A Genetic Algorithm Based Approach for Imputing Missing Discrete Attribute values in Databases

R.DEVI PRIYA
Research scholar and Assistant Professor
Department of Information Technology
Kongu Engineering College, Perundurai – 638 052, Erode
INDIA
scrpriya@gmail.com

S.KUPPUSWAMI
Principal
Kongu Engineering College, Perundurai – 638 052, Erode
INDIA
skuppu@gmail.com

Abstract: - Missing values create a noisy environment in almost all engineering applications and is always an unavoidable problem in data management and analysis. Many techniques have been introduced by researchers to impute these missing values. Most of the existing methods would be suitable for numerical attributes. For handling discrete attributes, only very few methods are available and there is still a necessity for good and sophisticated method. The proposed approach provides a solution for this need by introducing a new technique based on Genetic Algorithm and Bayes' Theorem to impute missing discrete attributes which often occurs in real world applications. The experimental results clearly show that the proposed approach significantly improves the accuracy rate of imputation of the missing values. It works better for datasets even with missing rates as high as 50% when compared with other existing methods. Rather than using highly complex statistical software, we use a simple procedure which does not demand much expertise of the user and still capable of achieving much better performance. The proposed approach not only imputes the missing values, it also provides information about the cases which behave similar to those with missing values.

Key-Words: - Missing values, Numerical attributes, Discrete attributes, Genetic Algorithm, Bayes' Theorem, Imputation

1 Introduction

The success of all statistical techniques used in engineering fields heavily relies on storage and retrieval of data. The databases used for this purpose are subjected to missing values either in data collection, storage or integration process. In real-world applications, information may be missing due to instrumental errors, optional fields and non-response to some questions in surveys, etc. Most of the data mining techniques need analysis of complete data without any missing information and this induces researchers to develop efficient methods to handle them. It is one of the most important areas where research is being carried out for a long time in various domains [1][11][12][13]. Even though many methods are available, different methods works better in different situations and different databases.

No common method is applicable in all missing instances since they differ in the way they are missing. Hence, the nature in which data is missing is a key factor in choosing the suitable method and the missing mechanisms are given below.

Mechanisms of Missing Data

According to Little and Rubin [10], missing data is categorized into three different mechanisms based on the nature of missingness and are given below.

(i) Missing Completely At Random (MCAR)

The probability that the value of an attribute missing does not have its influence on any other attribute in the dataset and it is entirely random.

(ii) Missing At Random (MAR)
The probability that the values of an attribute are missing depends on one or more other variables in the data set and does not depend on the same attribute.

(iii) Not Missing At Random (NMAR)
NMAR is a non-ignorable situation where the probability of missingness of an attribute's value is related to the missing variable's value itself and is independent of other variables in the dataset.

When important strategic decisions are to be made, the most critical challenge confronting researchers is applying the most appropriate method to handle missing data. When efficient high-end sophisticated procedures are more difficult to implement to handle them, simple methods like case deletion, random substitution can be effectively used. When inappropriate methods are used, sometimes they may create more problems by resulting in distorted estimates, hypothesis tests and introduction of more bias [13]. This problem restricts data analysts from making strategic decisions obtained from those inferences. Hence it signifies the need of effective missing data imputation to preserve valuable information by making right inferences [7]. The statistical procedures used should aim to make valid inferences from the available data rather than estimating missing values. The analysts should ensure that the techniques used to recover missing values do not negatively impact the inferences.

Most of the databases contain discrete attributes where its values are taken from the defined set of values. Missing values of those attributes are unavoidable in many situations [9]. Eventhough it looks to be a simple problem, using inefficient methods may produce biased results and hence always needs careful attention. Table 1 shows the complete dataset with discrete values and Table 2 shows the dataset where some values are missed out.

Table 1: Database with complete values

	A	B	C	D
1	a1	b1	c1	d1
2	a2	b2	c2	d2
3	a3	b2	c3	d3
4	a1	b3	c2	d4

Table 2: Database with incomplete values

	A	B	C	D
1	a1	?	c1	d1
2	a2	b2	?	d2
3	?	b2	c3	d3
4	a1	b3	c2	?

If the percentage of missing values is less, some simple methods like Litwise Deletion, Zero Substitution, Default value Substitution, Random value Substitution etc can be used. When the percentage of missing values is very low, they can even be ignored from further consideration. But if there is more number of records with missing values in multiple attributes, simple methods cannot be used and also the missing values cannot be ignored if strategic decisions have to be made from those attributes. It requires more efficient methods to handle them. Design and implementation of these methods are often complex and requires multifarious statistical background.

