A Genetic Algorithm Approach for Non-Ignorable Missing Data

R.Devi Priya
Assistant Professor
Department of IT
Kongu Engineering College

S.Kuppuswami
Principal
Kongu Engineering College
Perundurai

S.Makesh Kumar
Department of IT
Kongu Engineering College
Perundurai

ABSTRACT
The databases store data that may be subjected to missing values either in data acquisition or data storage process. The proposed approach uses the widely used optimization technique called genetic algorithm for the NMAR (Not Missing At Random) missing mechanism which prevails more in real life that are non-ignorable. Since the non-ignorable mechanism needs prior basic knowledge about the data that is supposed to be missing and have to make assumptions, Genetic algorithm (GA) suits well for this problem which derives solution based on the previously observed data. The empirical results show that Genetic Algorithm has better efficiency when compared with some of the traditional methods.

General Terms
Databases, data storage, Prior basic knowledge

Keywords
Data acquisition, Missing data, NMAR, Non-response, Genetic algorithm, Optimization.

1. INTRODUCTION
The serious quality problems that reduce the performance of data mining algorithms are due to incomplete, redundant, inconsistent and noisy data. Missing data is a common issue in almost every real dataset. If the rate of missing is less than 1%, missing data won't make any trouble for the discovery of Knowledge, 1-5% manageable, 5-15% requires sophisticated methods to handle and more than 15% may severely impact any kind of interpretation. Due to this reason, missing data has been an area of research in various disciplines for a quite long time. The missing data treatment methods have to be carefully chosen based on missing mechanism.
According to Little and Rubin [1], missing data can be divided into the following three mechanisms:

i. Missing At Random
Data are said to be Missing At Random (MAR) when the probability that Y is missing depends on the set of other observed responses X, but is unrelated to the specific missing values.

$$Pr(Y \text{ is missing} \mid X, Y) = Pr(Y|X) \text{ for all missing } Y$$

ii. Missing Completely At Random
Data are said to be Missing Completely At Random(MCAR) if the probability that Y is missing is unrelated to Y or other variables X (where X is a vector of observed variables).

$$Pr(Y \text{ is missing} \mid X, Y) = Pr(Y \text{ is missing})$$

iii. Not Missing At Random
Data are said to be Not Missing At Random (NMAR) when the probability that responses are missing depends on the specific missing values itself. NMAR is often referred as non-ignorable, non-random missingness. When missingness is non-ignorable, it means that future unobserved responses, conditional on past observed responses cannot be predicted. High levels of incompleteness usually falsify assumptions that missing data may be ignored [5]. But in some cases, data cannot be ignored and it is not amenable to use common missing data handling methods that are used for MAR and MCAR. Example – If the probability that income is provided in the survey varies within each age group (e.g., wealthy and poor subjects are less likely to answer the income question regardless of age), then the data are neither MAR nor MCAR, and hence NMAR (or NI). If the mental health of people is studied, the people who have been diagnosed as depressed are less likely to report their mental status than healthy people. Figure 1 shows a table with missing values in multiple attributes indicated by ?

Figure 1. Relational Table with missing values

Name	Age	Designation	Experience	Income
Aa	x	x	x	x
Bb	x	?	x	?
Cc	?	x	x	?
Dd	x	x	?	x

In this paper, Not Missing at Random (NMAR) condition is being dealt with using a simple optimization algorithm called Genetic Algorithm. Genetic Algorithm is an algorithm based on principles of Natural Selection and Genetics. Genetic Algorithm applies to a variety of problems and not works in a restricted domain. Genetic Algorithm works well in large search space problems, as better solutions tend to "grow old with time". They are used to find approximate solutions to difficult problems through application of the principles of evolutionary biology to computer science [6] [7]. They use biologically derived techniques such as inheritance, mutation, natural selection, and recombination to approximate an optimal solution to difficult problems [8].

Genetic algorithm is chosen to optimize NMAR condition because of its special feature of dealing with data by the prior knowledge about that data set in which the algorithm is going to deal with. This basic knowledge helps people in categorizing the data which are not missing at random. After initializing the population, several modules of genetic algorithm are applied to get the optimized result. The population concept in genetic algorithm will result in variety of output for varying population. Some traditional methods are compared with the GA in order to prove the efficiency of Genetic Algorithm. The results obtained from implementing GA show better performance than the available standard methods.

