



**“Differential Evolution Based Multi-objective Algorithm for Data Classification”**

A Dissertation submitted

By

**Sonia**

**(Regd. No. 11310337)**

To

**Department of Computer Science and Engineering**

In Fulfillment of the Requirement for the

Award of the Degree of

**Master of Technology in**

**Computer Science and Engineering**

**Under the guidance of**

**Mr. Baljit Singh Saini**

**(Asst. Prof., UID-15359)**

**(May 2015)**

School of: Computer Sc. & Engg.


**DISSERTATION TOPIC APPROVAL PERFORMA**

Name of the Student: Sonia Registration No. 11310337  
 Batch: 2013-2015 Roll No. RK2305B32  
 Session: Dec-2014 Parent Section: K2305  
 Details of Supervisor:  
 Name: Dr. Sushil Kumar Designation: A.P.  
 U.ID: 18923 Qualification: Ph.D.  
 Research Experience: 4.5yrs.

SPECIALIZATION AREA: Database (pick from list of provided specialization areas by DAA)

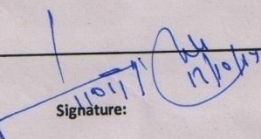
**PROPOSED TOPICS**

1. Applications of metaheuristic <sup>Algorithms</sup> for Database classification
2. Methodology development for reducing complexity of link list data structure
3. Development of Non-blocking data structure based on Nim's.

Signature of Supervisor  


**PAC Remarks:**  
Topic 1 is approved.

**APPROVAL OF PAC CHAIRPERSON:**

Signature: 

Date:

\*Supervisor should finally encircle one topic out of three proposed topics and put up for a approval before Project Approval Committee (PAC)  
 \*Original copy of this format after PAC approval will be retained by the student and must be attached in the Project/Dissertation final report.  
 \*One copy to be submitted to Supervisor.

## **ABSTRACT**

Data Classification is the process of organizing data into predefined categories so that it can be easily found and accessed. It makes the retrieval of data and data management more effective and efficient. Decision Trees, Support Vector Machines, Naïve Bayes Classifier are the basic methods for data classification. But we used Differential Evolution Optimization Algorithm for data classification, where two main objectives of our algorithm are to increase efficiency and decrease the time complexity than the Genetic algorithm approach for the same. Differential Evolution optimizes the problem by iteratively improving the candidate solution accordingly given measure of the quality. The proposed algorithm is multi-objective and objectives are to maximize the efficiency and to minimize the complexity. And the Algorithm we developed, using Differential Evolution algorithm is working properly and two main objectives are being fulfilled completely.

## **ACKNOWLEDGEMENT**

I would like to take this opportunity to express my deep sense of gratitude to all who helped me directly or indirectly during thesis work.

Firstly, I would like to thank my supervisor, Mr. Baljit Singh Saini, Assistant Professor (UID 15359) for being great mentor and best adviser I could ever have. His advice, encouragement and critics are sources of innovative ideas, inspiration and cause behind the successful completion of this dissertation. I am highly obliged to all faculty members of computer science and engineering department for their support and encouragement.

I would like to express my sincere appreciation and gratitude towards my friends for their encouragement, consistent support and invaluable suggestions at the time I needed the most.

Sonia

Reg. no. 11310337

## DECLARATION

I hereby declare that dissertation entitled **“Differential Evolution Based Multi-objective Algorithm for Data Classification”** is my own work conducted under the supervision of Mr. Baljit Singh Saini.

I further declare that to best of my knowledge this dissertation does not contain any of my work, which has been submitted for the award of any degree either in this university or any other institute.

Date: \_\_\_\_\_

Investigator: Sonia

Registration no: 11310337

## CERTIFICATE

This to certify that **Sonia** has completed M.Tech dissertation titled “**Differential Evolution Based Multi-objective Algorithm for Data Classification**” under my guidance and supervision. To the best of my knowledge, the present work is the result of his original investigation and study . No part of the dissertation has ever been submitted for any other degree or diploma.

This dissertation is fit for the submission and the partial fulfillment of the conditions for the award of M.Tech in Computer Science & Engineering.

Date:\_\_\_\_\_

**Baljit Singh Saini**

UID 15359

## TABLE OF CONTENT

<b>Chapter 1 Introduction.....</b>	<b>1</b>
1.1 Data Classification.....	1
1.1.1 A General Problem to solve data classification.....	2
1.1.2 Confusion Matrix.....	3
1.2 Soft Computing.....	4
1.2.1 Components.....	4
1.2.2 Evolutionary Algorithms.....	5
<b>Chapter 2 Review of Literature.....</b>	<b>7</b>
<b>Chapter 3 Present work.....</b>	<b>13</b>
3.1 Problem formulation.....	13
3.2 Objectives.....	14
3.3 Research methodology.....	15
<b>Chapter 4 Result and discussion.....</b>	<b>17</b>
<b>Chapter 5 Future scope and Conclusion.....</b>	<b>30</b>
<b>Chapter 6 References.....</b>	<b>33</b>

## LIST OF TABLES

Table 1. Confusion Matrix

3



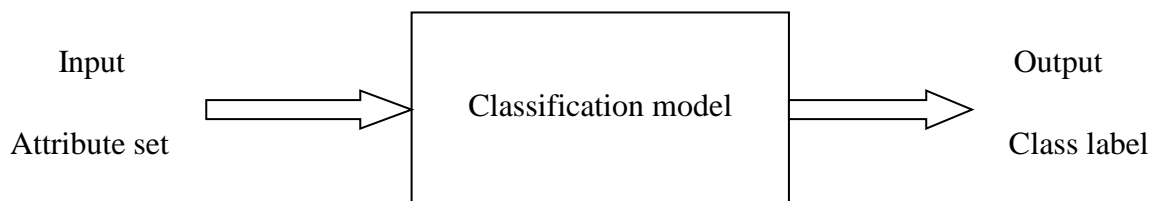
## LIST OF FIGURES

Figure 1.1. Classification Model	1
Figure 1.2. General Approach for Building Classification Model	2
Figure 1.3. Flow chart for Differential Evolution Algorithm	13
Figure 3.1 Methodology Flow Chart	17
Figure 4.1 Matlab Starting GUI	20
Figure 4.2 Rosen brock saddle Datasets Classification by GA	21
Figure 4.3 3D view of data clusters in Red by GA	21
Figure 4.4 Best Cost Value Calculation on basis of data classification in GA	22
Figure 4.5 Vector Distribution graph in GA	22
Figure 4.6 Rosen Brock Saddle by DE Result	23
Figure 4.7 3D views of cluster By DE	23
Figure 4.8 Best cost By DE	24
Figure 4.9 Different Vector Distribution by DE	24
Figure 4.10 Multiple Iterative results by DE	25
Figure 4.11 Final Result of vector distribution	25
Figure 4.12 Graph between fitness value of best and mean by GA	26
Figure 4.13 Time generation between mean and best fitness	26
Figure 4.14 Best and mean time by DE	27
Figure 4.15 Time Generation by DE	27
Figure 4.16 Comparison between time complexity of both algorithms	28

### 1.1 Data Classification

In this new era of technologies, terabytes and petabytes of data is being generated daily. For organizations this increasing volume of data is becoming more and more complex problem as finding specific files, information from such large data and its impacts on daily business operations is very much difficult. Data classification is solution for this problem.

Data classification is the process of organizing data into predefined categories (as needed by different organizations), so that data can be effectively and efficiently used. Data classification system makes it easy to find and access data.



**Figure 1.1.** Classification is the task of organizing data into predefined classes.

Classification model is useful for following purposes:-

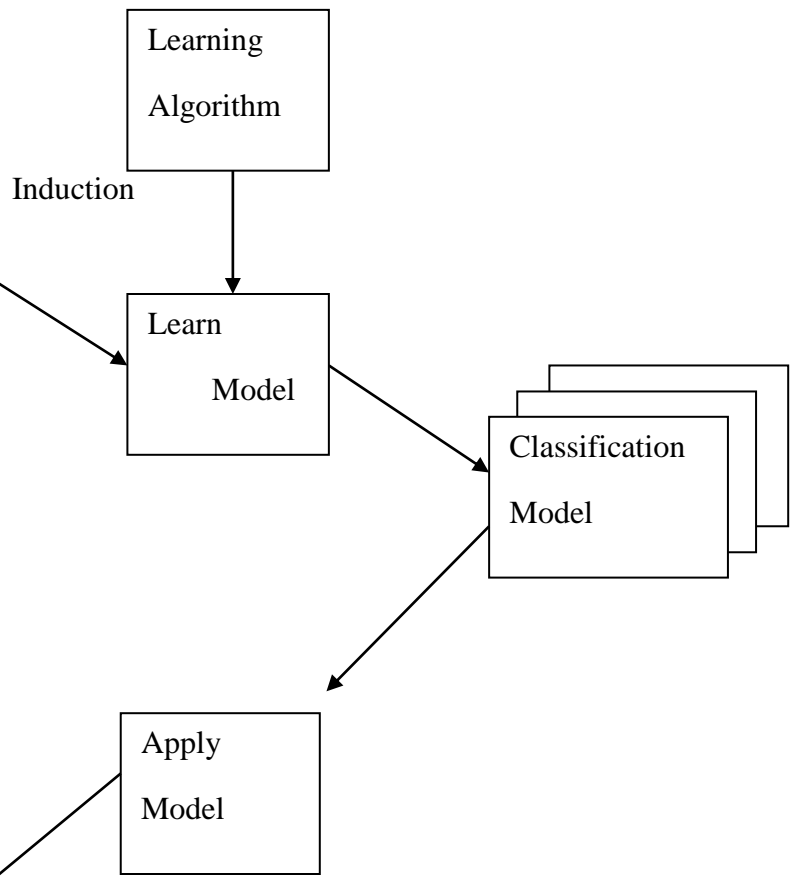
- 1. Descriptive Modeling** Classification model can serve as explanatory tool to distinguish data into different classes based on their attributes.
- 2. Predictive Modeling** Classification model can be used to predict class for unknown record. Classification techniques are best suitable for predicting data with binary and nominal class labels. It is less efficient for ordinal data. Superclass-subclass relationships are also not considered.

### 1.1.1 A General Approach to solve Classification Problem

A Classification Technique is an approach of building classification model from training data set.

Training Data Set

ID	Attrib1	Attrib2	Class label
1	Yes	High	Yes
2	Yes	Low	No
3	No	Medium	Yes
4	No	Medium	Yes
5	No	High	Yes
6	Yes	Low	Yes
7	No	High	No



Test Data Set

ID	Attrib1	Attrib2	Class label
1	Yes	Low	?
2	Yes	Medium	?
3	No	Low	?

**Figure 1.2.** General Approach for Building Classification Model

(a) Training data set consisting of records with their class label is provided, from which classification model is derived.

(b) Then that classification model is applied to Test Data without any class label to categorize that data into classes.

**1.1.2 Confusion Matrix** To evaluate the effectiveness of the classification model, confusion matrix is derived from its various results.

**Table 1.** Confusion Matrix

Actual	Predictive	
	Positive	Negative
Positive	True Positive(TP)	False Positive(FP)
Negative	False Negative(FN)	True Negative(TN)

(a) **True Positive** is the number of tuples selected from database by the model that belong to its class

(b) **True Negative** is the number of tuples not selected by the model that do not belong to its class.

