Missing Data Imputation Through the Use of the Random Forest Algorithm

Adam Pantanowitz and Tshilidzi Marwala*

School of Electrical & Information Engineering
University of the Witwatersrand, Johannesburg
Private Bag 3, Wits, 2050, South Africa
adam.pantanowitz@wits.ac.za

Abstract. This paper presents a comparison of different paradigms used for missing data imputation. The data set used is HIV seroprevalence data from an antenatal clinic study survey performed in 2001. Data imputation is performed through five methods: Random Forests; auto-associative neural networks with genetic algorithms; auto-associative neuro-fuzzy configurations; and two random forest and neural network based hybrids. Results indicate that Random Forests are superior in imputing missing data for the given data set in terms of accuracy and in terms of computation time, with accuracy increases of up to 32 % on average for certain variables when compared with auto-associative networks. While the concept of hybrid systems has promise, the presented systems appear to be hindered by their auto-associative neural network components.

Keywords: auto-associative, imputation, missing data, neural network, random forest.

1 Introduction

Real world studies are often impacted negatively due to data that are missing. This common problem creates difficulty with data analysis, study and visualisation [1,2]. Insights into characteristics of the data may be reduced due to missing information and, furthermore, the underlying cause for the missing data may make the missing data particularly interesting or of significance to the study. Cascaded systems, such as those responsible for decision making based on certain decision making policies, may be hindered by the missing information and rendered unusable. For these reasons, it is important to find effective and viable methods to impute missing data.

This paper evaluates the concept, classification, problem and treatment of missing data. A background on the methods and paradigms used is provided, followed by a description of the implementation. The data set is considered, and thereafter, comparisons are drawn between the implemented paradigms. Finally a discussion is presented and conclusions are reached.

* Tshilidzi Marwala has since become Executive Dean of the Faculty of Engineering and the Built Environment at the University of Johannesburg. P.O. Box 524, Auckland Park, 2006, Johannesburg, South Africa. tmarwala@uj.ac.za

W. Yu and E.N. Sanchez (Eds.): Advances in Computational Intell., AISC 61, pp. 53–62.
springerlink.com

2 Missing Data

Missing data are a problem inherent and common in data collection, especially when dealing with large, real world data sets. Missing data impact on decision support systems and make statistical evaluation methods difficult to perform. Results are degraded through the use of arbitrary or random assignment to the missing data elements [3]. In surveys in particular, information may be missing due to incomplete variable collection and non-response from subjects, poorly defined surveys, and data being removed for reasons such as confidentiality [1,2]. These explanations for missing survey data may provide insight into the large proportion of missing data in the set used in this study, as discussed in section 5.

2.1 Missing Data: Categorisation and Mechanism

Missing data can be categorised based on the pattern and mechanism of absence. The methods with which the missing data are dealt are dependent on this categorisation. Three broad categories for pattern absence are defined: *monotone missingness*, *file matching*, and *general missingness* [4,5].

Missing data are also often classified into one of three mechanisms, as defined by Little and Rubin [4]. The mechanisms, in order from least to most dependent on other information, are: missing completely at random (MCAR); missing at random (MAR); and the non ignorable case [1,4]. In the MCAR case, data cannot be predicted using any information in the set, known or unknown. For the MAR mechanism, there is a correlation between the missing data and the observed data, but not necessarily on the values of the missing data [6].

2.2 Dealing with Missing Data

A number of strategies have been devised for dealing with missing data. The simplest means is discarding the instances for which data are missing (a complete-case method), which is inefficient and potentially leads to biased observations and information waste [1]. Despite this, the method is used commonly in practice [2]. Other techniques include *available-case procedures*, *weighting procedures* and *imputation-based* procedures [6]. We consider imputation methods which involve predicting the values of the missing data and can be applied to MCAR and MAR cases [7]. Two categories of techniques exist, non-model based and model-based. Non-model based approaches include mean imputation and hot-deck imputation, techniques which are said to decrease the variance in statistical procedures, lead to standard errors and to result bias [6]. Model-based approaches include regression-based techniques, multiple imputation [7], expectation maximisation [8] and neural network (NN) based approaches [1,8,9].

