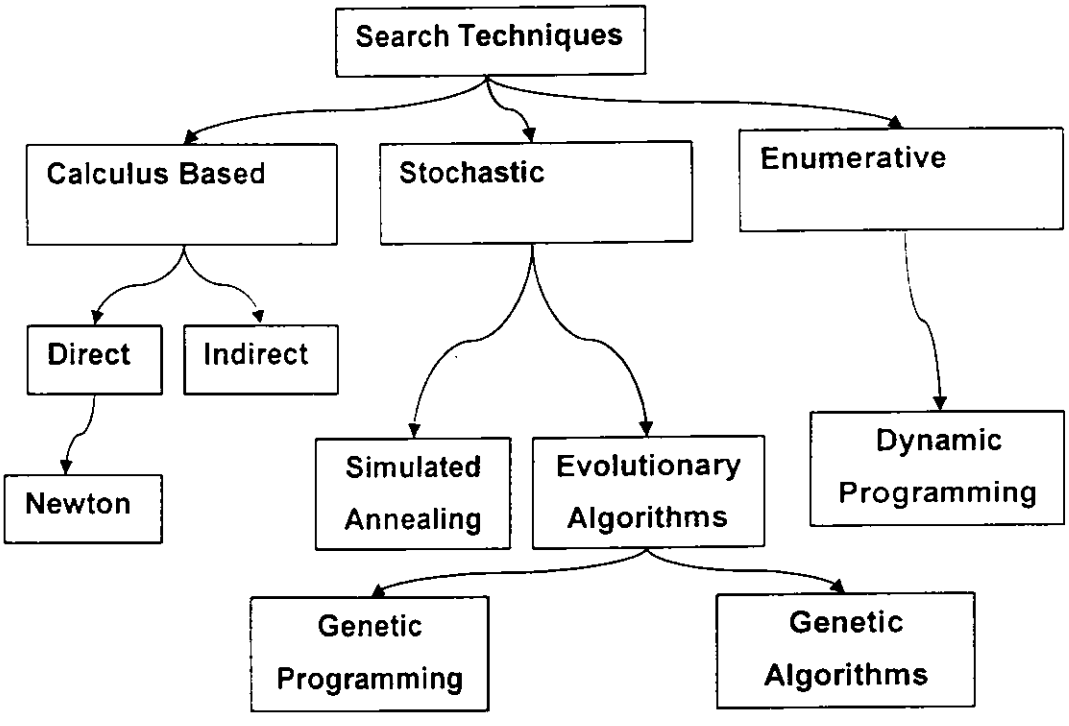


Figure 2.4 :Search Techniques

Simulated annealing is an probabilistic approach to find global optimization of the function that is supposed to have many **local** maxima or minima .This approach is inspired by the physical procedure where large **but** low energy soft crystal can be obtain by first

heating it then by slowly cooling .This is done to finally froze the structure this is happen at the stage when crystal left with minimum level of energy. SA is best for sparse data. But it is computationally expensive.

The classical methods failed to produce desired results because of certain in discrepancies mentioned above. So research shifted from the classical methods to the non-classical, un-orthodox and stochastic search and optimization algorithms.[45]

The evolutionary algorithms mimic the behavior and evolutionary principles of nature in the form of computer sciences and algorithms. These algorithms derive the search towards the optimum solutions in space.

The main advantage of these algorithms is that they keep solutions set and population instead of keeping one solution at a time. This results in the same value of the each individual and convergence to the same objective in case of single objective optimization. This is really great in multi objective where each individual can represent separate optimized objective value and this is a great outcome.

The evolutionary algorithms are divided into some main categories. These main categories are

1. Genetic Algorithms. An inspired model whose inspiration is taken from human genetics of selection and reproduction.

2. Genetic Programming. It can be thought as the specialization of genetic algorithms with the only difference that its representation is different from genetic algorithms and instead of using string representation. It uses trees representation mainly to represent computer programs.

3. Evolutionary programming. It is based on behavioral models and not the Genetic model. It is derived from the simulation of adaptive behaviors in evolution.

4. Evolutionary strategies. In this type instead of using string for representation of population, the real values are used. it also uses mutation as the only operator for evolution.

2.3 Genetic Algorithms

Genetic Algorithms are being commenced into use officially at University of Michigan in the 1970's by John Holland. Attractive, for some types of optimization problems because of price/performance improvements of current computational systems. GA considered as well suited for the combinatorial problems and shows good results on mixed discrete and continuous problems, and for the search spaces where little is known about the underlying search "space".

They are less chances of being hook in local optima as compare to the greedy approaches. However lean to be computationally pricey. In each hunt a fresh set of artificial offsprings (strings) are reproduced through bits and genes of the fittest of the previous generation; occasionally a new part tried. Randomized genetic algorithms are not pure random walk, employ historical information. Widely used in business, science and engineering.

2.3.1 Brief History

Genetic algorithms programmed on the computers by the biologist in the late 1950s and early 1960s, biologists who were clearly looking for to model aspects of natural evolution. At that time they don't have idea that this technique will be use in future as solution to artificial problems. By 1962, researchers namely G.J. Friedman, G.E.P. Box, H.J. Bremermann and W.W. Bledsoe all had developed evolution-inspired strategies for machine learning and function optimization, however work in this field concerned little follow up. In 1965, another more successful effort came into existence, through the introduction of the evolution strategy technique by Ingo Rechenberg, he was from the university of Berlin, This new technique was more similar to the hill-climbing than genetic algorithm. This new technique had no crossover and, to produce new offspring one parent was used to mutate. And the one that is better was consider as parent for the next time when mutation occurred. Concept of population was introduced later in next version. Scientist and engineers still use evolution strategies particularly in Germany.

These initial works recognized more common attention in evolutionary computation. By the near the beginning of 1980 to mid-1980s, GA had been widely used in the fields of mathematics in problems like graph coloring and bin-packing ,GA used in engineering

field in problems like flow control of pipeline , classification ,structural optimization and pattern recognition .

Nowadays, evolutionary computation is a flourishing field, GA are nowadays use in solving problems of daily interest such as biochemistry , aerospace engineering ,molecular biology ,scheduling airport ,prediction of stock market, planning portfolio .assembly line and microchip design. The influence of evolution has touched almost several fields individual cares to name, determining the world around us unnoticeably in innumerable ways, and novel applications keep on to be exposed while research is enduring. Moreover GA is nothing more than Charles Darwin theory of evolution: that the random chance of variant, joined with selection law, is a problem solving practice of massive power and almost limitless applications.[44]

2.3.2 Working of Genetic Algorithm

Genetic Algorithm is a technique to get solutions to difficult problem by evolving the population of the competing candidates .Genetic Algorithm methodology is inspired by biological evolution it's not logical but simply competing candidates are spawned to get population of better solutions.

In this evolution procedure, weaker candidates normally die and are discourage to evolve. While the stronger solutions are promise to survive. These are evolved to produce new solutions that are more competing than earlier generations.

GA begins with all possible candidate solution. But only the fittest survives and reproduce generation after generation. And this fitness measures is the potential of the candidate solution to solve problem. So evolution of the candidate solutions can be considered as learning. GA is heuristic in nature.

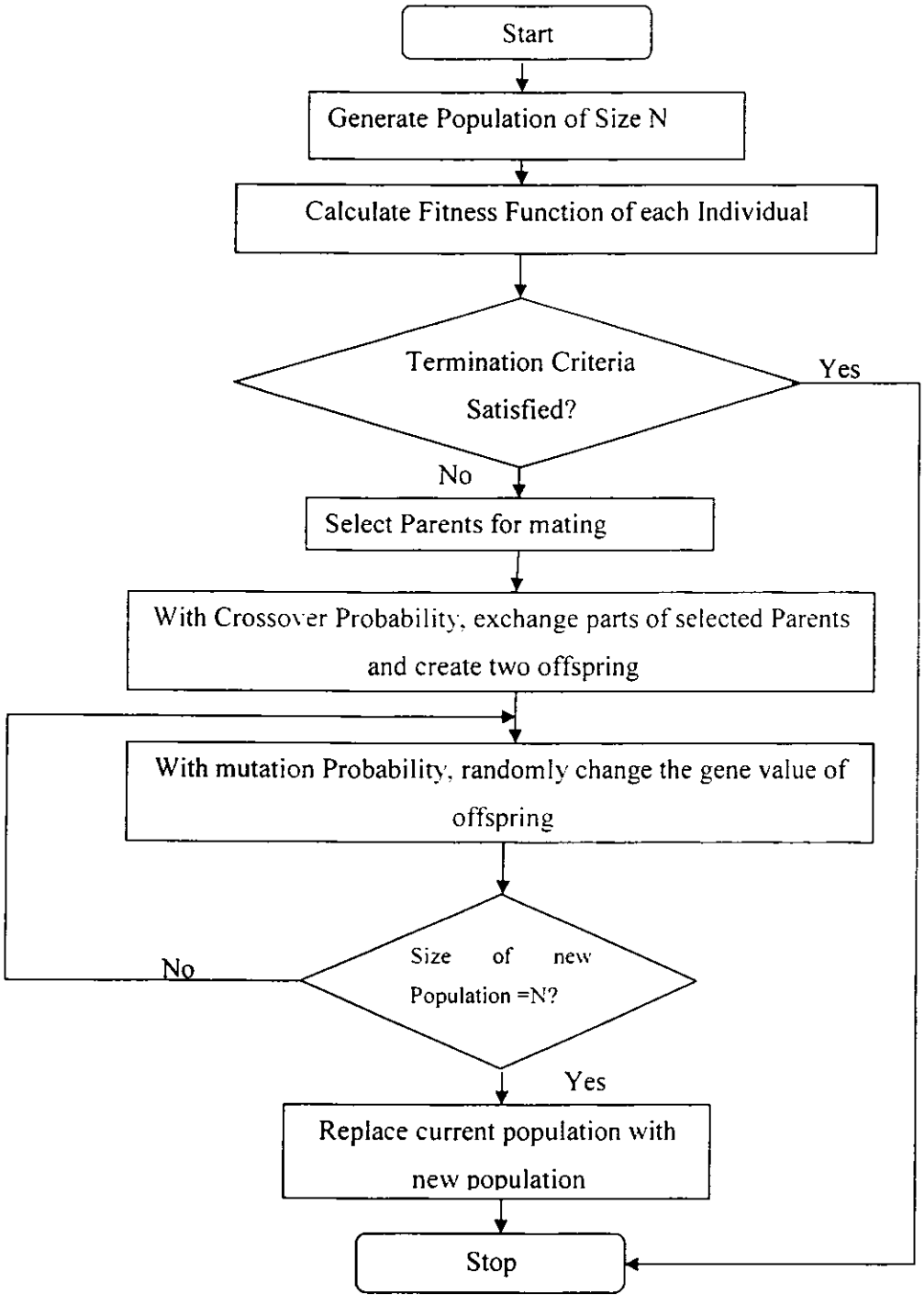


Figure 2.5 Genetic Algorithm

Algorithm 2.1 : SGA

```
Begin
Initialize gen=0
 Population initialization P (gen)
While termination criteria is not accomplish
    Begin
    Assess fitness of each candidate  solution of the current generation
    Select parents  to generate offspring
    Produce offspring by using genetic operators
    Replace some members from new offspring
    gen = gen+1
    End while
End while
```

Figure 2.6 : Algorithm for SGA

2.3.2.1 Chromosome

Representation of Solutions: The Chromosome

A single chromosome may compromise of different parts mention below.

Gene is a basic unit of chromosome it represents one single feature or attribute. A single chromosome is a collection of many genes or in other word a single chromosome may represents different attributes or features

Allele A single gene all possible values are called as allele.

Binary Coded Vector In case when Binary coded vector representation is used .A single allele may have two have values 1 or 0.So a chromosome may consist of strings of bit. And the length of the string will be equal total features or total Genes.

Real Coded Vector In Real coded vector a single gene may have any possible decimal value.

Population Group of chromosomes form a population. Where, one chromosome represents one individual, or one solution to a problem, or one point in the search space.

Solution Representation Importance Representation of the problem, effects the complexity and efficiency of overall algorithm

Example 2.3: Cookies Problem

In Cookie problem we need to try different ratios of sugar and floor in the range of 0 to 9 kg.

From 0 to 9 different combinations are possible here I will represent just two as example.

6	1	Chromosome # 1
2	4	Chromosome #2

Here two possible solution or two possible chromosomes has been shown. Chromosome consists of two genes. Each gene may have any value in the range of 0 to 9.

Learning of rules can be done through GA.

2.3.2.2 Fitness Function

Fitness Function is the evaluation measure of each individual in the population. Fitness function is so very important in GA .As all results will depends on this. On the basis of the fitness value we will decide who will be surviving and will die. Fitness Function Formulation is therefore a big issue in GA. In Fitness Function Formulation first of all we need to identify all possible objectives. Need to study the nature of objective. Either there is single objective or more than one. If more than one, are they dependent on each other or conflicting? Fitness Function varies with every problem. But the basic things that we need to consider in each domain are the deep study of the nature of the objective as it does matters a lot.

Example 2.4:

For the case study given in Table 2.3.1.2 fitness function can be the percentage of classification accuracy of the rule sets over the available training instances. Other criteria may also consider for example the magnitude of the rule set and the complexity of the rule set.

In case of cookie example fitness function may be consider as taste of the cookie that may be score in percentage, 1-100.

Table 2.1 : Example of Fitness Calculation

Number	Restaurant	Meal	Day	Cost	Reaction
1	Saira	Breakfast	Friday	Cheap	Yes
2	Saman	Lunch	Saturday	Expensive	No
3	Mammona	Lunch	Sunday	Cheap	Yes
4	Bushra	Breakfast	Friday	Expensive	No
5	Irum	Breakfast	Sunday	Cheap	Yes

2.3.2.3 Reproduction Operators

Genetic operators are being applied on the chromosomes that are selected as the best parent on the basis of fitness function to create offspring for next generation.

Reproduction operators are basically of two types

- crossover
- Mutation

2.3.2.3.1 Crossover Operator

In crossover operation we take two candidate solutions and swap their parts with each other to form two new offspring having good features of both parents in new created offspring.

Example 2.5:

Before crossover

6	1

Chromosome # 1

2	4

Chromosome #2

After Crossover

6	4	Offspring # 1

2	1	Offspring # 2

Depending on where to split have to made and how to split chromosome to share attributes. It can be dividing into different categories.

- Single Point Crossover
- Uniform Crossover

2.3.2.3.2 Mutation Operator

It takes a single candidate. For the diversity purpose small random changes are used to make in a single chromosome. Mutation will be helpful n the scenario when the whole population of current generation missing some attribute on the same location for example if initial population missing 1 at first bit location.

If the mutation rate is low, new trial will appear slow .If mutation rate is high each generation will be unrelated to each other.

2.3.2.4 Selection Process

Selection process is to choose the two best parents from the existing population on the basis of fitness function. So after fitness function it is critical to choose proper selection procedure.

Selective pressure is the rate at which a selection algorithm elects individuals with higher than average fitness .In case of low selective pressure population will not be abif there is not enough selective pressure, the population to converge towards a single solution. In alternate case of high pressure population will lose diversity and chances of premature convergence happen to rise.

- Random selection
- Proportional selection

- Tournament Selection
- Rank Based Selection
- Survival of the most diverse

2.3.2.5 Next Generation

The new offspring replaces the old. And some old better solutions also included in the next generation.New generation elitism describes the ratio to which old generation members will be added to new generation. It ensures that maximum fitness may not tend to decrease from one generation to next generation.

Generation gap means percentage of the gap of from one generation to the next generation. If new generation contains all the new individuals then this gap is 100%.