(c) **False Positive** is the number of tuples selected by the model that do not belong to its class.

(d) **False Negative** is the number of tuples not selected by the model that belong to its class.

Performance can be measured by **accuracy**(rate of true predictions) or **error rate**(rate of false predictions)

$$\text{Accuracy} = (\text{Number of true predictions}) / (\text{Total number of predictions})$$

$$\text{Error Rate} = (\text{Number of false predictions}) / (\text{Total number of predictions})$$

## **1.2 Soft Computing**

Soft computing is a term used in Computer Science to refer the problems which are uncertain, unpredictable. Techniques belonging to it can be tolerant of imprecise, incomplete or corrupt input data. Some of them can solve problems without requiring the solution steps or reasoning process to be explicitly stated. Some soft computing systems develop the capability to solve problems through repeated observation and adaptation. Some arrive at a solution through a process similar to evolution in nature.

In more than one ways, the human mind is the role model for soft computing techniques - for example, the ability to solve problems expressed in vague terms, or solving problems without making use of explicit solution steps. Arriving at a solution through an evolutionary process is commonplace in nature.

### **1.2.1 Components**

a) Neural Networks

b) Perception

c) Fuzzy Logic

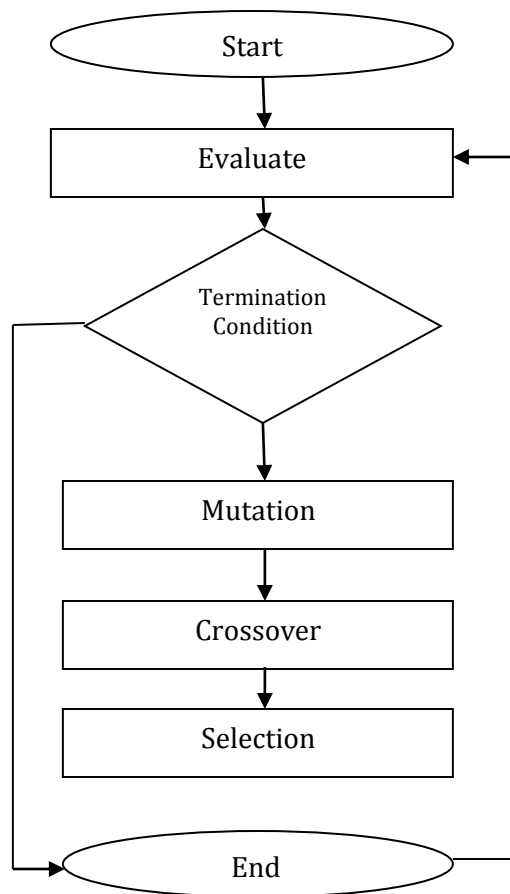
d) Evolutionary Algorithms

- Genetic Algorithm
- Differential Evolution
- Ant Colony Optimization
- Particle Swarm Optimization.

### 1.2.2 Evolutionary Algorithms

**Evolutionary algorithms** are based upon Darwin's Natural Selection theory of evolution, where a population is progressively improved by selectively discarding the worse and breeding new children from the better. There are many evolutionary algorithms but Differential Evolution Algorithm proposed for this study.

**Differential Evolution** is a method which optimizes the problem by iteratively improving its solution. DE can be used to find approximate solutions to such problems that are non- differentiable, non-continuous, non-linear, noisy, flat, multi-dimensional or have many local minima, constraints.



**Figure 1.3.** Flow chart for Differential Evolution Algorithm.

a) Differential evolution starts with the generation of random vectors together called as population. The vectors are generated as

$$x_i^G = x_{i(L)} + rand_i[0,1].(x_{i(H)} - x_{i(L)}) \quad (1)$$

$x_{i(L)}$  And  $x_{i(H)}$  represent the lower and upper limits of the dimensional vector  $x_i = \{x_{j,i}\} = \{x_{1,i}, x_{2,i}, \dots, x_{d,i}\}^T$ , respectively.  $rand_i[0,1]$  generates a random number in  $[0,1]$ .

b) Mutation

For every vector  $P_{i,G}$  in any generation G, a donor vector also called as the mutant vector is produced

$$x' = x_{r_3}^G + F.(x_{r_1}^G - x_{r_2}^G) \quad (2)$$

Where, the random integers  $r_1 \neq r_2 \neq r_3 \neq i$  are used as indices to index the current parent object vector. As a result, the population size  $N$  must be greater than 3.  $F$  is a real constant positive scaling factor and normally  $F \in (0,1+)$  controls the scale of the differential variation  $(x_{r_1}^G - x_{r_2}^G)$ .

c) Crossover

Crossover is the process of generation of the trial vector from the original population vector and the mutant vector. This is accomplished by shuffling the competing vectors.

Selection is the last step in Differential evolution algorithm. It lets us decide which vector  $(U_{j,i,G}, P_{j,i,G})$  should be the next generation member. It is based on the greedy approach as for the minimization problem a lower value vector is chosen.

The whole cycle of evolution is repeated till the termination criterion is satisfied.

# REVIEW OF LITERATURE

---

**Marconi de Arruda Pereira, et al. (2014)** proposed hybrid data classification based on genetic programming. Two algorithms are used to fulfill the two proposed objectives that are to maximize the efficiency and minimize the complexity. First algorithm is for rule extraction, in which confusion matrix is used to evaluate the effectiveness and complexity of rules generated by algorithm. And second algorithm is for diversity control, in which Mutation, Crossover, Elitism are used to keep the results in particular range. The proposed algorithm is suitable for unbalanced data also. Six different type of data (e.g. regular and non-regular data) has been used and results have been compared to basic data classification methods(e.g. Decision Tress, Support Vector Machines). The proposed algorithm has outperformed these basic methods in all instances.

**Mahesh Pal and Giles M. Foody (2012)** proposed relevance vector machine(RVM) and sparse multi-nominal logistic regression(SMLR) approaches to classify remotely sensed data that is equally efficient to support vector machine (SVM) classifier but require less training set. SVM, RVM, SMLR are supervised classification techniques in which results are dependent on quality and quantity of training data. In results to the proposed algorithm, RVM and SMLR have generated similar classified data with less training data.

**Chieh-Yuan Tsai and Chih-Jung Chena(2014)** proposed two-stage SPM based sequential classifier for complex sequential problems. Sequential Pattern Mining(SPM) is considered best for complex sequential problems but having unsolved issues such as pattern redundancy, inappropriate sequence similarity measures, and hard-to-classify sequences. The proposed algorithm removes redundant patterns in first stage and removes the partial similarities in sequences in second stage. Then it uses PSO-AB to optimize the results and to change the distributions of patterns that are hard to classify.



**Kung-Jeng Wang, et al. (2014)** enhanced the classification for 5-year survivability of breast cancer patients by integrating synthetic minority oversampling technique (SMOTE) and particle swarm optimization (PSO) with basic classifiers such as Decision Trees (C5), logistic regression, 1-nearest neighbor search. The proposed algorithm has improved the accuracy and has worked well for imbalanced data sets. For future, SMOTE+C5+PSO can be used to classify other data sets to improve classification.

**Bing Xue, Mengjie Zhang and Will N. Browne (2013)** proposed multi-objective algorithm based on particle swarm optimization for feature selection in classifications. As test data, for classification may contain large number of features, and most of features are irrelevant and redundant. This paper proposed an algorithm to use PSO for feature selection (to select relevant features for classification). Two algorithms have been used. First algorithm sorts the features into PSO and second algorithm applies mutation and dominance to PSO to search for solutions. In results, this algorithm shows better feature subsets than the conventional methods. The first algorithm has limitation of losing the diversity of swarm quickly, while second one works well.

**Nabila Nouaouria and Mounir Boukadoum (2014)** describes particle swarm optimization based classification algorithm for mixed attribute data (continuous and discrete data). The proposed algorithm introduced new interpretation mechanism before fitness evaluation, rest procedure is kept same as basic PSO and combined it with dispersion for improved search space coverage. The proposed algorithm is more accurate than the conventional methods and is handling continuous and discrete data at same time. Time complexity is higher than conventional methods.

**Isaac Triguero, Salvador Garcı and Francisco Herrera (2011)** proposed differential evolution based algorithm for nearest neighbor classification. The nearest neighbor classification is part of data mining, having lots of drawbacks such as such as low efficiency, high storage requirements and sensitivity to noise. Prototype selection and generation techniques are used to overcome these problems. This study proposed an algorithm which used differential evolution algorithm for searching and optimizing the positioning of prototypes in nearest neighbor classification. The algorithm focused on prototype generation problem and to maintain the good relation between prototype selection

(PS) and Prototype Generation (PG). Use of DE makes classification more accurate than the conventional techniques such as data reduction.

**Ivanoe De Falco (2013)** proposed a Differential Evolution based approach for data classification from medical databases using IF-THEN rules. A tool called DEREx is developed based on this study, which automatically extracts knowledge from databases. DE individual generates rules for each class which are connected to extract knowledge. DEREx provides explanation facility also, when there is need of knowing how a particular result has been generated. So it gives user friendly environment. The proposed work is the first one to implement DE for classification while in previous studies DE was being used for only optimizing results generated by other classifiers. DEREX has been tested on eight databases and the results show increase in performance compared to other methods. DEREx turns out to be the best performing tool in terms of highest classification accuracy. Also statistical analysis has confirmed that DEREx is the best classifier.

**Surendra Kumar and C.S.P. Rao (2009)** proposed use of ant colony optimization (ACO) algorithm, genetic algorithm (GA) and data mining techniques for extraction of knowledge from large set of schedules. ACO is meta-heuristic approach based on how almost blind ants find the shortest path to food source by communicating. In this proposed algorithm, ACO has been used for speedy and accurate results, GA operations such as mutation and crossover have been used for generating new range of results and data-mining technique (Decision Tress) has been used for extraction of rules from large databases. Results are more accurate and effective as compared to conventional methods.

**Fernando E.B. Otero, et al. (2012)** proposed the derivation of decision trees from ant colony optimization algorithm called ANT TREE-MINER. Basically decision trees are used for data classification (data mining) and ant colony optimization algorithm is used for optimizing the results. This paper proposed to induce decision trees from ant colony optimization algorithm. The proposed algorithm has been tested on 22 data sets. And results showed more effectiveness as compared to conventional classifiers such as decision trees(C4.5), CART, cACDT.Ant tree-miner is best suitable for predicting the class of unknowm data (predictive classification).

**Qazi Sami ul Haq, et al. (2012)** proposed l-minimization based classification approach for data with very less training data sample. With hyperspectral data, this problem becomes more complex as it contains high dimensional data. The proposed algorithm has explored the special features of hyperspectral data using l-minimization sparse based classification techniques. This paper proposes the use of Homotopy- based sparse classification approach, when data is highly sparse. The results have showed that the proposed approach offers better accuracy than conventional methods as well as it is time efficient. Unlike other classification approaches, the proposed approaches have no requirements of dimension reduction and model selection.

**Dewan Md. Farid, et al. (2014)** proposed two hybrid data classifier (combination of Decision Trees and Naïve Bayes Classifier). First hybrid data classifier used Naïve Bayes Classifier to remove noise from test data and then induced decision trees. In second algorithm, naïve bayes classifier was used after inducing decision trees from test data. Both hybrid data classifiers are time efficient and produced better results when compared to traditional Decision trees and Naïve Bayes classifier. In future, other classification algorithms such as Naïve Bayes Tree, Genetic Algorithm, Differential Evolution can be used to enhance the performance.