This paper proposes a new method which uses Genetic Algorithm, a widely used optimization method together with Bayes theorem to impute the missing values of discrete attributes. Genetic Algorithms and Bayes theorem are chosen because both of them are simple, easily understandable methods which are very effective in finding out the optimized solution from a pool of possible solutions.

Section 2 discusses about the previous works done related to the problem. Section 3 describes the proposed work and its methodology. Section 4 explains the implementation details. Section 5 discusses about the results and Section 6 concludes the paper.

2 Literature Survey

Many methods are widely used by researchers in imputing the values of missing discrete attributes and some common methods used are given below.

2.1 Mode Imputation

The value that has maximum number of occurrences for the attribute in the dataset is substituted for the missing value. This is a very blind method which is normally not suggested for databases which are

more sensitive. But because of simplicity, mode imputation is commonly used in places where complex techniques cannot be implemented.

2.2 Multiple Imputation

Multiple Imputation (MI) is the method which is used for imputing both discrete and continuous values. It was proved in many literatures that MI effectively imputes the values [17][21]. But the deficit in using this method is it needs more statistical knowledge and procedures which is extremely difficult for the normal users to understand and implement.

2.3 Bayesian Approach

Two methods based on Bayesian approach are introduced to impute missing discrete attributes [14]. In the first method, the value which has the maximum posterior probability (MaxPost) is used to replace the missing value. In the second method, the value which has the probability proportional to the posterior probability (PropPost) replaces the missing values. The above two techniques effectively use prior and posterior probability for imputing the missing values. But it is a prerequisite that the missing and the dependent attributes should be discrete. Bayesian approach is used for Not Missing At Random type of missingness [2].

2.4 Hybrid Soft Computing Techniques

In [20], a method which uses Genetic Algorithm together with Fuzzy logic was proposed. It uses fuzzy rules to define the fitness function for chromosomes. It suffers from the problem that it has to generate a large number of fuzzy rules. If the database is small, it may work better. If the database size increases, the number of rules to be generated also increases which will pose a great difficulty for the analysts. Another method which uses Genetic Algorithm combined with Neural networks to impute missing data was introduced in [15] [16]. In [18], the authors proposed to use Fuzzy Neural networks to impute nominal attributes. The difficulty faced by the users while implementing both these method is that they have to properly define the Neural network structure with appropriate hidden layers, neurons and error propagation mechanisms. The efficiency of the approach depends on the expertise of the designer. Multiple Imputation based method can also be used where decision trees [10] are used together with Genetic Algorithm to impute the missing discrete values [6]. Although the problem of discrete missing value replacement can be viewed as a classification problem, most of the classification techniques cannot be used as a missing value replacement method, because they themselves require a solution for missing values in the non-class attributes. Also, a Genetic Algorithm based method for imputing Non-ignorable missing data was suggested by [5].

3 Problem Formulation

To overcome the limitations of the existing methods, a new approach called Bayesian Genetic Algorithm (BGA) is proposed which combines the characteristics of both Genetic Algorithm and Bayes theorem. Genetic Algorithm is used to improve the accuracy of the imputed solution over the iterations. Bayes' theorem is used commonly for finding missing discrete attributes because it is very well known for its better performance and simpler form which attracts even a normal user with low statistical knowledge to effectively use them. Bayes' theorem is used as fitness function to fine tune those values. When values for the attributes are Missing At Random (MAR), Bayesian theorem performs better even when the distribution of missingness is not known [19]. In this paper, BGA is applied for MAR type of missingness.

3.1 Genetic Algorithm

Genetic algorithms (GA) are randomized, stochastic search and optimization techniques which work based on the concept of evolution [8]. Genetic Algorithms operate in simple to complex search spaces and can find optimal or near optimal solutions for the given problem when sufficient number of generations are given. It is widely used in many applications like medicine, design of machineries, banking, education, chemical industries, space research, scheduling etc. The pseudocode of the Standard Genetic Algorithm is given below in figure 1.

```
BEGIN
INITIALIZE population with random chromosomes
Calculate fitness function of all chromosomes
REPEAT until stopping criterion is satisfied
        1.SELECT best parents
        2. MATE the selected parents and
           produce new offsprings
        3. MUTATE the mated offsprings
        4. Next generation is populated with
           resulting chromosomes
END
```

Figure 1.Pseudocode of Standard Genetic Algorithm

First, structure of the chromosome which constitutes collection of genes (values) has to be defined. The collection of 'n' chromosomes is called as population where the value of 'n' is defined by the user. By default, the chromosomes (individuals) for the initial population are selected at random. *Fitness function* value of each chromosome has to be calculated which is the leading factor which determines the ability of GA in finding out the optimal solution. It varies for different kinds of problems.

