Missing Data Imputation using Genetic Algorithm for Supervised Learning

Waseem Shahzad
National University of Computer and Emerging Sciences, Islamabad, Pakistan

Qamar Rehman
National University of Computer and Emerging Sciences, Islamabad, Pakistan

Ejaz Ahmed
National University of Computer and Emerging Sciences, Islamabad, Pakistan

***Abstract*—Data is an important asset for any organization to successfully run its business. When we collect data, it contains data with low qualities such as noise, incomplete, missing values etc. If the quality of data is low then mining results of any data mining algorithm will also below. In this paper, we propose a technique to deal with missing values. Genetic algorithm (GA) is used for the estimation of missing values in datasets. GA is introduced to generate optimal sets of missing values and information gain (IG) is used as the fitness function to measure the performance of an individual solution. Our goal is to impute missing values in a dataset for better classification results. This technique works even better when there is a higher rate of missing values or incomplete information along with a greater number of distinct values in attributes/features having missing values. We compare our proposed technique with single imputation techniques and multiple imputations (MI) statistically based approaches on various benchmark classification techniques on different performance measures. We show that our proposed methods outperform when compare with another state of the art missing data imputation techniques.**

***Keywords*—genetic algorithm; information gain; missing data; supervised learning**

I. Introduction

Data is available in every sphere of life which is collected and used for various purposes. Processing and analysis of the collected data after being processed usually provides useful insights and knowledge about the system which has produced such data. The field of data mining basically deals with mining useful information from raw data instead of using all the data that also has some unimportant information. Data mining is a collection of techniques used for extracting or mining of previously unknown, useful and understandable patterns from large databases. Data mining integrates techniques from multiple disciplines such as database technology, machine learning, statistics, pattern recognition, neural networks, and image processing and data visualization. There is always a requirement for efficient and scalable data mining algorithms and it is a subject of ongoing research [1].

The process of data mining is to extract information from data. The first step is to extract data from the database and then perform preprocessing steps on it. Data mining techniques are used to extract data patterns. Evaluation and presentation mean to represent the knowledge which is understandable to users. The result is the empowerment of users with knowledge.

There are different data mining techniques including supervised classification, association rules mining or market basket analysis, unsupervised clustering, web data mining, and regression. One important technique of data mining is the classification of data. The objective of classification is to build one or more models based on the training data, which can correctly predict the class of test objects. There are several problems with a large scale of domains which can be cast as classification problems [1]. The classification has several important applications in our lives [2-5]. Examples include customer behavior prediction, portfolio risk management, identifying suspects, medical applications, sports and fraud detection etc. This research deals mainly with the data preprocessing evaluated on the basis of classification technique of data mining.

One of the challenging problems is to transform huge amount of data into an accessible and actionable knowledge. This knowledge is utilized by domain experts for decision making. Therefore, the core focus is on the knowledge discovery process in the databases (KDD). KDD is defined as a non-trivial process of identification and extraction of implicitly, previously unknown, and potentially useful information from the data [1].

The collected data may contain several states of the art deficiencies such as missing values, non-discredited data, inconsistent, incomplete and noise etc. If data is not of high quality it may hinder the discovery of useful patterns later in the process. The main purpose of the preprocessing step is to enhance the quality of data used in the experiment. All the data mining techniques are applicable once the data has been preprocessed and the objective of preprocessing is simple. Data collected from the real world is dirty and needs to be cleaned. The word dirty in data perspective means state of the art deficiencies described earlier. There can be various reasons due to which these issues arise, overcoming these problems is done by using KDD process, and there are different techniques that are proposed by various researchers which we will describe later in this paper.

In this paper, we address an important area of data preprocessing which is missing values imputation. Missing values in a dataset mislead the learning model. We have proposed a new approach based on GA and IG to impute the missing values. The proposed technique has been evaluated on different classification methods. The proposed technique has a higher accuracy rate and is well suited for large dimensional search spaces with a higher rate of missing values.

The rest of the paper is organized as follows. Section 2

describes the background of missing values, section 3 presents different classification algorithms, section 4 provides a detail description of proposed technique, section 5 presents experimentation results and finally section 6 concludes proposed technique and gives some future directions.

II. Background of Missing Data Imputation

A. Importance of Complete Datae

Basically, in data mining, the focus is on extracting useful information from a large amount of data that is collected from various sources and to take decisions using such data. Decisions are made on the basis of science, business and economic approaches on data available. As an example, sales and other information allow business class and investors to evaluate and make critical decisions regarding their investments with their future outcomes, whereas advances in research are based on the discovery of knowledge from various experiments and measured parameters.

During fault detection and identification, it is observed that most data is corrupt or incomplete. Predictive models that take observed data as an input are used for many decision-making processes, such models do not tolerate any incompleteness in data provided for prediction and as a result, such models are normally broken down. In many applications, simply ignoring the incomplete record is not an option. Most decision-making tools such as the commonly used neural networks, support vector machines, and many other computational intelligence techniques cannot be used for decision making if data is not complete. This is mainly due to the fact that ignorance can lead to biased results in statistical modeling or even damages in machine control [6]. For this reason, it is often essential to making the decision-based approach on available data [7].

The challenge missing data pose to the decision-making process is more evident in on-line applications where data have to be used almost instantly after being obtained. The biggest challenge is that the standard computational intelligence techniques are not able to process input data with missing values and hence, cannot perform classication or regression. Some of the reasons for missing data are sensor failures, omitted entries in databases and on- response to questions in questionnaires. There have been many techniques reported in the literature to estimate the missing data for some applications [7]. There are several reasons why data might be missing, and missing data may follow an observable pattern. Exploring the pattern is important and may lead to the possibility of identifying cases and variables that eect the missing data [7, 8]. A proper estimation method can be derived by identifying the variables that predict the pattern.

B. Missing Data Mechanisms

Missing data randomness is divided into three classes [9] such as missing completely at random missing at random, not missing at random [5] and missing data handling techniques (Ignoring data).

