Multiple Imputation of Missing Data with Genetic Algorithm based Techniques

Dipak V. Patil
Department of Computer Engineering
Sandip Institute of Technology and Research Centre,
Nashik, M.S., India.

R. S. Bichkar
Department of Electronics & Tele.
G.H. Raisoni College of Engineering and Management
Pune, M.S., India.

ABSRACT

Missing data is one of the major issues in data mining and pattern recognition. The knowledge contains in attributes with missing data values are important in improving decision-making process of an organization. The learning process on each instance is necessary as it may contain some exceptional knowledge. There are various methods to handle missing data in decision tree learning. The proposed imputation algorithm is based on the genetic algorithm that uses domain values for that attribute as pool of solutions. Survival of the fittest is the basis of genetic algorithm. The fitness function is classification accuracy of an instance with imputed value on the decision tree. The global search technique used in genetic algorithm is expected to help to get optimal solution.

Key words – missing data, genetic algorithm, decision tree.

1. INTRODUCTION

Missing data is the missing form of information about phenomena, which is important, and it is the information in which we are interested. The existence of missing data is one significant problem in data quality. Data quality plays major role in machine learning, data mining and knowledge discovery from databases. Machine learning algorithms handle missing data in a quite naive way. To avoid biasing in induced hypothesis missing data treatment should be carefully handled. Imputation is a process that replaces the missing values in instance by some reasonable values. The case substitution is the method developed for dealing with missing data in instances and it is having some drawbacks when applied to the data mining processes. The methods, such as substitution of missing values by the attribute mean or mode should be cautiously handed to avoid inclusion of bias.

1.1 Randomness of Missing Data

Missing data randomness is classified [1] in three classes.

Missing completely at random (MCAR): Missing values are scattered randomly across all instances. In this type of randomness, any missing data handling method can be applied without risk of introducing bias on the data. It occurs when the probability of an instance having a missing value for an attribute does not depend on either the known values or the missing data.

MCAR can be verified by separating instances into those with and without missing data, then using t-tests of mean differences on attributes establish that the two groups do not differ significantly.

Missing at random (MAR): Missing at random (MAR) is a condition, which occurs when missing values are not randomly distributed across all observations but are randomly distributed within one or more classes (ex. missing more among whites than non-whites, but random within each). The probability of an instance with a missing value for an attribute may depend on the known values, and not on the value of the missing data itself.

Not missing at random (NMAR): Not missing at random is the most challenging form, occurs when missing values are not randomly distributed across observations. It is also called as non-ignorable missingness. The probability of an instance with a missing value for an attribute might depend on the value of that attribute.

1.2 Handling Missing Data

Missing data handling methods are categoriesed as follows

Ignoring data: This method throw-outs all instances with missing data. There are two core methods to discard data with missing values. The first one is known as complete case analysis. It is available in every one of statistical packages and is the default method in many programs.

The next method is recognized as discarding instances or attributes. This method determines the level of missing data 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 is executed only if missing data are missing completely at random. The missing data that are not missing completely at random contain non-random elements that may prejudice the results.

Little and Rubin [1] stated that it is the dangers to delete instances. Instance deletion presumes that the deleted instances are a relatively small quantity of the entire dataset and when cases are missing completely at random. 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 a attributes have more than 5% missing values, cases are not deleted.

Imputation*:* In imputation-based procedures missing values are imputed with reasonable, probable values rather than being deleted totally. The objective is to use known associations that can be recognized in the valid range values of the data set to facilitate in estimating the missing values.

Imputation has several advantages such as efficiency and precision because no observations are discarded.
However, it suffers from implementation difficulties, especially in a multivariate database. In addition to that, some techniques can falsify data associations and distributions [2].