2. RELATED WORK

There are many traditional methods that are employed to solve this problem. A simple method is to substitute the missing values by 0(Zero) for integer and 'F' for Boolean data. The Mean Imputationtion (MS) method replaces the missing values by average of the other observed values in the same attribute [1, 9]. Litwise deletion method deletes the records with missing values where the vital records which should be used for analysis may be deleted. Random Substitution method will pick up a random value and substitute it for missing values. All these methods do not gain any significance as all are blind methods that may not result in required output and will change the characteristic of the original dataset ignoring the relationship among attributes and will bias the data mining algorithms [1]. K-Nearest neighbor approach [2] replaces the missing values with values of the k-nearest neighbors. 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.

Multiple imputation (MI) is one of the most attractive methods for general purpose handling of missing data in multivariate analysis. Rubin [3] described MI as a three step process. First, sets of plausible values for missing values are created using an appropriate model chosen that 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. Second, each of these datasets can be analyzed using complete-data methods. Finally, the results are combined. There are various ways to generate imputations. The implementation program for MI of continuous multivariate data (NORM) is available in [4] [18] and this is used in this experiment for analysis.

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, 12]. Recent overviews of NMAR modeling are given in [10,14 and 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, the MNAR model may perform even less well than standard MAR methods [17].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 adhoc approaches that may do more harm [13].

To bridge the gap between these complex statistical methods and the researchers in this field, simple and efficient genetic algorithm is suggested here. In our experiments, to show the efficiency of our algorithm, we compared the performance of Genetic algorithm with more commonly used approaches rather than these modeling methods and weighting methods.

3. SYSTEM DESIGN

As NMAR case of missingness depends on the values of the missing attribute itself, traditional techniques that is used for MAR and MCAR cannot be effectively used and simple yet efficient genetic algorithm is used.

3.1 Genetic Algorithm

In a genetic algorithm, a population of strings called chromosomes which encode candidate solutions called individuals to an optimization problem evolves toward better solutions. Traditionally, solutions are represented in binary as strings of 0s and 1s, but other encodings are also possible. The evolution usually starts from a population of randomly generated individuals and happens in generations. In each generation, the fitness of every individual in the population is evaluated, multiple individuals are stochastically selected from the current population (based on their fitness), and modified (recombined and possibly randomly mutated) to form a new population. The new population is then used in the next iteration of the algorithm. The algorithm terminates when either a maximum number of generations has been produced, or a satisfactory fitness level has been reached for the population [16].

Pseudocode of the Genetic Algorithm:

1. [Start] Generate random population of n Chromosomes (Suitable solutions for the problem)
2. [Fitness] Evaluate the fitness f(x) of each chromosome x in the population
3. [New population] Create a new population by repeating following steps until the new population is complete
   (i) [Selection] Select two parent chromosomes from a population according to their fitness (the better fitness, the bigger chance to be selected)
   (ii) [Crossover] With a crossover probability cross over the parents to form a new offspring (children). If no crossover was performed, offspring is an exact copy of parents.
   (iii) [Mutation] With a mutation probability mutate new offspring at each locus (position in Chromosome).
   (iv) [Accepting] Place new offspring in a new population
4. [Replace] Use new generated population for a further run of algorithm
5. [Test] If the end condition is satisfied, stop , and return the best solution in current population
6. [Loop] Go to step 2

3.2. Genetic algorithm for NMAR data

In this paper, the genetic algorithm is executed against the real datasets collected from engineering college students. We suppose that the overweight students hesitate to specify their weights in the survey. The proposed genetic algorithm tries to

fill the missing values to draw a conclusion to estimate the percentage of overweight students so that they can be put under physical training to reduce their weights in order to improve their health. The experiments are also conducted with Adult and Housing dataset from UCI repository[19].

4. EMPIRICAL RESULTS

The genetic algorithm is executed with different genetic operators. The algorithm efficiency will critically vary depending on the objective function and the genetic operators chosen. Hence the algorithm is being run with varying genetic operators and the results are compared. For the experiments, the crossover probability is chosen to be 0.80 and the mutation probability is 0.01.For mutation, a variant of the standard approach of changing the bits are used. The first i out of n bits in the chromosomes are kept silent and they will not participate in the mutation. This is done to retain the best values across the iterations.

The root mean square error (RMSE) [17] is a frequently-used measure of the Square root of differences between Original value and the estimated value. The Accuracy is calculated by subtracting it with 100. The formula becomes:

$$\text{RMSE}= \sqrt{\frac{\sum_{i=1}^{n}(x_{1,i}-x_{2,i})^2}{n}}.$$

Where x1,i is the Actual value and X2,i is the predicted value and n is the count of missing values.