Duplicates allow If duplicates will be removed it may increase the efficiency of search and reduces premature convergence but can increase the processing time in case of large population.

Population size: Number of the solutions in current population.

2.3.2.6 Termination Requirement

The GA continues generation after generation some predefine criteria is met.

- Find solution that have fitness that met the predefine threshold value
- Predefined Number of generation have reached
- No father improvement observe, shows stable results

2.4 Particle Swarm Optimization

Phrase ALife or Artificial life is employed to portray human created systems that mimic the behavior of life to solve the computational problems. It's a biological system more specifically social system that is inspired by that is inspired by the single individual behavior with other in its environment called as swarm intelligence. Another popular

method of swarm optimization is Ant Colony Optimization. Here I will give the brief overview of just PSO.

Particle Swarm Optimization (PSO) [23] is population based optimization technique inspired by social behavior of the fish coaching and bird flocking. PSO was developed by Dr .Kennedy and Dr. Eberhart in 1955.

PSO shares many similarities with other evolutionary techniques such as Genetic Algorithm. In PSO likewise in GA Population is initialized with population of random solution. Then Search is made to find optima. However PSO have no reproductive operators' PSO candidates or potential solutions are termed as particle fly through the search space by following the existing optimum solutions.

As compare to the GA PSO is easy to implement and few parameters needs tuning.PSO can be applied to the problem as function optimization, weights optimization, fuzzy system control etc.

Suppose the scenario a flock of bird is searching for food. There is only one piece of the food in area under search. The birds don't know where actually food is what they know is just how far they are from food all birds will follow a bird that is more near to food. In PSO each bird is a single solution in a search space i.e. termed as particle. All the particles have velocities which direct the particles for flying and associated fitness function that needs to be optimized. A particle flies through the space by following the current optimum particle.

PSO initiate, by initializing particles randomly. Search procedure set in motion by updating generations. In all generations particles are updated by following the superlative two values. The primary one is the most excellent solution or fitness that it has yet attained. This best value is make out as **pbest** .One more superlative value track by the swarm optimizer is the best value till then now attain by any particle and is determine as global best(**gbest**).When particle takes part of population as its topological neighbor now the best value is the local best(**lbest**).

Algorithm 2.2 : PSO
for all particle

```
    Initialize particle
end
do
  for all of particle
Compute fitness of particle
      If the fitness of present particle is better than the best fitness
            Set it as fresh pbest
End
Pick the particle with the top fitness as compare to all the particles as the gbest
For all Particles
   Compute velocity of particle
   Renew the position of particle
End
```

Figure 2.7: Algorithm for PSO

2.5 Summary

This chapter presents the categories of optimization techniques. In the perspective of given problem i.e. optimized features selection for biomedical domain , as biomedical data is particularly known for its high dimensionality with relatively few instances, calculus base method assumes that objective function should be an analytical expression as well as this function should be differentiable. But decision variables in the problem under taken are inversely proportion to each other thus it is impossible to formulate biased free analytical expression. Consequently it is clear that they are not applicable for given problem.

If we go for enumerative techniques, it inquires about objective function value at every point in the search space. Therefore due to dimensionality curse it is not appropriate for given problem.

Random methods or Random walks on average are not efficient in accuracy as enumerative techniques.

Random Search Strategies use randomness to guide the search through the huge search space. Booming example of randomized search technique is genetic algorithm.GA works

with the coding of decision variables not the values of decision variables.GA not initiate the search from single point like hill climbing but by the population of points.GA make use of induce information not the derivatives.GA employ transition rules that are probabilistic not the deterministic ones. Because of the multi objective problem, extension of genetic algorithms presented knows as multi objective genetic algorithm. The main difference between simple genetic algorithms and multi-objective genetic algorithm is the multi-objectives handling. The best method for multi-objective selection is pareto-optimality. Chapter 4 includes the details of multi-objective genetic algorithms and pareto-optimality.

Additional advance Randomize search existing techniques are particle swarm optimization. ant colony optimization and honey bee optimization etc. For the at hand problem the randomized search method of opposition based multi-objective genetic algorithms has chosen. [46]

Chapter 3

3. Literature Review

Biomedical data endures with huge amount of data having lot of redundant and irrelevant features that slows down not only training and testing process, but also utilizes higher resources as well as poor diseased detection rate. Therefore feature selection is important issues for biomedical field. This chapter introduces basic concept of feature selection techniques. Surveys existing feature selection techniques in biomedical field. Groups and compare feature selection algorithms under two broad categories.

- Filter Approaches
- Wrapper Approaches

Last section of the chapter concludes the survey by identifying the current trends and challenges in research for feature selection in biomedical field.

3.1 Feature Selection Procedure

This section presents the details of four key steps in feature selection.

- Generate Subset
- Evaluate Subset
- Termination criteria
- Results validation

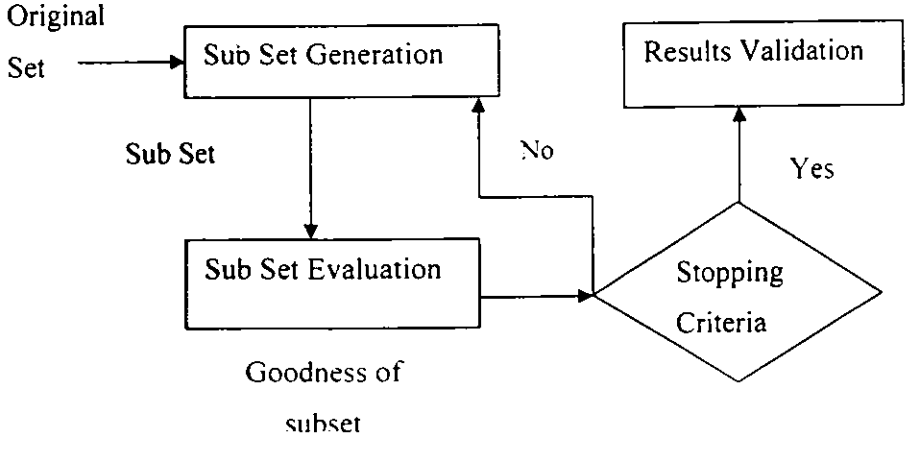


Figure 3.1 Feature selection procedures

Subset generation is heuristic search process. Each state in search space represents a candidate solution. In generating subset, two most important issues considered. First of all searching point or points from where the search (to the optimal solution) in the search space will be start. It is very important, as it guides the direction of the search that can influence the final results. Search can be start from an empty set and then in succession add features to the set, this is called as forward search. If search starts from full set and then removing the features in succession is termed as backward search. If search starts from both ends and keep on removing and adding features simultaneously is called as bidirectional search. Another way to start the search is by randomly selecting the subset in order to avoid the local minima or maxima. Secondly one must select a suitable search strategy. If data set contains 'n' features then the possible candidates are 2n.Exhaustive search strategies are not even suitable for moderate 'n' value. Different search strategies are random , sequential and complete.

Each **subset generated** requires to be evaluated using some criterion function. Evaluation function can be broadly classified into two categories depending on the dependency of learning algorithm. If the evaluation of the features not depends on learning algorithm then the evaluation criteria's are normally distance measures, consistency measures, information measures and dependency measures and the model that uses this evaluation criteria is called as filter mode. The model that uses classification accuracy to evaluate the accuracy of selected features is called as wrapper mode.

There are different criteria to **stop Feature Selection process** .It may stops after certain generation reach, the output solution may not be the best but will be improved one as compared to the previous generations. Or it stops when search is complete. Another possibility is to set the threshold fitness that when achieved should stop the procedure (minimum features or iterations), or it can be stop if further iterations don't shows improvement in results. Finally it presents selective Subset feature.

Results validation is normally done on the basis of some prior knowledge about data. Prior knowledge about data usually not available. In general results are validated using performance measures of the classifier or by comparing the results using all the features for classification task with using selected features.

3.2 Filter Approaches

Filter approach is a method of features selection that is independent of any machine learning algorithm (while evaluating the features for selection as candidate solution), and search the basic properties of the individual feature is called filter method.

From this definition we can conclude that in filter approach individual characteristic of the features usually being observed and their collective behavior is ignored.

3.2.1 RELIEF

It works[5] in filter mode .it randomly select an instance from the sample and finds its near miss and near hit by finding another instance from the same class that is close to the selected instance .and finds the another instance that is close to the selected instance from other classes. The underlying idea is that a feature is more relevant to I the more it separates I and its near miss, and the least it separates I and its near hit .The advantage of the relief was it can analyzed the irrelevant feature but poor in the case of removing the redundant features.

3.2.2 Distance based Re-sampling

The aim of this presented work [30] was to indentify the significance features effecting diabetes control, by the application of feature selection techniques to the current working patient management system. With the help of classification model we can identify the individual in the population with poor diabetes control position with the help of two factors that are physiological and examination

The data of the patient was collected by Ulster Community as part of clinical management. To improve the online computational efficiency offline feature selection technique (filter) was applied through supervised model construction. An optimized relief method is used to rank the important features that effects diabetes control. Afterwards three different classifiers are used to evaluate models prediction accuracy.

To deal with the categorical problem handling, they used encoding scheme that was frequency based .Scalability of the data was optimized by distance based re-sampling. This technique was applied to select the few instances that are uniformly distributed and hence

not cause biasness. The accuracy achieve was although similar to the old relief technique but it has been observed that it implies less computational complexity for large databases. Features were assessed by exercising FSSMC that gauge s the importance by comparing its value to the patient's outcome. And if more patients have this range of value the feature is consider as useful otherwise not. The optimal features selected were 15 out of 47.

3.2.3 Correlation

The idea of this paper [6] was to find the attributes that were extremely correlated with the target class but are uncorrelated with other features. To search such type of feature search used was the best first search. That was too much computationally expensive .and they not deals with to find higher order dependencies if this too done it will be even more computationally expensive.

3.2.4 Fast Correlation

This approach [29] illustrates the importance of selection of features techniques in medical databases for diagnosing the diseases. Basic theme is to explore the small set of informative features that are helpful in diagnosing diabetes disease from PIMA Indian diabetes database. Optimal feature subset is determined make use of Symmetrical Uncertainty Attributes Set Evaluator and filter that is correlation-based. Results are evaluated using Lib SVM classifier

Given 'N' instances of data and total number of 'N' features and Class C. Problem definition was to find optimal features subset from m dimensional observation space S^m Total sub spaces available are 2^m.

Feature Selection process has two modules. Compute correlation between each feature 'i' and class 'C' . The attributes whose values exceed the threshold value are chosen and are called as best subset feature S_B . In the Second module S_B is further processed and unneeded features are removed.

F_i is chosen if F_i is SU(F_i ,C))≥SU(F_j ,C) and SU(F_i , F_j)≥ SU(F_j , C)

$$\text{SU=Symmetrical Uncertainty} = 2\left[\left(IG(\frac{X}{Y}) \Big/ H(X)+H(Y)\right)\right] \quad (3.1)$$

Where IG is information gain

$$IG(X/Y) = H(X)-H(X/Y) \quad (3.2)$$

And

$$H(X) = -\sum p(x_i)\log_2(p(x_i)) \quad (3.3)$$

H(X/Y) is the Entropy of X conditioned that Y has been observed already

$$H(X/Y)= -\sum p(y_i)-\sum p(x_{i/yi})\log_2(p(x_{i/yi})) \quad (3.4)$$

ROC and AOC are used as performance measures. Accuracy using all features (77.4%) and selected feature (77.9%) though improved but just in point so there is no difference in both.

All these above mentioned feature selection techniques used different measures to remove the irrelevant features and their evaluation criterion was distance, correlation and inconsistencies even after all these measure we have no confidence about the group performance of the features and it can be only done by involving some extra evaluation measure that is to select the features based on their prediction accuracy on some classifier. So to remove this filter drawback in literature some wrapper approaches were proposed that select the features based on their prediction accuracy performance on some classifiers.

3.3 Wrapper Approaches

The wrapper approach applies a specific machine learning algorithm to evaluate the featues. Performance of current selected feature will be compared with the best feature subset obtained during the previous steps. This searching process will be continued until a pre-define criterion accomplished. Wrapper approach has high probability of producing

classifiers with better classification performances than the filter approaches and features collectively contribution to model generation. Whereas its associated drawbacks are higher risk of over-fitting , computationally intensive (large number of features) .

3.3.1 Fusion of GA with NN

The basic theme of the paper [9] was to reduce the input feature for Dist all neural network classifier in order to increase classification accuracy, minimize cost and improving generalization.

At front end for feature selection they have used GA for exploring the search space for optimal features to be selected .To evaluate the selected offspring they have used Dist all neural network instead of traditional neural network. The best individual obtained after the last generation.

Feature evaluation is done using equation 3.5

$$\text{Fitness}(x) = \text{accuracy}(x) - \text{cost}(x) / (\text{accuracy}(x) + 1) + \text{max cost} \quad (3.5)$$

Dist all is a fast and simple neural network learning algorithm for pattern classification. The key feature of the Dist all is to add one hidden neuron at one time using greedy algorithm which guarantees maximum correct classification.

Experimentation was carried out on both real world dataset as well as synthetic data .real world data set was obtained from UCI [15]. Real world data set belongs to different domains. My main emphasis was on biomedical data. So I here mention the results of just biomedical domain in experimentation chapter.

Limitation of the work was that although the features were reduced but with this reduction number of the hidden neuron was also increasing. That is increase in the computational complexity. Other limitations were .this fitness function will not work well in the case as zero cost solution with low accuracy will be preferred rather than higher accuracy solutions with moderate cost. So it is not good to combine multiple conflicting optimization criteria's into one fitness function.

3.3.2 Fusion of GA with SVM

In this paper[10] feature reduction is carried out for support vector machine classifier .The basic theme of the paper was same as before increasing classification accuracy to achieve this goal they not only optimized feature subset by GA but also kernel parameters .scope of their work was as follow.

- Data Preprocessing (Scaling)
- Converting genotype to phenotype
- Feature Subset: At front end for optimizing feature subset and kernel parameter they have used GA
- Fitness Evaluation: To evaluate the selected offspring they have used Kernel based Support vector machine.
- Termination Criteria
- Genetic Operation

The chromosome compromises three parts features mask Penalty parameter and Gamma

Fitness Function

$$W_A \times SVM\_accuracy + W_F \times \left(\sum_{i=1}^{nf} C_i \times F_i \right)^{-1} \tag{3.6}$$

Kernel based support vector machine is used to evaluate the selected features. Experimentation is done only on real world data set taken from the UCI repository. It has the same limitations of not dealing with conflicting objective in a satisfactory way. Same is the case with all other literature e.g. [11] Feature Subset Selection Problem using Wrapper.

3.3.3 Integer-Coded GA with SVM

The basic theme of the work [25] was to diagnose the heart disease using heart disease database. The wrapper based approach is used in this work for feature selection task. The main contribution of this work was that they classify heart disease not only as absence or presence of disease unlike in the literature work but they further also classify it as 5 classes where '0' represent as absence of disease and other 4 classes are different types of heart

diseases. Used integer coded representation. And feature evaluation was through just classification accuracy. They used Simple GA for optimal feature subset selection and used support vector machine algorithm for selected feature evaluation.

Scenario 1

Multi class SVM: One against One and One against all

Optimal C =150

Scenario 2

RBF kernel: RBF shows good performance with value of parameter 0.0.25 with respect to classification accuracy. As a binary class accuracy achieved is 90.5%.and as a multiclass problem accuracy achieved is 72.55% with 6 features out of 13.

