EDM Classification Using Genetic Algorithm: A Review

Jaskaranjit Kaur, Navneet Kaur
Dept. of Computer Science & IT
Lyallpur Khalsa College, Jalandhar, India

Abstract - Education Data mining has made a great progress in the area of research & technology. Education Data mining is used to find out knowledge out of data and present it in a form that human can easily understand. With the evolution in the field of Information Technology and Computer Science, high capacity of data appears in our lives. Today educational institutions stores and compile large amount of data such as student enrolment and attendance records, their scholarship criteria as well as their examination results. Education Data Mining helps us to find out useful information from large dataset. This paper provides a brief review on classifying student's data in order to predict their performance on the basis of features extracted from the data logged in an Education System using GA. This paper studied the classification techniques like 1-NN, K-NN and Decision Tree (C4.5, C5.0, CART, Random Forests Algorithm) and Genetic algorithms which is used to predict previously unknown class of objects in order to determine whether student is eligible for scholarship or not. Also a brief study of WEKA & MATLAB has been done. WEKA is a data mining tool which is used to perform various data mining operations like clustering, classification and associate data. And MATLAB tool is used to further optimize the results.

Keywords- Education Data Mining(EDM);1-NN; K-NN; Decision Tree; C4.5; C5.0; CART; Random Forests Algorithm; Genetic Algorithms

I. INTRODUCTION

Data mining also known as ''knowledge discovery in databases" (KDD), is the way of retrieving meaningful information from large set of data. It is the way toward changing over the low-level information into high-level knowledge. Data mining is a strategy of analyzing very large data sets to extract and discover previously unknown structures and relations out of such colossal loads of points of interest. It is an innovation which is utilized with extensive potential to help organizations and huge ventures to discover their client's practices [11][12]. Education Data mining has made a great progress in the area of research & technology. It is concerned with discovering different ways for exploring the distinctive types of data that come from educational environments. Its main motive is to improve educational system. It manages dissecting, creating, inquiring about, and applying most recent computerized methods to recognize designs from extensive accumulations of educational dataset. In recent years, EDM has become a prominent research area which aimed at analyzing the exclusive kinds of data that occur in educational system to reconcile educational research issues [18].

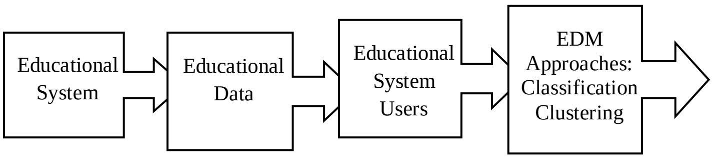


Fig.1: Educational Process

Classification is one of the most well-known and most operational data mining techniques which is used to identify unknown values and then further classify and predict it. There are many various methods of classification and approaches used in Knowledge Discovery and data mining. Every method or approach has its own advantages and disadvantages. In this paper we discuss various data mining classification techniques and Genetic algorithms. The objective of this survey is to classify the Educational data so that we can obtain the best optimized results using GA. The rest of the paper is divided into 6 sections. Section 2 contains the detail of data mining classifiers process followed by section 3 in which we discussed about Genetic Algorithm. Section 4 contains dataset attributes & class labels in which collected data is represented, explored and further visualized. In next section a brief introduction about WEKA and MATLAB Tool is given. In last section the work is concluded and insights about future work are included.

II. CLASSIFICATION AND ITS TYPES

Classification is a data mining technique that assigns items in a group to target categories or classes. It is a method of identifying a model that well defines the classes of dataset. In Classification models categorical class labels are predicted. The Data Classification process includes two steps [13].

- Building the Classifier
- Using that Classifier for Classification

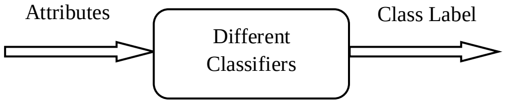


Fig.2: Classifier Input/output

The major subject of consideration in classification is preparing the data for Classification and Prediction. Preparing the data involves the following actions −

- Data Cleaning − Data cleaning means eradicating the noisy data and treatment of omitted values. The noise is discarded by applying smoothing techniques. We can resolve the problem of missing values by replacing a missing data value with the most frequently arising value for that particular feature.