To discard data with missing values two core methods are used. One is called complete case analysis. It is available in every one of statistical packages and is the default method in many programs. The other method is discarding instances or attributes called listwise deletion. In this method, the level of missingness is determined on each instance and attribute and deletes the instances or attributes with high extents of missing data. Prior to deleting any attribute, it is vital to evaluate its connotation to the investigation. The methods, complete case analysis and discarding are executed only if missing data is missing completely at random. The missing data that are not missing completely at random contain non-random elements that may prejudice the results [9]. The deletion can bring in significant bias into the experimentation. In addition, the reduced sample size can significantly hamper the analysis. The thumb rule for deletion instances is, if attributes have more than 5

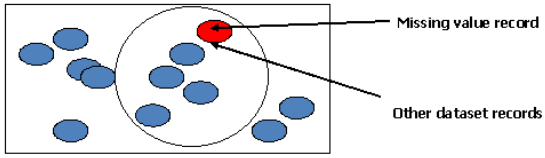

Fig. 1. KNN search space

1) Mean-fill approach: Most common technique in missing data imputation is finding the estimates of the values and then these estimates are replaced with the missing entries, the focus of our work is related to the estimation of values and its comparison to proposed technique. These estimates include statistical calculation i.e., means, zero filling, min replacement and max replacement.

These estimation techniques are used in datasets with missing values as observed values results are observed in the form of classifiers accuracy and other output measures like precision, recall, f-measure and Area under ROC. The main reason of calculating other results is just because if the classifier does not satisfy the accuracy reported. Then these measures can also be observed in the case of finding a better result.

Mean-fill approach is one of the most common statistical estimation approaches that is actively used as filling up missing values attributes of data with missing values, which is provided by various open source data mining toolboxes or packages. Also in latest researchers are using comparison technique and their majority cases research provides promising results. But it is observed that for data with a large amount of missing information this approach do not work very well.

Mean of the attribute values (in case, of numeric values, for discrete values MODE is taken) set is taken and all the missing values are replaced by the mean value in that particular attribute, similar is the case for all attributes for any dataset. Min fill approach, Max fill approach, Zero fill approach and K-Nearest Neighbor approach [10] are most common approaches being used.

K-Nearest Neighbors are determined on the bases of some kind of distance between points. It has the biggest disadvantage since it looks for the most similar instances, the whole dataset should be searched. On the other hand, how to select the value k and the measure of similar will impact the result greatly.

2) Multiple imputations (MI): It is one of the most attractive methods for general purpose handling of missing data in

the multivariate analysis. Rubin [9] described MI as a three step process, imputation, analysis, and pooling.

The most challenging step is imputation, that is, the construction of the m-completed datasets. This step accounts for the process that causes the creation of the missing data. First, sets of plausible values for missing values are created using an appropriate model chosen, reflects the uncertainty due to the missing data. Each of these sets of plausible values is used to fill-in the missing values and creates a completed dataset. Typical problems are:

- Missingness could be related to the value of information (e.g., people with higher incomes tend to skip income questions more often).
- Missing entries can appear anywhere in the data.
- The method used in the imputation step must foresee the intended complete-data analysis.

The repeated ANALYSIS step on the imputed data is actually somewhat simpler than the same analysis without imputation because there is no need to bother with the missing data. Each of these datasets can be analyzed using complete data methods.

The POOLING step consists of computing the mean over the m repeated analysis, its variance, and its confidence interval or P value. Results are combined finally. In general, these computations are relatively simple.

There are various ways to generate imputations. The implementation program for MI of continuous multivariate data (NORM) is available in [12]. However, it is not necessarily true that any particular method will perform better for any particular empirical study. It is well known that methods for handling nonignorable data require the analyst to make assumptions about the model of missingness [11]. Recent overviews of NMAR modeling are given in [13, 14, 15]. Selection and Pattern mixture models are used for NMAR data Models need more statistical formulas to impute the data. If the chosen model is incorrect then MNAR model may perform even less well than standard MAR methods [9]. Different types of weighting methods are also used for non-ignorable missing data. Even though many methods are available, they could not be used by researchers due to lack of familiarity and computational challenges and researchers often opt for ad-hoc approaches that may do more harm [7].

3) Auto-associative Neural Networks: An auto-associative referred to as autoencoder neural network is a specific neural network, trained to recall its inputs [19]. Given a set of inputs, the network predicts these inputs as outputs and thus has the same number of output nodes as there are inputs. However, the hidden layer is characterized by a bottleneck, with fewer hidden nodes than output nodes.

The smaller hidden layer projects the inputs onto a smaller space, extracting linear and non-linear interrelationships such as covariance and correlation, from the input space and also removes redundant information [19]. This means that they can be used in applications to recall the inputs and missing data estimation applications.

III. Classification

Data mining learning models are categorized into two, the one in which class to which training sample is known while there is a learning stage; it is called labeled training data. The predictive models are built on the basis of supervised learning data, whereas unlabeled data is used to test the model. One example is the classification method in which class labels are known. Other is unsupervised learning method where the class label for the training data is unknown. Here the training data is grouped according to their similarities, clustering is the example of unsupervised learning where data is unlabeled.

A fundamental aim of this research work in the field of classification is to perform preprocessing on data available and to make clean data available to the classifiers highly accurate models from the available data that can be learned. Other objective includes verification of correctness of proposed technique on the basis of classification results. Decision Tree (C4.5), PART, NB-Tree and RIPPER are the most common classifiers used in the field of machine learning and these are also used in this research [23, 24, 25, 26].

IV. Proposed Technique

We have used GA with IG for imputation of missing values. Following subsections will describe the proposed technique.

A. Genetic Algorithm

GAs are basically evolutionary ideas of natural selection and genetics [16, 17]. GAs are adaptive heuristic search algorithm. Inspired by Darwins theory of evolution survival of the fittest, it is common in nature that in a competition where individuals are looking for resources fittest individuals dominate over weaker ones. Evolutionary computing today holds GAs as one of the important parts. Among random search methods employed to solve optimization problems, GAs represent an intelligent structure which is easy to implement.

For any particular problem GAs works for solving it is by mimicking processes nature use, like selection, crossover, mutation and acceptance, to evolve a good solution for that problem.

1) Operators of GA: GAs use genetic operators to maintain genetic diversity. It is important that genetic diversity or variation is maintained for the process of evolution. Inspired by natural genetic structure, genetic operators are the same. Following are operators used in genetic algorithms.