3 Background

A number of learning paradigms are considered which form networks and hybrid networks for comparative purposes in this work. These are generally connected in auto-associative configurations [9], as discussed in section 4.

3.1 Random Forests

Ensemble or network committees are algorithms in machine learning that combine individual paradigms to form combinations that are often more accurate than the individual classifier alone [10]. In the classification case, overall predictions can be obtained from such a network using a weighted or an unweighted voting system; in the regression case, overall predictions can be chosen through an averaging technique. Obtaining a general understanding of why such methods succeed is an active area of research [10,11].

A *decision tree* is a tree with nodes that contain information corresponding to attributes in the input vectors. This information is used to follow a decision path for a given set of input attributes, depending on either thresholding nodes (as in the case of a continuous variable) or categorical nodes (as in the case of categorical data) [12]. Even though decision trees have appeal for being straightforward and fast, they are prone to being overly adapted to the training data or to a loss in accuracy for generalisation through tree pruning [13].

"Random Forest" (RF) is an algorithm that generalises ensembles of decision trees [14], with the ability to perform both regression and classification. RFs make use of bagging (bootstrap aggregation) to combines multiple random predictors in order to aggregate predictions [15], allowing for high complexity without over-generalising and over-fitting to the training data [13]. RF has been used with success in the context of missing data [12]. RFs were first introduced in 2000 and are a trademark of Cutler and Breiman [10].

If Θ is the possible variables, and $h(x, \Theta)$ denotes a tree grown using Θ to classify a vector x, then an RF can be defined as $f = \{h\left(x, \Theta_k\right)\}, k = 1, 2, ..., K$, in which $\Theta_k \subseteq \Theta$ [12,15]. Each tree contains an individually selected *subset* of the overall attribute collection. The algorithm for RF growth is presented in [14].

RFs are a good candidate for a missing data study [10,14] and have recently been an area of active research since they have advantageous features and high success rates [10]. They are fast and have accuracy greater than that of single classification and regression trees (CART). They are furthermore impervious to over-fitting the data and do not have dimensionality problems - running effectively on thousands of variables. RFs give a self-assessment and have built in variable importance assessment capabilities [14].

3.2 Other Paradigms

Multilayer Perceptron (MLP) Artificial Neural Networks (ANNs) are ANNs that consist of an interconnection of the processing elements, generally placed in three classes: the input layer, the output layer and the hidden layer [16]. A process of supervised learning allows the weights of the network to be adjusted yielding a feed-forward network capable of modelling complex and non-linear relationships [17]. A number of different optimisation strategies are available in training the network, such as conjugate gradient descent [16].

Fuzzy Inference Systems (FISs) operate through a process of fuzzification, operation and implication [18]. FISs are well suited to knowledge of linguistic if-then rules, offering an advantage in terms of learning capability, while data based

learning is better suited to ANNs [19]. Neuro-fuzzy systems provide the benefit of both subsystems and consist of rule sets and inference systems combined with or governed by a connectionist structure for optimisation and adaptation to given data. Adaptive neuro-fuzzy inference system (ANFIS) implements a Takagi Sugeno (TS) FIS and consists of five layers, as indicated by the schematic architecture of the system presented in [20,21].

Genetic Algorithms (GAs) are search methods that are widely used to solve optimisation problems. Genetic algorithms employ their heuristic search by modelling properties of biological evolution including: crossover; inheritance; mutation; and selection, to solve optimisation problems [22]. Convergence exists due to the fitness of an individual (representing an element in the search space that may be an appropriate solution to the problem) in a given population dominating over another individual [22,23]. Through an evolutionary process, survival of the fittest ensues.

Auto-Associative NN Encoder Networks are system models which are trained to recall an input. The number of outputs is thus equal to the number of inputs [9]. The networks usually have fewer hidden nodes than inputs (or outputs), creating a *bottleneck*. The auto-encoder network detects missing data by forward propagating known elements and a predicted value for the unknown elements, and minimising the overall error between the input and the output using an intelligent search method, such as a GA [8,9].