- Relevance Analysis − Database may likewise contain unimportant qualities. The Irrelevancy is expelled by Correlation investigation to know whether any two given characteristics are connected or not.
- Data Transformation and reduction – With the help of these two methods the data can be transformed:
  - **Normalization–** The information required for classification can be changed utilizing Normalization. Normalization is scaling method or a mapping strategy or a preprocessing procedure in which we can discover new range from a current range.
  - **Generalization**– Transformation of data can be done by generalizing it to the higher concept. The concept of hierarchies is used for generalization.

The various classifiers used in this paper are:

A. 1-NN for Classification

Let's see how to use 1-NN for classification. For this situation, we are given a few data points for training and furthermore another unlabeled point. The algorithm has distinctive behavior based on k. KNN is a nonparametric lazy learning algorithm. Here non parametric implies that it doesn't make any suspicions on the underlying data distribution. It is a lazy learner in which training dataset is put away. On querying similarity between test data & training dataset, records are calculated to predict the class of test data. The input to this algorithm is the K closest training example and output is the class membership. When K=1 (where k is the number of neighbors) it means object is assigned to the class of single nearest neighbor. It means we are considering first immediate neighbor. This number determines how many neighbors (where neighbors are defined based on the distance metric) influence the classification. This is usually a odd number The similarity between test data & training data is mostly calculated using the Euclidean distance.

B. K-NN Algorithm

K-Nearest neighbor is a supervised learning classification algorithm in which the consequence of new data query is classified based on the basis of majority of K-Nearest neighbor. The motivation behind this calculation is to classify a new data based on attributes & training data. K-Nearest neighbor algorithm used neighborhood classification as the prediction value of the new query instance. Let m be the number of training data samples. Let p be an unknown point.Store the training samples in an array of data points arr []. This means each element of this array represents a tuple (x, y).

1. for i=0 to m:
2. Calculate Euclidean distance d (arr [i], p).
3. Make set S of K smallest distances obtained. Each of these distances corresponds to an already classified data point.
4. Return the majority label among S.

C. Decision Tree Algorithms

Decision Tree algorithm belongs to the family of supervised learning algorithms. The general thought process of utilizing Decision Tree is to construct a training model which we can use to find out the class of the target variables by learning decision rules deduced from prior data (training data). Table 1 depicts the various features of Decision tree algorithm in terms of type of attribute it can handle (Numeric or Categorical), Splitting criteria they follow for decision tree, whether they can handle missing values and outliers and which pruning strategy they follow.

TABLE 1. Basic features of Decision Tree Algorithms

Features **Algo's**	**Type of Attribute**	**Splitting Criteria**	**Missing values**	**Detection of Outlier And Pruning Strategy**
CART	Handles both Numeric and Categorical value	Towing Criteria	Handle missing values.	Can handle outliers Cost-Complexity pruning is used
C4.5	Handles both Numeric and Categorical value	Gain Ratio	Handle missing values.	Susceptible on outliers Error Based pruning is used
C5.0	Handles both Numeric and Categorical value	Inform-ation Gain (Entropy)	Estimate missing values as a function of other attributes or apportions the case statistically among the results.	Binomial Confidence Limit method

The decision tree algorithm tries to solve the problem, with the help of tree representation method. Each internal node of the tree relates to an attribute of the dataset and each leaf node relates to a class label. This method used tree structure to build the classification models. It divides a given dataset into littler subsets. Leaf node signifies a decision. The decision trees classify the instances on the basis of the feature values of various instances. In a decision tree, each node indicates a feature of an instance which is to be classified and each branch indicates a value. Classification of the data instances starts from the root node and sorting is done in light of their feature values. Categorical and numerical data can be handled by decision trees.