1) *Reproduction/ Selection:* Usually, the first operator applied on population is a reproduction, from the population the chromosomes are selected to be parents for the crossover step and producing offsprings.
   According to Darwins theory survival of fittest, the best ones should survive and create new offsprings. Reproduction operator is also called selection operator because it is basically extraction of genes subset from existing population based on some quality criteria or definition. The fitness function is the quality measurement that can be performed to select best genes subset, as every gene contains some meaning.
2) *Crossover/ Recombination* This genetic operator is called crossover because it mates (combines) two

parents (chromosomes) to produce a new offspring (chromosome).
Most commonly used methods for the selection of parents to crossover are:

- Roulette wheel selection.
- Rank selection.
- Boltzmann selection.
- Steady state selection.
- Tournament selection.

The idea behind crossover is that after mating any chromosomes (parents) that are selected based on some function, offsprings (chromosomes) will be fitter as they are derived as a result of best characteristics of their parents. According to user-defined crossover probability, it takes place during evolution stage.

3) *Mutation:* During the evolution stage mutation occurs where the user defines mutation probability, this probability is usually set to a fairly low value, like 0.01 is a good first choice. Mutation is the genetic operator used to maintain genetic diversity from one generation of a population of chromosomes to the next generation.

B. Proposed Technique

This section provides detail of the proposed technique along with fitness function used

1) General Description: GA is used for missing data imputation, the importance of missing data imputation varies from problem to problem, and we use this technique to clean the dirty data for classification problem. The missing values are imputed in the datasets using GA and GA is run for each attribute which is treated as a chromosome. We divided these chromosomes into frames for further accurate measures; these frames are explained by the example in the following section. Frames are dependent upon the no of classes in the dataset. i.e., there is n number of class labels in a dataset.

The flow chart describes the working of proposed technique as shown in figure 2. Using attribute instances first we create an initial solution of population size defined in parameter section. Evaluate the fitness of each solution. Check termination criteria for a maximum number of generations. For generation number 1 initial size of the new population is 0. Select individuals randomly from the population according to tournament size for selection using tournament selection. Select genetic operator to be applied to the selected individuals probabilistically. Perform crossover or mutation on the selected individual's bases on the probability of selection for crossover or mutation. The resultant of the genetic operator is inserted in the new population. Check for population size on every iteration, if population size is equal to maximum population size then start a new generation and check for termination criteria else continue to select new individuals from the current population. When the population size is reached maximum new generation become started, if the current population has the fitness of individual then how previous populations best fitted then we keep that individual from the previous population in current population (Elitism= keep best).

To illustrate how GA works in improving data quality by imputing missing values based on estimation, the following is

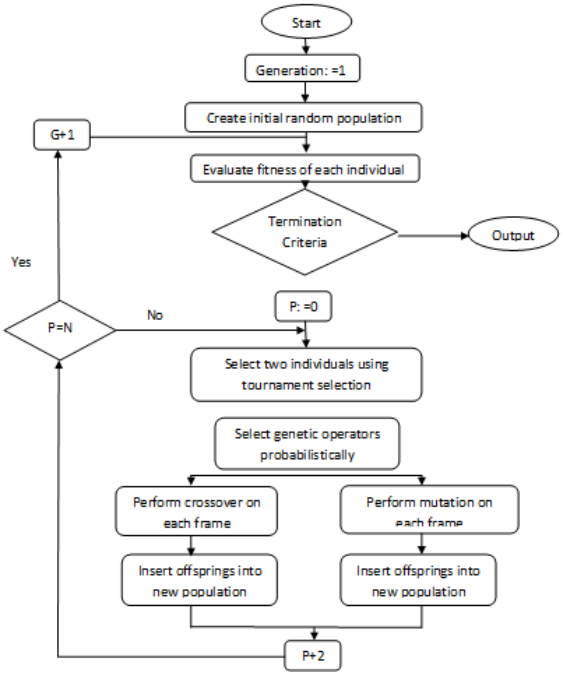

Fig. 2. Flow Chart of proposed method

Attr#1	Attr#2	Attr#3	Attr#4	Attr#5	Class
5	-1	3	5	9	1
4	43	1	1	-1	1
5	58	4	5	3	1
-1	28	1	1	3	0
5	74	1	5	-1	1
-1	59	-1	4	5	1
4	-1	2	1	-1	0
5	51	4	-1	-1	1
4	39	1	1	-1	0
11	44	4	4	3	0
4	53	-1	3	4	0
2	-1	1	3	0	1
9	22	3	0	6	0
2	54	4	-1	-1	1
2	45	1	6	-1	0
1	65	-1	4	-1	1

Fig. 3. Sample Datasets

a simple example that describes the technique. This sample dataset is given in figure 3 is a part of dataset named Memographicmasses. Remember -1 represents missing values.

A chromosome split into n number of frames (sub-chromosome). N is the number of classes in the dataset. Each frame initialized independently to another frame, within restricted range that it must contain values obtained by the attribute to a specific class, in the first generation. Merging all frames into one makes the valid structure of a chromosome in population.

Each frame is treated as full fledged independent chromosome at the time of applying genetic operators on it. One point crossover used on each frame so n number of cross points are used for every chromosome. Mutation operator mutates randomly n number of genes depending on the probability of mutation criteria. Each gene belongs to a specific frame so, during mutation of a gene, gene value is replaced by a specific set of domain values of class from the dataset.

1) *Structure of Chromosome:* The data illustrated above belongs to two class problem so N = 2. The number of the frame will be two in each chromosome as shown

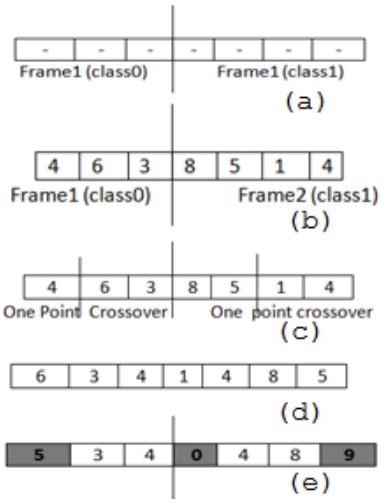

Fig. 4. Chromosome Structure

in figure 4 (a-Chromosome Structure). The size of the frame depends on missing values related to specific class as explained in the previous section.
According to figure 3 let us take the case of attribute # 5, the range of selecting values for frame1 will be min gene max. According to dataset values of the gene are assigned between 3 gene 6. Similarly for frame 2 the values between 0 gene 9. As shown in figure 4 (b- Values of genes assigned).