3.3.4 Feature Selection using Wrapper Approach

Feature selection technique presented in this [26] work is wrapper based. Feature selection is done on biomedical data from four different domains Pima Indian Diabetes database, Wisconsin Breast Cancer, Heart Stat log, Breast Cancer

- For optimal feature selection Wrapper based GA is used
- Four pattern recognition techniques are used to evaluate selected features

Pattern Recognition Techniques Naive Byes, C4.5, RBF, Bayesian and DT

Results show that there is no one standard technique that is best for all databases. But another outcome of the experimentation is that surely by apply feature selection technique we can improve the classification accuracy.

3.3.5 Pharmacogenomics Approach

Often Chronic Hepatitis patients stop the treatment (interferon-Alfa and ribavi-rin (IFN-Alfa/RBV)) because of high cost of the treatment and its adverse effects. So it is required both economically and clinically to have tool that can distinguish among responders and non responders and to estimate the outcome of IFN-Alfa/RBV treatments. The basic aim of this study [27] was to build a predictive model based on a Pharmacogenomics approach.

- Clinical factors for data gathering

- Wrapper based feature selection where at the front end for exploring the search space for good sub set feature selection they used best-first search. Best fist search starts with empty set of the features and continues by choosing the feature through greedy hill climbing augmenting with backtracking.
- Feature evaluation by Multi Layer Feed Forward Neural Network(MFFN) and logistic regression

Data Set was collected from 523 HCV patients at National Taiwan University Hospital. From the collected samples virologic responders (SVRs) of the treatment and NRs are defined on the basis of serum HCV RNA test. Afterwards Genomic DNA has been extracted from each blood sample. In their work they focused on 24 SNPs.

As one locus constitute of three genotypes, similarly one single allele have represented by three values where 0 is for homozygote (major allele), 1 for heterozygote, 2 for homozygote for minor allele. To evaluate the generality of predictive model they used 10 cross validation and to evaluate the performance of predictive model they used accuracy of true predictive patients.

Average accuracy obtained for 1-4 hidden layers was 80.4%,80.4%,80.0%,and 79.7% with Genetic factors identified are 4 out of 24 after applying wrapper approach. Whereas the accuracy obtained through logistic regression was 75.3% with 5 genetic factor identified as best. So results gain by MLNN is better.

Sample size was small that can effects the drawn results problem. SNP's could not be the only factor to identify the responders more factors are required.

3.3.6 Fuzzy Preprocessing Weighted Approach

Basic aim [28] was to diagnose hepatitis using machine learning system. Their main contribution was their own new fuzzy pre processing weighted technique.

- Feature Selection is performed by means of C4.5 decision tree

Table 3.1 Summary of the related work for hepatitis

Author or year	Method	Classification Accuracy in percentage
Karol Grudzin ´ski	Weighted 9-NN	92.9
Karol Grudzin ´ski	18-NN, stand. Manhattan	90.2
Karol Grudzin ´ski	15-NN, stand Euclidean	89.0
Rafał Adamczak	FSM with rotations	89.7
Rafał Adamczak	FSM without rotations	88.5
Rafał Adamczak	RBF (Tooldiag)	79
Rafał Adamczak	MLP + BP (Tooldiag)	77.4
Stern & Dobnikar	LDA, linear discriminant analysis	86.4
Stern & Dobnikar	Naive Bayes and Semi-NB	86.3
Stern & Dobnikar	QDA, quadratic discriminant analysis	85.8
Stern & Dobnikar	1-NN	85.3
Stern & Dobnikar	ASR	85
Stern & Dobnikar	Fisher Discriminent Analysis	85.5
Stern & Dobnikar	LVQ	83.2
Stern & Dobnikar	CART(decision tree)	82.7
Stern & Dobnikar	MLP with BP	82.1
Stern & Dobnikar	ASI	82.0
Stern & Dobnikar	LFC	81.9
Norbert Jankowski	IncNet	86.0
Ozyıldırım and Yıldırım et al. (2003)	MLP	74.37
Ozyıldırım and Yıldırım et al. (2003)	RBF	83.75
Ozyıldırım and Yıldırım et al. (2003)	GRNN	80.0
Work under discussion	FS-Fuzzy-AIRS (10 · CV)	94.12

- Data Set is scaled in the range of [0,1]and weighted by means of fuzzy weighted preprocessing
- Then through AIRS classifier system the weighed input was being classified
  For the detail of their proposed technique read [28].

The accuracy obtained by their proposed method was 94%.Other comparative studies of feature selection and their result obtained are mentioned in table 3.2.

In the above literature we focus on just three diseases that are heart disease, hepatitis and diabetes where as the application of feature selection is not limited to theses three but a tremendous work in biomedical data mining is done as a pre-requisite step here we mention some of them in table briefly.

Table 3.2: Summary of other related work

Diseases	Feature Selection Approach	Optimization technique	Classifier	References
Lung Nodule CAD	Wrapper	GA	SVM	[31]
Gene Array Cancer	Hybrid	GA+ Tabu	SVM	[32]
Thyroid dysfunction & Acute Appendices	Wrapper	GA	KNN	[33]

All papers discussed here under wrapper having the limitation of not properly solving multi-objective optimization problem in proper way. Objective function need to calculate in a way so that it does not produced biased results.

3.3.7 Hill Climbing Approaches

In this work [7] author proposed two algorithms to reduce the computation al cost of k nearest neighbor here I will discuss only hill climbing based approach. Here problem is represented by bit of string with '0' indicates the lack of the features and '1' denotes the

selection of the feature. The algorithm starts with randomly choosing a binary string and assumes it well evaluated .It apply mutation randomly on this string in result produces a new string its evaluation is measured and compared with the previous one. This procedure is carried on until some stop criteria is met. Limitation of this work was that hill climbing approach is poor in performance in the case of global maxima. It finds only local maxima.

3.3.8 Tabu Search Approach

For finding the optimal feature subset heuristic applies was the Tabu search [8]. Tabu Search promises to find in the unique search space means it remembers what features it had already analyzed for this purpose it maintains a tabu list in the memory for the candidate solution it already visited. The cost of the feature subset is measured by leave ne out strategy of k nearest neighbor. Its limitation was the extra complexity to maintain memory of already visited solution in case of large size sample data.

Simple Genetic Algorithm on the other hand instead of starting from a single randomly chosen string starts with many randomly chosen candidate solutions that will be more confidential as compare to the single chosen solution . from these handful solution optimal solution is searched that is guided by some fitness function or objective function. Following are the some examples of the wrapper techniques that used GA as front end.

3.3.9 MOGA Approaches

In this part of the survey I will discuss about the new trends in the feature selection and this survey will be domain independent later I will try to shift new trend toward biomedical domain that will be my part of research. Now on some data bases Multi objective Genetic Algorithm is applied to overcome the disadvantages of simple genetic algorithm. In Some work GA is replaced with other new heuristics e.g. particle swarm optimization.

3.3.9.1 Fusion of MOGA with ID3

Main aim of the paper [11] was to overcome the limitations of SGA. In Order to so they used NSGA

- Feature Selection through NSGA
- Features Evaluation through ID3

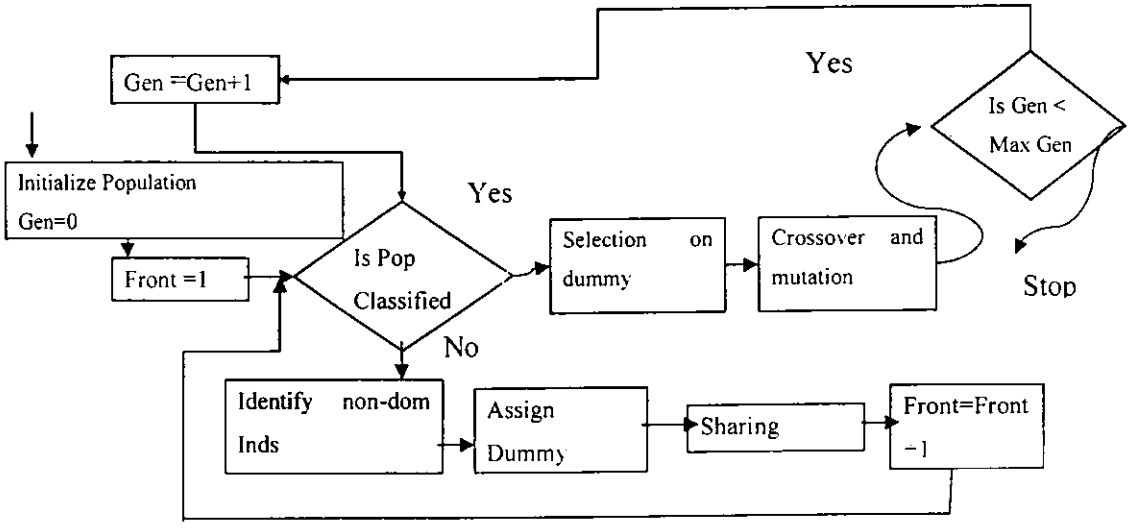


Figure 3.2 : NSGA

Number of objectives in this work is equivalent to number of Classes in data set. Number of correctly classified instances. Dataset used was domain independent .Results obtained were good only in the case when we are interested in some specific class. Results obtained were even bad than 1997 work [11].

3.3.9.2 Fatigue Fracture Image Identification

The theme of the paper [29] was the identification of the fatigue fracture image. Feature selection technique used in it was the amalgam of linear prediction and MOGA. Problem identified in this paper was that the Simple GA deals with single objective function although feature selection problem is not single objective problem.

- Feature Extraction
- Feature Selection
- Classification is done through quadratic distance classifier

Objective Function

1. Minimize error identification rate
2. Undetected identification rate
3. Number of selected feature

Firstly in LP-MOGA linear prediction is defined. Then in all generations of GA Predicted solutions are obtained through pare to optimal solution one by one from each non dominated front. Linear Prediction is performed using the following formula $\hat{x} = \sum w_i x_i$ where $x_1, x_2, x_3, x_4, \ldots\ldots x_k$ are the solutions in ordered in non dominated single front and $\hat{x}$ is a predicted solution, where wi is the weight on the non dominated elite solutions, and kl is the order of the elite solutions.

Algorithm: LP MOGA

Randomly generate initial population p
Generate children population q
Combine children and parent population R=pUq
Evaluate all the individuals in population
Divide combined population into non dominated fronts
*predict p1 new chromosome vt1 from front 1 and predict p2 new chromosome vt2 from front 2, where p1 = 0.2*h1 and p2=0.2*h2 rt = pUqUvt1 Uvt2*
Generate fronts of rt population.
Sort non dominated fronts
Choose the best solution to complete the population.
Create new population
If maximum generation reach switch to 3 else go to 12
End

Figure 3.3: Algorithm for LP-MOGA

Training 42 fatigue image 80 non fatigue rest of data is used as testing.MO give more better results than SO. When comparison is done with NSGA-II number of selected feature is same and other two objectives are less.

3.4 Problem Statement

Literature survey prepared in this chapter was in the context of our research topic. On the basis of the objective of literature can be categorize the literature into five different dimentions.

- Feature Subset Selection using Non GA based Methods
- Filter Approach
- Wrapper Approach
- Feature Subset Selection using GA in biomedical domain
- Feature Subset Selection using Multi Objective Genetic Algorithm

The first one branch was reviewed to know the classic trends to solve feature selection problem. The method discussed here in that section were non GA based methods for example simulating annealing, Tabu, Hill Climbing and Fuzzy methods. These all methods are obsolete as compare to the simple Genetic Algorithm with respect to performance and convergence to the solution. Tabu Search fails in the scenario when sample data is large data as in it have to maintain the memory for last visited solution. Hill Climbing deals with only local maxima.

Filter approach also comes under the classic method to solve feature selection methods. Filter approach is independent of learning algorithm and look at the basic properties of the individual features to remove the irrelevant features. Hence not consider the group performance of the features. That is necessary to consider in many domains especially in biomedical domain.

Whereas Wrapper approaches using Simple Genetic Algorithm covers the limitations of the filter approach. But the methods usually employing the wrapper approach calculate the fitness of the candidate's features by only considering the classifier accuracy on the selected features. Dataset or pattern available usually has more samples for one class and less for the other one class. So Due to this uneven distribution of the sample data it is not adequate to consider classifier accuracy as a fitness of selected features. There is another problem with the dataset is that when we found any missing class label we include that missed labeled too in high frequency class. So all these factor suggest considering some more factors to evaluate the fitness of the selected features. In some works multiple objectives have been consider But they try to plug all the conflicting objectives into one function and evaluate the features on the basis of that single value. Fitness Function Formulation is most tedious and important task in GA.As it is that which guide the search towards the solution. So we should be very careful about its formulation.

Although there are many applications of the feature selection almost in all fields now feature selection is used as preprocessing step in bioinformatics as real prerequisite. In literature for symptoms selections or clinical test selection wrapper based technique has been employed to diagnose the disease. Symptoms selections are important both economically, clinically and with respect to the efficiency of the classifiers that models to predict the disease.

Above Discussions reveals that objective function or fitness function formulations is more critical. For efficient objective formulation first of all one should must have the deep domain knowledge. Domain knowledge is necessary to identify all the objectives and properties of objectives. Either identified objectives are dependent on each other or independent. In case of dependent features or conflicting features they cannot be fit into one function.

Usually while formulating the objective function only classifier performance parameters are consider as objectives and objective from the domain is ignored. Although it should also be consider. Symptoms of the disease are mostly dependent on each other and conflicting in nature. So the feature selection is a conflicting multi objectives problem.

3.5 Summary

This chapter covers the detail history of the feature selection techniques. Different approaches to solve the problems were described thoroughly along with their merits and demerits filter approach is not consider well because features evaluated on their individual merits and ignores the interaction between features.

Wrapper using simple genetic algorithm is dependent of learning algorithm but fitness evaluation is biased in the case of multiple objectives optimization whereas feature selection is not a single objective problem, more specifically for biomedical. Consequently there is need for the algorithm that can better formulate objective function. In wrapper representation biases of the classifier are considered. Hence for the problem at present wrapper with multi –objective GA is selected.

Chapter 4

4. Proposed Technique: Opposition based MOGA

This chapter provides the detail description of proposed technique opposition based MOGA. historical highlights of MOGA , architecture of proposed technique and detail description of its each module. The last section provides the details of SVM classifier both theoretically and mathematically. For the best feature selection at front end we used opposition based multi- objective genetic algorithm NSGAII.

In the history many efforts had been made to solve multi objective problems with the help of some mathematical functions. But fails to do so and history shows the need of some heuristic to solve such problems. Here some historical highlights are presented to show why we move toward evolutionary algorithms to resolve the multi-objective issues.

The main objectives of the classical methods to solve the multi objective optimization problem was to reduce the multi objective problems into the single objective and then this is worked by the normal single objective optimization algorithms to produce results. This whole process can be done in the evolutionary algorithms as well but when we talk about the multi objective evolutionary algorithms then things are different.

Some of the important classical techniques that have been used are [13]

Stochastic - (e.g. simulated annealing ,random walk, & tabu). All these methods are very general, but unproductive .**Linear Programming** - speedy, but limited to solutions that are linearised .**Gradient Based(Hill Climbing)** - nonlinear, appropriate for only differentiable functions.**Simplex Based** is nonlinear for the functions that are discontinuous .**Sequential Optimization** - (lexicographic). Objectives are ranked by preferences and optimizes objectives in order .**Weighting Objectives** – function optimization is done by generating a scalar vector, manifold runs required.**Constraint** ,To optimize objectives with respect to others constraint has been used**Global Criterion** –Distance to an vector that is ideal is minimizes **Goal Programming** – Deviation from the mark constraints is need to minimizes. Employs topology math, multi criteria polyhedral dynamics.**Dynamic**

Compromise Programming - uses functions that are state transition, parameters that alter over time.