**Kemal Polat and Salih Gunes (2009)** proposed hybrid data classification technique by combining Decision Trees(C4.5) and one-against-all approach to classify multi-data class problems. The study showed that the accuracy of results was more for hybrid data classifier than individual Decision Tree Classifier and other traditional methods. This method is suitable for unbalanced data sets and can be used in pattern recognition techniques. In future, other classification algorithms such as Naïve Bayes Tree, Genetic Algorithm, Differential Evolution can be used to enhance the performance.

**Serafeim Moustakidis, et al. (2012)** proposed SVM-Based Fuzzy Decision Trees for Classification of High Spatial Resolution Remote Sensing Images. The paper purposed a fuzzy decision tree where binary SVM's are used to implement node discriminations. An effective feature has been added which can be used to select various features which was not feasible in other systems. It has advanced features like enhanced classification accuracy, interpretable hierarchy and low model complexity. It has low computation and data storage

demands. This algorithm has been tested on two different tasks: natural forest classification using a Quick Bird multispectral image and urban classification using hyperspectral data. The experimental investigation has concluded that FDT-SVM is favorably compared with six existing methods, including traditional multiclass SVMs and SVM-based binary hierarchical trees. Comparison was done in testing rates, computational time requirement and architecture complexity

**Pugalendhi Ganesh Kumar, et al. (2014)** proposed Hybrid Ant Bee Algorithm for Fuzzy Expert System Based Sample Classification Fuzzy expert systems are used to maximize the accuracy and minimize the complexity. Genetic swarm Algorithm approach reduces the interpretability but enhances the classification accuracy. If then rules used in this system are very complex in nature. It became very difficult to understand these rules by a physician. To remove this difficulty Ant Colony optimization algorithm are used. After Ant Colony Optimization, Ant Bee Algorithms are used. Performance of these was checked on 6 gene expressions data sets. Results shows that the Ant Bee Algorithm has generated an accurate fuzzy system which has compact rule sets as compared to other approaches

**David Martens, et al. (2007)** proposed Classification with Ant Colony Optimization to extract rule based classifiers ACO are used in the vast field of data mining. We have compared Ant based approaches and compared them with C4.5,RIPPER and has conducted a bench mark study. A new technique named as Ant Miner + is also proposed. The basic difference between the newer technique and the older ACO's is about the use of the MAX-MIN ant system which has a defined environment for the ants to work which includes the ability to handle multiclass problems. By using default parameter s problem faced in ACO of setting system parameter is resolved. We have found that the accuracy provided by the Ant Miner+ is much more than the any other version of the Ant Miner. Thus we have concluded that if the proper environment is provided then the ants can choose their path and can implicitly construct a rule also. Ant based systems give comprehensive results and these results are in a rule based format.

**Fernando E. B. Otero, et al. (2013)** proposed a new Sequential Covering Strategy for Inducing Classification Rules with Ant Colony Algorithms. One of the effective methods for discovering the classifications is ACO that is ant colony optimization. It is done by using single rule in a sequential manner at each iteration so that a set of rule can be build. We have carried out some studies to eliminate the problem of iteration which will be discussed here. We have conducted a lot of experiments using available data sets which shows that the new aco has considerable accuracy. They have higher accuracy than the induction classification algorithms. Here we are talking about the limitations and the strategy used by the ant colony classifications algorithms. Here we have provided a list of rules to enhance the search capacity of the ACO thus eliminated the old problem of discovering rules which ultimately leads to rule interactions. What we have done is implementation of sequential covering strategy named as cAnt-Miner. Our results were very positive which shows that cAnt-Miner is a better option because it is more reliable as its accuracy is enhanced.

## Chapter 3

# PRESENT WORK

---

### **3.1 Problem Definition:**

Strategy parameter for data classification is one of the emerging problems in evolutionary algorithms. So here, we are trying to modify an algorithm which may resolve strategy problem, performance in data classification and retraining should be depending on number of iterations with regarding to Differential Evolution. In order to achieve more efficiency more parameters must be included.

### **3.2 Objectives:**

- Modifying algorithm which resolves the problem of data classification.
- Maximization of algorithm efficiency and minimize the complexity of rules generated.
- Optimized CPU Time complexity for the modified algorithm.
- Differential Evolution algorithm and our algorithm comparison are done.
- Simulation of both the algorithms with efficient results has been shown.

### 3.3 Research Methodology:

Research methodology is the systematically defined steps to carry out any research work. It also identifies the methods to be used in research. It aims to give the work plan of the research. In other words, methodology defines the set of methods to be used in study. We are using Differential algorithm for our proposed work. In future implementation, we will work according to the following steps

- Resolving the problem by the help of DE and our Algorithm, Problem is Rosen burg where both algorithm works for classification of data using R-function. We create a database containing different types of data. This database will be initial population for proposed algorithm.
- Use DE to generate rules using training data set and predefined categories for classification using mutation and crossover
- Use fitness function to extract the best rules. The matrices that will be used to evaluate the performance of the algorithm are

➤ To check the efficiency of the classification rule generated

$$F_1(I, X) = \text{sensitivit}_y(I, X) * \text{specificit}_y(I, X) \quad (1)$$

Where,

$$\text{Sensitivit}_y(I, X) = TP / (TP + FN)$$

$$\text{Specificit}_y(I, X) = TN / (TN + FP)$$

TP, TN, FP and FN are explained in Table1.

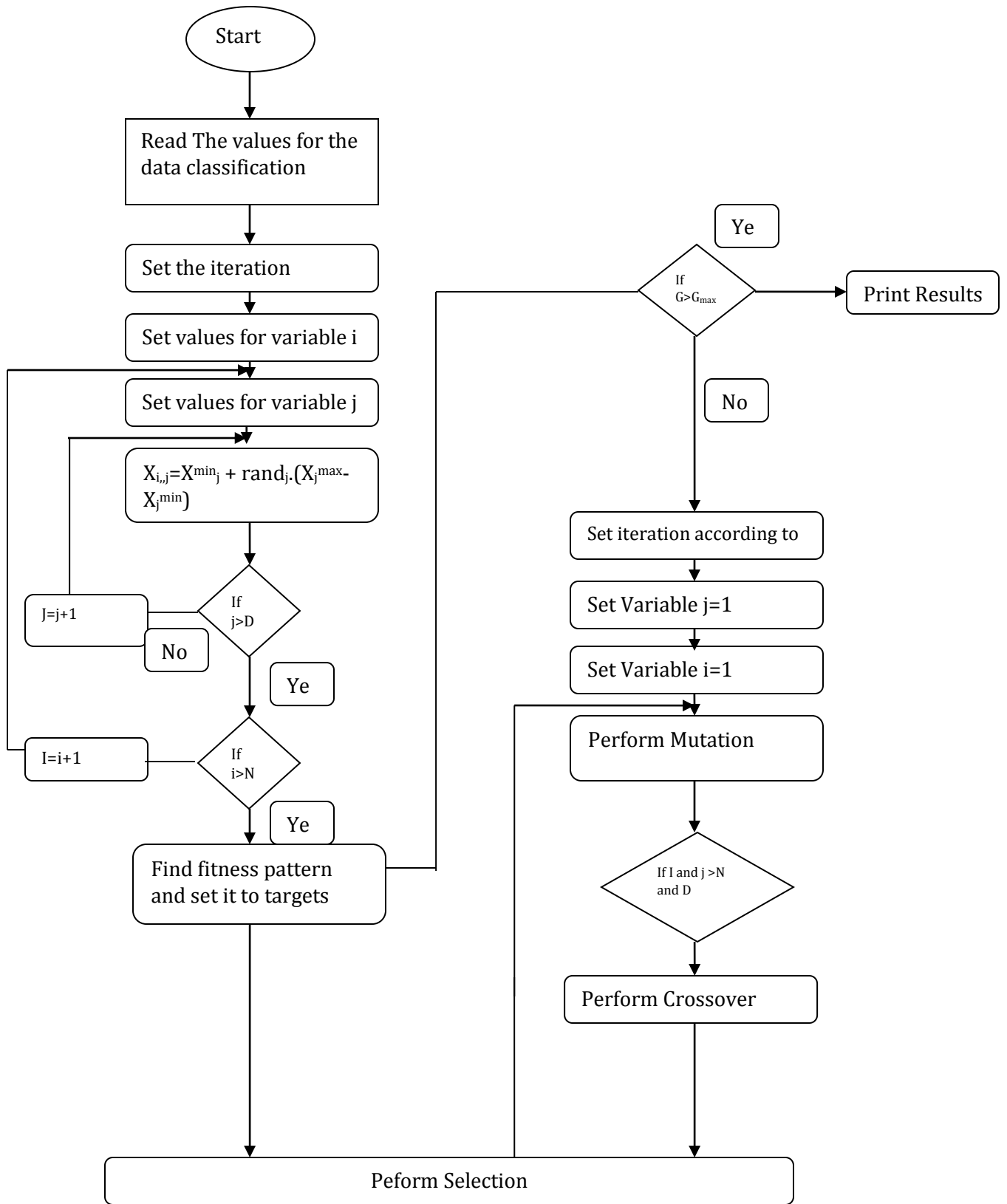
➤ To check the complexity of the rules

$$F_2(I) = 1 / \text{number\_of\_nodes} \quad (2)$$

- The best rules will be applied to test data.
- Mutation and crossover will be applied to generate new rules from the existing ones.

*The above algorithm is implemented using MATLAB.*





**Figure 3.1. Methodology Flowchart**

# RESULTS AND DISCUSSION

---

Under this heading the results and discussion part has been taken place. It describes about the different aspects of our algorithm for minimizing the complexity and maximizing the efficiency of our algorithm. In this heading different things are discussed like tool used working and figures of our results.

This dissertation focuses on evolutionary algorithm for the data classification with the comparison of differential evolution and our algorithm by taking the problem of Rosen Brock. Rosen Brock is a kind of optimization problem in which objective functions are inexpensive to compute and there derivatives cannot be identified efficiently. So with the help of Differential evolution function we can optimize the data.

Comparison of both the algorithm is done on the basis of Rosen Brock function because on the basis of this function we can generate optimized coordinate system which can iteratively repeated without using any gradient function and without building any data sets. By which the classification is already is done under Rosen Brock function.

With the help of these things we can prove that our algorithm is faster than DE in doing the data classification with more accuracy and maximum efficiency with minimum time complexity.

### **4.1 Tool Used:**

**Matlab:** As we all know in now a day's computing is moving very fast at very age. So here is one of the most important and useful tool is going to discuss. This tool helps in understanding the concept of programming very easily and can simulate the result according to you. Name of this tool is MATLAB. It is a very high level language, basically for the numerical computation and visualization. By using this tool we can analyze the data, modify the algorithms and creates different applications. By the help of this tool we can perform all the mathematical functions which enables you to explore the mathematics world by the help of programming and you can visualize the things in front of you by using simple syntax.

MATLAB has several advantages over other methods or languages:

- Matrix formulation is done on the basis of data element. Matrix generation is done in rows and column basis. The Mathematical function required for the different operation on matlab is given by built in functions to matlab environment.
- For the generation of two arrays for single operation that is for addition needs only single command instead of using looping condition.
- We can plot the different graph on the basis of results we get in numeric form. It's so easy to change the colours, size, scales gesture and many more of the graph on the need for differentiation the methods
- Functionality of this tool is greatly experienced by the user; it is only possible with the help of toolbox used. It provides all kind of tools which is specified for different function. It gives more accurate results according to it.