2) Operate on genetic operators:

1) *Crossover:* One point crossover is performed. Figure 4 (c- Crossover Performed) shows the crossover points on the chromosome. As a result of crossover, offspring is created that is shown in figure 7. The chromosome in figure 4 (d- Result of crossover) shows the result of crossover operation performed on it in the previous step.
2) *Mutation:* In mutation gene values mutate according to the domain of each frame defined according to the range of distinct values that are available in the particular data attribute, here figure 8 illustrates the outcome of the mutation operation performed on chromosome shown in figure 4 (e- Mutation performed). After mutation is performed the fitness of the resultant chromosome is calculated if it is greater than the previous data fitness the values are saved and next iteration takes its place, until the termination criteria are met that can be the end of condition or some value which achieved terminates GA.

C. Fitness function

In the proposed GA for missing data imputation, we are using IG as Fitness function that is based on the entropy of each attribute regarding its class label in the given dataset. Remember that when we calculate the fitness of an attribute then the whole attribute is used for the calculation of fitness after imputing missing values.

Following is the brief description of Fitness function used in the proposed work.

$$\mathbf{H(X)} = -\sum_i \boldsymbol{P(x_i) log_2}\ (\boldsymbol{(P(x_i))} \qquad \textbf{(1)}$$

$$\mathbf{H\ (X|Y)} = -\sum_j \boldsymbol{P(y_j)} \sum_i \boldsymbol{P(x_i|y_j) log_2 (P(x_i|y_j))} \qquad \textbf{(2)}$$

$$\mathbf{IG\ (X|Y)} = \mathbf{H(X)} - \mathbf{H\ (X|Y)} \qquad \textbf{(3)}$$

IG is a correlation-based measure. It is based on an information theoretical concept of entropy i.e. a measure of the uncertainty of a random variable. Following is the equation of entropy of X eq(1).

The entropy of X when the value of another random variable Y is known, following is the conditional entropy eq(2).

In the above equation (2), P(x_i) is the prior probability for all the values of X. P(x_i—y_j) is the posterior probability of X after the values of Y are known. The amount of which the entropy of X decreases, it depicts the decrease in uncertainty level. This is achieved through the additional information regarding X provided by Y. This measure is called IG. Following is the formula for information gain eq(3).

V. Experimentation and Analysis

A. Experimentation Framework

The population size is defined as 500, for 100 generations and tournament size is kept 6. These parameters setting have been chosen after performing several experiments. All the other parameters used are defined in the following table 1. Experimentation has been performed with different combinations of these parameters and best values are kept same for all the experimentation as shown in table 1.

In the experimentation, worth of a missing data imputation through GA is evaluated on 5 key measures, i.e. predictive accuracy, along with precision, recall, f-measure, and ROC.

Following table 2 elaborates about the datasets used in the experimentation. All the datasets used are publicly available and taken from UCI repository [31]. We have used standard implementation of MI which is available as NORM [12], and classifiers like NB tree, PART, JRIP, j48, NAVE bases and K-Nearest Implementation of these algorithms is provided by data mining software Weka [30]. All the algorithms are used with their default values and no tweaking is done over the methods. Since these algorithms are implemented by their authors, therefore, it is assumed that parameter setting is already incorporated.

Following table 2 describes datasets used for experimentation along with total number of features, the number of instances and percentage of missing values in these datasets.

The techniques that are used for comparison with the proposed method are Multiple Imputation, Mean filling, Min filling, Max fills and Zero fill. Table 3 shows a comparison of classification accuracies and their standard deviations after being imputed by various techniques including proposed technique of GA fill.

TABLE I. PARAMETERS USED IN GA

Parameters	Values
Population Size	500
Generation	100
Chromosome size	Missing values in attributes
Fitness Function	IG
Selection	Tournament
Tour Size	6
Crossover	One point crossover
Probability of crossover	0.8
Probability of mutation	0.2
Probability of gene mutate	0.5
Elitism	Keep-best
Runs	20

TABLE II. DATASETS USED FOR EXPERIMENTATION

Dataset	Total Attributes	No of Instances	%age missing
Labor	17	57	33.82
Echocardiogram	12	132	5.16
Cylinderbands	34	540	5.26
Memographicmasses	6	961	4
Colic-Horse	22	368	23.90

B. Comparison with other techniques

For comparison, four single imputation techniques have been adopted; filling missing values using mean, min, max and zero by replacing all missing data by 0, Multiple Imputation (MI) of missing data is also used for comparison. The results are compared for NB-Tree, JRIP, PART, NAVE Bayes, IBK (Lazy) and j48 (C4.5) classifiers. For performing experimentation with these classifiers we used Weka machine learning tool [30]. We used supervised discretization filter of Weka-3.4 machine learning tool [30] to discretize continuous attributes as a preprocessing step. The GA has seven user-defined parameters. The values of these parameters are given in table 1. The predictive accuracies of the compared algorithms are shown in tables 3 and 4. Ten-fold cross validation is used to obtain the results. The Bold value represents the highest accuracy achieved.

From tables 3 and 4, it is observed that missing data imputation using GA clearly out marks by 70% of the datasets than other estimation and predictive model techniques. These predictive accuracies show worth of the proposed approach. A genetic algorithm is an evolutionary algorithm and has much diversity when imputing missing values.

Our method performs better in three datasets for NAVE Bayes algorithm, similarly, for K-Nearest neighbors classifier, it achieves better accuracies on three datasets and better in four datasets for the j-48 classifier.

Table 5 shows results of precision results of proposed algorithm and single, and MI technique is presented. Different classifiers are used for classification with 10-fold cross validation. The Bold value represents the highest accuracy achieved.

In the above-mentioned table, our method performs comparable and/or better in 70% of the datasets. It can be observed that our proposed method achieves better/ comparable classification precisions as compared to single and MI techniques n most of the cases. It is observed that missing data imputation using GA clearly out marks/ comparable with other estimation and predictive model techniques.

In table 6, recall measures of proposed algorithm and single and MI technique is shown.