Next, *Selection* operation is performed to improve the global convergence and computational efficiency which selects best chromosomes that are to be used in the subsequent steps. To prevent the algorithm from premature convergence, some low fit chromosomes are also retained in the population thereby maintaining large diversity in the population. Nowadays, Elitism is used to retain the best individuals in a generation without being changed in the next generation which is proved to produce better results. All the selection mechanisms attempt to choose best chromosomes to be carried over to the successive generations. But different mechanisms work better for different kinds of problems and different fitness functions. Hence it is the prime responsibility of the user to choose the appropriate selection mechanism for the defined problem.

Crossover (mating) is the recombination process where the selected parents in the population are mated with each other in order to produce offsprings for the next generation. Based on the crossover probability (Pc), chromosomes in the population are selected at random for participating in crossover. *Mutation* is used in Genetic Algorithm to search a broader space and it helps in creating genetic diversity among chromosomes in population of one generation to the next generation. This step is included in the algorithm in order to prevent the phenomenon of premature convergence. The above steps from fitness calculation to mutation are repeated until some defined termination criterion is met. The criteria can be either fixed number of iterations or convergence of chromosomes in the population to a single optimized value. Even though GA looks to be a random procedure, it does not produce random results. It results in better solutions for the given problem since the objective function value gets improved over the generations resulting in optimized value at termination.

3.2 Bayes' Theorem

Bayes' theorem follows a probabilistic approach which is used to link conditional probability and its inverse i.e., it demonstrates the relationship between P (E|C) and P (C|E). It has its wide application in all engineering fields. If actual cause and its effects are known for an event, Bayes rule estimates the possible causes from the effects which have been observed. Graphically this can be represented as in figure 2.

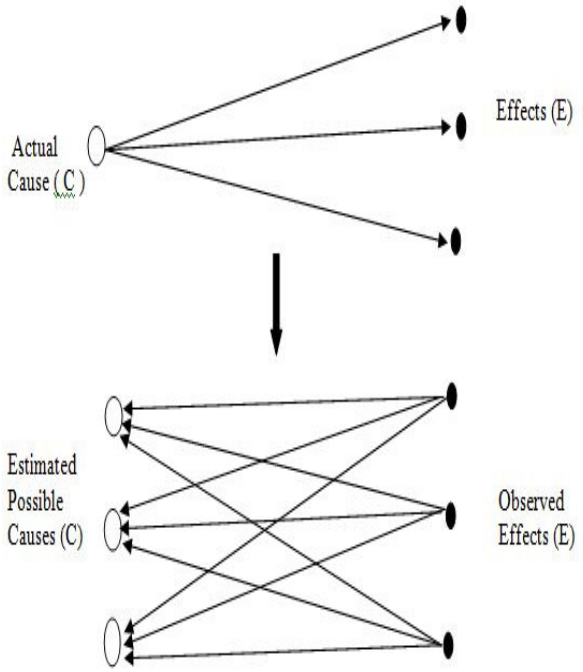

Figure 2. Methodology of BGA

In general, this can be represented mathematically as in Eqn (1)

$$P(E|C) = \frac{P(C|E).\,P(E)}{P(C)} \quad \text{----------} \quad (1)$$

where

P (E) = Probability that the event E has occurred
P(C) = Probability that the event C has occurred
P (E|C) = Probability that the event E has occurred given that C has occurred
P (C|E) = Probability that the event C has occurred given that E has occurred

3.3 Bayesian Genetic Algorithm

The Standard Genetic Algorithm procedure is being used for our experiments with inclusion of Bayes theorem to impute the values of missing discrete attributes. Suppose if the value of attribute

A is missing and it depends on other attributes B, C and D, Bayes rule can be formulated as in Eqn(2)

$$P(A|B, C, D) = \frac{P(B, C, D|A).\ P(A)}{P(B, C, D)} \quad \text{------} \quad (2)$$

The probability of A whose value is missing is conditioned on their dependent attributes B, C and D. The conditional probability value obtained can be taken as the fitness value of the corresponding chromosome with given categories of attributes B, C and D. The chromosomes with high Bayes probability will be carried over to the next generation and the ones with low Bayes probability are removed from the next generation population. That is, high fit individuals are preferred over low fit individuals to undergo crossover operation. After selecting best parents, genetic operators like crossover and mutation are performed and the algorithm is iterated until the termination criterion is satisfied. Rather than using complex techniques and formula to calculate the fitness values, the proposed method uses a simple approach which follows the conditional probability. The structure of the proposed Bayesian Genetic Algorithm is given below in figure 3.