***Multiple Imputations*:** It is a method by Rubin [1] for making multiple simulated values for each incomplete information, and then iteratively examining datasets with each simulated value substituted in every turn. The intention is, possibly, to generate estimates that better indicate true variance and uncertainty in the data than do regression methods. This permits expert staff and software to be used to create imputed datasets that can be analyzed by relatively naive users equipped with standard software. It can be very effective particularly for small to moderate levels of missingness, where the missing data mechanism is organized, and for datasets that are to be placed in the public domain.
Successful multiple imputation postulates three conditions. Firstly, that all associations among variants are properly reflected in the imputations. Unless the missingness follows a clear pattern of comparatively small consecutive blocks of variables so that we are mainly able to predict imputed values by using variants that are either known or have already been imputed, then the imputation task turns complex. Secondly, each imputed value should indicate both the full range of organizations with other variants and the full degree of our uncertainty. This may necessitate many forecasters to be enclosed, but Bayesian considerations would propose that the single constants should be shrunk. Thirdly, to pass on the uncertainty in our imputed values to the analysis stage, several datasets are created. In every dataset the determined values stays same but the imputed values may changes as a result of using different imputations. These different imputations should indicate the full uncertainty in the approximated models used for imputation (e.g. should draw parameter values from their sampling organization and not just re-use the same point estimates of the constants from the imputation regression) and not just add various 'residuals'. The preceding has been represented in the context of model based imputation but related thoughts apply where the imputation is based on some kind of matching and replacement of incomplete data records by complete data records – so called hot-deck methods. The last result is that instead of a single file with incomplete evidences we end up with several files but each one is complete. With modest amounts of missing data 5 imputation replicates are often sufficient. As the proportion and uncertainty of the missing data increases, so more replicates are required. Each of the, say m, files is then analyzed using naive methods that know nothing about any missing data, and the desirable statistic, be that a mean, a difference or a model coefficient, is estimated as if the data were truly all complete. The results from each of the m analyses are then united. The average of the computed statistics is reported as the point of estimation. The standardized error of this estimate is calculated from the simple variant of the point estimates over the m replicates and the average of the estimated variances being combined using what has become known as Rubin's Formula by equation (1).

$$\text{Overall standard error} = \text{sqrt}\,\{(1-1/m)\,B + W\} \qquad \mathbf{(1)}$$

Where m is the number of replicates, B is the variance of the imputations, and W is the average of the estimated variances.
Other than these methods there are more estimations i.e. replacement of missing values with the series mean, by the mean or median of nearby points, or linear interpolation between prior and subsequent known points, interpolating between the adjacent valid values above and below the missing one, or substitution of the linear regression trend value for that point i.e. missing values are replaced with their predicted values.
Mean substitution, once the most common method of imputation of missing values is no longer preferred or used. In this case the substitution of mean will reduce the variance of variables. If same cases are missing for two variables and if means are substituted, correlation can be inflated. This method creates a spiked distribution at the mean level in the frequency distribution that causes attenuation in correlation of item with others and underestimates the variance. Taking these effects into consideration, these will carry over in a regression context to lack of reliability of beta weights and related estimates of the relative importance of independent variables. It means, in the case of one variable mean substitution can lead to biased estimates of others or all variables presents in the regression analysis, since bias in one correlation effects the beta weights of all variables. The better solution to this is substitution of group mean for a categorical (grouping) variable known to correlate highly with the variable, having missing values. The mean substitution is no longer recommended

To predict the values of missing data, multiple regressions can be used. But it has to be noted that this may over-correct by introducing unrealistically low levels of noise in the data. The regression method has the problem that all instances with the same values on the independent attributes are imputed with the same value on the missing attributes, as a result of which same part of the problem used as mean substitution. Accordingly, preferred method for this is stochastic substitution that uses the regression method but adds a random value to the predicted result. The random value generated from this is the regression substance from a randomly selected case from a set of cases with no missing values. In regression estimates, it adds the residual of a randomly picked case to every estimate, even the user can select residuals, normal variants, Student's variants, or no adjustment. Irrespective of this, residual a guess and it is likely that the standard errors (and hence confidence intervals and probability values) will be smaller than they should be. The regression method expects that missing values are MAR (as opposed to MCAR). It also assumes that the same model explains the data for non-missing cases as for the missing cases, which, of course, is not need fully true. Ultimately**,** the user can set a maximum extent on the number of predictor variants used to estimate, as larger the number of predictors, greater the chance that the imputed estimation is molding noise in the data rather than an actual abstraction of model variants to missing data.
Little & Rubin [1] described, different methods available for handling missing values and their categorization. List wise or pair wise deletion, in the list wise deletion approach, also known as complete-case analysis, all instances with one or more missing values are deleted from the analysis. Pair wise

deletion or available-case analysis uses diverse sets of sample instances for each statistic. This approach conserves more information.

2. RELATED WORK

Kuligowski and Barros [3] proposed a use of a back propagation neural network for estimation of missing data by using concurrent rainfall data from neighboring gauges. Brockmeier et al [4] experimented on various missing data handling techniques and the authors have provided empirical comparative analysis of deletion and imputation Techniques.

Abebe et al. [5] proposed a use of a fuzzy-rule-based model for substitution in missing rainfall data using data from neighboring stations. The authors have provided empirical comparative analysis of results using the fuzzy-rule-based model and results using an ANN model and a traditional statistical model. The fuzzy-rule-based model performs slightly better.