TABLE III. COMPARISON OF GA WITH DIFFERENT TECHNIQUES BASED ON CLASSIFIERS ACCURACIES ALONG WITH STANDARD DEVIATIONS

Datasets	NB-Tree						
	With missing value	GA filled	MI filled	Mean filled	Min filled	Max filled	Zero filled
Labor	80.80±15.66	**95.43±8.80**	87.53±13.38	85.47±13.99	81.07±16.70	80.73±14.86	90.07±13.20
Echocardiogram	87.29±7.70	**90.90±7.89**	89.17±7.01	86.10±8.00	88.62±7.85	86.58±8.33	88.10±7.63
Cylinderbands	67.41±6.90	68.78±6.14	66.24±6.26	68.63±6.23	**69.57±6.14**	68.39±6.10	68.37±5.76
Memographicmasses	81.85±3.04	**83.08±3.19**	82.74±3.39	81.98±3.30	82.24±3.28	82.84±3.34	82.35±3.29
Horsecolic	83.53±6.51	83.85±5.64	83.02±6.97	83.74±6.45	**83.98±5.56**	83.96±6.17	83.36±6.28
				PART			
Labor	76.77±16.28	**93.30±9.58**	87.70±15.29	85.57±15.37	81.53±15.01	79.13±12.09	83.43±12.72
Echocardiogram	88.00±7.54	**89.59±8.38**	88.09±7.85	86.92±8.71	85.55±7.86	85.82±7.97	86.35±8.47
Cylinderbands	69.56±5.67	69.93±6.57	68.81±6.11	72.39±6.76	72.52±5.66	70.89±6.34	**73.09±5.25**
Memographicmasses	81.53±3.36	**82.23±3.31**	82.43±3.53	80.56±3.83	81.46±3.66	81.15±3.32	81.46±3.34
Horsecolic	84.77±6.40	**85.34±5.70**	79.76±6.58	78.83±6.16	79.36±7.24	78.33±6.93	78.56±6.43
				JRIP			
Labor	**88.40±15.86**	86.97±11.1	85.59±8.90	81.60±15.95	80.33±16.73	84.00±14.21	79.03±14.86
Echocardiogram	83.86±8.67	**86.26±8.46**	83.58±3.50	84.23±8.29	84.47±9.21	83.67±8.88	84.03±8.47
Cylinderbands	67.02±6.29	68.37±5.99	67.59±6.51	69.67±5.90	69.91±6.34	69.98±6.11	**70.63±6.00**
Memographicmasses	82.38±3.22	**83.81±3.37**	82.99±5.64	81.89±3.41	82.60±3.36	82.55±3.43	82.79±3.31
Horsecolic	83.28±6.55	**85.15±5.28**	84.73±13.98	82.01±6.47	84.18±5.98	81.79±5.99	83.06±6.18

TABLE IV. COMPARISON OF GA WITH DIFFERENT TECHNIQUES BASED ON CLASSIFIERS ACCURACIES ALONG WITH STANDARD DEVIATIONS

Datasets	NAIVE Bayes						
	With missing value	GA filled	MI filled	Mean filled	Min filled	Max filled	Zero filled
Labor	89.67±13.38	**98.20±5.45**	91.93±11.83	92.07±11.72	90.53±10.43	87.97±13.56	65.60±18.24
Echocardiogram	89.59±7.03	**92.18±6.64**	89.05±7.24	85.19±8.62	88.92±7.85	86.56±8.01	74.64±9.77
Cylinderbands	65.67±6.35	66.24±5.91	65.44±5.00	63.72±6.93	67.09±5.75	**68.74±4.54**	67.63±4.50
Memographicmasses	**82.40±3.18**	79.09±3.49	79.16±4.17	77.26±4.19	81.12±3.24	81.29±3.32	80.84±3.39
Horsecolic	79.73±6.71	**85.83±4.67**	82.60±5.76	80.73±7.04	77.69±6.57	79.41±6.53	76.67±6.92
				K-Nearest			
Labor	82.33±16.57	90.10±10.74	**98.93±4.88**	92.57±10.30	85.87±13.38	82.83±16.68	75.63±17.15
Echocardiogra	84.03±7.83	**86.14±7.92**	86.04±7.93	85.20±8.48	85.97±7.70	82.34±9.00	84.37±8.25
Cylinderbands	68.22±5.57	70.04±5.84	68.35±5.97	70.85±5.75	75.09±5.84	75.07±5.29	**77.48±5.82**
Memographicmasses	74.69±3.78	**76.57±3.74**	74.64±3.69	75.10±3.76	75.34±4.07	73.01±3.90	75.39±4.02
Horsecolic	77.25±5.98	**81.64±5.90**	79.64±6.44	77.77±5.56	77.59±6.14	76.34±7.64	73.46±5.77
				J-48			
Labor	71.67±15.38	**89.20±11.50**	84.60±13.59	83.33±13.83	86.83±14.58	81.40±14.10	87.80±14.6
Echocardiogra	85.77±7.17	**89.31±7.09**	86.81±7.66	85.40±8.28	86.53±8.37	83.70±7.74	86.14±8.14
Cylinderbands	68.52±6.10	70.13±5.83	68.37±6.34	71.96±5.85	71.76±6.10	72.33±5.95	**73.43±6.07**
Memographicmasses	81.38±3.14	**83.00±3.03**	82.36±3.42	81.10±3.24	81.85±3.31	81.13±3.33	82.08±3.40
Horsecolic	85.15±5.78	**86.36±5.30**	84.54±6.22	83.94±6.17	82.80±6.22	83.85±5.75	82.74±5.86

In the below-mentioned table, our method performs comparable or better in most of the datasets.

In table 7, F-Measure of proposed algorithm and single, and MI technique is presented. The proposed approach also has better F-measure values in most of the datasets.

In table 8, AREA under ROC of proposed algorithm, single and MI techniques are presented. The AREA under ROC approach is high on most of the datasets.

These experimentation results of different data sets are evaluated on different benchmarks evaluation methods. This has shown the worth of proposed approach when compared with well-known missing data imputation algorithms. These results indicated that GA is a suitable method for the imputation of missing values.