Decision Tree Algorithm Pseudocode

1. Place the best characteristic of the dataset at the root of the tree.
2. Split the training set into subsets. Subsets should be build in a way that for an attribute each subset include data with the same value.
3. Repeat step 1 and step 2 on each subset until you find leaf nodes in all the branches of the tree.

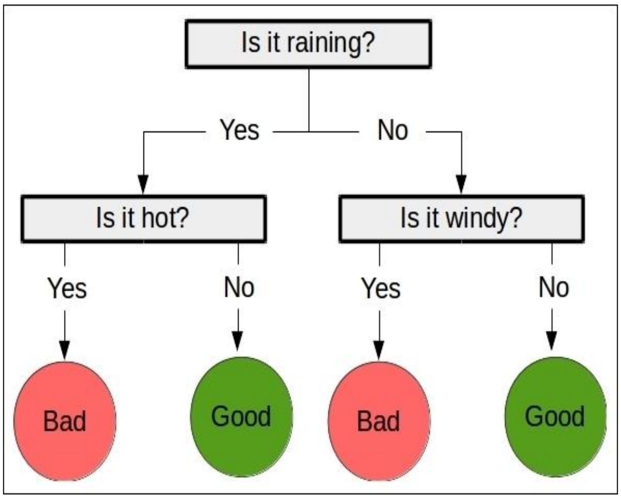


Fig.3: Decision Tree Example [16]

Figure 3 example depicts a weather forecasting process which deals with predicting whether is rain or not and the day is hot or windy. This above hierarchical tree can be applied to decide about the weather conditions.

C 4.5:

C4.5 is collection of algorithms which are used to perform classifications in data mining and machine learning. It creates the classification model in the form of decision tree. Decision tree are constructed only using those attributes that are best able to differentiate the concept of learned. C4.5 is categorized into three groups of algorithm: C4.5, C4.5-no-pruning and C4.5-rules. In this paper we use basic C4.5 algorithm.C4.5 implements greedy (i.e., non-backtracking) approach in which decision trees are build in a top- down recursive divide and conquer manner. The tree building starts with a training set of tuples and their corresponding class labels. The training set is recursively partitioned into smaller subsets as the tree is being built.

Pseudo Code:

```
1. Check for base cases.
2. For each attribute att, calculate the normalized Information Gain for splitting an attribute.
3. Out of this select the best attribute att which has the highest information gain.
4. Find a decision node that splits the best att, as root node.
5. Recurs on the sub lists obtained by splitting on best of att and add those nodes as children node [17].
```

C 5.0:

C5.0 algorithm is an enhancement of C4.5 which follows the rules of C4.5 algorithm. C5.0 is the classification algorithm which we can apply on large data set. C5.0 is better than C4.5 based on efficiency and the memory. C5.0 works by splitting the sample dataset based on the field which gives us the maximum Information Gain value. The C5.0 algorithm split samples on premise of the greatest Information Gain field. The sample subset that we get from the previous split will be split afterward. This procedure will proceed until the point that the sample subset cannot be further split and is usually according to another field. At last we look at the lowest level split, those sample subsets which don't have significant contribution to the model will be discarded. The splitting in this algorithm is done on the basis of Information Gain. This parameter is used to find the gain originated by a split over an attribute [14].

C5.0 algorithm has features like:

1. We can view a large dataset as a set of rules which we can easily understand.
2. In C5.0 algorithm we get the acknowledgement on noise and missing data.
3. C5.0 algorithm solves the problem of over fitting and error pruning.
4. In C5.0 classifier can foresee the relevant and irrelevant data. [15].

CART Algorithm:

Classification and Regression Trees (CART). At the point when the value of the target attribute is ordered, it is called regression tree and when the value is discrete, it is called classification tree. This algorithm influences utilization of binary tree to divide the forecast space into various subsets. Tree's leaf nodes compare to various division areas which are dictated by Splitting Rules relating to each internal node. CART uses GINI Index to determine in which attribute the branch ought to be generated. The technique is to pick the attribute whose GINI Index is minimum after splitting.The decision-tree generated by CART algorithm is a simple structured binary tree[14].