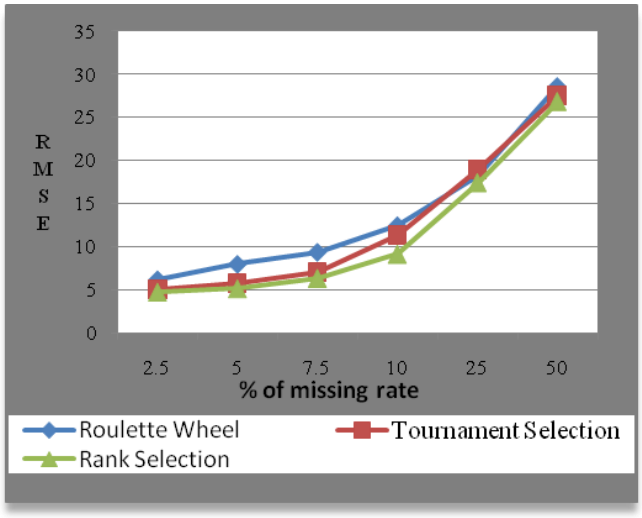
4.1 Effect of varying selection mechanisms in the approximation of data

Selection mechanism is the process of selecting the parents to be mated which has its major impact in the final efficiency of the algorithm. *Elitism* is the process of retaining the few best parents and processing only the rest of the population. It will always replace the worst, so that the population will converge quicker. This means all the individuals will be more or less the same.

Roulette Wheel selection is the process of spinning the wheel 'n' times and the individuals with highest probability calculated according to fitness value is selected for mating. *Tournament selection* involves running several "tournaments" among a few individuals chosen at random from the population. The winner of each tournament (the one with the best fitness) is selected for crossover. *Rank selection* is the process of assigning ranks to the individuals based on the values of fitness function. The individuals with the highest rank are selected as parents. The crossover mechanism chosen here is uniform crossover for the population size 50.

The incomplete data in the survey are approximated using different selection mechanisms and their results are compared in the figure 2. In many studies, selection mechanism is proved to be efficient when combined with elitism. Elitism solution will be usually sub-optimal or near-optimal. Hence in all our experiments the selection mechanisms are combined with elitism in order to retain the best individuals over the iterations.

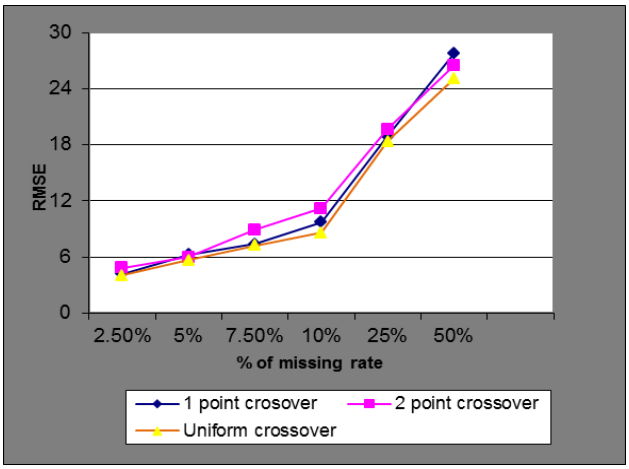
Figure 2. Effect of various selection mechanisms



4.2. Effect of varying crossover mechanisms in the approximation of data:

By applying crossover, the algorithm creates a new population of candidate solutions [8].The different crossover mechanisms are implemented and the results are analyzed. In One point Crossover, any one random point is chosen and the bits are exchanged between the parents after that bit position. In Two point Crossover, two random points are chosen and the bits are exchanged alternatively between the parents after the bit positions chosen. In Uniform Crossover, the mask is chosen randomly and if the bit in the mask is 1, the bits for the offspring is chosen from the first parent, if it is 0 in the mask, the bits for the offspring are chosen from the second parent for offspring 1 and the reverse of this for Offspring 2.

Figure 3. Effect of various crossover mechanisms



New generations of solutions are produced containing, on average, better genes than the solutions in previous generations. Each successive generation will contain more good `partial solutions' than previous generations. Appropriate Crossover

mechanism chosen will preserve the objective function characteristics and will also help to escape from the local optima. As claimed in many studies, the results as shown in figure 3 reveals that the uniform crossover has less RMSE value and hence the highest predictive accuracy. So the uniform crossover is used for the following experiments.