There are also disadvantages:

- As we all know matlab is very huge software, it needs large amount of memory which makes our system to slow for working.
- It take all the as much CPU time from windows because it is too much process running under it.

You can enter inputs into MATLAB through different ways:

- 1 Enter an explicit list of elements.
- 2 Load matrices from external data files.
- 3 Generate matrices using built-in functions.
- 4 Create matrices with your own functions in M-files.

## 4.2 RESULTS:

Here now we are going to discuss about the results which we generated from our hypothesis. The graph shows the positive results according to me which we expect from it. First of all we are going to introduce the very first figure.

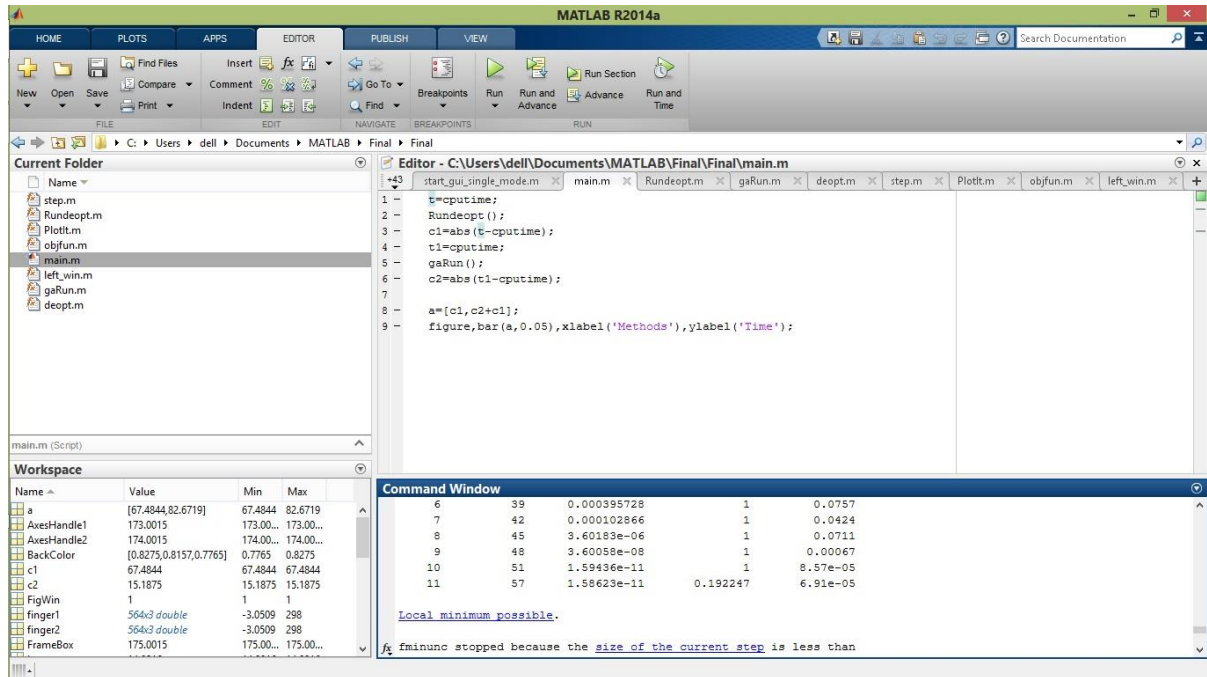


Figure 4.1. Matlab Starting GUI

Here in figure 4.1. descriptions of main files has shown. As we click on run button the simulation begins. It starts from datasets uploading by the help of Rosen Brock function. Afterwards it switch to next figure where plotting of different saddle block takes place under which data classification is done which can we visualize under two ways that is in 3D mode and clustering mode.

The best cost of data classifier is also noted down at different iterations which is done by genetic algorithm and also by differential algorithm.

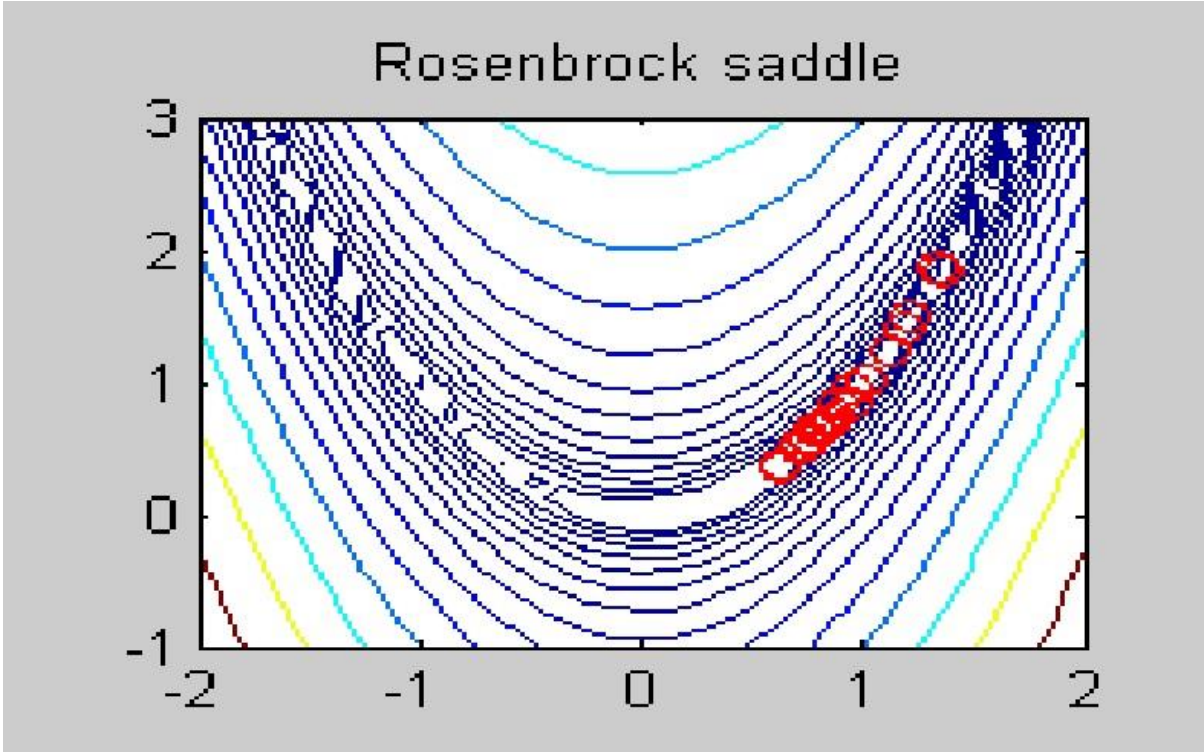


Figure 4.2. Rosen brock saddle Datasets Classification by GA

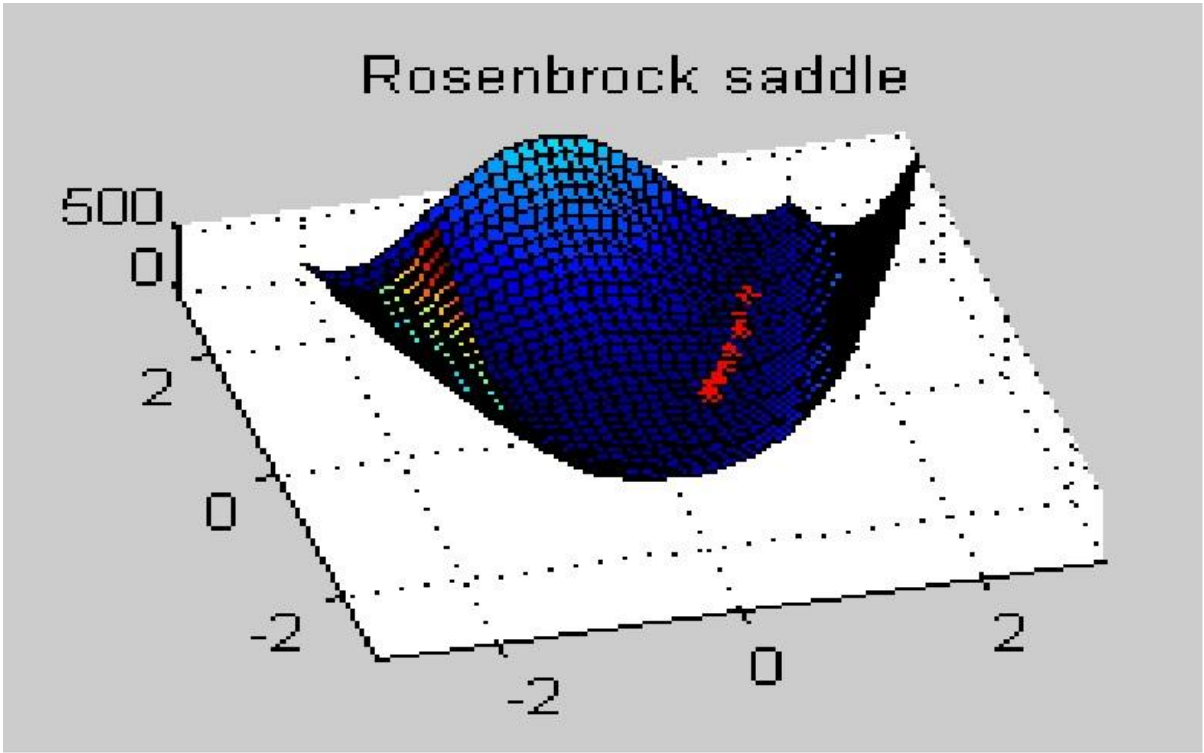
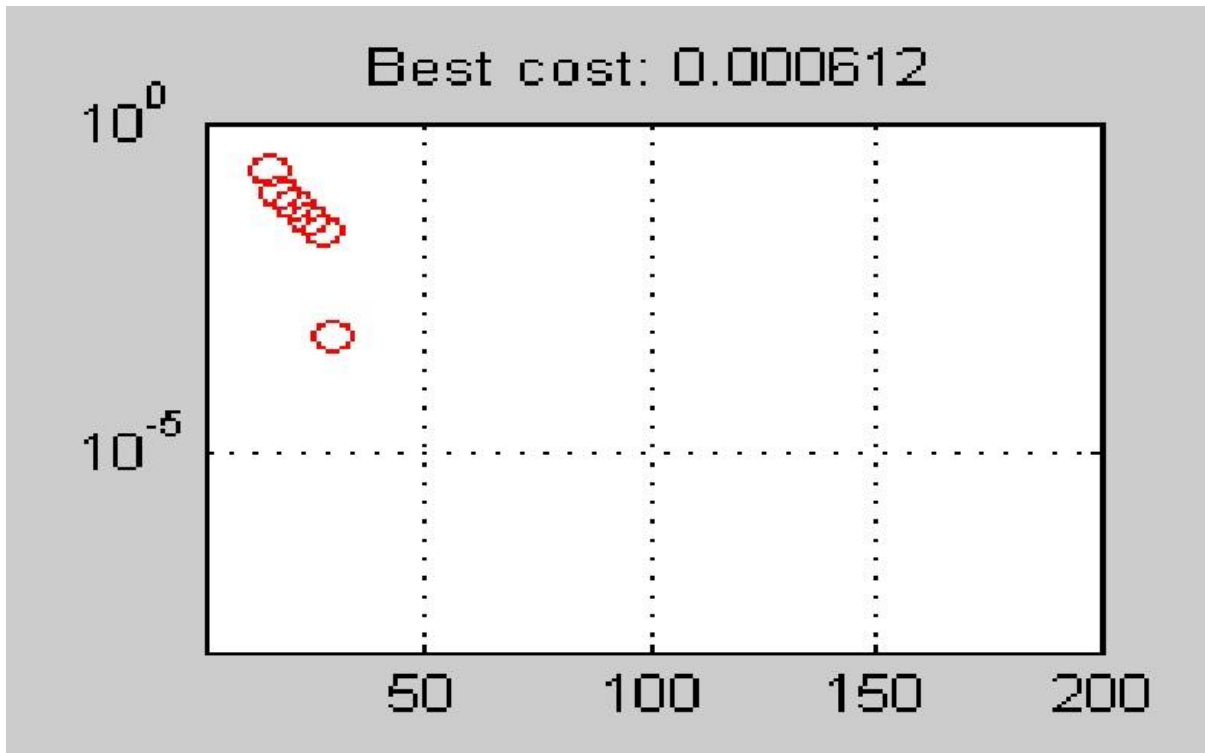
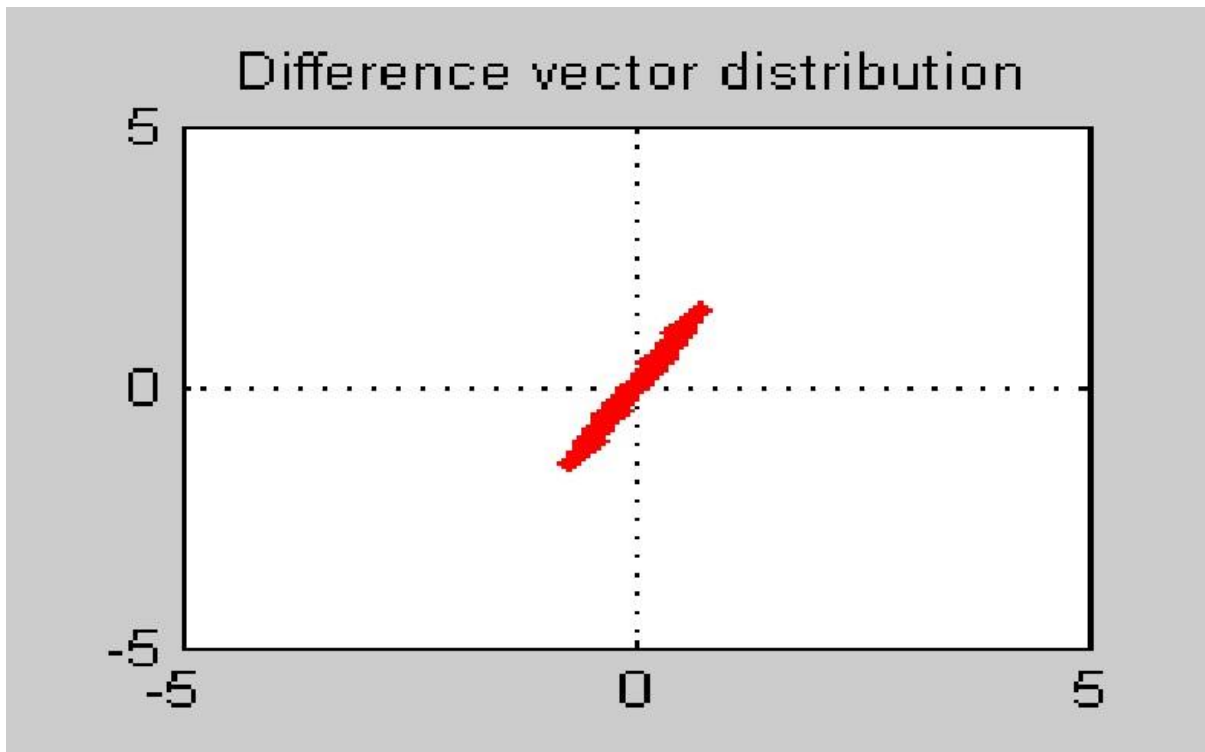