Random Forests Algorithm:

Random Forests is a new approach to explore the data, data analysis and to do predictive modeling. It has its roots in CART. Random Forests provide visualization of data for high dimensional dataset. It also offers anomalies, outlier, error detection and automated identification of important predictors. A random forest is a collection of CART-like trees which follows specific rules for tree growing, tree combination, self-testing, post-processing. Random Forests use binary partition in which each parent node is split into no more than two children. Following are the steps for Random Forests split selection:

First of all randomly select a small subset of available variables

1. It is a bootstrap subsample which is like considering a 50% sample from the original training data. Mostly we select square root of (K) where K is the aggregate number of predictors available. That implies If we have 520 columns of predictors we will choose just around 23

2. We split our node with the best variable among the 23, not the best variable among the 520
3. Radically accelerates tree developing procedure

The best splitter from the picked random subset is utilized to split the node in question.

RF tree evolution:

- Once a node is split on best eligible splitter the process is repeated in its entirety on each child node
- For each node, select randomly a new list of eligible predictors.
- With a large number of predictors the eligible predictor set will be quite different from node to node [19].

III. GENETIC ALGORITHM

Genetic algorithm is an adaptive heuristic method which is based on the “survival-of-the-fittest” principle; genetic search proves particularly effective when the search space is very vast for classical search methods to examine efficiently. The genetic algorithms attempt to locate a best or great answer to the problem by genetically breeding the population of individuals. The genetic algorithm convert a population of individual objects, each with an associated fitness value, into a new generation of the population using the Darwinian principle of reproduction and survival of the fittest and naturally occurring genetic operations such as crossover and mutation[1]. Each individual in the population represents a possible solution to a given problem[2][3].Before running Genetic algorithms, we will define a relevant encoding of chromosome with different attributes to solve a problem, select an objective function for fitness, and construct genetic operators. In order to run Genetic algorithms, we have generated an initial population consisting of chromosomes having different attributes and evaluated these chromosomes using the defined objective function. And then we select two chromosomes randomly and apply crossover and mutate them and replace a low quality chromosome with a new one of high quality. Higher the fitness value, higher will be the chance that it will survive in next generation. With the recurring of this process, the population will comprises of great quality chromosomes.

Genetic Algorithm Pseudocode:

1. The algorithm begins by creating initial population of random chromosome strings..

2. The algorithm then creates new populations, or generations by applying various genetic operators. At each step, the algorithm uses the individuals in the current generation to create the next generation. To create the new population, the algorithm performs the following steps:

a. Find fitness value of each chromosome in population.

b. Find average fitness value of population.

c. Select strings based on their fitness for reproduction.

d. Select two strings for crossover. Also select crossover site for mating.

e. Mutation is the third operator used in GA process. Mutation involves the modification of few bits of a chromosome with some mutation probability.

3. The algorithm process terminates when one of the stopping criteria is met [17].

A. GA Operations

Selection and reproduction operator copies the individuals with the best fitness value. The method used for selecting individual string for next generation is roulette wheel reproduction [5]. Crossover is one of the genetic operators that mix two chromosomes strings together to form new offspring. Purpose of crossover operator is exploration of a new solutions and exploitation of old solutions. GA constructs a better solution by applying crossover operator on strings. The candidate’s string having upper fitness value has more priority to be nominated than lower fitness value, so good solution always alive to the next generation. Single point crossover is used to interchange the weights of attributes of two chromosomes, which are candidate for this genetic process [6].