TABLE V. COMPARISON OF GA WITH DIFFERENT TECHNIQUES BASED ON CLASSIFIERS PRECISION RESULTS

Datasets	NB-Tree						
	With missing value	GA filled	MI filled	Mean filled	Min filled	Max filled	Zero filled
Labor	0.88	**0.98**	0.94	0.90	0.86	0.87	0.93
Echocardiogram	0.62	**0.81**	0.73	0.63	0.69	0.61	0.65
Cylinderbands	0.60	0.66	0.60	0.68	0.68	0.68	**0.74**
Memographicmasses	0.81	0.80	0.82	0.80	0.81	**0.84**	0.81
Horsecolic	0.84	**0.89**	0.87	0.85	0.86	0.85	0.85
PART							
Labor	0.86	0.93	**0.94**	0.89	0.88	0.87	0.89
Echocardiogram	0.68	**0.74**	0.64	0.67	0.61	0.60	0.63
Cylinderbands	0.66	0.66	0.64	0.69	0.69	0.67	**0.70**
Memographicmasses	0.81	0.81	**0.83**	0.81	0.82	0.82	0.82
Horsecolic	0.86	**0.87**	0.85	0.83	0.84	0.83	0.83
JRIP							
Labor	0.91	**0.93**	0.92	0.91	0.84	0.91	0.86
Echocardiogram	0.55	**0.63**	0.64	0.58	0.58	0.58	0.55
Cylinderbands	0.62	0.64	0.63	0.67	0.67	0.67	**0.68**
Memographicmasses	0.82	0.79	**0.84**	0.81	0.83	0.83	0.83
Horsecolic	0.87	0.87	**0.89**	0.85	0.86	0.85	0.85
NAIVE Bayes							
Labor	0.93	**0.98**	0.94	0.91	0.93	0.92	0.83
Echocardiogram	0.69	**0.76**	0.87	0.60	0.68	0.62	0.14
Cylinderbands	0.63	0.66	0.69	0.59	0.67	**0.73**	0.57
Memographicmasses	**0.78**	0.73	0.75	0.72	0.78	0.77	0.78
Horsecolic	0.84	**0.89**	0.86	0.84	0.85	0.82	0.84
K-Nearest							
Labor	0.96	**0.96**	0.99	0.93	0.89	0.89	0.89
Echocardiogram	0.56	0.62	0.59	0.57	**0.64**	0.51	0.59
Cylinderbands	0.66	0.69	0.64	0.69	0.74	0.74	**0.76**
Memographicmasses	0.72	**0.73**	0.72	0.73	0.73	0.70	0.73
Horsecolic	0.78	**0.85**	0.85	0.83	0.82	0.81	0.80
J-48							
Labor	0.79	0.88	0.91	0.88	**0.89**	0.88	0.89
Echocardiogram	0.63	**0.70**	0.65	0.61	0.65	0.57	0.61
Cylinderbands	0.66	0.68	0.65	0.70	0.70	0.71	**0.72**
Memographicmasses	0.81	0.81	0.82	0.82	0.83	0.82	**0.84**
Horsecolic	0.85	**0.87**	0.89	0.87	0.86	0.87	0.86

TABLE VI. COMPARISON OF GA WITH DIFFERENT TECHNIQUES BASED ON CLASSIFIERS RECALL MEASURES

Datasets	NB-Tree						
	With missing value	GA filled	MI filled	Mean filled	Min filled	Max filled	Zero filled
Labor	0.84	**0.96**	0.87	0.90	0.88	0.85	0.94
Echocardiogram	0.79	**0.92**	0.81	0.75	0.87	0.72	.86
cylinderbands	**0.66**	0.56	0.62	0.53	0.56	0.53	0.46
Memographicmasses	0.80	**0.85**	0.82	0.82	0.81	0.79	0.82
Horsecolic	**0.91**	0.88	0.89	0.90	0.90	0.91	0.90
PART							
Labor	0.81	**0.98**	0.89	0.93	0.86	0.84	0.89
Echocardiogram	0.68	**0.73**	0.64	0.64	0.60	0.61	0.63
cylinderbands	0.59	0.60	0.61	**0.64**	**0.64**	0.63	0.64
Memographicmasses	0.80	0.81	0.79	0.76	**0.82**	0.77	0.79
Horsecolic	**0.92**	0.89	0.85	0.84	0.84	0.82	0.83
JRIP							
Labor	0.90	**0.92**	0.86	0.87	0.89	0.86	0.85
Echocardiogram	0.81	**0.90**	0.80	0.82	0.85	0.81	0.83
cylinderbands	0.58	0.57	0.59	0.57	0.59	0.59	**0.60**
Memographicmasses	0.80	**0.84**	0.80	0.80	0.83	0.79	0.80
Horsecolic	0.88	**0.90**	0.89	0.88	0.89	0.87	0.89
NAIVE Bayes							
Labor	0.92	**1.00**	0.95	0.98	0.95	0.91	0.58
Echocardiogram	0.87	**0.91**	0.87	0.78	0.90	0.79	0.15
cylinderbands	0.48	0.42	0.33	**0.51**	0.45	0.33	0.30
Memographicmasses	**0.86**	0.76	0.84	0.83	0.78	0.86	0.83
Horsecolic	0.85	**0.89**	0.86	0.86	0.79	0.87	0.79
K-Nearest							
Labor	0.75	0.90	**0.99**	0.98	0.92	0.85	0.74
Echocardiogram	0.57	**0.72**	0.59	0.59	0.62	0.64	0.57
cylinderbands	0.52	0.54	0.58	0.58	0.64	0.65	**0.70**
Memographicmasses	0.74	**0.76**	0.75	0.75	0.73	0.73	0.75
Horsecolic	**0.89**	0.87	0.85	0.82	0.83	0.83	0.78
J-48							
Labor	0.81	**1.00**	0.88	0.89	0.94	0.85	0.96
Echocardiogram	0.73	**0.83**	0.65	0.68	0.68	0.62	0.66
cylinderbands	0.55	0.57	0.55	0.59	0.59	0.58	**0.62**
Memographicmasses	0.79	**0.83**	0.79	0.77	0.83	0.77	0.77
Horsecolic	**0.93**	0.92	0.89	0.88	0.87	0.88	0.88