Evolutionary algorithm is on the whole suitable to resolve multi-objective optimization problem, since they treat concurrently with a set of promising solutions that forms population. This permits us to locate numerous portions of the Pareto optimal set in a single iteration. Furthermore, evolutionary algorithms are not as much liable to the shape or continuity of the Pareto front means they are easy to deal with discontinuous fronts or concave fronts .The classical methods failed to produce desired results because of certain in discrepancies mentioned above. So, research shifted from the classical methods to the non-classical, un-orthodox, stochastic search and optimization algorithms.

4.1 Proposed Architecture

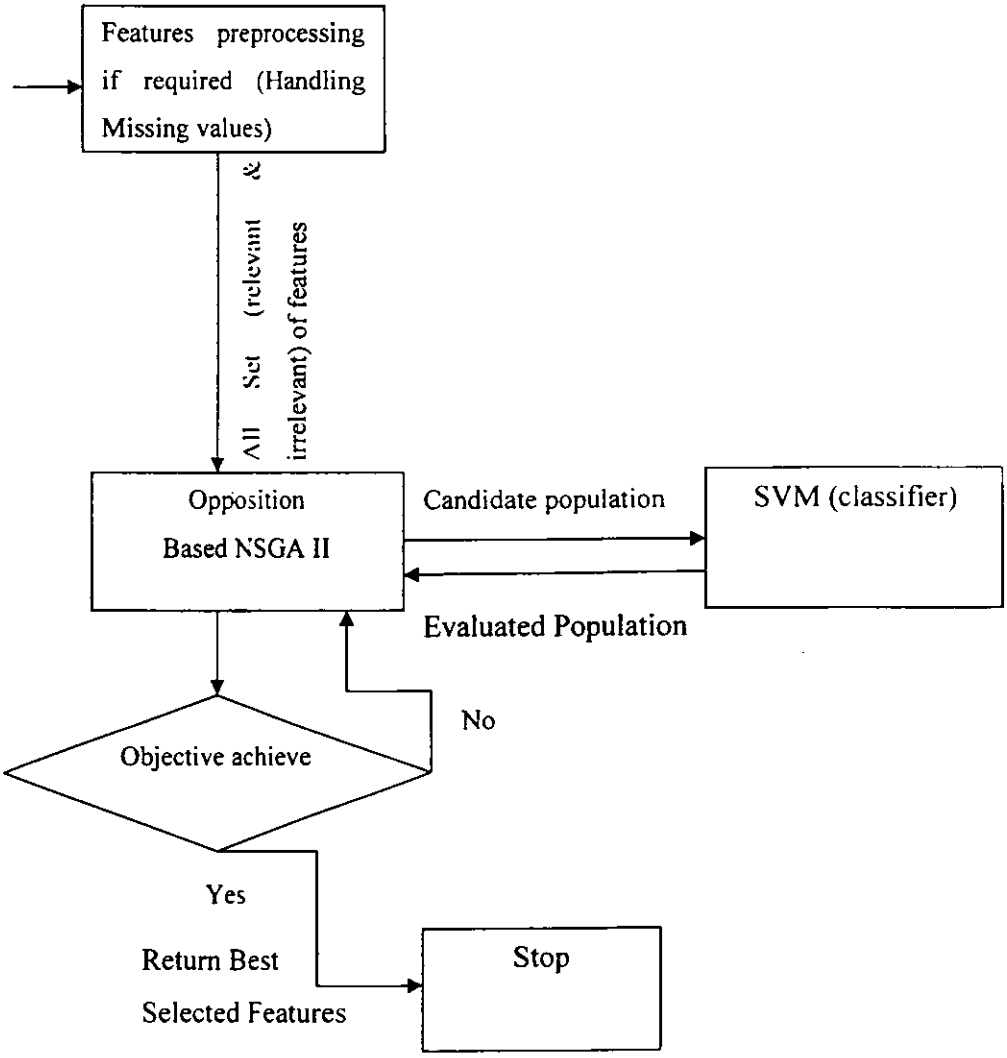


Figure 4.1: Proposed Technique Architecture

4.1.1 Features Preprocessing

Features preprocessing involves many steps as Feature discretization, normalization, handling missing values .The data set used in proposed technique has been taken from UCI [15] was already in discrete form. Problem of missing values was only to face in the case of hepatitis dataset. Missing values are filled by the value that was most frequent in other available samples.

4.1.2 Opposition based NSGAII

After the features preprocessing is done they are ready to give as input to the opposition based NSGAII. Here we incorporate OGA[41] opposition based GA to multi objective NSGAII to achieve fast convergence toward best sub set of optimal features.

4.1.2.1 Overview of Simple NSGA II

NSGA is a well known technique used in literature but it is mostly criticized because of its limitations of complexity, it is non elitist and for choosing the value of the sharing parameter. NSGAII overcomes the above mentioned limitations of the NSGA.

First of all it initializes the population in the same way as simple GA do. Initialized population is sorted into different fronts on the basis of non dominance. First Front contains the individuals that are dominated over all the population. Subsequent Front contains the individuals that are non-dominated just for individuals of the front 1.in the same way other fronts continue the procedure until each member of population get belongs to its relative front.

All entities in a front is allocated a rank for example every member of first front is assigned a rank 1.Each member of second front is assigned a rank or fitness value 2.

Furthermore to maintain the diversity distance between each member in a front is also calculated and this calculated distance is referred to as crowding distance or Density estimation. Large average distance between individual points to the fact of better diversity in population.

Afterward the parents are chosen from the population using Roulette wheel selection. This selection is based on rank and crowding distance of the individuals (Lower rank and greater crowding distance will be prefer). Selected parents will form new generation by creating offspring through mutation and crossover operator.

4.1.2.2 Explanation of the population based NSGAII

Algorithm: Wrapper Based Feature Selection using opposition based NSGAII for SVM

1. *Initially, create a random population (P_0) of size N, where each chromosome is representing different symptoms to target disease.*

   1.1 Calculate opposition of population

2. *Evaluate the population using SVM(RBF)*
3. *Using non-dominant sort, sort the parent population last created.*
4. *Divide whole sorted population into different fronts. Such that each member of front 1 is dominated by all other subsequent fronts members. And ranks the individual according to the front. For example all the individuals in front 1 has rank 1 and it's the best rank. Rank all other fronts accordingly.*
5. *Using Roulette wheel selection genetic operators and recombination, and opposition calculation create child population (Q_0) of size N.*
6. *Including first generation and all subsequent generations will follow the following steps.*
   a. *Create intermediate population (R_t) of size 2N by concatenating the parent population (P_t) and the child population (Q_t).*
   b. *Evaluate intermediate population (R_t) using SVM(RBF)*
   c. *Sort (R_t) and divide it into fronts as we did in step 4.*
   d. *Produce the new parent population (P_{t+1}) of dimension N. Sort the intermediate population. Selection of the best parent will be done based on rank and crowding distance. All front is filled in ascending order until the addition of population size is reached. The final front is incorporated in the population depending upon on the individuals with slightest crowding*

distance.

e. to produce new offspring population (Q_{t+1})) Perform the selection, genetic operations on the newly generated parent population (P_{t+1})

7. Repeat Step 5 until last generation reached.

Figure 4.2: Wrapper Based Feature Selection using opposition based NSGAII for SVM

4.1.2.3 Opposition based Population Initialization

Here each individual is a possible combination of the feature from the available set of the features. In other words each individual in population is subset of the original set of the feature. Size of the population is given by the user as it is tunable parameter. So if the population size is N .So randomly N subset of the original population will be generated. Large population size can take larger processing time and hence solution will converge slowly towards optimal solution. On the other hand if the population size is too small it can cause pre mature convergence toward solution.

Because of poor diversity in this case GA can be stuck in local optima. To improve this factor of the Genetic algorithm we incorporate opposition based GA to multi objective technique. Improved convergence and success rate in [41], motives to incorporate it to proposed technique.

In current case studies of biomedical data we take the population size as 50 and initialize it by using uniform population methods then opposites of these chromosomes are calculated using same chromosome size. Our proposed method converges faster as compare to the techniques used in the literature.

$$x' = a + b - x \tag{4.1}$$

Opposite of a candidate solution can be generated by using equation 4.1.Where x is the original value of the gene and x' is the opposition of the original gene value .a +b is the sum of minimum and maximum value of i^{th} decision variable(features).For the problem under study maximum possible value of the gene is 1 and minimum possible value is 0. As 0,1 indicates the presence or absence of the given feature in the generated subset.

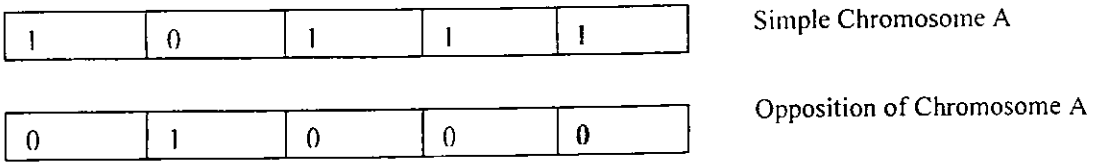


Figure 4.3 : Opposition chromosome calculation

4.4.2.4 Problem Encoding

Representation of the solutions is in the form of chromosome.

Gene is the basic building block of the chromosome. It represents one characteristics of single chromosome or single solution. In our case study (biomedical data) single symptom of disease is gene.

Value of the gene is called as **allele.** In current case studies there are two possible values of allele 0 or 1.'0' value for any symptoms shows that we are trying to diagnose disease without using that symptom or feature. If the allele is set to 1 it depicts that symptom (gene) is selected to diagnose the disease.

Chromosomes are collections or string of the genes. Collection of the genes forms single solution of the problem. Each chromosome is represented as a point in the search space.

Set of the chromosome is called **population**. All possible chromosomes or all possible solutions form population. The way the chromosomes are represented effect the efficiency and complexity of the genetic algorithm.

4.1.2.5Non-dominated Sort

To sort a population of size N on the basis of level of non-domination. Domination of each solution in population will be check against all other solution in the population. Before explaining the algorithm of non-dominated sort, some important concepts of multi objective genetic algorithm needs to be explained.

Above diagram shows some fundamental concepts of Multi objective optimization. The curve in the objective space is pareto- front. The solution on this curve are collectively called as pareto-set and single solution on this curve is said to as pareto-point.

Depending on objective either it aim to maximize or minimize points below the curve here in this example are infeasible solution .The points above the curve are feasible solutions.

As the name indicate multiple objective optimization deals with more than objectives at a time. In the past due to the absence of the proper solution to multi objective problems it has been usually treated as single objective problem. However, there are a lot of differences between working strategies of both of them. [22]Besides having multiple objectives other basic difference are two goals instead of one, two search spaces, no artificial fix up.

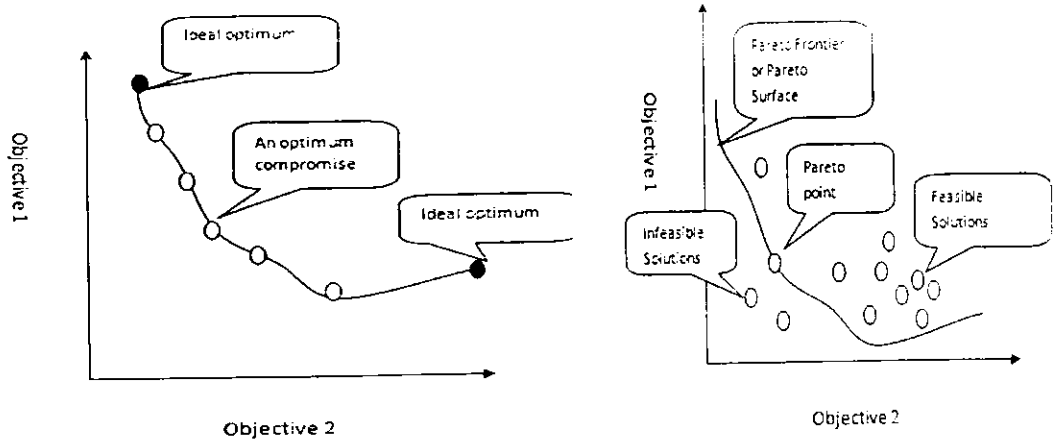


Figure 4.4 :Multi-objective Optimization

4.1.2.6 Dominant Non-Dominant and Pareto-optimality

For the assortment of multi objectives the most excellent and most vastly applied approach is pareto-optimal or non-domination approach

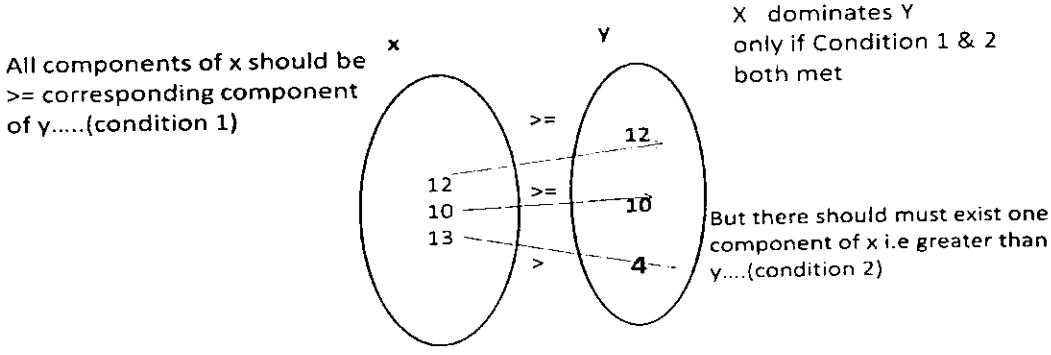


Figure 4.5 : Dominant, non Dominant and Pareto Optimality

First, for each candidate solution in current population, compute two things: Firstly the amount of solutions which dominates the given solution n_i and secondly a set of solutions which the current solution dominates i .e S_i. To calculate these two quantities we would have to do $O(mN^2)$ comparisons. In the same we have to discover all those solutions which have $n_i = 0$, and insert all such solutions to front F_i. Now this front is current front.

Now, for all solution in the existing front traverse each member (j) in its set S_i and decrease its n_j count by one. If count reaches zero for any member j, Add that member to a separate list H. When all individuals of the existing front have been verified, State the members in the list F_i as individuals of the first front. Continue this procedure using the newly acknowledged front H as existing front. All these iteration needs O(N) computations. This procedure will be repeated until all fronts are identified. As at most total fronts are N, $O(N^2)$ is the worst case complexity of this loop. The on the whole complexity of the algorithm is $O(mN^2)$.+$O(N^2)$ or $O(mN^2)$.

Algorithm: Fast non-dominated Sort

```
fast-nondominated-sort(P)
for each p ∈ P
  for each q ∈ P
    if (p ≺ q) then                    if p dominates q then
      S_p = S_p ∪ {q}                  include q in S_p
    else if (q ≺ p) then               if p is dominated by q then
      n_p = n_p + 1                    increment n_p
  if n_p = 0 then                      if no solution dominates p then
    F_1 = F_1 ∪ {p}                    p is a member of the first front
i = 1
while F_i ≠ ∅
  H = ∅
  for each p ∈ F_i                     for each member p in F_i
    for each q ∈ S_p                   modify each member from the set S_p
      n_q = n_q - 1                    decrement n_q by one
      if n_q = 0 then H = H ∪ {q}      if n_q is zero, q is a member of a list H
  i = i + 1
  F_i = H                              current front is formed with all members of H
```

Figure 4.6 :Pseudocode of nondominated Sort

It is important to mention here that even though the computational complexity has reduced from $O(mN^3)$to $O(mN^2)$ by carrying out orderly book-keeping, the storage space has enlarged from $O(N)$ to $O(N^2)$ in the worst case. Hence when the population undergoes the fast non-dominated sorting procedure it returns the list consisting of non-dominated fronts F.