5. RESULT ANALYSIS

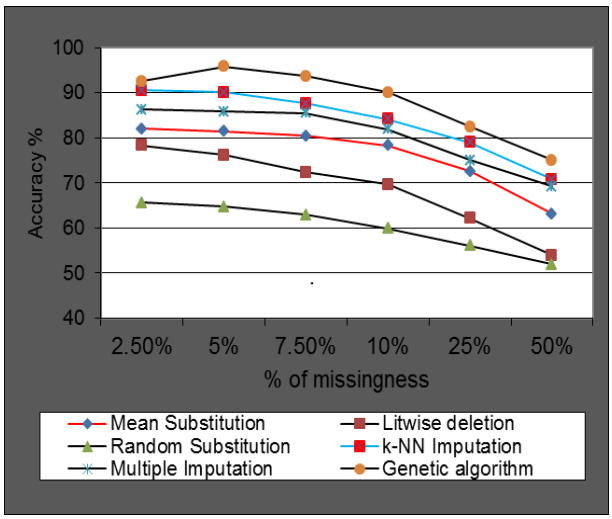
The initial population for the GA run is selected depending on the basic knowledge about the dataset. The values in the missing column are sorted and the range of elements that may be eventually missed may be in the top of the population. The resulting value from the run of GA is substituted for the missing instances. Rather than using complex methods for non-ignorable mechanism, Genetic algorithm proves to be a simple one to implement. The population size is varied proportional to the percentage of missing records. The genetic operators most suited for this problem are obtained from the above results (Parameter Encoding- Binary Encoding, Selection Mechanism-Elitism, Crossover-Uniform Crossover, Crossover Probability-0.80 and mutation probability-0.01) and are implemented .If the algorithm has terminated due to a maximum number of generations, a satisfactory solution may or may not have been reached. So in the proposed approach, the individuals are checked to get saturated over the iterations.

The results obtained from the genetic algorithm are compared with the commonly used techniques like Litwise deletion, Mean Imputation, Random Substitution, 3-NN Imputation, and Multiple Imputation and are shown in the figure 4. Artificially the records are deleted with varying percentages to check the performance of the existing methods and Genetic algorithm. Percentage of accuracy is taken as the measure for comparison.

From the results, it can be observed that the Genetic algorithm has superior performance than other methods. The solution of the methods like Litwise deletion, Random substitution is not optimal and it will lead to loss of original characteristics of the dataset. As the initial population chosen for genetic algorithm is based on the model and basic knowledge of the problem, it will lead to optimal solution when the population is run across iterations with different genetic operators. The incomplete results are approximated using Genetic algorithm and it is discovered that about 20% of the girls and 25% of the boys are overweight for their age and they must be put under some physical exercises to reduce the weight.

The characteristics of the original individuals in the initial population are retained over the generations even though the genetic operators like crossover and mutation are used to introduce diversity of values. This is the special characteristic of genetic algorithm and that is used in approximating the values for the non-response attributes in the surveys.

Figure 4. Comparison of GA with other missing data handling methods for student dataset



The genetic algorithm is executed with 2 real datasets from the UCI repository. The missing data are synthetically generated by deleting the values which are assumed to be not missing at random and the results are analyzed. The genetic algorithm is run with best genetic operators chosen for the given dataset. Figure 5 shows the comparison of GA with other traditional methods for Adult dataset. The accuracy percentage for the approximated data is superior to other methods which use highly complex statistical formula like Multiple Imputation.

Figure 5 Comparison of GA with other missing data handling methods for Adult dataset