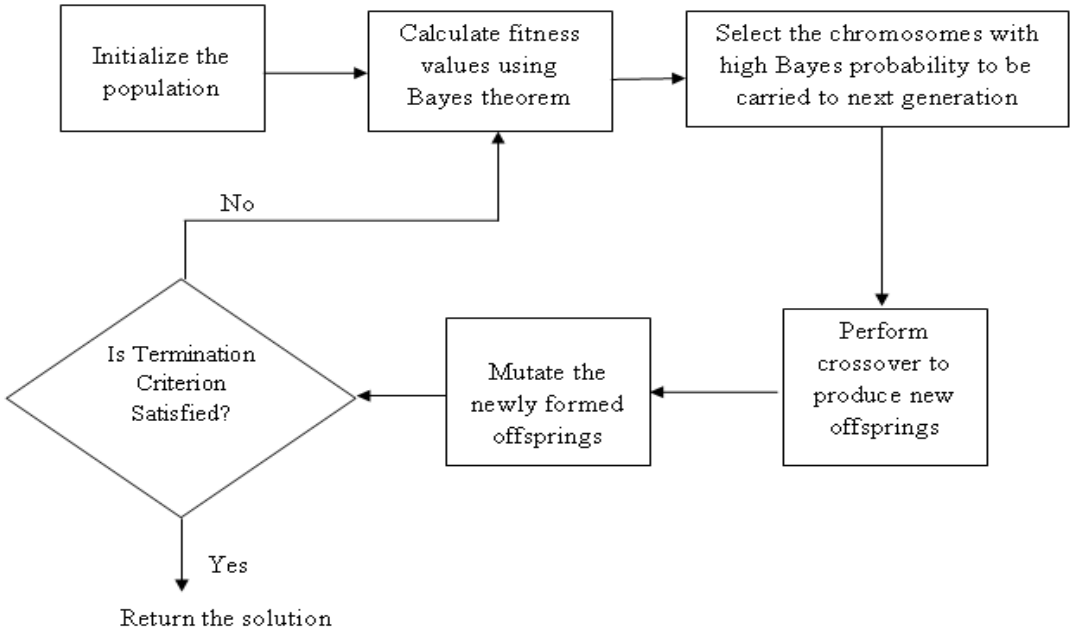

Figure 3. Proposed BGA

4. Implementation

Here two kinds of cases are considered where the first case involves the imputation of a missing discrete attribute which depends upon other discrete attributes. The second case involves the imputation of a missing discrete attribute which depends on both continuous and discrete attributes.

Without knowing the original values, it is impossible to assess how well the implemented method imputes the missing values. Since it is easier to examine the effectiveness of the proposed approach when their true values are known, some records are randomly selected in the dataset and missing values are introduced artificially. The databases used in our experiments are taken from the UCI repository [3] and their details are given in Table 3.

Table 3. Databases used in experiments

Dataset	Type of Attributes	Total Number of attributes	No of instances
Adult	Mixed	14	48842
Breast Cancer	Discrete	9	286
Heart Disease	Mixed	14	303
Iris	Mixed	4	150
University	Mixed	17	285
Zoo	Mixed	17	101

The genetic parameters chosen for this implementation are:

Encoding Scheme : Real value Encoding
Selection method : Rank Selection
Crossover : One Point Crossover
Elitism : 10%
Mutation Probability : 0.2%

Real value encoding is chosen because it will be more suitable to represent the category numbers of the discrete attributes in the chromosome structure. Rank selection suits better for this problem because it ranks the chromosomes in the order of their fitness values. Choosing small population size which runs for more number of generations will generate individuals with high fitness value rather than choosing large population size which runs for less number of generations. Hence population size should be carefully chosen for the problem.

The Root Mean Square Error (RMSE) is the performance measure used to find the predictive ability of the algorithm [4].

$$RMSE = \sqrt{\frac{1}{m}\sum_{i=1}^{M}\sum_{k=1}^{L_i}(\tilde{m}_{ik} - m_{ik})^2},$$

where the former gives the number of missing values which originally belong to the k_{th} category of the i_{th} attribute and latter is the corresponding count estimated using the imputed values and m refers to the total number of missing values.

Case 1: Imputation of missing discrete attributes dependent on other discrete attributes

In some databases to impute values of missing discrete attributes, values of other dependent discrete attributes may be used. Mostly in the standard Genetic Algorithm, the initial population is chosen randomly. Genetic Algorithms work better if prior knowledge is incorporated in initializing the population. Hence in this approach, this step is further improved by selective initialization where some defined criterion is used to select the chromosomes in the initial population itself. Here, only the records which have similar categories as the missing record are chosen for the initial population. If any of the dependent attributes contains ordinal values, then the category of nearest neighbours is also included in the population.

For example in the Adult dataset taken from UCI repository, Education attribute is ordered and the categories are assigned in the same order. In such cases, for imputing the missing values based on education category 6, categories 5 and 7 are also included in the population with the assumption that category 6 tends to behave more similar to the categories 5 and 7. The categories of the Education attribute which have ordered values are given below in Table 4.