4.1.2.7Density Estimation

To find an approximate of the density of solutions adjacent to a particular point in the population calculate the average distance of the two points on both sides of this point beside all objectives. Density estimation is actually an estimate of the largest cuboids cover by particular point(solution) without considering any other point. This is known as crowding distance or density estimation. The pseudo code for crowing distance is mentioned in figure 4.7.

Algorithm: Crowding Distance

- *N is the total individuals for all the fronts Fi .*
  - *initialize the distance value equivalent to zero for all the members of the front i.e. Fi(d)$_{ij}$ = 0, Where' i' is the index for front number and j is the index for jth individual in front Fi.*
  - *Against all the objective function m*
    - *Sort all the members of front Fi based on all the objective functions m*
    - *To the the individuals at extreme boundaries assign the value infinity*
    - *for k = 2 to (n - 1)*
      - *I(dk) = I(dk) + (I(k + 1).m - I(k - 1).m)/fmax(m) - fmin(m)*
      - *I(k).m is representing the value of of mth objective function of kth member in I*

Figure 4.7 : Crowding Distance

4.1.2.8 Crowded Comparison Operator

The crowded comparison operator directs the selection process at different stages of the algorithm in the direction of consistently spread out of Pareto-optimal front. In a way that if we have to made selection between the individuals that both belongs to different rank then we will prefer the one with lower rank. Another scenario is that both points belongs to the same front then selection will be base on crowding distance operator, then the solution

that has less solution in its surrounding will be considered (the mass of the cuboids inclosing it is bigger).

4.1.2.9 Recombination and Selection

The union of the child population and parent population will be taken and then selection will be performed on this new combined population. As on this stage population contains the best solutions of both the child population and parent population elitism is ensured. Sort the population based on non-dominated sort. This process will be further continue until the last generation.

4.2 SVM

A support vector machine (SVM) [42, 43] is an algorithm that is being learned by training data to assign target classes to specific features. So that on new test data it can correctly predict the expected label or class. For Example SVM can be learn to diagnose whether a person is diseased or not by examining the thousands of patient's data that are diseased or not diseased.SVM now has been using in many applications including bioinformatics applications. Microarray gene expression profiles automatic classification is a common biomedical application of SVM .Current case study involves three different diseases data sets (Heart stat log ,diabetes, hepatitis).SVM is used in Current case study is to evaluate the selected features fitness in term of accuracy and number of features.

SVM is aigorithm for maximizing particular criterion function with respect to the given available data. There exist two categories of classification, binary classification and multi-class classification. In binary classification we divide data into two groups while in multi-class classification we divide data into more than two groups. SVM classification can be explained using following four concepts.

- The separating hyper plane
- The maximum-margin hyper plane
- The soft margin
- The kernel function

4.2.1 The Separating Hyper Plane

For example in case we have only one attribute to classify data then the space where data resides of two class classification problem is a one dimensional line and that one dimensional line. This line can be dived into half using one point. In case if data resides in two dimensional space two dimensional space can be divide into half using straight line. And in case of three dimensional there is need of plane to divide the space. So the general term for the straight line is plane in higher dimensional space. Hence the separating hyper plane will be the line that separates two types of samples.

4.2.2 The Maximum-margin Hyper Plane

Considering each sample as points in higher dimensional space and dividing them into two groups by using hyper plane is not the only job of SVM .SVM is different from other classifier in the sense it select the best hyper plane out of many possible hyper plane. If the distance between separating hyper plane and the nearest sample point is consider as margin then SVM selects the hyper plane that have maximum margin to the closest data point.SVM not assume sample data forms the normal distribution.

4.2.3 The Soft Margin

We expect from the SVM that misclassification rate should be very less to deal with this problem SVM algorithm is extended by adding the soft margins. Soft margin provides the control to the user to specify to how many sample points they allow to violate the separating hyper plane .and a point can go how much far from the line in wrong region of the space .Hence Soft margin declares the size of the margin and the violation rate of separating hyper plane.

4.2.4 The Kernel Function

The problem is that, it is not possible that only single point can separate the data into two classes. Adding soft margin can even not solve this problem.SVM added one new additional dimension to the original dimension of the data point. To get additional dimension just original expression will be squared. Kernal function is a mathematical trick that allow SVM to perform two dimensional classification on the feature space that was originally of one dimension. The SVM by using some non linear mapping $\langle : D \rightarrow \Omega$. Map the input vector x to high dimensional feature space Ω . One should choose the kernel function that can separate the hyper plane with maximum margins and rather low dimensions.

4.2.5 Linear SVM

From known training set , samples labeled pairs (x_i, y_i), Where i=1 to m, where m=total number of training patterns xi are training input patterns and y_i are the corresponding target class of the pattern. If the problem in hand is linearly separable case it can be classified using following equation.

$$\langle w . x_i \rangle + b \geq +1 \; for \; y_i = +1 \tag{4.2}$$

$$\langle w . x_i \rangle + b \geq -1 \; for \; y_i = -1 \tag{4.3}$$

Equation 4.1 and 4.2 can be written in the following inequality form.

$$y_i \{ \langle w . x_i \rangle + b \} - 1 \geq 0 \; \forall \; i = 1, \ldots, m \tag{4.4}$$

The hyper plane defined by b and w is called the separating hyper plane.

Optimal separating hyper planes and maximum margins can be found by the following equations

$$\underset{w,b}{Min}\ \frac{1}{2} w^{t} w \quad subject\ to: y_i\{\langle w.x_i\rangle + b\} - 1 \geq 0 \tag{4.5}$$

To solve this optimization problem one should must find the saddle point of Lagrange function.

$$L_p(w,b,\alpha) = \frac{1}{2} w^{t}.w - \sum_{i=1}^{m} \left(\alpha_i y_i \left(\langle w.x_i\rangle + b\right) - 1\right) \tag{4.6}$$

Here α_i is the Lagrange multiplier that must be greater than are equal to zero. Saddle point is must be need to search because L_p must be minimized with respect to w and b and maximized with respect to α_i.

By differentiating the 4.5 with respect to w and b following equations are obtained.

$$\frac{\delta}{\delta w} L_p = 0, \quad w = \sum_{i=1}^{m} \alpha_i y_i x_i \tag{4.7}$$

$$\frac{\delta}{\delta b} L_p = 0, \quad \sum_{i=1}^{m} \alpha_i y_i = 0 \tag{4.8}$$

To find the maximum of the equation (4.5) the Karush Kuhn–Tucker (KTT) conditions are necessary and sufficient.

$$\alpha_i \left[y_i \left(\langle w.x_i \rangle + b \right) - 1 \right] = 0 \ \forall_i \tag{4.9}$$

Substitute equation (4.6) and (4.7) into (4.5)

Now the L_p will be transformed into dual Lagrange multiplier

$$Max_{\alpha} L_{D(\alpha)} = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, y_j \rangle$$
$$subject\ to: \alpha_i \geq 0\ i = 1, \ldots, m\ and \sum_{i=1}^{m} \alpha_i y_i = 0 \quad (4.10)$$

Hence the optimal decision hyper plane is

$$f(x, \alpha^*, b^*) = \sum_{i=1}^{m} y_i \alpha^* \langle x_i, x \rangle + b^* = \sum_{i \in SV} y_i \alpha_i^* \langle x_i, x \rangle + b^* \quad (4.11)$$

Mostly in classification process only the small subset of Lagrange multiplier approaches greater than or equal to zero. Geometrically these are the closest vector to optimal hyper plane .The training vectors that have non zero Lagrange multiplier are called as respective vectors .Decision hyper plane as mentioned above depends on these supporting vectors.

4.2.6 Soft Margin

The goal is to make the optimal hyper plane in a way there is minimum number of misclassification error. To solve this problem a non negative slack variable is introduced in following equations.

$$\langle w . x_i \rangle + b \geq +1 - \xi_i for\ y_i = +1 \quad (4.12)$$

$$\langle w . x_i \rangle + b \leq -1 + \xi_i for\ y_i = -1 \quad (4.13)$$

Now in term of these slack variable finding the optimal hyper plane has the following equation.

$$Min_{w,b,\xi} \frac{1}{2} w'w + C\sum_{i=1}^{m} \xi_i$$
$$subjectto : y_i(\langle w.x_i \rangle + b) + \xi_i - 1 \geq 0, \xi_i \geq 0 \quad (4.14)$$

Using Lagrange Multiplier

$$Max_{\alpha} L_{D(\alpha)} = \sum_{i=1}^{m} \alpha_i - 1/2 \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i, x_j \rangle$$
$$subjectto : 0 \leq \alpha_i \leq C, i = 1, \ldots, m \; and \sum_{i=1}^{m} \alpha_i y_i = 0 \quad (4.15)$$

To find optimal hyper plane dual Lagrange multiplier needs to be maximized with respect to α_i and the constraint $\sum \alpha i yi = 0$ and $0 <= \alpha_i <= C$ where i=1 to m.The Penalty parameter p will be defined by the user that is the upper bound on α_i.

4.2.7 Non Linear SVM

Non linear SVM Maps the input feature vector to high dimensional feature space using some non linear mapping 'φ' functions that are called kernel function in equation 4.9 inner products are replaced by kernel function 4.15

$$(\varnothing(x_i).\varnothing(x_j)) := k(x_i, xj) \quad (4.16)$$

$$L_{D(\alpha)} = \sum_{i=1}^{m} \alpha_i - 1/2 \sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j k \langle x_i, x_j \rangle$$
$$subjectto : 0 \leq \alpha_i \leq C, i = 1, \ldots, m \; and \sum_{i=1}^{m} \alpha_i y_i = 0 \quad (4.17)$$

Now the form of optimal hyper plane is of the form of following expression.

$$f(x,\alpha^*,b^*) = \sum_{i=1}^{m} y_i\alpha^* \langle \varnothing(x_i).\varnothing(x) \rangle + b^* = \sum_{i\in sv} y_i\alpha_i^* k\langle x_i, x\rangle + b^* \quad (4.18)$$

With the bias term b non linear SVM classifier can be expressed as

$$f(x,\alpha^*,b^*) = \sum_{i\in sv} y_i\alpha_i^* \langle \varnothing(x_i).\varnothing(x)\rangle = \sum_{i\in sv} y_i\alpha_i^* k\langle x_i, x\rangle \quad (4.19)$$

Different types of the kernels are

Polynomial Kernal

$$k\langle x_i.x_j\rangle = (1 + x_i.x_j)^d \quad (4.20)$$

Redial Bias Kernal (RBF)

$$k\langle x_i.x_j\rangle = \exp(-\gamma \square x_i - x_j)^2 \quad (4.21)$$

Sigmoid Kernal

$$k\langle x_i.x_j\rangle = \tanh(kx_i.xj - \delta) \quad (4.22)$$

Kernal function used in proposed method is RBF.