Single Point Crossover Example

```
        String 1      1 0 0    1 0 0 1 0 1 0

        String 2      0 0 1    0 1 1 0 1 1 1
After Crossover:
        String 1      1 0 0     0 1 1 0 1 1 1
        String 2      0 0 1     1 0 0 1 0 1 0
```

The crossover and mutation probability is set by the user [8].Mutation is the third operator used in our GA systems. Mutation encompasses the change of the gene values of a solution with low mutation probability. Mutation operator restores lost information or adds information to the population as shown in given example.

String 1 1 0 **0** 0 0 0 0 0 1 0 0 1 1

New string after mutation 1 1 0 **1** 0 0 0 0 0 1 0 0 1 1

Chromosome may be better or poorer than old chromosome. If they are poorer than old chromosome then they are eliminated in selection step. The objective of mutation is restoring lost and exploring variety of data [6].In genetic algorithms mutation is randomly applied with low probability, typically in the range 0.001 and 0.01, and it modifies elements in the chromosomes. Mutation probability used in the system is 0.02 [4].

IV. DATASET ATTRIBUTES AND CLASS LABELS

We consider the dataset of educational system consisting of 7 attributes which includes GENDER, NCAT, LANG, TLANG, PPER, CPER, and SCH.

Table 2 labels the different attributes of the data and their possible values [9]. Training Dataset for Education System has been shown. Also different parameter for CPER has been shown.

This is based upon internal and external assessment. Internal assessment is dependent on MST, online assignments, attendance and class behavior. Also, we labeled the students in relation to their percentage [7][10]. And group them into three classes, “high” representing high level scholarship for the students who scores 80-99%, “middle” representing middle level scholarship who scores 60-79.9%, and “low” representing low level scholarship who scores less than 60%.We also labeled the students in relation to their percentage and group them into two classes “yes/no” depending upon whether the student is eligible for scholarship or not. Classification process involves these steps: Students data collection, Pre-Processing, Indexing, feature Selection and Classification of data with class labels and Optimization with GA as depicted in Figure 4

TABLE 2. Training Dataset for Educational System

Dataset Attributes	Description	Possible values
GENDER	Student's gender	{Male, Female}
NCAT	Nationality category	{Indian, NRI }
LANG	First Language	{Hindi, Punjabi, English, Other}
TLANG	Teaching language in the university	{Hindi, Punjabi, English}
PPER	Previous class %age	{Excellent (90% to 100%), Very Good (80% to 89.9%), Good (70% to 79.9%), Average (55% to 69.9%) Poor(below 55) }
CPER	Current semester Percentage	{Excellent (90% to 100%), Very Good (80% to 89.9%), Good (70% to 79.9%), Average (55% to 69.9%) Poor (below 55) }
Parameters for CPER: **Internal(40)** MST(15) Online Assignments(10) Online Attendance(10) Class Behaviour(5) **External Exams:60**		
SCH	Does the student have any scholarship on the basis of **PPER+CPER**	{Yes, No} If yes then SCH= **High**(50% of Admission fee) **Medium** (35% of admission fee) **Low**(25% of Admission fee)

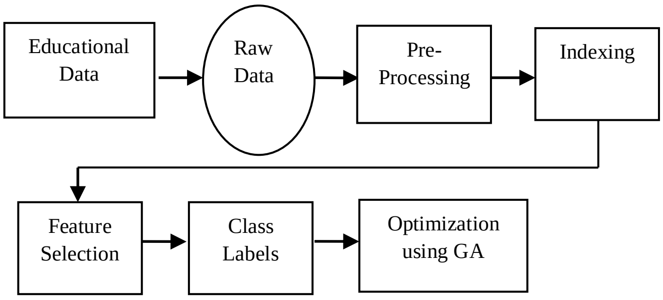
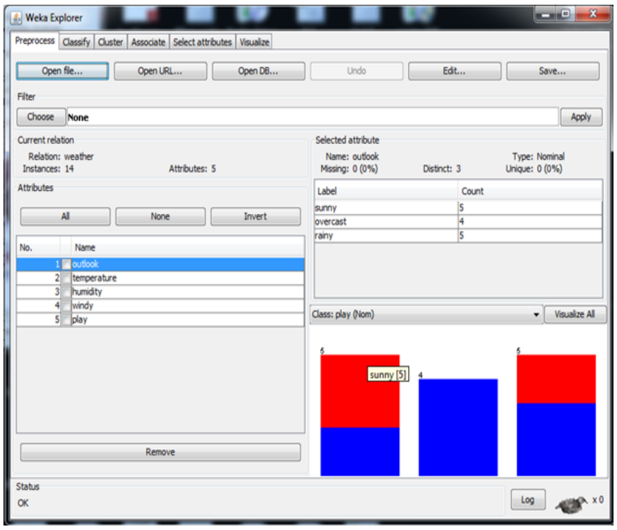