Table 4. Categories of Education

S.No	Education	Category
1	Preschool	1
2	1^{st}-4^{th}	2
3	5^{th}-6^{th}	3
4	7^{th}-8^{th}	4
5	9^{th}	5
6	10^{th}	6
7	11^{th}	7
8	12^{th}	8
9	HS-Grade	9
10	Bachelors	10
11	Masters	11
12	Doctorate	12

The sample chromosomes are given below in figure 4 where 'a' represents the category of attribute A whose value is missing and 'b' and 'c' represents the categories of dependent ordinal attributes B and C.

1	**5**	1
2	**6**	2
	:	
3	**7**	11
b	c	a

Figure 4. Sample chromosomes for ordinal attributes (Case 1)

But if the nominal attributes are considered for our investigation, the initial population is seeded in such a way that all chromosomes should have either of the categories of the dependent attributes that are missing. This is because of the fact that the nearest neighbouring categories used for ordinal attributes have no meaning for nominal attributes since categories are just defined at random and no ordering is maintained among them.

The sample chromosomes for this case are given below in figure 5 where 'a' represents the category of attribute A whose value is missing and 'b' and 'c' represent the categories of dependent nominal attributes B and C. For example, values are missing for the record with categories 5 and 7 respectively for the attributes b and c, the initial population is loaded with chromosomes with any of the categories for b and c. This will help GA to search for the solution in the reasonable search space.

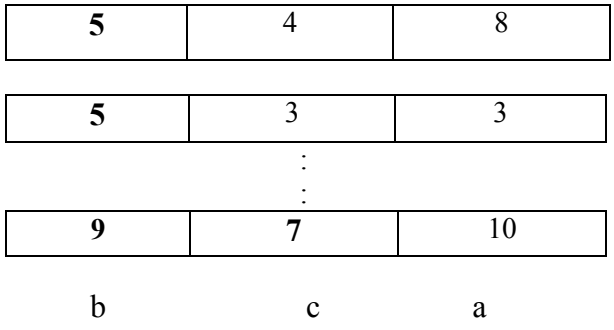

Figure 5. Sample chromosomes for nominal attributes (Case 1)

Case 2: Imputation of missing discrete attributes dependent on other Continuous and Discrete attributes

Sometimes, there are possibilities that approximation of a discrete attribute may be dependent on some other combination of continuous and discrete attributes. For example in the adult database taken from UCI repository, if Occupation is missing in a record, the proposed approach tries to fill that value based on the values of Age(continuous) and Education(Discrete). Consider for a record with Age = 47, Education = Bachelors, value for the attribute Occupation is missing. It is assumed that this record is likely to behave similar to those individuals with Age group between 43 to 50 and Education = Bachelors and hence they are all chosen as individuals in the initial population. If there are more number of records satisfying this condition, then the proposed approach chooses best 'n' records among them. This kind of approach will enhance the possibility of providing better individuals in the beginning phase itself. After the initial population is chosen, they are then subjected to application of genetic operators like selection, crossover and mutation.

If any dependent attribute is continuous and if there are many possible set of values, they can be converted into ordinal values. If there are only few values that can be taken for the continuous attribute, then dicretization need not be done and the numerical value can be directly taken as the category of the corresponding attribute.

Even though it is normally said that discretization leads to loss of actual information; it does not have much effect in the final inference in imputing the missing values. In highly sensitive databases, some standard discretization techniques can be used. For example, numerical values of age attribute are grouped into 'n' categories as in Table 5.

Table 5.Categories of age

S.No	Age	Category
1	21-25	1
2	26-30	2
3	31-35	3
4	36-40	4
5	41-45	5
6	46-50	6
7	51-55	7
8	56-60	8

The sample chromosomes for this case are given below in figure 6.

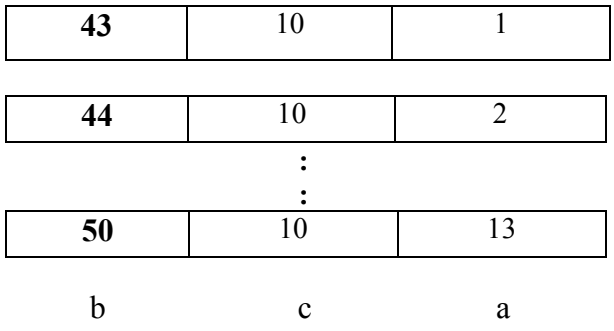

Figure 6 .Sample chromosomes for case 2

5 Results and Discussion

The experiments are conducted for both the cases and the results are discussed below.