TABLE VII. COMPARISON OF GA WITH DIFFERENT TECHNIQUES BASED ON CLASSIFIERS F-MEASURES

Datasets	NB-Tree						
	With missing value	GA filled	MI filled	Mean filled	Min filled	Max filled	Zero filled
Labor	0.84	**0.96**	0.89	0.89	0.85	0.84	0.92
Echocardiogram	0.67	**0.79**	0.71	0.64	0.73	0.63	0.71
Cylinderbands	**0.62**	0.60	0.60	0.58	0.60	0.57	0.53
Memographicmasses	0.80	**0.82**	0.81	0.81	0.81	0.81	0.81
Horsecolic	0.87	0.87	0.87	0.87	**0.88**	0.88	0.87
PART							
Labor	0.81	**0.95**	0.90	0.89	0.85	0.83	0.87
Echocardiogram	0.65	**0.70**	0.63	0.62	0.57	0.59	0.60
Cylinderbands	0.62	0.63	0.62	**0.66**	0.66	0.64	0.66
Memographicmasses	0.80	**0.81**	0.81	0.78	0.80	0.79	0.80
Horsecolic	**0.88**	**0.88**	0.84	0.83	0.84	0.83	0.83
JRIP							
Labor	**0.90**	**0.90**	0.88	0.85	0.85	0.87	0.83
Echocardiogram	0.63	**0.71**	0.66	0.64	0.66	0.64	0.64
Cylinderbands	0.59	0.60	0.60	0.61	0.62	0.62	**0.63**
Memographicmasses	0.81	**0.82**	0.82	0.80	0.81	0.81	0.81
Horsecolic	0.87	**0.88**	0.87	0.86	0.88	0.86	0.87
NAIVE Bayes							
Labor	0.92	**0.99**	0.94	0.94	0.93	0.91	0.65
Echocardiogram	0.75	**0.81**	0.74	0.65	0.75	0.67	0.13
Cylinderbands	**0.54**	0.51	0.44	0.54	0.53	0.46	0.43
Memographicmasses	**0.82**	0.77	0.79	0.77	0.80	0.81	0.80
Horsecolic	0.84	**0.89**	0.86	0.85	0.82	0.81	0.81
K-Nearest							
Labor	0.83	0.92	**0.99**	0.95	0.89	0.85	0.78
Echocardiogram	0.53	**0.64**	0.57	0.56	0.60	0.54	0.54
Cylinderbands	0.58	0.60	0.60	0.62	0.68	0.69	**0.72**
Memographicmasses	0.73	**0.74**	0.73	0.73	0.74	0.71	0.74
Horsecolic	0.83	**0.86**	0.84	0.82	0.82	0.81	0.79
J-48							
Labor	0.78	**0.93**	0.88	0.87	0.90	0.85	0.78
Echocardiogram	0.53	**0.72**	0.60	0.61	0.63	0.56	0.59
Cylinderbands	0.59	0.61	0.59	0.64	0.64	0.64	**0.66**
Memographicmasses	0.80	**0.82**	0.81	0.79	0.80	0.79	0.80
Horsecolic	0.89	**0.90**	0.88	0.87	0.86	0.87	0.86

TABLE VIII. COMPARISON OF GA WITH DIFFERENT TECHNIQUES BASED ON AREA UNDER ROC

Datasets	NB-Tree						
	With missing value	GA filled	MI filled	Mean filled	Min filled	Max filled	Zero filled
Labor	0.88	**1.00**	0.95	0.89	0.84	0.86	0.92
Echocardiogram	0.92	**0.94**	0.92	0.88	0.93	0.89	0.92
Cylinderbands	0.70	**0.74**	0.70	0.73	0.73	**0.74**	0.73
Memographicmasses	0.89	**0.90**	0.89 |	0.89	0.89	0.88	0.85
Horsecolic	0.84	**0.89**	0.84	0.84	0.85	0.84	0.83
PART							
Labor	0.79	**0.91**	0.89	0.87	0.78	0.81	0.81
Echocardiogram	**0.93**	0.90	0.90	0.88	0.88	0.89	0.89
Cylinderbands	0.75	0.72	0.73	**0.76**	0.75	0.75	0.76
Memographicmasses	0.88	**0.89**	0.88	0.87	0.88	0.87	0.88
Horsecolic	**0.87**	0.85	0.79	0.78	0.79	0.78	0.77
JRIP							
Labor	**0.88**	0.86	0.85	0.81	0.76	0.85	0.76
Echocardiogram	0.83	**0.88**	0.83	0.83	0.85	0.83	0.84
Cylinderbands	0.66	0.69	0.67	0.70	0.71	0.71	**0.72**
Memographicmasses	0.84	**0.86**	0.85	0.84	0.84	0.84	0.84
Horsecolic	0.82	**0.85**	0.81	0.80	0.82	0.81	0.81
NAIVE Bayes							
Labor	0.89	**1.00**	0.96	0.94	0.99	0.90	0.80
Echocardiogram	0.96	**0.97**	0.96	0.91	0.96	0.93	0.71
Cylinderbands	0.70	0.70	0.71	0.69	0.71	**0.74**	0.73
Memographicmasses	**0.89**	0.86	0.86	0.85	0.89	0.89	0.89
Horsecolic	0.85	**0.93**	0.86	0.84	0.84	0.83	0.81
K-Nearest							
Labor	0.89	0.91	**0.99**	0.91	0.81	0.83	0.75
Echocardiogram	0.83	**0.97**	0.79	0.82	0.84	0.82	0.72
Cylinderbands	0.65	0.69	0.73	0.69	0.74	0.74	**0.75**
Memographicmasses	0.79	0.79	0.79	0.79	**0.80**	0.77	**0.80**
Horsecolic	0.73	**0.80**	0.78	0.77	0.77	0.74	0.71
J-48							
Labor	0.71	0.85	**0.88**	0.82	0.84	0.83	0.85
Echocardiogram	0.93	**0.93**	0.88	0.88	0.87	0.87	0.87
Cylinderbands	0.70	0.72	0.69	0.72	0.72	**0.74**	0.74
Memographicmasses	0.87	**0.87**	0.86	0.85	0.87	0.86	0.87
Horsecolic	**0.85**	0.84	0.83	0.81	0.81	0.82	0.79

VI. CONCLUSION AND FUTURE WORK

Data mining is an active area of research and in this area data is the most vital and valuable asset. Without applying automatic data mining techniques and preprocessing methods it is difficult to effectively analyze large amounts of data. Researchers are interested in finding efficient and accurate technique/method that cleans dirty and noisy data so that achieve higher accuracy rate, are comprehensible and can be learned in reasonable time, even for large databases.