Sinharay et al [6] experimented on the use of multiple imputations for the analysis of missing data. Khalil et al. [7] proposed cyclic federation of data intended for budding ANN models to estimate missing values in monthly surplus datasets. Bhattacharya et al [8] used ANN models to substitute the missing values of wave data. Fessant & Midenet [9] proposed use of a self-organizing map (SOM) for imputation of data along with the multilayer perceptron (MLP) and hot deck methods.

Musil et al. [10] provided empirical comparative analysis on list wise deletion, mean substitution, simple regression, regression with an error term and the EM algorithm. Junninen et al. [11] experimented on univariate linear, spline and nearest-neighbor interpolation algorithm, multivariate regularized expectation–maximization algorithm, nearest-neighbor, self-organizing map, multilayer perceptron (MLP) as well as hybrid methods where combining the best features of univariate and multivariate methods are combined in air quality datasets.

M. Subasi, et al [12] proposed new imputation method for incomplete binary data. Amman Mohammad Kalteh & Peder Hjorth [13] experimented on imputation of missing values with self organizing map, multilayer perceptron, multivariate nearest neighbor, regularized expectation maximization algorithm and multiple imputation for precipitation runoff process data set.

3. GENETIC ALGORITHMS

While dealing with larger, potentially huge search space Genetic algorithms provide global search through space in many directions simultaneously, thereby improving the probability of finding the global optimum to obtain optimal combinations of things and solutions [14]- [16].

Genetic algorithm initializes with a set of possible solutions and altering them during several generations, the Genetic Algorithm expects to meet on the most 'fit' solution. A set of possible solutions or chromosomes in the form of bit strings that are randomly generated or selected. The entire set of these chromosomes comprises a *population*. Genetic algorithms combine survival of the fittest among string structures with a structured yet randomized information exchange. To improve performance genetic algorithm efficiently uses the past information with randomized search on new search points. The offspring are evolved using the crossover and *mutation* technique. The chromosomes are then *evaluated* for a certain fitness values and the best solution is accepted while the remaining solutions are not needed. This process continues until final chromosome with best fitness value and thus is taken as the best solution of the problem.

Genetic Algorithms works with several advantages. It works sound for global optimization problems having the object function discontinuous or with several local minima. It can work without using auxiliary information such as gradients. The Genetic Algorithm can be used for combinatorial optimization problems.

4. THE DECISION TREE CONSTRUCTION

Decision tree [17]-[20] is a classifier in the form of tree data structure that contains a decision node and leaves. A leave specifies a classification. A decision node specifies a test to be carried on single attributes value. A solution is present for each probable outcome of the test in the form of child node. A performance measure of a decision tree over a set of cases is called classification accuracy. It is defined as the percentage of correctly classified instances.

5. THE EVOLUTIONARY DECISION

A hybrid learning methodologies that integrates genetic algorithms (GAs) and decision tree learning in order to evolve optimal decision trees has been proposed by different authors. Although the approaches are different the objective is to obtain optimal decision trees. The GAIT algorithm proposed by Z. Fu [21] generate a set of diverse decision trees from different subsets of the original data set by using a decision tree algorithm C4.5, on small samples of the data. These decision trees are taken as inputs (the initial population) to genetic algorithm. The fitness criterion for evaluation is the classification accuracy on test data. A. Papagelis, and D. Kalles proposed GAtree, a genetically evolved decision trees [14]. The Genetic Algorithms is used to directly evolve binary decision Trees, without using binary string i.e. actual decision trees (A decision trees that have one decision node with to two different leaves) are operated by genetic operators and not on strings. Thus constructed Trees are called GAtree Genetically Evolved decision Trees. The next section explains problem definition and proposed algorithm.

6. PROBLEM DEFINITION & ALGORITHM

Let T_f be a set of all available n training instances. Let the training instance be denoted by t. An instance denotes values for set of attributes and a class. Let the attributes be denoted by $\{a_1, a_2, \ldots., a_n\}$ and the classes be denoted by the set values $\{C_1, C_2, \ldots, C_n\}$. Let some attribute a_n are with missing values and let T_e be a set of instances that includes set of instances with missing attribute values with normal data instances. The proposed algorithm for imputation of missing data is a genetic algorithm based method that uses global search techniques to find value of missing attribute. This experimentation is limited to categorical attribute values only. We use domain values for a particular attribute as a population; a pool of solution to impute the missing value and combination of all possible values provides crossover operation. The classification accuracy on ensemble of decision tree classifiers is the fitness function. Details are explained in next sections. The proposed algorithm to replace missing values with some plausible values works as follows..

```
    Imputation GA Algorithm
1. Find attribute with missing value.
2. For every attribute find set of
   domain values of missing attribute.
3. Substitute missing attribute with
   all available values from domain and
   make possible set of instances with
   available domain values of missing
   attributes.
4. Repeat above steps for all
   attributes with missing values.
5. From this pool of solution do
   selection of instances.
6. Do crossover on selected instances.
7. Do validation with fitness function
   i.e. Classification accuracy on
   decision tree.
8. If instance is classified, values
   substituted are validated else
   delete that unclassified instance.
   Classified instances substituted
   attribute values are successfully
   imputed values.
9. Repeat this procedure for expert
   vote.
10.      End.
```