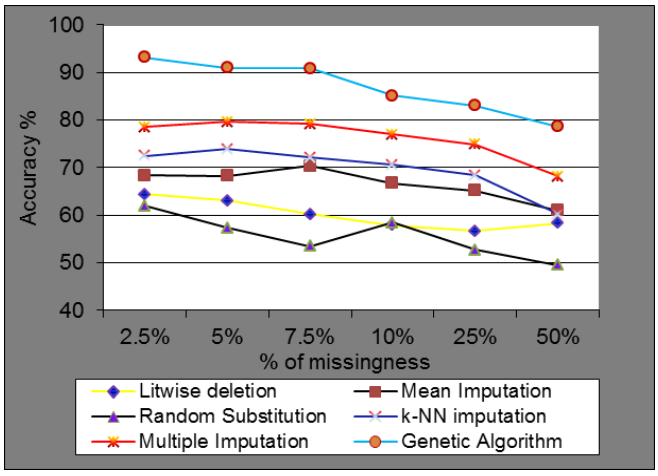
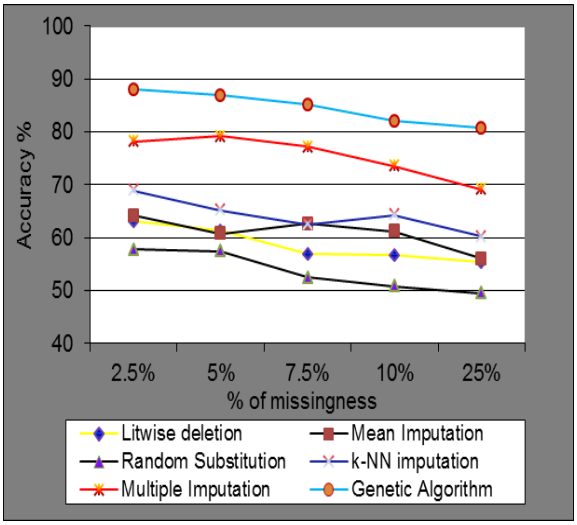


Figure 6 shows the comparison of Genetic algorithm performance with other methods for Housing dataset in UCI repository.

Figure 6.Comparison of GA with other missing data handling methods for housing dataset



All the results clearly show that Genetic algorithm outperforms other missing data handling methods in all the datasets considered. As the best individuals are retained across the generations in genetic algorithm, the final solution is found to be optimal.

6. CONCLUSION

In this paper, Genetic Algorithm has been proposed for estimating non-random non-ignorable missing values which needs some model to impute the missing data. The approach does not demand more complex statistical analysis and are easier to understand and implement. However, appropriate genetic operators for the problem have to be chosen for implementing GA since the choice of genetic operators will significantly impact the output. The proposed method can get a lesser RMSE value and hence higher predictive accuracy rate than the existing methods for estimating missing values in databases.

7. REFERENCES

[1] R. J. A. Little & D. B. Rubin, Statistical analysis with missing data – Second edition, Wiley - Interscience, New Jersey, 2002.

[2] 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.

[3] Rubin, D.B . (1996). "Multiple Imputation After 18+ Years", Journal of the American Statistical Association, 91, pp. 473- 489.

[4] Graham JW, Cumsille PE, Elek-Fisk E. 2003. Methods for handling missing data. In Research Methods in Psychology, ed. JA Schinka, WF Velicer, pp. 87–114.Volume 2 of Handbook of Psychology, ed. IB Weiner.New York:Wiley.

[5] Yang X, Shoptaw S. Assessing Missing Data Assumptions in Longitudinal Studies: An Example Using a Smoking Cessation Trial. Drug and Alcohol Dependence, 77, 213-225, 2005.

[6] Z. Michalewicz. Genetic Algorithm + Data Structures = Evolution Programs. Berling Heidelb NY: Springer- Verlag. third ed,1996.

[7] S. Forrest. "Genetic algorithms." ACM Compur, Sum., vol. 28, no.1.pp 77-80 1996.

[8] W, Banzhaf, P. Nordin, R. Keller, and E Francone Genetic Programming- on the automatic evolution of computer program and Its applications. Califomia: Morgan Kaufmmn Publishers, fifth ed., 1998.

[9] J.R. Quinlan, " Unknown Attribute values in Induction," Proc. Sixth Int'l Workshop Machine Learning, pp. 164-168, 1989.

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

[11] Rubin, D.B. (1987). Multiple Imputations for Non – response in Surveys. New York: John Wiley and Sons.

[12] McKnight, P.E. et al. (2007) Missing Data: A Gentle Introduction, Guilford Press.

[13] Schafer, J.L . AND Graham, J.W. (2002). Missing data: Our view of the state of the art. Psychological Methods, 7 (2), 147-177.

[14] Little, R.J. (2009). Selection and pattern – 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] David E. Goldberg (2005) Genetic Algorithms in Search, Optimization, and Machine Learning.

[17] Demirtas & Schafer 2003 " On the performance of random- coefficient pattern- mixture models for non - ignorable drop - outs". Statistics in Medicine 22, 2553-2575.

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

[19] UCI repository www. ics. uci. edu/ mlearn/ MLRepository. Html