Figure 4.3. 3D view of data clusters in Red by GA



**Figure 4.4.** Best Cost Value Calculation on basis of data classification in GA



**Figure 4.5.** Vector Distribution graph in GA

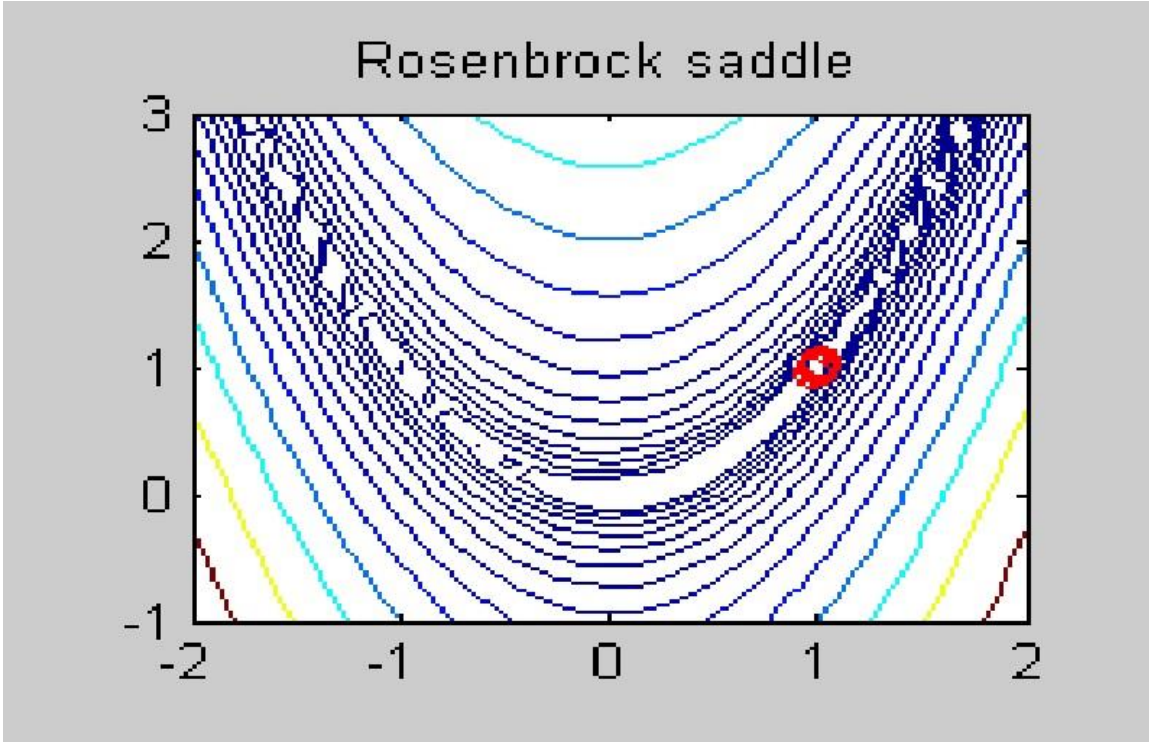


Figure 4.6. Rosen Brock Saddle by DE Result