Case 1

BGA is applied on different datasets with discrete attributes mentioned in Table 2 to impute the missing attribute values and the empirical results are studied. The experiments are conducted for different population sizes 40, 50 and 60 and the results are

averaged. The Root Mean Square Error(RMSE) obtained by applying BGA and other existing methods like Mode Imputation, MaxPost, PropPost and BGA are calculated and the results are shown in figure 7 for adult dataset.

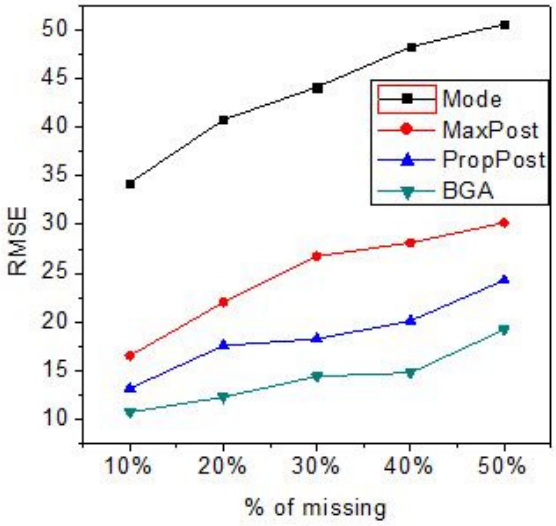

Figure 7. RMSE values for Adult dataset at different missing rates

The same experiments were conducted for Breast Cancer and Iris datasets and the results obtained are shown in figures 8 and 9 respectively.

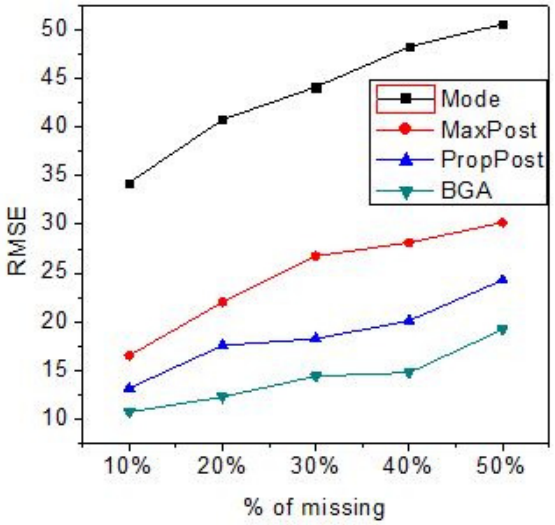

Figure 8. RMSE values for Breast Cancer Dataset at different missing rates

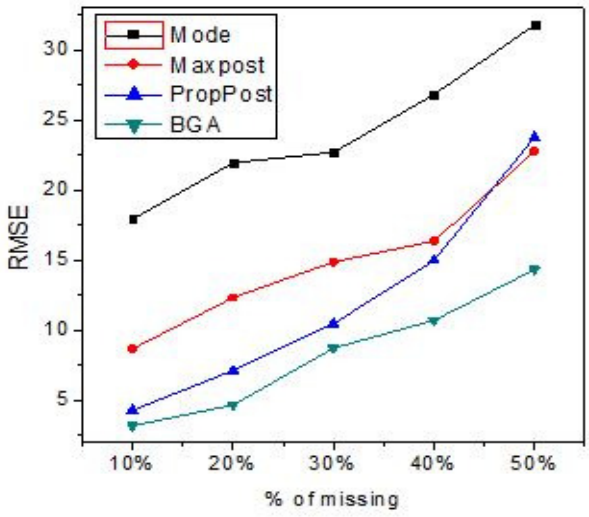

Figure 9. RMSE values for Iris dataset at different missing rates

Case 2

Mode imputation produced results with high RMSE (less accuracy). Maxpost and PropPost results in average performance and BGA produced results with high predictive accuracy. It can be observed that BGA outperforms other methods in all three different datasets. This is due to the fact that the Bayes' rule drives the Genetic Algorithm towards better solution in every successive generation.

The experiments are conducted with Adult, Heart Disease, University and Zoo datasets and the results obtained are given in figures 10, 11, 12 and 13 respectively.