Fig.4: Classification Process

V. WEKA AND MATLAB TOOL

WEKA makes learning applied machine learning easy and efficient. It is a Graphical user interface tool that provides us various options like loading the datasets, running the algorithms, designing and running experiments. WEKA provides a number of small common machine learning datasets that we can apply to experiment our algorithms. After loading a dataset, it's time to select a machine learning algorithm to model the problem and make analysis. Steps for classification in WEKA Tool are [20]:

- Preparing the data
- Data is loaded
- Choose Algorithm
- Evaluating the output.



Filter: None

Relation: weather
Instances: 14 Attributes: 5

No.	Name
1	outlook
2	temperature
3	humidity
4	windy
5	play

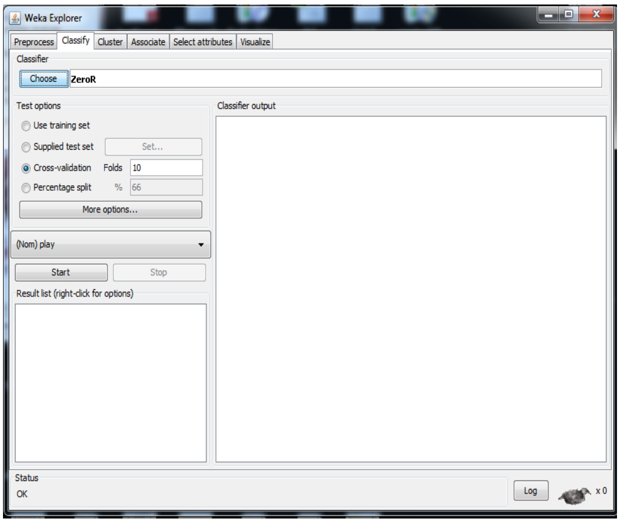
Name: outlook Type: Nominal
Missing: 0 (0%) Distinct: 3 Unique: 0 (0%)

Label	Count
sunny	5
overcast	4
rainy	5

Class: play (Nom)

OK

Fig.5: WEKA Classification Tool



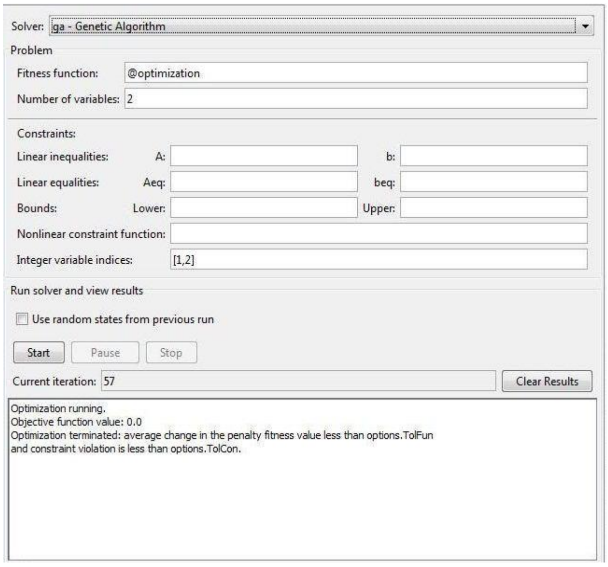
ZeroR

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) play

OK

Fig.6: WEKA Classification Tool



Solver: ga - Genetic Algorithm

Problem

Fitness function: @optimization

Number of variables: 2

Constraints:

Linear inequalities: A: b:

Linear equalities: Aeq: beq:

Bounds: Lower: Upper:

Nonlinear constraint function:

Integer variable indices: [1,2]

Run solver and view results

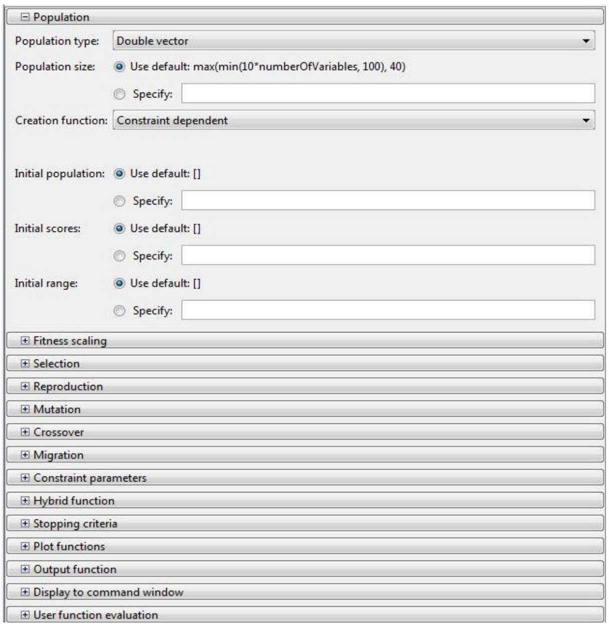
☐ Use random states from previous run

Current iteration: 57

```
Optimization running.
Objective function value: 0.0
Optimization terminated: average change in the penalty fitness value less than options.TolFun
and constraint violation is less than options.TolCon.
```

Fig.7: MATLAB Tool for GA Optimization

The MATLAB tool comes with large numbers of libraries for performing various matrix operations, numeric methods and plotting of data. MATLAB become the first preference of every software developer to work upon their GUI, scientific and mathematical based applications. For GA implementation MATLAB is come with tool that is GA-tool [21]



⊟ Population

Population type: Double vector

Population size: ● Use default: max(min(10*numberOfVariables, 100), 40)

○ Specify:

Creation function: Constraint dependent

Initial population: ● Use default: []

○ Specify:

Initial scores: ● Use default: []

○ Specify:

Initial range: ● Use default: []

○ Specify:

Fig.8: MATLAB Tool for GA Operators

VI. CONCLUSION

We studied various classification algorithms and Genetic Algorithm. The presented discussion on Educational Data Mining (EDM) from Educational databases is a review of the ongoing research in this area. It point to interesting directions of our research, where the aim is to apply hybrid approach of classification schemes and Genetic Algorithms using WEKA and MATLAB. The approaches in review are diverse in data mining classification methods and Genetic Algorithms.

VII. FUTURE SCOPE

The future scope of our study is to evaluate the performance of Educational data mining using different classification methods and to optimize the result using Genetic Algorithms. We can also include more classifiers like Naive Bayes, QUEST and collaborate multiple classifiers together. Also we can apply other evolutionary algorithms for best optimization of results.

VIII. REFERENCES

[1]. Donald H. Kraft, Frederick E. Petry, Bill P. Buckles,

[2]. ThyagarajanSadasivan,“The use of genetic programming to build queries for information retrievals”, IEEE, pp 468-473, 1994.

[3]. Goldberg, D. E, “Genetic Algorithms in Search, Optimization and Machine Learning”, Addison-Wesley Publishing Co, 1989.

[4]. Navneet Kaur, Jaspreet Singh Budwal, “Search Optimization Using Genetic Algorithms”, Fifth International Conference on Neural Networks and Artificial Intelligence, (ICNNAI-08) at Minsk, Belarus, pp. 302-305, 2008.