Table 1 Comparison classification accuracy on classifiers

No.	Data Set	X_e	X_i	ΔX
GATree				
1	Breast	71.79	80.57	08.78
2	Weather	60.00	85.00	25.00
3	Lymph	72.86	84.21	11.35
J48				
1	Breast	74.13	79.27	5.14
2	Weather	68.97	73.91	4.94
3	Lymph	76.35	79.79	3.44
CART				
1	Breast	69.23	80.67	11.44
2	Weather	65.52	73.91	8.39
3	Lymph	75.00	79.79	4.79
			Average ΔX	9.25

7. EXPERIMENTATION AND RESULTS

Using three different dataset [22] from University of California Irvine repository experiments were performed. The missing data was introduced in some data instances so that we have the actual values of missing attribute. The missing values were substituted as per proposed algorithm. The original data set T_f was partioned in four exclusive datasets T_0, T_1, T_2 and T_3 for expert vote In this implementation we have used implementation of GA Tree [14] as evolutionary decision tree classifier to validate the fitness function. At the same time to check the validity of proposed algorithm on other classifiers, we have done testing on J48 and Simple CART [23].

The four different hypotheses H_0, H_1, H_2, H_3 on partioned data were induced and the instance with substituted values was validated on all four hypotheses. The instance with highest classification score was considered as correctly imputed instance. The test instance data if classified on H the score was flagged as 1 otherwise 0. The total score on four trees were added to get final score as a fitness function. The final imputed instances achieved from this algorithm were added to data set with missing attribute values T_e and the instances with missing values were excluded from the set. Let us call the resulting imputed data set as $T_{i.}$ The datasets T_i and T_e were tested on all classifiers mentioned above for classification accuracy. Readings of classification accuracy of tree were obtained using 10 fold classification method this validation method gives accurate results.

The results for proposed algorithm as per experimental method explained above are summarized in Table 1. Let X_i be classification accuracy of the trees build on data set T_i and similarly Xe on T_e. The result table summarizes the absolute difference in accuracy (ΔX) between the tree build on imputed instance data set $T_{i.}$ and the trees build on the training instances with missing data T_e.

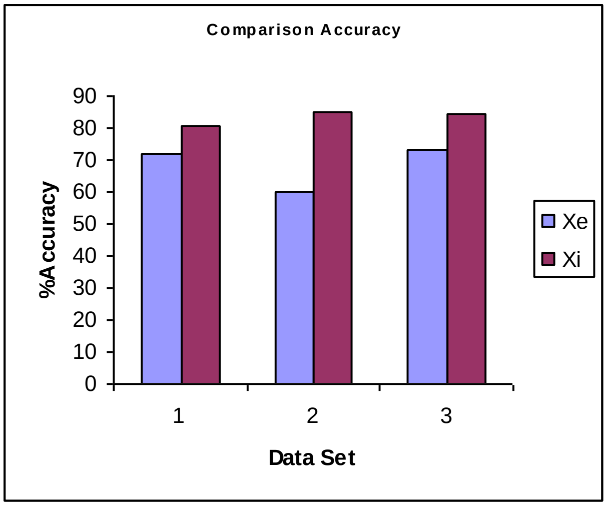

Fig.1 Comparison accuracy on GATree

It has been observed that percentage enhancement in tree classification accuracy on imputed data set is significant as compared to data set with missing data in all thee data sets on all classifiers. The improvement in classification accuracy is minimum of 3.44% for Lymphography data where as it is maximum of 25% for weather data set. The Average ΔX for all experiments on GATree, J48 and Simple CART jointly is 9.25%. The improvement in accuracy is significant. Figure 1 provides graphical visualization of results on GATree.