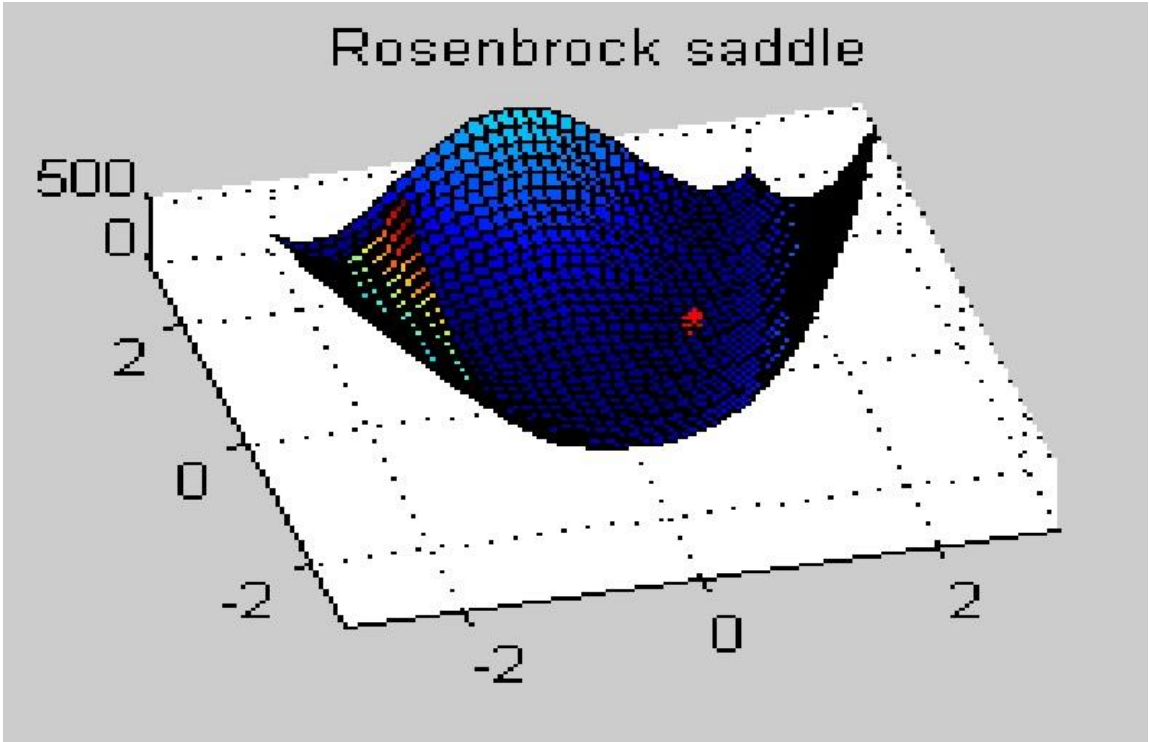


Figure 4.7. 3D views of cluster By DE

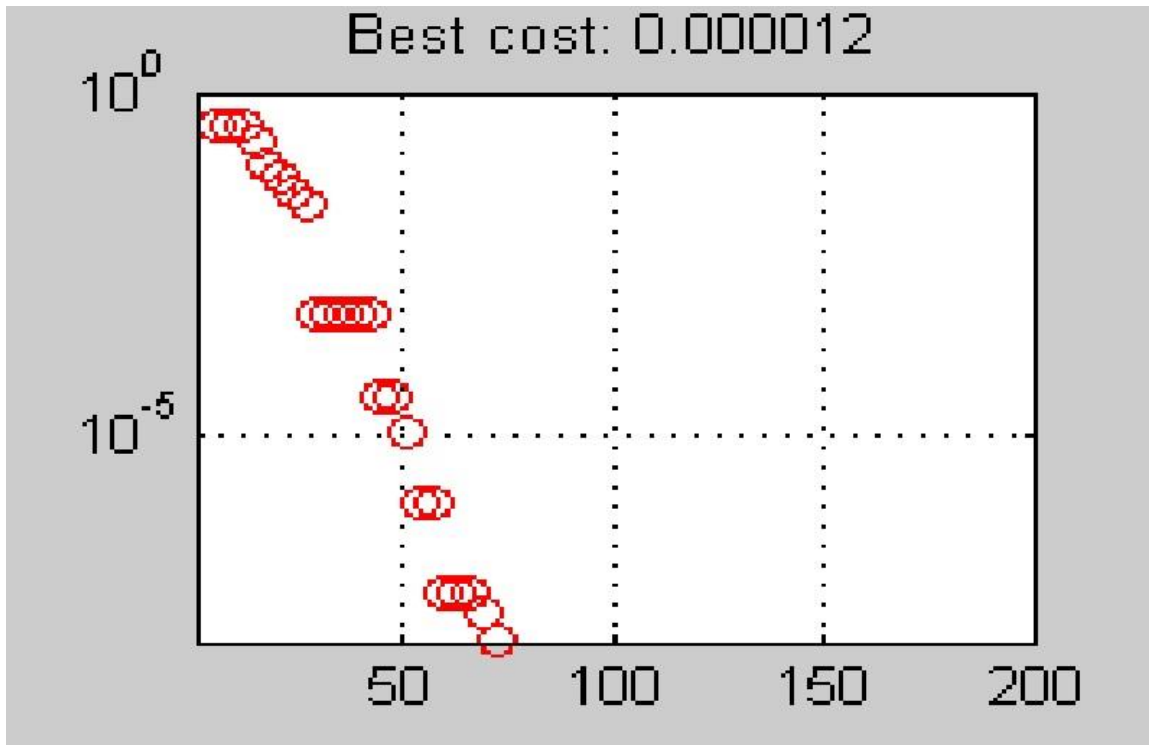


Figure 4.8. Best cost By DE

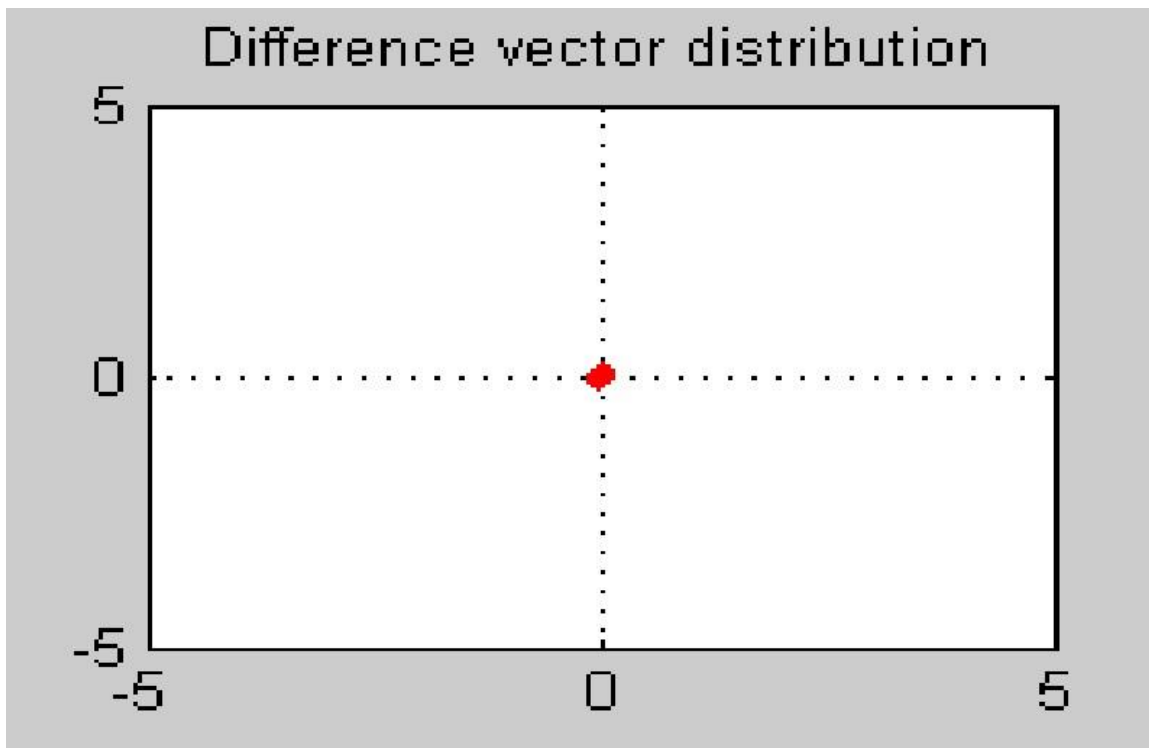
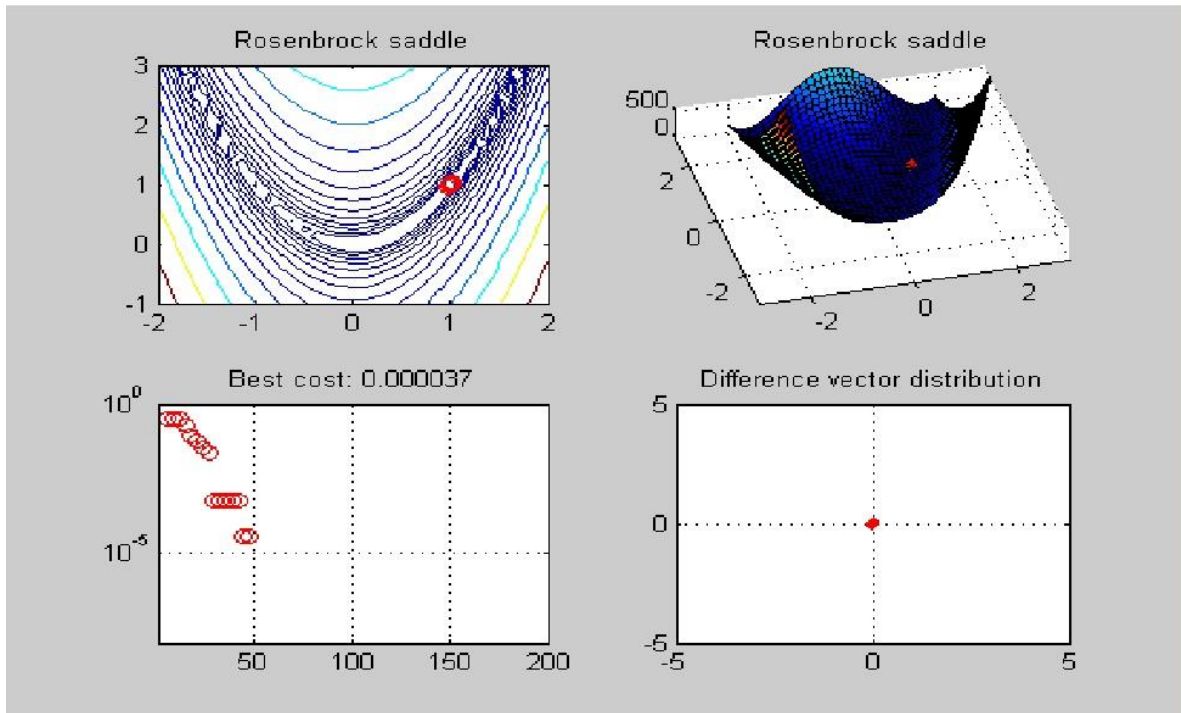
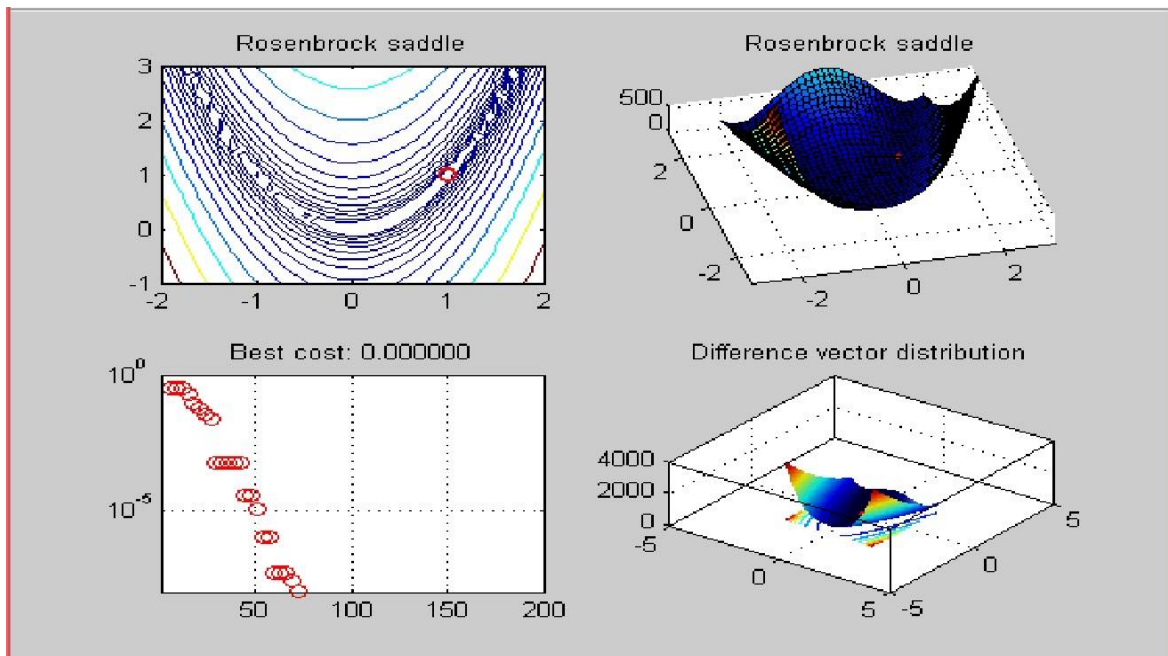