In this paper, we addressed the problem of missing data imputation. First, we have elaborated on the importance of clean data (complete) in KDD. We have proposed an evolutionary technique for filling missing data on the basis of good estimation using GAs. Our main objective was to embed population-based search mechanisms to explore more search

space along with exploitation. The datasets used are standard datasets having by default missing values. We have also demonstrated that proposed technique works well for datasets with a greater percentage of missing values also for datasets where attributes are having a large range of distinct values, as GA gets into real play where there is space for more and more combination of different values. In future, we like to extend our algorithm to the domain of Noise reduction/removal.

REFERENCES

[1] J. Han, and M. Kamber, Data Mining: Concepts and Techniques, 2nd ed., Morgan Kaufmann Publishers, 2006.

[2] M. J. Berry, and G. Linoff. Data Mining Techniques for Marketing, Sales, and Customer Support. New York: John

[3] J. Pesce, Stanching hospitals, Financial hemorrhage with information technology, Health Management Technology, Vol. 24, No. 8, pp. 6-12, 2003.

[4] W. Ceusters, Medical natural language understanding as a supporting technology for data mining in healthcare Chapter 3 in: Cios K.J., eds. Medical Data Mining and Knowledge Discovery, Heidelberg: Springer-Verlag, pp. 32-60, 2000.

[5] A.C. Tessmer, "What to learn from near misses: an inductive learning approach to credit risk assessment," Decision Sciences, Vol. 28, No. 1, pp. 105-120, 1997.

[6] Roth, P. L. and Switzer III, F. S.: 1995, A Monte Carlo analysis of missing data techniques in a HRM setting, Journal of Managment 21, 10031023.

[7] Schafer, J. L. and Graham, J. W.: 2002, Missing data: Our view of the state of the art, Psychological Methods 7(2), 147177

[8] Allison, P. D.: 2002, Missing Data: Quantitative Applications in the Social Sciences, Thousand Oaks, CA: Sage.

[9] Little, R. J. A. and Rubin, D. B.: 1987, Statistical Analysis with Missing Data, Wiley, New York

[10] Batista, G. and Monard, M.C. (2003). An Analysis of Four Missing Data Treatment Methods for Supervised Learning, Applied Artificial Intelligence, 17, pp. 519-533.

[11] Graham JW, Cumsille PE, Elek-Fisk E. 2003. Methods for handling missing data. In Research Methods in Psychology, ed. JA Schinka, WF Velicer, pp. 87114.Volume 2 of Handbook of Psychology, ed. IB Weiner.New York:Wiley

[12] Multiple Imputation [Online], www.multiple-imputation.com

[13] Beunckens, C., Molenberghs, G., Verbeke, G, and Mallinckrodt, (2008). A latent- class mixture model for incomplete longitudinal Guassian data. Biometrics, 64, 96- 105.

[14] Little, R.J.(2009). Selection and patter mixture models. In Fitzmaurice, G., Davidian, M, Verbeke,G. & Molenberghs, G42 (eds.), Longitudinal Data Analysis , pp. 409-431. Boca Raton: Chapman & Hall/CRC Press

[15] Albert, P.S. & Follman, D.A. (2009). Shared parameter models.

[16] Papagelis A. and Kalles D. 2000. GAtree: Genetically Evolved Decision Trees, Proceedings 12th International Conference on Tools with Artificial Intelligence 13-15 November 2000 pages 203-206..

[17] Goldberg D.1999. Genetic Algorithms in Search, Optimization, and Machine Learning, Addison-Wesley.

[18] Nelwamondo F V, Mohamed S, Marwala T. Missing Data: A Comparison of Neural Network and Expectation Maximisation Techniques, eprint arXiv:0704.3474, April 2007

[19] Betechuoh B L, Marwala T, TetteyT, Autoencoder networks for HIV classification, Current Science, Vol 91, No 11, December 2006, pp 1467-1473.

[20] I.H. Witten, and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, 2nd ed. Morgan Kaufmann, 2005.

[21] J.R. Quinlan, Generating production rules from decision trees, in Proceedings of International Joint Conference of Artificial Intelligence, pp. 304-307, San Francisco, USA, 1987.

[22] P. Eklund, and A. Hoang, A performance survey of public domain machine learning algorithms, Technical Report, School of Information Technology, Griffith University, 2002.

[23] R. Rastogi, and K. Shim, A decision tree classifier that integrates building and pruning, Data Mining and Knowledge Discovery, Vol. 4, pp. 315344, 2000.

[24] Eibe Frank, Ian H. Witten: Generating Accurate Rule Sets Without Global Optimization. In: Fifteenth International Conference on Machine Learning, 144-151, 1998.

[25] W. Cohen, Fast effective rule induction, in Machine Learning: Proceedings of the Twelfth International Conference (ML95), pp. 852-857, 1995.

[26] Ron Kohavi: Scaling Up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In: Second International Conference on Knoledge Discovery and Data Mining, 202-207, 1996.

[27] M.L. Zhang and Z.H. Zhou, A k-nearest neighbor based algorithm for multi-label classification, in 1st IEEE International Conference on Granular Computing, Vol. 2, pp 718721, 2005.

[28] N. Friedman, D. Geiger, and M. Goldszmidt, Bayesian network classifiers, Journal of Machine Learning Research, Vol. 29, pp. 131163, Dec. 1997.

[29] A. Hedar, J. Wang, and M. Fukushima, "Tabu search for attribute reduction in rough set theory", presented at Soft Comput, 2008, pp.909-918.

[30] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten, The WEKA Data Mining Software: An Update, SIGKDD Explorations, Volume 11, Issue 1, 2009.

[31] S. Hettich, and S.D. Bay, The UCI KDD Archive. Irvine, CA: Dept. Inf. Comput. Sci., Univ. California, 1996 [Online]. Available: http://kdd.ics.uci.edu.