4.3 Objective Space

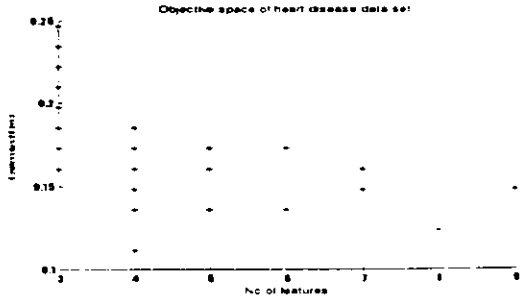


Fig (a)Objective Space for heart disease data set in iteration 1

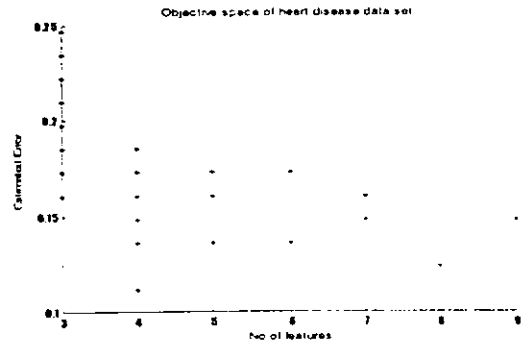


Fig (b)Objective Space for heart disease data set in iteration 2

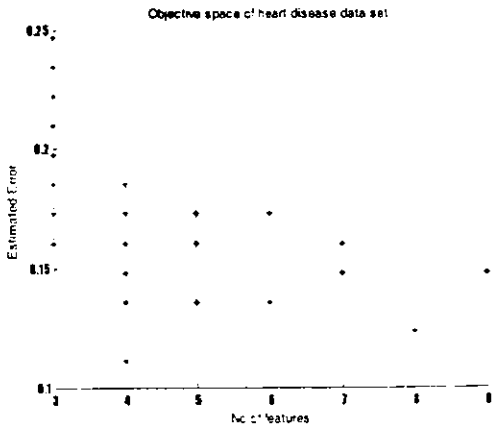


Fig (c)Objective Space for heart disease data set in iteration 3

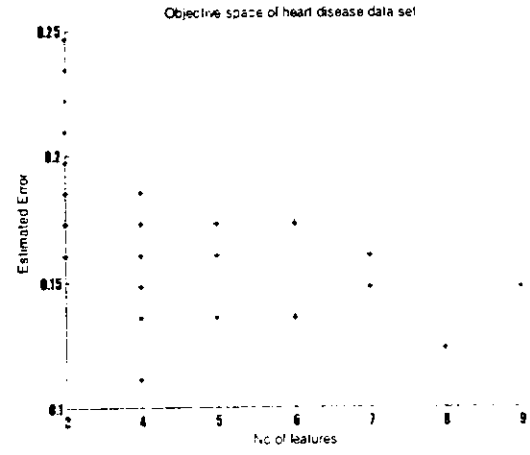


Fig (d)Objective Space for heart disease data set in iteration 4

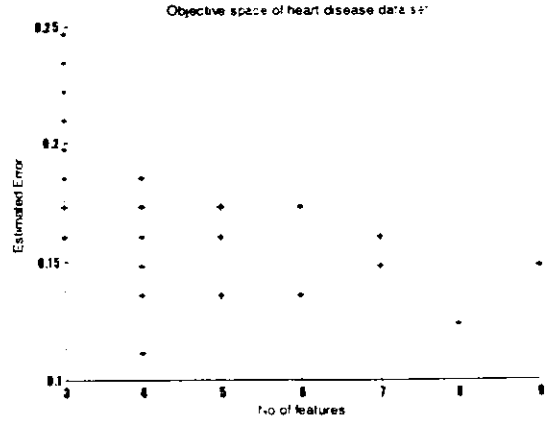


Fig (e)Objective Space for heart disease data set in iteration 5

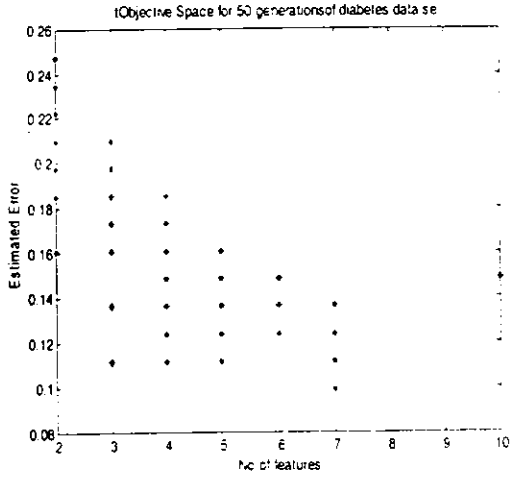


Fig (f)Objective Space for diabetes disease data set

Figure 4.8: Objective Spaces

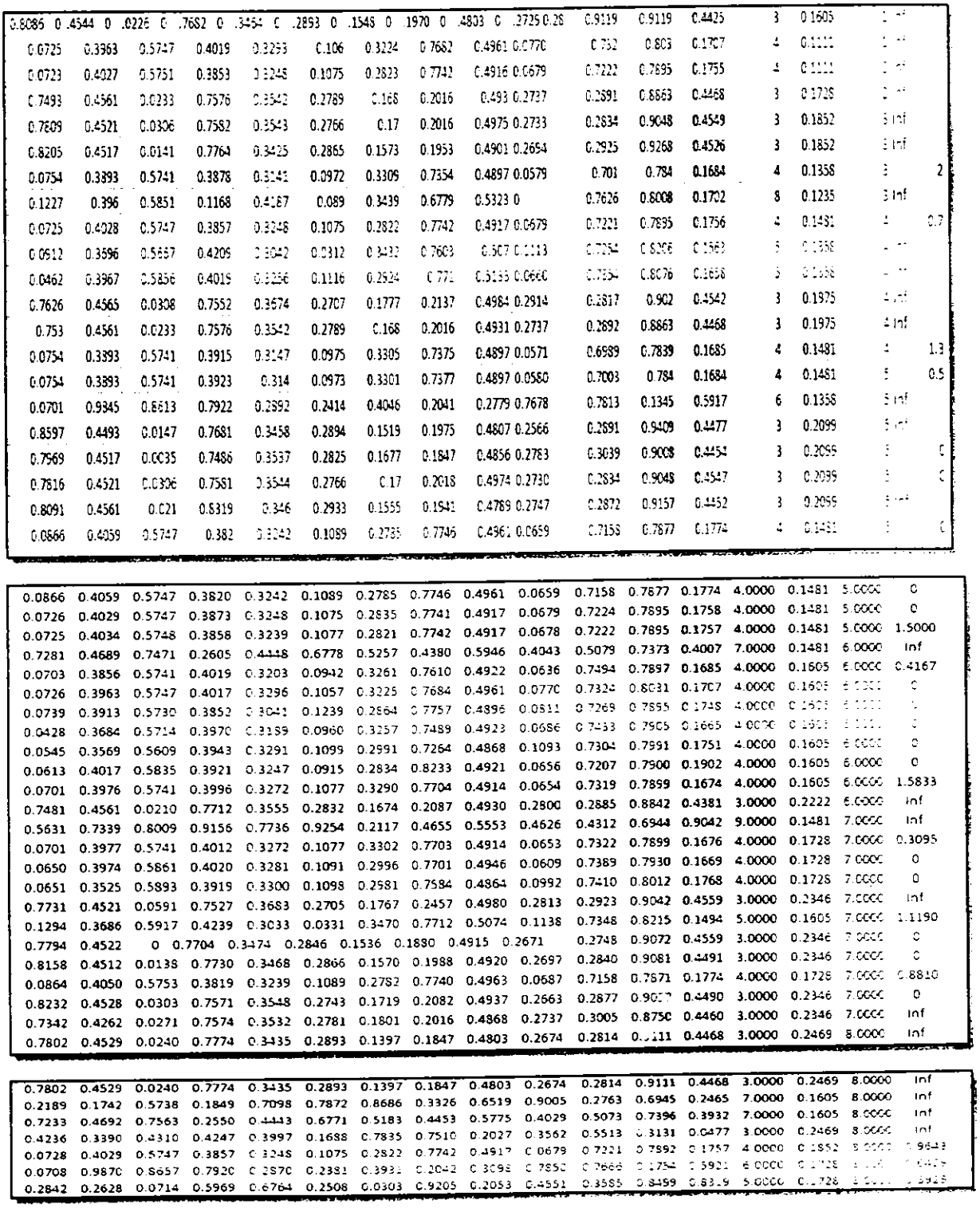


Figure 4.9 Snap shot of last generation individuals their objectives values and crowding distance

Figure 4.7 presents the objective space for heart disease data set and diabetes data set that is two dimensions. Dimensions of the objective space are equals to the number of the

objectives. Figure 4.8 shows the population of the last generation both objectives values rank and crowding distance respectively.

Here this figure 4.8 is taken from the heart disease case study. Where 13th column presents the chromosome that are generated randomly we treat them as zero for the absence of the feature and 1 as the presence of the feature. Next two columns show the evaluation of this set of feature in term of number of features and estimated error. The next column that is 16th column presents the rank of that specific candidate solution. The last column shows the crowding distance.

4.4 Summary

This chapter presents the importance of the multi objective genetic Algorithms its background and its effectiveness in biomedical sciences. As in the late history there was present no mathematical solution to the multi objective optimization. So there was the need of some heuristic that can provide solutions to the optimization of the conflicting objective.

It has been observed from the discussion that evolutionary algorithm can serve better for such type of problems. Here in this chapter we have discussed in detail about proposed technique that is multi objective in nature. It's a feature selection task for biomedical data for SVM Classifier. We not only test biomedical feature optimization using NSGAII but we also improved it by opposition based population spread that makes the application more effectives in term of fast convergence and all objectives under consideration.

Chapter 5

Experimental Results

This chapter presents the experimentations under different scenarios. This chapter gives the brief overview of different performance measures and statistical analysis, some of them will be used further in experimentation to evaluate the performance in different scenarios.

We will analyze the performance of new feature selection technique with SVM classifier using different biomedical data. We will compare proposed technique with other state of art techniques. We will also analyze how the multi objective technique improves the performance.

5.1 Performance Measures

After building a classifier or model, an organization or even a single developer wish to determine whether the classifier works accurately on future data on which it is not trained. Or it is also possible that some one develops more than one classifier and wants to compare their accuracy level or their error level. They are tool to help us understand, manage and improve what we want to do. Following are the various performance measures to evaluate the classification techniques:

5.1.1 ROC

ROC is a curve which gives the graphical plot of the sensitivity against false positive rate and is used for a binary classification.

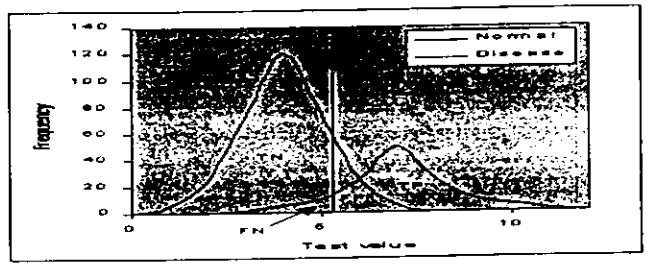


Figure 5.1 : ROC

The specificity and sensitivity of an indicative test depends on more than just the "quality" of the test. The position of the cut point between the curves will determine different measures such as number of true positive, false positives, true negatives and false negatives. When estimated test gives positive means in actual disease also present then it will be "true positive". Test positive and in actual there is no disease then it will be false positive. Test negative, but in actual disease is present it will false negative. Test negative, and in actual no disease it will be true negative. The overlap area of the curve indicates that the test cannot distinguish normal from disease.

$$Sensitivity = \frac{NumberofTP}{(NumberofTP+NumberofFN)} \quad (5.1)$$

Sensitivity is the probability that when the disease is present the test result will be positive.

$$Specificity = \frac{NumberofTN}{(NumberofFP+NumberofTN)} \quad (5.2)$$

Specificity is the probability that when there is disease the test results will be negative [18]

5.1.2 ROC Convex Hull

The receiver operating characteristic (ROC) curve is used to review the performance of binary classifier due to its ease of analysis, but it does not take account of misclassification cost information Provost and Fawcett have developed the ROC Convex Hull (ROCCH) method by using techniques from ROC curve analysis, decision analysis, and computational geometry .that searches for the optimal classifier that would be robust in case of skewed or inaccurate class distributions and distinct misclassification costs.[19]

5.1.3 K –fold cross

K-fold cross validation used in the field of machine learning which determines that at which rate a learning algorithm will predict the data; which not originally trained. It is In k-fold method, data is partitioned randomly into k groups. Using all the training set instances except those in the k^{th} group so, the classifier is trained k-1 times. Count number of errors; calculate mean error over all k test sets.

If appropriate value of the k is chosen then this approach is very effective. It is less extravagant of data than test set cross validation, and less costly than leave-one-out cross validation. K-fold gives the best estimate of cross validation error. [18]

5.1.4 Precision

Precision provides measurements when experiment is performed under unchanged conditions and gives the same results. Precision is the proportion of all the positive results (both true positives and false positives) and the true positives [17].

$$\Pr ecision = \frac{NumberofTP}{(NumberofTP+NumberofFP)} \qquad (5.3)$$

5.1.5 Leave-one-out cross-validation

Leave one out is a cross validation technique which is used to measure the accuracy of a classifier. This method is used in the research paper [4] for checking the performance of SVM for distinguishing the patients data from normal persons data by making two classes, one for Alzheimer diseased patients and other for normal persons.

The process has the following steps.

1. The initial data is partitioned randomly into k number of samples having equal size
2. Performed training and testing for k number of times.
3. In the first iteration, some data is kept as test set and the remaining is used to train the classifier.
4. The second set is kept as a test set and the remaining is used to train the model in the second iteration.
5. When mean is obtained in i^{th} iteration then the i^{th} sample reserved as a test set and the remaining samples are used as training set
6. Each sample is used to train and test the model for same number of time
7. At the end, for checking the accuracy of overall number of accurate classification is from the k iteration is separated by the total number of tuples in initial data. No data loss.

As training process is repeated for large number of times so this technique is usually very expensive from a computational point of view.

5.1.6 T-test

T-test is a statistical significance test, simply use to find the difference between averages of two classes. This performance measure is used in [5] to find the difference between mean values of MDD (major depressive disorder) group and Normal persons group.

The process has the following steps:

1. First, assume a null hypothesis: H_0
2. By setting the value for level of significance 1 and degree of freedom 2 we can find a critical region limit from the t-table given in any standard statistical book.

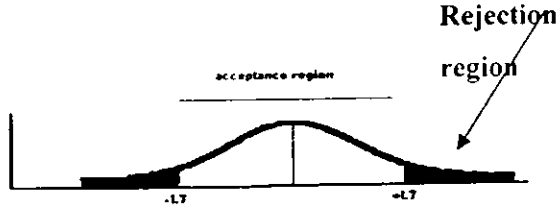


Figure 5.2: t-distribution show the acceptance and rejection region

1. Level of significance: also known as confidence level is set by the user and measure in percentage. Usually 1% and 5% are commonly use. Actually, this is the value of accuracy of the model that user wants that classifier must show. For example, a manager of an organization wants that a system must give him approx. 90% benefit, in this case the level of confidence is 0.09.

2.Degree of freedom: if we have n sample then we will have n-1 degree of freedom. Simply we can say this is the no of choices of randomly selecting any number of samples. For example, if we want to select 5 samples from a data set then we have 4 degree of freedom because till selecting 4^{th} sample we have choice for sample data but for the 5^{th} sample we have no more choice as we are restricted to select all the remaining point as a

last sample. Level of significance and degree of freedom are arranged in t-distribution table as shown in figure 5.3.

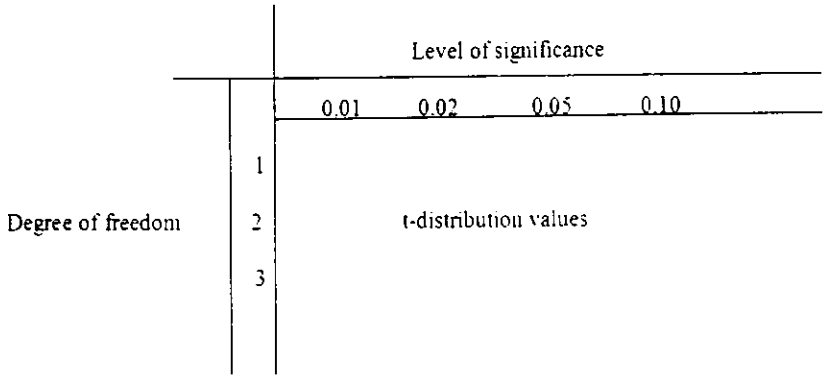


Figure 5.3: t- test value

By using the formula in equation (5.6), we find value for t-statistic

$$t = \frac{err(M1) - err(M2)}{\sqrt{var(M1 - M2)/K}} \qquad (5.4)$$

Where,

err(M1)= error in first model
err(M2)= error in second model
Var(M1-M2)= variance difference between two models

If value of t-statistic lies within the acceptance region we accept null hypothesis otherwise reject it and make a new hypothesis accordingly.

It's Simple, easy to compute. Just calculate difference between two class averages, for more than two classes we should implement pair wise test in iterations.

5.1.7 Metric for Convergence

Equation 5.5 computes the distance measures between the non-dominated points find by MOGA and global pareto-optimal front, so lower d value symbolizes superior convergence ability. Suppose P* contains the list of the target set of points on the Pareto-optimal front and let 'f' represents the ultimate non-dominated set attained by an NSGAII, di is the smallest distance between 'i ' to 'f'.

$$d_i = \min_{j=1}^{|P^*|} \sqrt{\sum_{K=1}^{M} \left(\frac{f_k(i) - f_k(j)}{f_k^{max} - f_k^{min}} \right)^2} \qquad (5.5)$$

Here , f_i^{max} and f_i^{min} are the possible maximum and minimum function values of k-th objective function in mention points. By taking the average of all the distances of the points in F convergence metric can be computed. [20]

5.1.8 Pareto-optimal front

Ideally, we expects from MOGA to converge to global pareto-optimal front. If MOGA converges to global pareto-optimal front .We divides objective space into grid according to the global pareto-optimal front, then one point in each grid represents the best possible diversity value. This is termed as diversity matrix 1. [20]

5.1.9 Converged front

But if the algorithm isn't able to converge to the global PO front then the above metric will not be able to measure the diversity of non-dominated solutions produced by the MOEA. In such cases where an algorithm is stuck in a local PO front, Diversity matrix should be calculated based on the actual converged front instead of global PO front. We will call this diversity metric (obtained by splitting the converged PO region into grids) diversity metric2.To compare two algorithms (for diversity) the use of diversity metric2 should be preferred because even if one of them has converged to true PO front and the other hasn't,

the diversity metric2 of the former will be almost equal to its diversity metric1 value, which would not be the case with latter. [20]

5.2 Experimental Setup

5.2.1 Datasets

To demonstrate the performance of the proposed method, the data sets have been taken from UCI [15].

5.2.1.1 Heart Disease (Stat Log Project)

Table 5.1: Heart Disease Data Set Attributes

Data Set Attributes	
Total Attributes	13
Missing values	No
Instances	270
Attribute types	Real, Ordered, Nominal ,Binary
Variables to be predicted	1)Absence2)presence(of disease)

Heart disease database finally contains thirteen attributes these attributes has been extracted from total 75 attributes.Labels for these thirteen attributes are This database contains 13 attributes (which have been extracted from a larger set of 75) The data set contains the attributes are that how much grown old patient , gender ,type of the chest pain (4 values) , blood pressure(resting), serum cholesterol in mg/dl fasting blood sugar > 120 mg/dl , results of resting electrocardiographic (values 0,1,2) , utmost heart rate attained , angina persuade by exercise,old peak = S``T depression persuaded by exercise comparative to rest , the gradient of the peak exercise ST segment, total major vesseles (0-3) that are colored by fluoroscopy , thal; 6 = permanent fault; 3 = normal ;7 = reversible defect.

5.2.1.2 Diabetes Database

Table 5.2: Diabetes Data Set Attributes

Data Set Attributes	
Total Attributes	8+1(class)

Missing values	Yes
Instances	768
Attribute types	Numeric

Title of the database is Pima Indians Diabetes Database, Sources from where the data set is obtained are

(a) Original owner of the database are National Institute of Diabetes and Digestive and Kidney Diseases

(b) Benefactor of database are Vincent Sigillito (vgs@aplcen.apl.jhu.edu)

Research Center, RMI Group Leade.Applied Physics Laboratory .The Johns Hopkins University,Johns Hopkins Road , Laurel, MD 20707(301) 953-6231

The indicative, '0' or '1' valued variable indicate either patients shows signs of diabetes as per to world health organization criteria (that is as a minimum 200 mg/plasma post load in two hour at any survey examination). The population lives near Arizona , USA and Phoenix: Predicion made by algorithm was in the range of 0-1. This was transformed into a binary decision using a cutoff of 0.448. All patients includes in database are women of minimum 21 year old.