[5]. Navneet Kaur, Jaspreet Singh Budwal,“Intelligent WebSearch Optimization with reference to Mutation Operator of Genetic and Cultural Algorithms Framework”, 2014 IEEE International Conference on Advanced Communication, Control and Computing Technologies, pp. 619-623, 2014.

[6]. Abdul Kadar Muhammad Masum, Mohammad Shahjalal, Md. Faisal Faruque, Md. IqbalHasanSarker, “Solving the Vehicle Routing Problem using Genetic Algorithm”, International Journal of Advanced Computer Science and Applications, Vol. 2, No. 7, pp. 126-131, 2011.

[7]. Ahmed A. A. Radwan, Bahgat A. Abdel Latef, Abdel Mgeid A. Ali, and Osman A. Sadek, “Using Genetic Algorithm to Improve Information Retrieval Systems”, World Academy of Science, Engineering and Technology 17, pp. 6-12, 2006.

[8]. BehrouzMinaei-Bidgoli, William F. Punch III, “Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System”.

[9]. Cristóbal Romero, Sebastián Ventura, Carlos de Castro, Wendy Hall, and Muan Hong Ng, “Using Genetic Algorithms for Data Mining in Web based Educational Hypermedia Systems”.

[10]. Amjad Abu Saa , “Educational Data Mining & Students’ Performance Prediction”, International Journal of Advanced Computer Science and Applications, Vol. 7, No. 5, pp. 212-220, 2016.

[11]. SrečkoNatek, MotiZwilling, “Data Mining for Small Student Data Set – Knowledge Management System for Higher Education Teachers”, Management, Knowledge and Learning International Conference, pp. 1379-1389, 2013.

[12]. Gurpreet Singh, Jaskaranjit Kaur & MD. Yusuf Mulge, “Performance Evaluation of Enhanced Hierarchical and Partitioning Based Clustering Algorithm (EPBCA) in Data Mining”, 2015 International Conference on Applied and Theoretical Computing and Communication Technology.

[13]. Jaskaranjit Kaur and Gurpreet Kaur, “Clustering Algorithms in Data Mining: A Comprehensive Study”, 2015 International Journal of Computer Science and Engineering, Vol.-3(X), pp57-61, July 2015, E-ISSN: 2347-2693.

[14]. I.Bhuvana, Dr.C.Yamini,“ Survey on classification algorithms for data mining: (comparison and evaluation)” International journal of Advance Research in Science & Engineering, Vol.N0.4, Special Issue(01), August 2015.

[15]. Prof. Nilima Patil, Prof. Rekha Lathi, Prof. Vidya Chitre, “Comparison of C5.0 & CART Classification algorithms using pruning technique”, International Journal of Engineering Research & Technology (IJERT) Vol. 1 Issue 4, June 2012.

[16]. Rutvija Pandya, Jayati Pandya, “C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning”, International Journal of Computer Applications, Vol 117 -No. 16, May 2015.

[17]. https://www.safaribooksonline.com/library/view/learning-data- mining/9781784396053/ch03s02.html

[18]. S. Behzadi, Ali A. Alesheikh, “A Pseudo Genetic Algorithm For Solving Best Path Problem”, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B2. Beijing 2008.

[19]. Cristobal Romero and Sebastian Ventura, “Data mining in education”, WIREs Data Mining and Knowledge Discovery, Volume 3, January/February 2013.

[20]. Eesha Goel, Er. Abhilasha, “Random Forest: A Review”, International Journal of Advanced Research in Computer Science and Software Engineering, Vol 7, Issue 1, January 2017

[21]. Kulwinder Kaur1, Shivani Dhiman , “Review of Data Mining with Weka Tool” ,International Journal of Computer Sciences and Engineering, Vol4, Issue-8

[22]. Manish Saraswat, Ajay Kumar Sharma, “Genetic Algorithm for optimization using MATLAB”, International Journal of Advanced Research in Computer Science, Vol 4, Issue 3, 2013.