Figure 4.9. Different Vector Distribution by DE

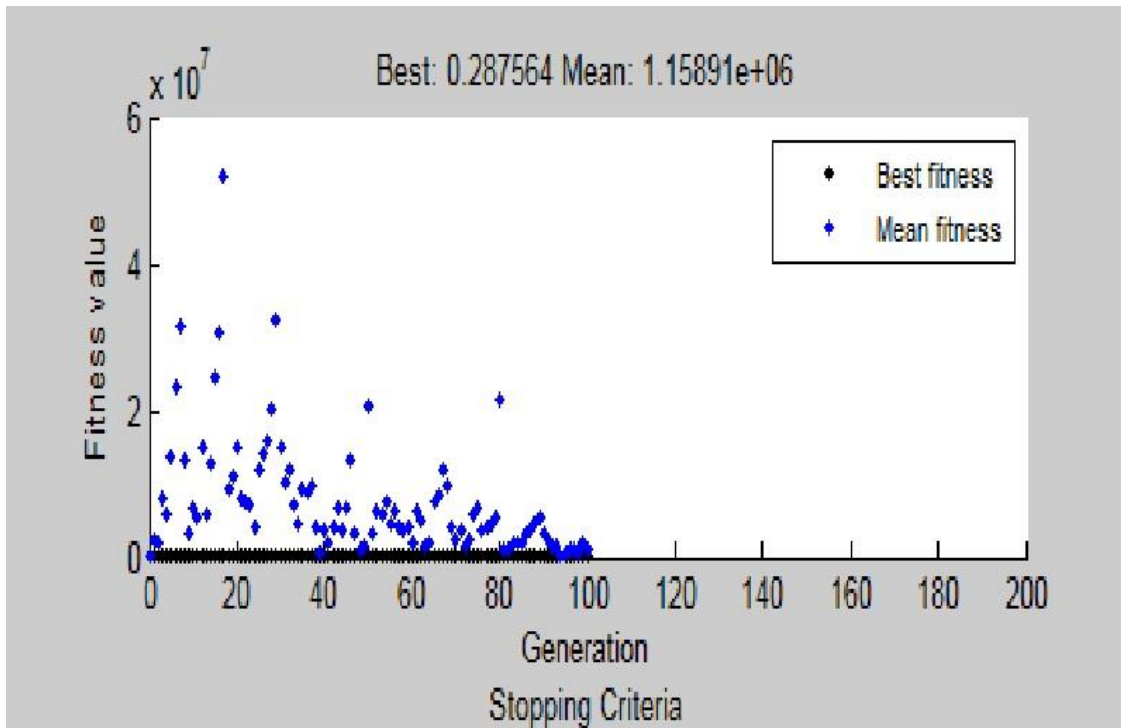




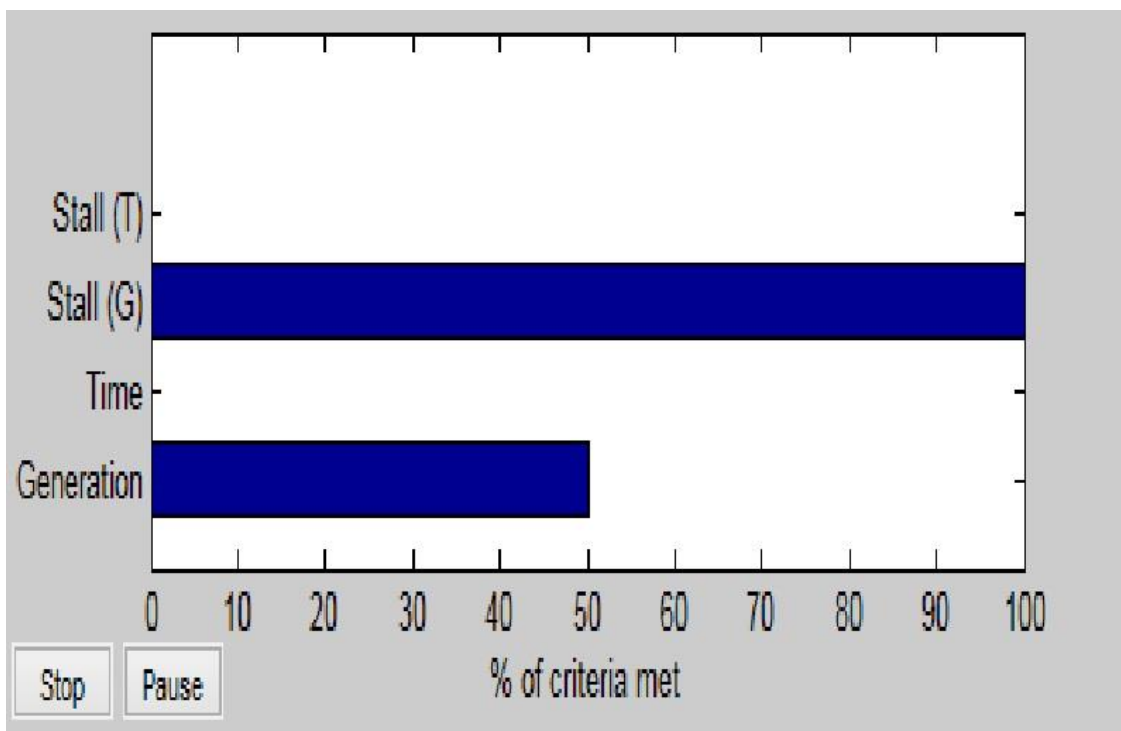
**Figure 4.10.** Multiple Iterative results by DE



**Figure 4.11.** Final Result of vector distribution



**Figure 4.12.** Graph between fitness value of best and mean by GA



**Figure 4.13.** Time generation between mean and best fitness

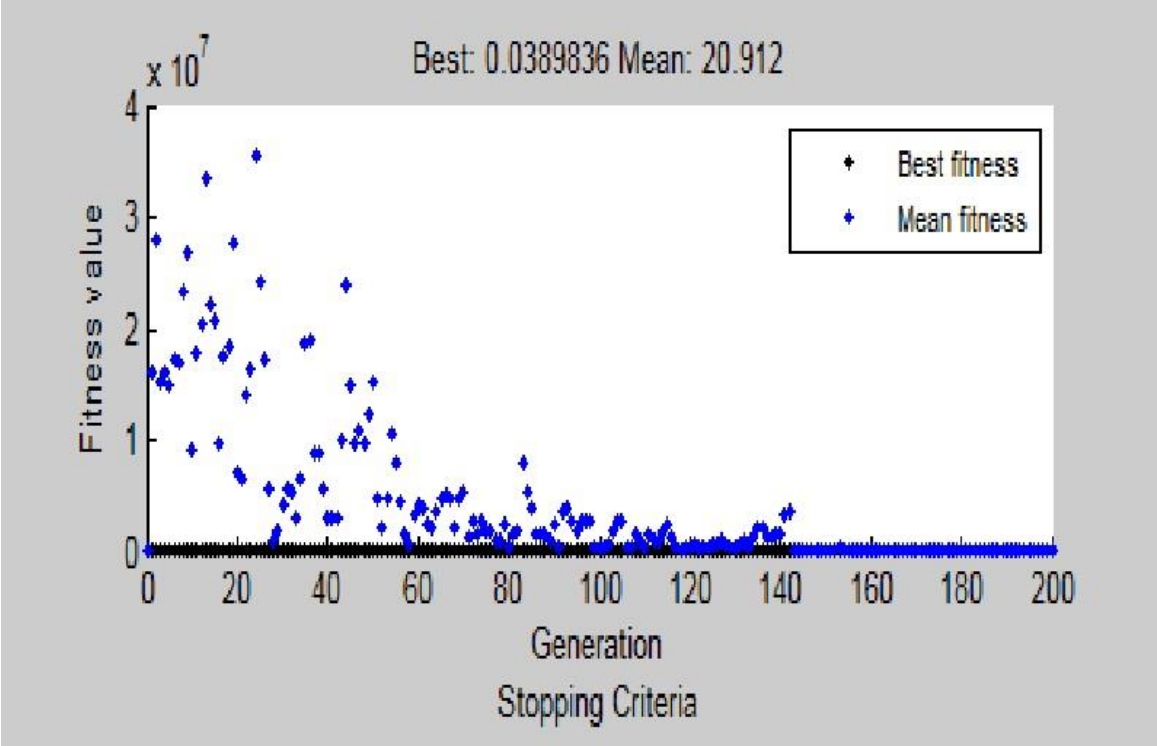


Figure 4.14. Best and mean time by DE

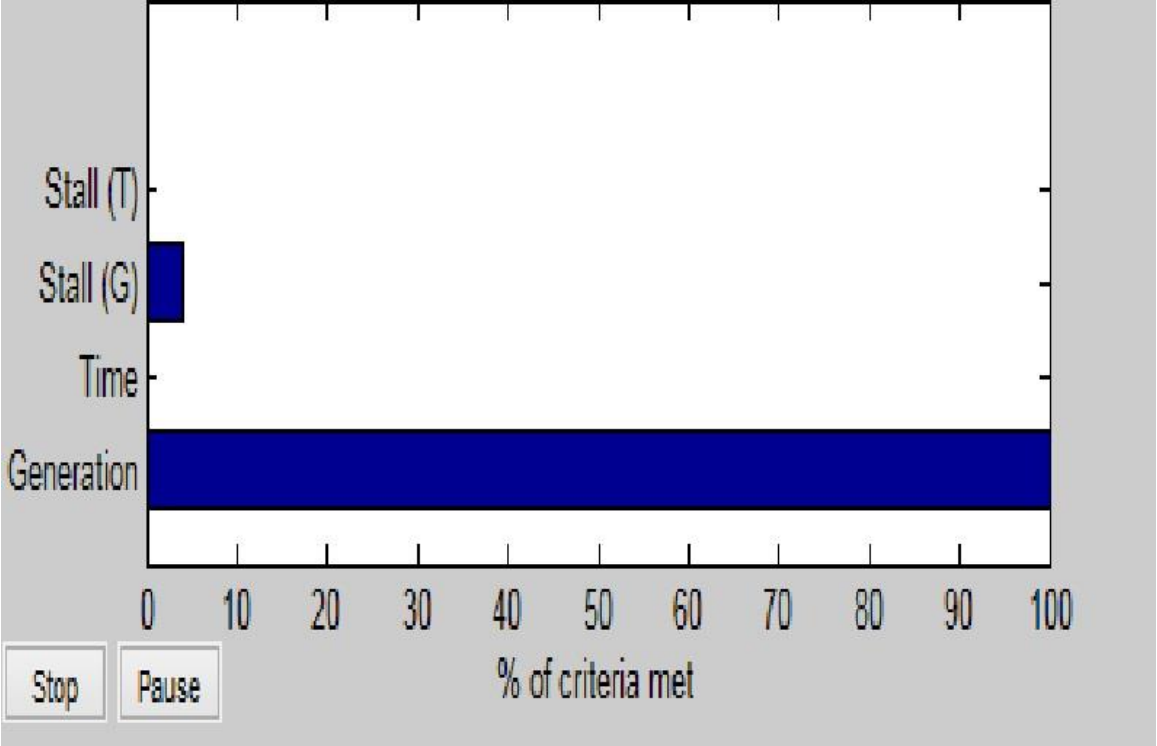
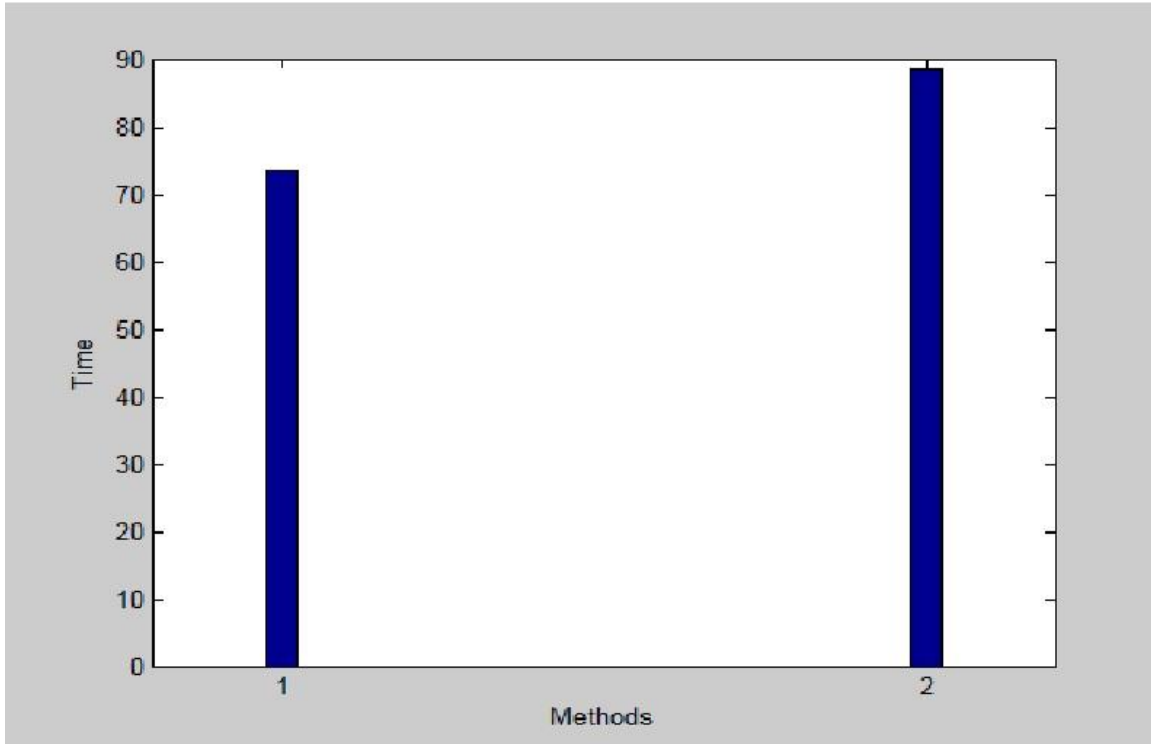


Figure 4.15. Time Generation by DE



**Figure 4.16.** Comparison between time complexity of both algorithms

In above figure 4.2, 4.3, 4.4, 4.5 show the result using genetic algorithm. Figure 4.2 shows the value which is given by datasets and it is plotted according to function Rosen Brock. Under that genetic algorithm is placed so that the complexity of time is calculated on the basis of best cost and different fitness values.