Attributes label of diabetes database are frequency of pregnancy,concentration of the plasma glucose is tested in 2 hour using oral glucose tolerance test. 3. blood pressure (mm Hg)(Diastolic),skin fold thickness of the triceps, serum insulin ,index f the body mass, and pedigree function of diabetes ,age in years of patient, and last is the class variable for diagnosing sign of diabetes.

5.2.1.3 Hepatitis disease

Table 5.3: Hepatitis Data Set Attributes

Data Set Attributes	
Number of Instances	155
Number of attributes	20
Missing Attribute	Yes
Classes	2(Die or Live)

5.2.2 Performance measures used

5.2.2.1 2x2 Contingency table

Table 5.4 : 2x2 Contingency table for subset of original set of features (Heart Data Set)

True Labels	Estimated Labels		Totals
	Presence disease	Absence Disease	
Present Disease	(TP)	(FP)	105
Absence Disease	(FN)	(TN)	84
Totals	116	73	189

TP= (True positive) is the case when test was positive and the prediction on the test case was also positive. FP = (False positive) when test is positive and prediction is made it is negative.(FN =False negative)is the case when true label is test is negative but predictive label is it is negative.(TN = True negative)When true label and predictive label both are negative.

Table 5.5: 2x2 Contingency table for subset of original set of features (Pima Indians Diabetes Database)

True Labels	Estimated Labels		Totals
	Presence disease	Absence Disease	
Present Disease	314	36	350
Absence Disease	93	95	188
Totals	407	131	538

5.2.2.2 Sensitivity Calculation

Sensitivity is also refer as true positive rate as it is the measure of the proportion of true positive rate, formula for sensitivity calculation is

$$Sensitivity = TP/(TP + FN) \tag{5.6}$$

5.2.2.3 Specificity Calculation

Specificity is referring to as negative hit rate as it measures the proportion of True negative rate. Formula to calculate specificity in term of probability is Specificity or True negative rate in term of probability = TN/TN+FP

5.2.2.4 Average Overall hit rate calculation

For binary classification performance measure sensitivity specificity and average overall hit rate is used. But for multi class classification problems only average over all hit rate is used.

Overall Average =(TP+TN)/(TN+FP+FN+FN) (5.6)

5.2.3 Parameters Setting

Detail parameter settings of the experiments are given in table 5.3.2

5.2.4 Training and Testing data set

Gendat function is used to generate data randomly for testing and training 70% data is used randomly for training and 30% for testing. Random selection of the instances follows the prior probability of the class. So the estimation of the sample that it would be selected from the particular class is equal to P*N, where P is the prior probability of the class. And N is the percentage of the data for training.

5.2.5 Statistical analysis

From 10 runs the mean and standard deviation of the AUC and estimated error are listed in the table 5.4.1.

5.3 Results and Discussions

5.3.1 Classification using complete set of attributes (Scenario 1)

Following table summarizes the classification performance of SVM using polynomial kernel of order 2.Table shows the result of the scenario when complete set of the features

or clinical test are used to diagnose the disease. To handle the biasness of randomness experiment is repeated 11 times and its average and standard deviation is considered to report the rate of change in results of all iterations.

Table 5.6: Results using all the set of feature for SVM (heart data set)

Iteration #	Estimated error	AUC
1	0.2963	0.9146
2	0.3210	0.9121
3	0.1975	0.9364
4	0.3580	0.8916
5	0.2963	0.9120
6	0.2469	0.9409
7	0.2469	0.9319
8	0.3951	0.8827
9	0.2716	0.9319
10	0.3333	0.8953
11	0.3210	0.9194
Average	**0.298536**	**0.915345**

5.3.1.1 Estimated Error or Accuracy calculation

Estimated error is the probability of error and accuracy =1-probability of error (5.7)

5.3.1.2 AUC calculation

Area under the ROC curve (this is an error and not a performance!). For multi class problems this is the weighted average (by class priors) of the one-against-rest contributions of the classes

Table 5.7 : Results using all the set of feature for SVM(diabetes data set)

Iteration number	Estimated error	AUC
1	0.2478	0.8592
2	0.2696	0.8537
3	0.2609	0.8574
4	0.2565	0.8591
5	0.2826	0.8431
6	0.2261	0.8681
7	0.2435	0.8606
8	0.2478	0.8569
9	0.2609	0.8564
10	0.2435	0.8667
11	0.2609	0.8577
Average	0.254555	0.858082
Standard deviation	0.015109	0.006571

5.3.2 Classification using Selected Attributes (Scenario 2)

Following table report the performance of SVM classifier when wrapper based feature selection technique with Multi objective Genetic Algorithm is used. Table also shows the comparison of our MOGA techniques with other state of art simple GA technique for SVM (RBF).

Table 5.8: parameter setting of three feature selection methods including MOGA

Parameters	SVM(polynomial,2) +MOGA	SVM(RBF)+SGA	NN+SGA
Population size	50	50	50
Total generations	20	20	20
Cross over rate	0.8	0.8	0.8
Mutation rate	0.05	0.02	0.05
Selection method	Roulette Wheel	Roulette Wheel	Roulette Wheel

Table 5.9: Comparison of MOGA and SGA using SVM

GA with SVM	Heart	13	5	89.30	90.50	89.90
	Diabetes	8	3	76.10	83.50	79.30
MOGA with SVM	Heart	13	4	80.2	83.6	81.5
	Diabetes	8	3	77.1	50.5	70.6
Proposed MOGA with SVM	Heart	13	4	91.2	90.5	90.9
	Diabetes	8	3	81.7	80.4	81.1
GA with NN	Heart	13	5	Null	Null	86.5
	Hepatitis	19	7	Null	Null	88.5
MOGA with NN	Heart	13	4	Null	Null	88.0
	Hepatitis	19	6	Null	Null	94.0
Proposed MOGA with NN	Heart	13	4	Null	Null	91.4
	Hepatitis	19	6	Null	Null	93.2

5.4 Discussion

5.4.1 Classification using All Features

Using all the features as input to the classifier result of 11 iterations shows the average estimated error that is 0.298536 it means 30% error and 70% accuracy. And standard deviation 0.05 shows that change in accuracy due to randomness is less. So less standard deviation depicts less biasness due to randomness.

5.4.2 Classification using Selected Features

In case of selected features when MOGA with SVM using polynomial of order 2 proportion of true positive rate is 80% and true negative hit rate is 83% and average over all hit rate is 88%.and features reduced from 13 to 4 it out perform as compare to the

results of experiments when all the features were used. And this increase in performance is 8% in case of accuracy.

Table 5.4.2.0summarizes the results of the experiments using GA based approach for SVM classifier using RBF kernel and proposed techniques results using MOGA for SVM .Results shows proposed technique performs very well under all performance parameters especially it performs excellent in the case of average true positive hit rate. For the binary class classification accuracy can be demonstrated through average positive hit rate (Sensitivity),average negative hit rate(Specificity)and average over all hit rate. Whereas, In case of multiclass classification only average overall hit rate can be consider. Data set used by us was the case of binary classification. Hence we used all three parameters to determine the accuracy of the proposed technique.

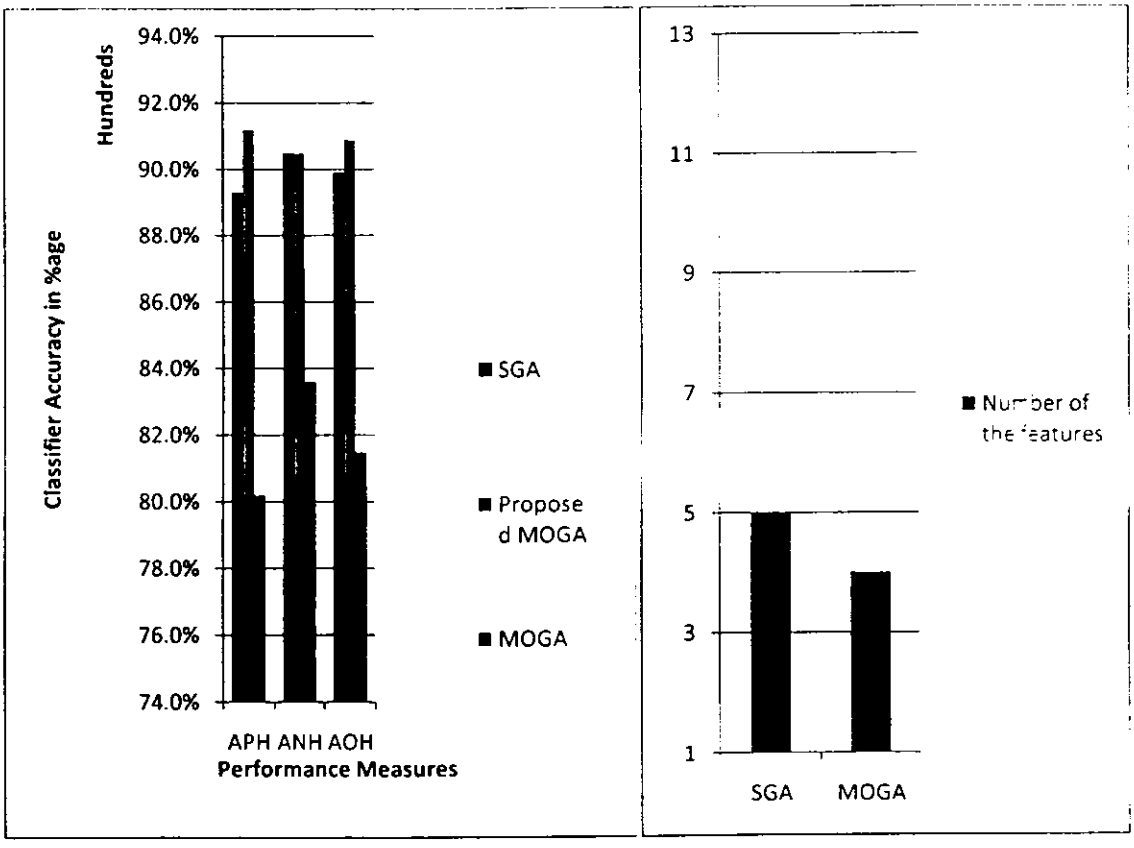
Accuracy of the classifier using heart disease data set improved when featured selected by MOGA is used. As proportion of true positive rate is 91.2% and proportion of true negative rate is 90.5%.and average over all hit rate is 90.9% with feature reduction 4 from total of 13.as compare to the Simple GA for SVM that shows the results under same parameter as 89.3%, 90.5%and 89.9% respectively. And feature reduces to 5 out of 13.

Accuracy of the classifier under same experimental setup for diabetes data set is as 81.7%(Average positive hit rate),80.4%(Average negative hit rate),81.1% (Average overall hit) and features reduces from 8 to 3 using proposed MOGA based feature selection for SVM. Comparative technique of feature selection for SVM using SGA shows results as 75.1%(Average positive hit rate),83.5%(Average negative hit rate),79.3%(Average overall hit rate) with feature reduction 3 out of 8.So results shows although the number of the features remain same as to the state of the art technique but even though accuracy improved that's mean MOGA is efficient in selecting more fit features then simple genetic algorithm.

For the comparison purpose third study used by us is used as benchmark in many literature studies as well[11].They used simple GA for feature selection for Dist all NN (Neural Network).GA parameter setting is mentioned in Table 5.4.2.Table 5.4.2.1shows the

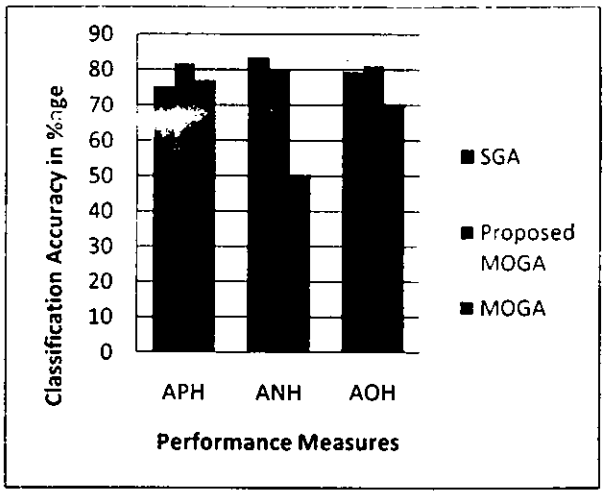
comparison of results of proposed technique with the simple GA for NN technique. Results shows proposed technique perform very well.

Proposed technique out performs in both objectives accuracy and feature reduction as NN reduces the feature to 7 with accuracy of 86.5%.for heart disease data set. And for hepatitis accuracy achieved by NN is 88.3% and features reduce to 9 out of 13.Whereas proposed technique achieve accuracy of 91.4% with 4 features out of 13 using heart disease dataset. And when dataset is replaced by hepatitis results obtained are 93.2% accuracy with 6 features out of 9. Proposed method is giving better accuracy with even more reduction in features. Figure 5.5 and 5.4 shows the experimental results details with the help of bar charts.

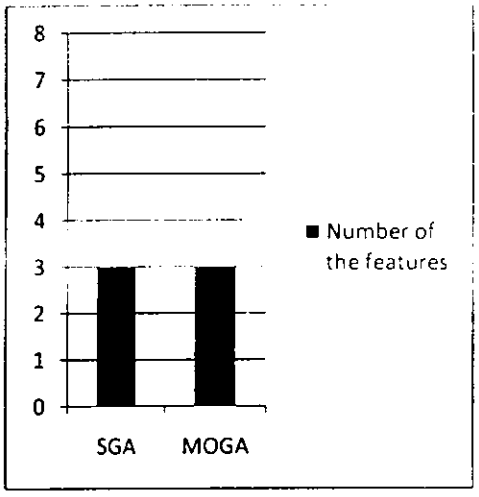


(a): Accuracy comparison for SVM(heart)

(b): Number of feature reduced comparison for SVM (Heart)

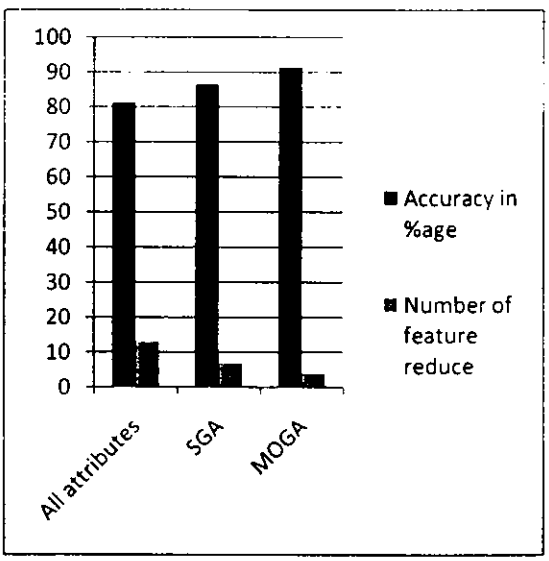


(c) : Accuracy comparison for SVM(Diabetes)

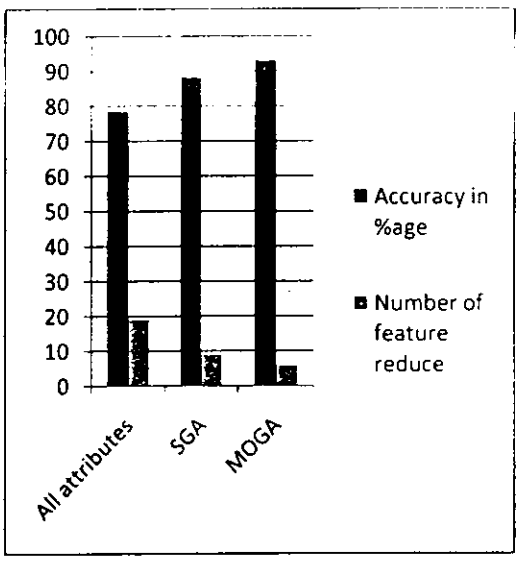


(d) : Number of feature reduced comparison for SVM (Diabetes)

Figure: 5.4 : Accuracy comparison



(a): comparison of SGA based Feature selection technique for NN with MOGA based for SVM (Heart Disease)



(b): comparison of SGA based Feature selection technique for NN with MOGA based for SVM (Hepatitis Disease)

Figure 5.5 :Comparison of SVM and Neural Network

5.5 Summary

This chapter presents the details of experimental setup and analysis on the output of the experiments. Results shows that proposed technique is more efficient in selecting optimal features with better convergence rate as compare to the any present good technique. It is prove that with same control parameters for all comparisons their results are comparable on the basis of accuracy and features reduction, opposition based MOGA performs better than SGA and simple MOGA in all scenarios.SGA performs better than simple MOGA.