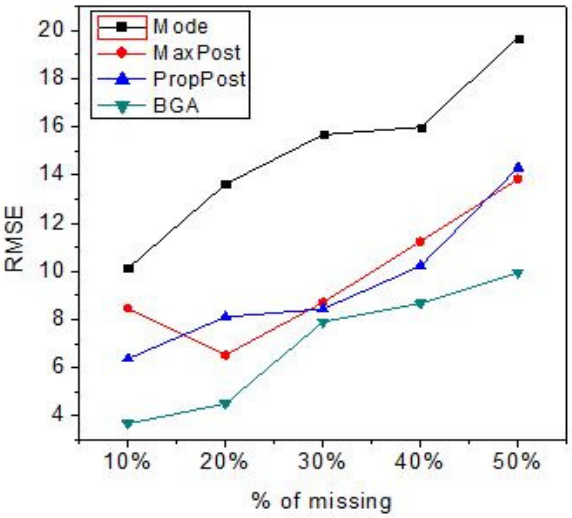

Figure 10. RMSE values for Adult dataset at different missing rates

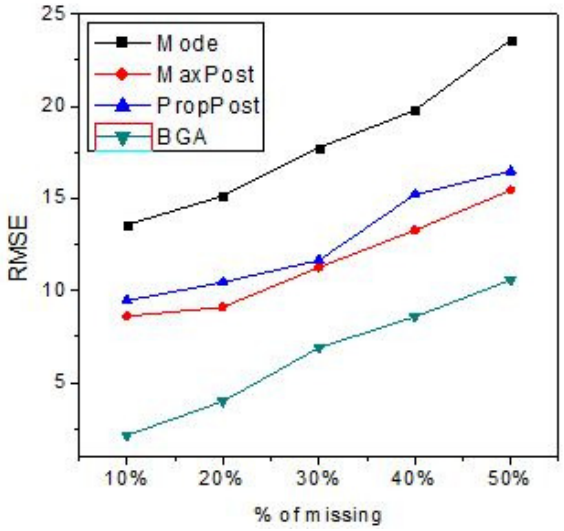

Figure 11. RMSE values for Heart Disease dataset at different missing rates

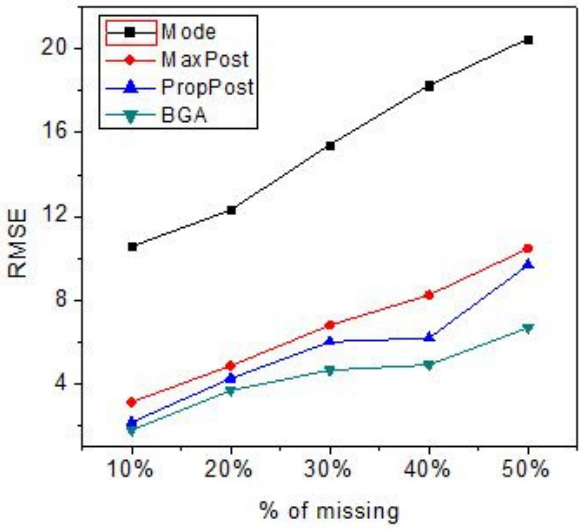

Figure 12. RMSE values for University dataset at different missing rates.

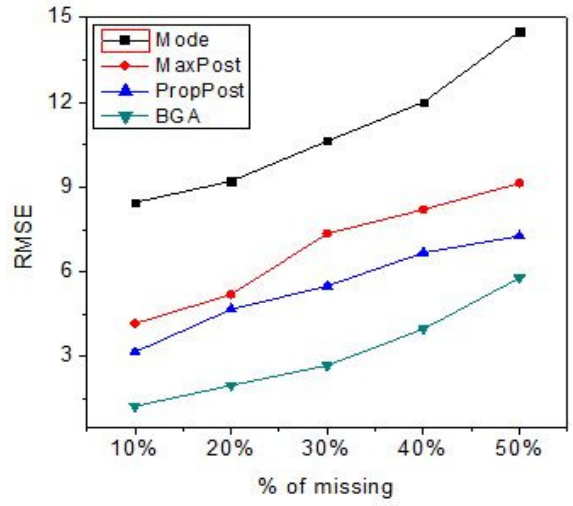

Figure 13. RMSE values for Zoo dataset at different missing rates

It can be observed from the above results that BGA shows better performance. In small datasets, the proposed procedure always uses all the available data in an efficient manner. Also in large datasets like Adult dataset, BGA results in highest predictive accuracy. If there are many variables with missing values, the number of samples to be taken for consideration may decrease dangerously and may result in biased results. Even though RMSE increases when missing rate increases, BGA tends to result in more appropriate results with different missing rates varying from 5% to 50% when compared with other methods. It is also to be noted that if the number of categories for the missing attribute is small, the results produced are more accurate. But if there are too many categories, then the performance is considerably lower which is the case with any other imputation methods. The important added advantage in using BGA is that it helps in knowing the records which behave similar to the records with missing values. The chromosome which has the highest fitness value is chosen as the final solution and its structure depicts the categories of relevant records which greatly helps the analysts not only in filling the missing values but also in making other strategic decisions. .