Figure 4.3 shows the 3D view of datasets which are plotted in scattered form by GA. The red dots which are shown in figure shows that data are classified on the basis of classifier presents in genetic algorithms and it takes too much time to compute it.

Figure 4.4 shows the best cost value which is calculated on the basis of data classification.

Figure 4.5 generated the different vector visualization so that we can see the different clusters are placed in area on the basis of genetic algorithm.

In above figure 4.6, 4.7, 4.8, 4.9 show the result using Differential evolution algorithm. Figure 4.6 shows the value which is given by datasets and it is plotted according to function Rosen Brock. Under that differential algorithm is placed so that the complexity of time is calculated on the basis of best cost and different fitness values.

Figure 4.7 shows the 3D view of datasets which are plotted in scattered form by DE. The red dots which are shown in figure shows that data are classified on the basis of classifier presents in differential algorithms with mutation and crossover functions and it takes too much time to compute it.

Figure 4.8 shows the best cost value which is calculated on the basis of data classification.

Figure 4.9 generated the different vector visualization so that we can see the different clusters are placed in area on the basis of DE algorithm.

Figure 4.10 shows the final value which is calculated by DE algorithm at multiple iterations, which is its basis functionality. Under this figure we can watch the best cost value is 0.000037 and in DE it is 0.000012, so by this we can say that by increasing the iteration we get more classified data with fewer complexes in time.

Figure 4.11 generates a different vector graph on the basis of best cost and cluster formation is done in 3D view.

Figure 4.12 shows the best mean and best fitness value which are 0.283 and mean 1.2876+e132 etc. At the time of cluster formation genetic algorithm generated fitness values on the basis of their mutual function which is not enough efficient in comparison to DE.

Figure 4.13 plotting of these values in graph have been done, it shoes the time generation between best and mean fitness values.

Figure 4.14 shows the best mean and best fitness values which are 0.0382 and mean 20. At the time of cluster formation genetic algorithm generated fitness values on the basis of their mutual function like crossover and mutual which is enough efficient in comparison to GA.

Figure 4.15 plotting of these values in graph have been done, it shoes the time generation between best and mean fitness values.

Figure 4.16 shows the time complexity graph between DE (1) and GA (2). It shows the graph of differential evolution and 2 shows genetic algorithm. We can see that De takes 75 and GA take 89 second to completing the process. Hence we can prove that DE is more efficient than GA.

# Future Scope and Discussions

---

### 6.1 Scope of Study:

Data classification is used in the ILM (Information lifecycle management) to classify the data so as to know the types of data, the data implementation, various access levels in the data etc. While the data classification has wide variety of applications, in this section the main corporately used applications will be discussed. These applications, while covering the wide scope of the data classification, will also give us the way in which we have to head while analyzing the data classification architectures and techniques using the Differential equations.

The first and foremost application of data classification is the use of this technique in tagging the data. When the data has been tagged, the data search quickly and thus it saves the time too. We see in our real world scenarios that there is huge amount of duplication in the data due to which there is a chaos in the structure of the information. With the help of data classification techniques, we can do the de-duplication of our data. Suppose we are given a timeframe to act upon a certain set of data thus the data classification comes in the picture by helping the searching-and-acting time to be reduced effectively so as to meet the legal requirements in the corporate sector. Continuously handling the classified data gives our dataset stability and thus gives stability to the structure of the information that has been stored. The data classification is the only technique that can be used to handle various data domains such as multimedia, text, time-series, network, discrete sequence, and uncertain data etc. Various aspects of the classifiers such as ensembles

, rare-class learning, distance function learning, active learning, visual learning, transfer learning, and semi-supervised learning as well as evaluation aspects of classifiers can be used in depth to handle the data effectively.

## **6.2 DISSCUSSION**

As we discuss about the genetic algorithm is used to optimize the problem and moved toward the better solutions. Heuristic search is done for the generalization of problem; a set of properties can be mutated and altered according to it.

In differential evolution algorithm is a novel approach for minimization the methods. It has been designed for handling the multimodal function and non differentiable functions.



- [1] Pal, M., & Foody, G. M. (2012). Evaluation of SVM, RVM and SMLR for accurate image classification with limited ground data. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 5(5), 1344-1355.
- [2] Tsai, C. Y., & Chen, C. J. (2014). A PSO-AB Classifier for Solving Sequence Classification Problems. *Applied Soft Computing*.
- [3] Wang, K. J., Makond, B., Chen, K. H., & Wang, K. M. (2014). A hybrid classifier combining SMOTE with PSO to estimate 5-year survivability of breast cancer patients. *Applied Soft Computing*, 20, 15-24.
- [4] Carrano, E. G., Wanner, E. F., & Takahashi, R. H. (2011). A multicriteria statistical based comparison methodology for evaluating evolutionary algorithms. *Evolutionary Computation, IEEE Transactions on*, 15(6), 848-870.
- [5] Bazi, Y., Alajlan, N., Melgani, F., AlHichri, H., Malek, S., & Yager, R. R. (2014). Differential Evolution Extreme Learning Machine for the Classification of Hyperspectral Images.
- [6] Kaya, G. T. (2013). A hybrid model for classification of remote sensing images with linear SVM and support vector selection and adaptation. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal of*, 6(4), 1988-1997.
- [7] de Arruda Pereira, M., Davis Júnior, C. A., Gontijo Carrano, E., & De Vasconcelos, J. A. (2014). A niching genetic programming-based multi-objective algorithm for hybrid data classification. *Neurocomputing*, 133, 342-357.
- [8] Xue, B., Zhang, M., & Browne, W. N. (2013). Particle swarm optimization for feature selection in classification: a multi-objective approach. *IEEE transactions on cybernetics*, 43(6), 1656-1671.
- [9] Nouaouria, N., & Boukadoum, M. (2014). Improved global-best particle swarm optimization algorithm with mixed-attribute data classification capability. *Applied Soft Computing*, 21, 554-567.
- [10] De Falco, I. (2013). Differential Evolution for automatic rule extraction from medical databases. *Applied Soft Computing*, 13(2), 1265-1283.

- [11] Kumar, S., & Rao, C. S. P. (2009). Application of ant colony, genetic algorithm and data mining-based techniques for scheduling. *Robotics and Computer-Integrated Manufacturing*, 25(6), 901-908.
- [12] Otero, F. E., Freitas, A. A., & Johnson, C. G. (2012). Inducing decision trees with an ant colony optimization algorithm. *Applied Soft Computing*, 12(11), 3615-3626.
- [13] Sami ul Haq, Q., Tao, L., Sun, F., & Yang, S. (2012). A fast and robust sparse approach for hyperspectral data classification using a few labeled samples. *Geoscience and Remote Sensing, IEEE Transactions on*, 50(6), 2287-2302.
- [14] de Arruda Pereira, M., Júnior, C. A. D., & de Vasconcelos, J. A. (2010). A niched genetic programming algorithm for classification rules discovery in geographic databases. In *Simulated Evolution and Learning* (pp. 260-269). Springer Berlin Heidelberg.
- [15] Farid, D. M., Zhang, L., Rahman, C. M., Hossain, M. A., & Strachan, R. (2014). Hybrid decision tree and naïve Bayes classifiers for multi-class classification tasks. *Expert Systems with Applications*, 41(4), 1937-1946.
- [16] Polat, K., & Güneş, S. (2009). A novel hybrid intelligent method based on C4. 5 decision tree classifier and one-against-all approach for multi-class classification problems. *Expert Systems with Applications*, 36(2), 1587-1592.
- [17] Martens, D., De Backer, M., Haesen, R., Vanthienen, J., Snoeck, M., & Baesens, B. (2007). Classification with ant colony optimization. *IEEE Transactions on Evolutionary Computation*, 11(5), 651-665.
- [18] Ganesh Kumar, P., Rani, C., Devaraj, D., & Victoire, T. (2014). Hybrid Ant Bee Algorithm for Fuzzy Expert System Based Sample Classification.
- [19] Moustakidis, S., Mallinis, G., Koutsias, N., Theocharis, J. B., & Petridis, V. (2012). SVM-based fuzzy decision trees for classification of high spatial resolution remote sensing images. *Geoscience and Remote Sensing, IEEE Transactions on*, 50(1), 149-169.
- [20] Otero, F. E., Freitas, A. A., & Johnson, C. G. (2013). A new sequential covering strategy for inducing classification rules with ant colony algorithms. *Evolutionary Computation, IEEE Transactions on*, 17(1), 64-76.

## Abbreviation

RVM	Relevance Vector Machine
SMLR	Sparse Multi-Nominal Logistic Regression
SVM	Support Vector Machine
SPM	Sequential Pattern Mining
SMOTE	Synthetic Minority Oversampling Technique
PSO	Particle Swarm Optimization
PS	Prototype Selection
PG	Prototype Generation
ACO	Ant Colony Optimization
GA	Genetic Algorithm
ILM	Information Lifecycle Management

## **Glossary of Term**

- Descriptive Modeling-Classification model can serve as explanatory tool to distinguish data into different classes based on their attributes
- Predictive Modeling -Classification model can be used to predict class for unknown record.
- Neural Networks- It is a machine learning and cognitive science the neural network part of artificial neural network, the member of statistical learning algorithm it is inspired by the biological neural network.
- Fuzzy Logic- fuzzy logic generally based on computer based on degree of truth usually true or false the fuzzy logic used generally modern based computer.
- Evolutionary Algorithms- it is a part of artificial intelligence, the subset of evolutionary computation a generic population-based metaheuristic optimization algorithm. where a population is progressively improved by selectively discarding the worse and breeding new children from the better
- Differential Evolution- DE can be used to find approximate solutions to such problems that are non- differentiable, non-continuous, non-linear, noisy, flat, multi-dimensional or have many local minima, constraints
- Adaptation- it is process of adjustment outer condition.