8. CONCLUSION

This paper proposes a new methodology for imputing the missing attribute values, the methodology that integrates genetic algorithms (GAs) techniques and decision Tree learning for imputation of the missing attribute values. The proposed imputation algorithm uses domain values for missing attribute as possible solution and the set of instances with imputed attribute values is used as pool of solutions. This pool of solution is used as chromosomes in genetic algorithm. The decision trees, the evolutionary decision tree GATree are used to evaluate the fitness function. The method incorporates genetic algorithm to pursue global search in the problem space with classification accuracy as fitness function without being biased towards a local optimum that gives us best classification accuracy. The proposed method is also tested on J48 and Simple CART to check the validity of proposed algorithm and was found doing well. The method is experimented on categorical values only. The classification accuracy of decision tree is improved. The proposed algorithm when applied on missing data provides us sufficient instances with imputed data values for knowledge acquisition. We can use knowledge in instances with missing data attributes data that may be most important in seeking some decision. The data imputation on training data set helps in induction of enhanced and accurate hypothesis.

10. REFERENCES

[1] Little R. J. and Rubin D. B. 1987. *Statistical Analysis with Missing Data*. John Wiley and Sons, New York.

[2] Schafer J. L. and Graham J. W. 2002. Missing data: our view of the state of the art Psychol. Methods 7(2), 147–177.

[3] Kuligowski R. J. & Barros A. P. 1998. Using artificial neural Networks to estimate missing rainfall data. Journal AWRA 34(6), 14.

[4] Brockmeier L. L., Kromrey J. D. and Hines C. V., 1998. Systematically Missing Data and Multiple Regression Analysis: An Empirical Comparison of Deletion and Imputation Techniques. *Multiple Linear Regression Viewpoints*, Vol. 25, 20-39.

[5] Abebe A. J., Solomatine D. P. & Venneker R. G. W. 2000. Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events. Hydrological Sciences Journal.45 (3), 425–436.

[6] Sinharay S., Stern H.S. and Russell D. 2001. The use of multiple imputations for the analysis of missing data. Psychological Methods Vol.4: 317–329.

[7] Khalil K., Panu M. and Lennox W. C. 2001. Groups and neural networks based stream flow data infilling procedures. Journal of Hydrology, 241, 153–176.

[8] Bhattacharya B., Shrestha D. L. & Solomatine D. P. 2003. Neural networks in reconstructing missing wave data in dimentation modeling. In the Proceedings of 30th IAHR Congress, Thessaloniki, Greece Congress, August 24-29 2003 Thessaloniki, Greece.

[9] Fessant F. & Midenet, S. 2002. Self-organizing map for data imputation and correction in surveys. Neural Comput. Appl. 10, 300–310.

[10] Musil C. M., Warner C. B., Yobas P. K. & Jones S. L. 2002. A comparison of imputation techniques for handling missing data. Weston Journal of Nursing Research 24(7), 815–829.

[11] Junninen H., Niska H., Tuppurainen K., Ruuskanen J. & Kolehmainen M. 2004. Methods for imputation of missing values in air quality data sets. Atoms. Environ. 38, 2895–2907.

[12] M. Subasi, E. Subasi and P.L. hammer, 2009. New Imputation Method for Incomplete Binary Data, Rutcor Research Report, August 2009.

[13] Amman Mohammad Kalteh & Peder Hjorth, 2009. Imputation of Missing values in precipitation-runoff process database. Journal of Hydrology research.40.4, pages 420—432.

[14] 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.

[15] Rajasekaran G.A., Vijayalakshmi Pai, 2004. Neural Networks Fuzzy Logic and Genetic Algorithms Synthesis and Applications, Prentice-Hall of India.

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

[17] Quinlan J. R. 1993. C4.5: Programs for machine learning. Morgan Kaufman, San Mateo.

[18] Salvatore Ruggieri, 2002. Efficient C4.5, IEEE Transaction On Knowledge and Data Engineering, Vol. 14, No. 2 March/April.

[19] Endou T. and Qiangfu Zhao, 2002. Generation of comprehensible decision trees through evolution of training data, Proceedings of the 2002 Congress on Evolutionary Computation, 2002. Volume 2, 12-17 May.

[20] Quinlan J. R., 1990. Decision Trees and Decision making IEEE Transaction On Systems, Man, And Cybernetics vol. 20, No. 2, March/April.

[21] Zhiwei Fu, Fannie Mae, 2001. A Computational Study of Using Genetic Algorithms to Develop Intelligent Decision Trees, Proceedings of the 2001 IEEE congress on evolutionary Computation.

[22] Newman D.J. & Hettich S. & Blake, C.L. & Merz C.J. UCI Repository of machine learning databases [http://www.ics.uci.edu/].