6 CONCLUSION

Missing data is always considered as a tough unavoidable problem which raises many conceptual difficulties and computational challenges in various domains. Before any dataset is used for analysis and strategic decision-making, they must be imputed. The main advent of the proposed method is that it combines the advantageous features of Genetic Algorithm and Bayes' theorem where the former is used to produce new combination of chromosomes and the latter is used to evaluate the worthiness of those chromosomes in finding the missing values. The empirical results clearly show that BGA produces far better results for imputing missing discrete attributes for which only limited number of techniques are already available. It is also found that the proposed method imputes the missing data with more accuracy for different percentage of missing rates. BGA can be applied to all kinds of datasets with missing discrete values provided it is Missing At Random.

References:

[1] Allison P.D.Multiple imputation for missing data: A cautionary tale. *In Sociological*

Methods and Research, Vol. 28, 2000, pp. 301-309.

[2] Alexina Mason, Sylvia Richardson, Ian Plewis and Nicky Best. Strategy for modelling non-random missing data mechanisms in observational studies using Bayesian methods. *Monographs on statistics and applied probability,* Vol.109, Chapman & Hall/CRC, 2010.

[3] Blake, C. L. and Merz, C. J. *UCI Repository of machine learning databases,* Irvine, University ofCalifornia,1998.[http://www.ics.uci.edu/~mlearn/ MLRepository.html]

[4] Chen, G. and Astebro, T. How to deal with missing nominal data: Test of a simple Bayesian method. *Organ. Res. Methods* Vol.6, No.3, 2003, pp. 309–327.

[5] Devi Priya. R, Kuppuswami. S and Makesh Kumar S. A Genetic Algorithm Approach for Non- Ignorable Missing Data. *International Journal of Computer Applications* Vol 20,No.4 2011,pp. 0975 – 8887.

[6] Dipak V. Patil and R. S. Bichkar. Multiple Imputation of Missing Data with Genetic Algorithm based Techniques. *IJCA Special Issue on "Evolutionary Computation for Optimization Techniques"* ECOT, 2010.

[7] Fogarty, D. J. Multiple imputation as a missing data approach to reject inference on consumer credit scoring. *Interstat.* [URL http://interstat.statjournals.net /YEAR/ 2006/ articles/ 0609001.pdf]

[8] Goldberg, D.E. *Genetic Algorithms in Search, Optimization and Machine Learning*. New York: Addison-Wesley, 1989.

[9] Gryzmala-Busse, J.W. and W.J.Gryzmala-Busse. Handling missing attribute values . In O.Maimon and L.Rokach, editors, *Data Mining and Knowledge Discovery Handbook,* Springer-Verlag, Berlin, Heidelberg, pp.37-57, 2005.

[10] Han, J and Kamber, M. *Data Mining: Concepts and Techniques"* Morgan Kaufmann Publishers, Second Edition, 2006.

[11] Huisman, M. Post-stratification to correct for non-response: classification of zip code areas. *Proceedings of the 14th Symposium on Computational Statistics,* 2000, pp.325-330.

[12] Linda G. DeMichiel. Resolving Database Incompatibility: An approach to Performing Relational Operations over Mismatched Domains. *IEEE Transactions on Knowledge and Data Engineering,* Vol.1,No.4, 1989, pp.485-493.

[13] Little, R., Rubin, D. *Statistical Analysis with Missing Data.* Wiley & Sons, New York, 2002.

[14] Li, X.-B. A Bayesian approach for estimating and replacing missing nominal data. *ACM J. Data Inform. Quality* Vol.1, No.1, Article 3, 2009.

[15] Mussa Abdella and Tshilidzi Marwala. Treatment of Missing Data Using Neural Networks and Genetic Algorithms. *International Joint Conference on Neural Networks,* Montreal, Canada, 2005.

[16] Mussa Abdella and Tshilidzi Marwala. The use of Genetic Algorithms and Neural Networks to approximate missing data in database. *IEEE 3rd International Conference on Computational Cybernetics,* 2005, pp. 207 – 212.

[17] Pearson,R.K., The problem of disguised missing data. *ACM SIGKDD Explorations* Newsletter 8(1), 83-92.

[18] Pilar Rey-del-Castillo and Jesus Cardenosa. Nominal Missing Data Imputation Using Fuzzy Neural Networks with Numerical and Nominal Inputs. *World Academy of Science, Engineering and Technology,* 2005.

[19] Schafer, J., Graham, J. Missing data: Our view of the state of the art. *Psychological Methods* Vol.7, 2002, pp.147 -177.

[20] Shyi-Ming Chen and Chung-Ming Huang. Generating Weighted Fuzzy Rules from Relational Database Systems for Estimating Null Values Using Genetic Algorithms. *IEEE Transactions on Fuzzy Systems,* Vol. 11, No. 4, 2003.

[21] Yang Yuan. Multiple imputation for missing data: Concepts and new development. *In SUGI Paper* 267-25, 2000.