It is very simpler and easier in MOGA to formulate the unbiased Objective function. Opposition based MOGA is effective for feature selection for biomedical databases as it achieves better convergence and more optimal solutions than baseline methods.

This work can be further extended by introducing more objectives , E.g. clinical cost can be considered as third objective ,Moreover Constraint can be define for the objectives to make preferences flexible according to the requirement of the decision makers.

Chapter 6

Conclusion & Future Work

This chapter concludes the whole report by summarizing its main points. pointing to the several new techniques and problems that should be explored to make feature selection using multi-objective genetic algorithm extends more robustly and efficiently. Chapter 6 sketches the importance of proposed technique by enlightening its applications in other fields.

6.1 Conclusions

So from the whole discussion it is prove that feature subset selection is not a single objective problem but it is a multi objective problem. In Biomedical domain these multiple objectives are conflicting in nature with each other and not the ideal case. So there is need to identify all alternative solutions. Single Objective optimization can identify just one best solution in a single run while multi objective optimization can identify all possible optimal solutions in a single run. In other words we can say it can identify the whole trade off surface. To Lump multiple objectives into one fitness function can produce biased results. In Single Objective optimization it is very critical to formulate the objective functions through complex mathematical tools .And if it is not formulated properly it can detract the whole search from being approaching to true optimal solution. Moreover it is very simpler and easier in MOGA to formulate the Objective function. From the results of the experiments it is being observed that opposition based MOGA approach used by us is effective for feature selection for biomedical databases.

6.2 Future Work

6.2.1 Augment Objectives

This work can be further extended by boosting up more objectives, Supplementary objectives from the domain or case study undertaken will enhance the worth of the proposed opposition

based MOGA .Because with the increase in objective vector length, effectiveness of MOGA can be viewed more clearly. For example, clinical test cost can be considered as an additional objective. More objectives can be explored according to the application requirement.

6.2.2 Use of Related Techniques

Other heuristics with multi objective handling can be tried, some other possible heuristics are mentioned here.

Swarm algorithms, including:

- Ant colony optimization
- Bee's algorithm
- Particle swarm optimization

Other population-based met heuristic methods:

- Differential evolution
- Firefly algorithm
- Invasive weed optimization algorithm
- Harmony search
- Gaussian adaptation

6.2.3 Multi-objective Clustering Approach

Feature selection for unsupervised leaning in this area yet very less work has done. This field can be explored more using Multi objective for clustering techniques.

6.3 Applications

proposed technique covers all those application areas that are covered by feature subset selection technique too as it is enhancement step of it. It covers not only the case studies of this report but all the other biomedical data too. Not only biomedical domain it can be used any where data

mining is done for relevant knowledge discovery is required as pre-requisite step. Following are different application other than biomedical domain.

6.3.1 OCR Recognition

Character Recognition systems[39] deals with hundreds of the features and it is not unusual Feature subset selection when applies to practical applications such as character recognition appears to be a present s a multi-criterion optimization problem, For example accuracy of classification and number of features [40]. Proposed approach offer a particularly attractive approach for this kind of problems since they are generally quite effective for rapid global search of large, complex and nonlinear spaces.

6.3.2Diagnose Business Crisis

Business crisis[37,38] means bank corruption, shutdowns, failures, distress and bond default. With the advancement of Artificial intelligence and database technologies, data mining techniques can be used to predict the business failure before alarming situations occur. Such system can be very helpful for the decision maker to evaluate and select the firm with which collaboration is beneficial. Usually the financial crises are of three types that can be read from .Bank carupcy prediction is also a multi criterion problem that can be best solved by proposed technique.

6.3.3 Rapid Image Information Mining

In disaster situations there is required to retrieve image information in real or near real time to out of the vast amount of the data that is coming from the various remote sensors. For example, in flooding disasters classes may have been trained dealing with roadways, forest and undamaged buildings. Though, after a disaster, it is now necessary to classify flooded roadways, damaged buildings, and knocked down forests from limited training data. Thus, there is a need for rapid IIM (RIIM). Generally Image Information mining techniques generates large amounts

of the features that are computationally very expensive and incompetent before the useful information is discovered. So, it is very necessary to extract the relevant features that are close to the hypothesis space.[36]

6.3.4 Evaluation of Modeling Music Similarity

Due to the advancement in digital storage technology and large availability of the digital data and digital music there is a need to arrange the music in the order in the database so that when user queries for some music, his/her query should be entertained very quickly and accurately. Present applications that manage music, utilizes the textual Meta information. To be able to search through different available digital music first need to convert raw data to the some level of information granularity. From Information granularity we here meant that data contains only important information that involves feature selection. So here proposed technique is applicable [35].

6.3.5 Emotion Recognition using Feature Subset Selection

Study [34] of human's emotions in human computer interaction is currently a growing research area. Research is being performed on automatic emotion recognition that is speech and facial gesture recognition. Using large set of parameters of speech studies are being made to analyze the results of the emotion recognition. Development of the effective systems for detecting and responding to the human system is a great challenge that involve as a part analyzing different multi model data resources. Proposed technique can be proof as a good application in researching relevant speech parameter in automatic speech recognition parameter as affective resources like effective databases provide a chance for training affective application for any affective classifier model.

This chapter sums up the conclusion and the applications of the proposed techniques. Due to improved results the effectiveness of the method can be tested in other field as well where the multi objective optimization is require and there is need to explore the huge search space. This approach can be further extended to any multi criterion optimization problem.

References

[1] Kononenko I. "Inductive and Bayesian learning in medical diagnosis". Appl Artif Intell 1993;7(4):317–37.

[2] Wolberg W-H, Street W-N, Mangasarian OL. "Machine learning techniques to diagnose breast cancer from fine-needle aspirates". Cancer Lett 1994;77:163–71.

[3] Wolberg W-H, Street W-N, Mangasarian OL. "Image analysis and machine learning applied to breast cancer diagnosis and prognosis". Anal Quant Cytol Histol 1995;17(2):77–87.

[4] Liu, H. and Setiono, R. 'Incremental feature selection .Applied Intelligence, 1998a. 9(3):217–230.

[5] Kononenko, I. 'Estimating Attributes: Analysis and Extensions of Relief'. In Proc. of the European Conference on Machine Learning, 1994. pages 171–182, Vienna.

[6] Lei Yu leiyu ,Huan Liu, Feature Selection for High-Dimensional Data: "A Fast Correlation-Based Filter Solution".2003.

[7] David B Sakalak,Prototype, " feature selection by sampling and Random Mutation hill climbing algoritham",Febrauary 1994.

[8] M. A. Tahir, A. Bouridane, F. Kurugollu, and A. Amira,"Feature Selection using Tabu Search for Improving the Classification Rate of Prostate Needle Biopsies", Proceedings of the 17th International Conference on Pattern Recognition (ICPR'04)2004.

[9] Jihoon yang and Vasant Honavar Iova State University, "Feature subset selection using GA". : Intelligent Systems and their Applications, IEEE Mar/April 1998

[10] Cheng-Lung Huang , Chieh-Jen Wang ,"A GA based feature selection and Parameter optimization technique for support vector machine", Expert Systems with Applications 31 ,2006 .231–240.

[11] Mr. Kashif Waqas Dr. Rauf Baig1and Dr. Shahid Ali,"Feature Subset Selection Using Multi-Objective Genetic Algorithms",INMIC 2009 IEEE 13 International Conference. 2009.

[12] Jianjiang Lu , Tianzhong Zhao, Yafei Zhang, "Feature selection based-on genetic algorithm for image annotation", Knowledge-Based Systems 21, 2008 .887–891

[13] Sérgio Francisco da Silva , Marcela Xavier Ribeiro , João do E.S. Batista Neto a,Caetano Traina-Jr. a, Agma J.M. Traina, Ranking evaluation functions to improve genetic feature selection in content-based image retrieval of mammograms, 22nd IEEE International Symposium on Computer-Based Medical Systems, CBMS, 2009. pp. 1–8.

[14] Sarojini Balakrishnan,Feature Selection Using FCBF in Type II Diabetes Databases , Special Issue of the International Journal of the Computer, the Internet and Management, Vol.17 No. SP1, March, 2009

[15] UCI database Hettich, Blake, & Merz, 1998.

[16] Abdullah Konak, David W. Coit, Alice E. Smith: Multi-objective optimization using genetic algorithms: A tutorial, Reliability Engineering and System Safety 91 (2006) 992–1007

[17] http://en.wikipedia.org/wiki/Accuracy_and_precision

[18] http://everything2.com/title/k-fold+cross+validation

[19] Ross Bettinger ,SAS Institute: Cost-Sensitive Classifier Selection Using the ROC Convex Hull Method

[20] K. Deb and Sachin Jain, Running PerformanceMetrics for Evolutionary Multi-Objective Optimization, Kanpur Genetic Algorithm Labratory (KanGAL), Report No.2002004

[21] Deb, K. "Multi-Objective Optimization Using Evolutionary Algorithms".

Reading, John Wiley & Sons, Ltd, Reprinted April 2002.

[22]Kalyanmoy deb,"Multiobjective Optimization using Evolutionary algorithm",2002,pp 1-24.

[23] PSO Tutorial: http://www.swarmintelligence.org tutorials.php

[24] Sarojini Balakrishnan , Ramaraj Narayanaswamy "Feature Selection Using FCBF in Type II Diabetes Databases". Special Issue of the International Journal of the Computer, the Internet and Management, Vol.17 No. SP1, March, 2009.

[25] Sumit Bhatia, Praveen Prakash, and G.N. Pillai," SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features",Proceeding of the world congress on Engineering and Computer Science 2008.

[26] Asha Gowda Karegowda, M.A.Jayaram, A.S. Manjunath ."Feature Subset Selection Problem using Wrapper Approach in Supervised Learning " ,International Journal of Computer Applications ,2010

[27] Wan-sheng Ke,Yuchi hwang,eugene Lin."Pharmacogenomics of drug effcacy in the interferon treatment of chronic hepatitis C using classifcation algorithms" Advances and Applications in Bioinformatics and Chemistry,2010.

[28] Kemal Polat , Salih Gu "nes" Medical decision support system based on artificial immune recognition immune system (AIRS), fuzzy weighted pre-processing and feature selection" Expert Systems with Applications 33 (2007) 484–490.

[29] Li Ming, Lu YuMing, Zhang YongLiang "A New Multi-Objective Genetic Algorithm for Feature Subset Selection in Fatigue Fracture Image Identification" . JOURNAL OF COMPUTERS, VOL. 5, NO. 7, JULY 2010.

[30] Yue Huang , Paul McCullagh , Norman Black , Roy Harper "Feature selection and classification model construction on type 2 diabetic patients' data", Artificial Intelligence in Medicine (2007) 41, 251—262

[31] Lilla B" or " oczky, Member, IEEE, Luyin Zhao, and K. P. Lee" Feature Subset Selection for Improving the Performance of False Positive Reduction in Lung Nodule CAD". IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 10, NO. 3, JULY 2006

[32] Jiexun Li, Hua Su, Hsinchun Chen, Fellow, IEEE, and Bernard W. Futscher," Optimal Search-Based Gene Subset Selection for Gene Array Cancer Classification", IEEE TRANSACTIONS ON INFORMATION TECHNOLOGY IN BIOMEDICINE, VOL. 11, NO. 4, JULY 2007.

[33] Raymer, M. L., Punch, W. F., Goodman, E. D., Kuhn, L.A., And Jain, A. K. "Dimensionality Reduction Using Genetic Algorithms", IEEE Transactions On Evolutionary Computation, Vol. 4, No. 2, July 2000.

[34] Aitor Álvarez, Idoia Cearreta, Juan Miguel López, Andoni Arruti, Elena Lazkano, Basilio Sierra, and Nestor Garay," A Comparison Using Different Speech Parameters in the Automatic Emotion Recognition Using Feature Subset Selection Based on Evolutionary Algorithms" V. Matoušek and P. Mautner (Eds.): TSD 2007, LNAI 4629, pp. 423–430, 2007.

[35] D.N. Sotiropoulos, A.S. Lampropoulos, and G.A. Tsihrintzis" Evaluation of Modeling Music Similarity Perception Via Feature Subset Selection", Conati, K. McCoy, and G. Paliouras (Eds.): UM 2007, LNAI 4511, pp. 288–297, 2007.

[36] Surya S. Durbha, Member, IEEE, Roger L. King, Senior Member, IEEE,and Nicolas H. Younan, Senior Member, IEEE," Wrapper-Based Feature Subset Selection for Rapid Image Information Mining", IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, VOL. 7, NO. 1, JANAURY 2010.

[37] Chih-Fong Tsai," Feature selection in bankruptcy prediction", Knowledge-Based Systems 22 (2009) 120–127.

[38] Liang-Hsuan Chen , Huey-Der Hsiao," Feature selection to diagnose a business crisis by using a realGA-based support vector machine: An empirical study", Expert Systems with Applications 35 (2008) 1145–1155.

[39] M. Soryani, and N. Rafat" Application of Genetic Algorithms to Feature Subset Selection in a Farsi OCR ",world academy of science Engineering and technology 18 ,2006.

[40] Kudo M, Sklansky J. , "Comparison of Algorithms that Select Features for Pattern Classifiers" Pattern Recognition, Vol.33, pp.25–41, 2000.

[41] M. Amjad Iqbal,, Naveed Kazim Khan2, M Arfan Jaffar, M. Ramzan, A. Rauf Baig5," Opposition based Genetic Algorithm with Cauchy Mutation for Function Optimization", 978-1-4244-5943-8/10/$26.00 ©2010 IEEE

[42] William S Noble," What is a support vector machine?", Natural Biotechnology volume 24 number 12 december 2006.

[43]Lukas ," Least Squares Support Vector Machines Classification Applied to Brain Tumor Recognition using Magnetic Resonance Spectroscopy" December 2003.

[44] http://www.talkorigins.org/faqs/genalg/genalg.html

[45] http://www.scribd.com/doc/50858466/3/Search-Techniques

[46] D.T. Tsahalis, S.K. Katsikas, D.A. Manolas ," A genetic algorithm for optimal positioning of actuators in active noise control: results from the ASANCA project", Aircraft Engineering and Aerospace Technology, Vol 72 Iss: 3, pp(252 – 258).