4 Methodology and System Topologies

The RFs used throughout the analysis generally have 70 trees. The parameter for minimum size of terminal nodes set at 7, and number of variables to be sampled randomly at each split (m) is set to 3 (much less than the total number of inputs M which is 14 [24]). This combination was determined experimentally to be the optimal set of parameters through maximisation of the number of hits (i.e. the number of correct predictions). Regression RFs are used when predicting ordinal variables, which are encoded to be continuous values ranging from 0 - 1, and classification RFs when predicting categorical variables, such as HIV status (indicated in table 1). Since each RF predicts a single variable, an attempt was made to form RFs to predict each of the fourteen variables, and combine this with a GA to form an RF based auto-associative network. This method does not yield favourable results, but the resulting RFs are used to impute the different missing variables. For multiple missing values, the methodology presented in figure 1 is employed.

The **Auto-associative Neural Network** combined with a **GA** (**AANN-GA**) is used as in [9]. For the 14 input/output system, the optimal number of hidden nodes, determined experimentally, is 11. The number of training cycles is 400, based on the minimum point of the validation curve and a linear activation function is used with scaled-conjugate descent training.

The **Auto-associative Adaptive Neuro-Fuzzy Inference System** with a **GA** (**AANF-GA**) implements a network of 14 ANFIS networks. Since each

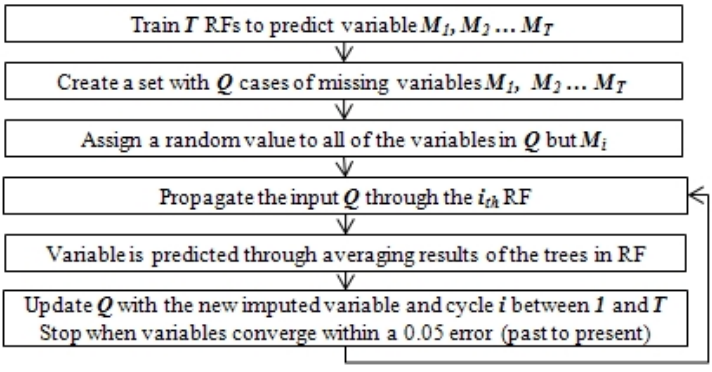

Fig. 1. Flow chart indicating the concept used in imputing multiple missing variables with regression and classification RFs

network predicts a single value, a system of 14 is set up in order to minimise the error between the input and the output in an auto-associative configuration. Each of the ANFIS systems uses subtractive clustering to train, with a training radius of 0.2, 20 training epochs, and a stopping criterion of 0.01.

In the **Random Forest & Auto-associative Neural Network Hybrid** topology, the RF is placed in cascade with an AANN-GA, forming the **RF-AANN-GA**. The RF is used to predict a set of missing variables in an experiment set, and the predictions are recorded. These predictions are then used as limits for the search space of the GA in the AANN-GA system. Since the variable range is 0 - 1, a tolerance of 10 % is added the variable such that the GA has a slightly broader search space. The principle is that by limiting the search space, the AANN-GA will have improved efficiency. A similar principle is successfully applied in [1], in which C4.5 decision trees are used to limit the search space. The AANN-GA and RF have the same structures and parameters as the aforementioned standalone optimised structures.

The **Auto-Associative Neural Network & Random Forest Hybrid** topology combines the AANN-GA system in cascade with an RF, forming the **AANN-GA-RF**. The principle behind the operation is that the RF learns the underlying problems with the AANN-GA system and compensates for them. In order to achieve this, the data are divided into four sets: training; validation; testing and experimental. The training and validation data are used to train and obtain the best model for the AANN-GA using *early stopping* [25]. Thereafter, data are removed from the testing and experimental sets to yield artificially incomplete sets, and these incomplete sets are propagated through the AANN-GA to obtain missing data predictions from the AANN-GA. The testing data and imputed values are made to form a complete set. This testing set is then used as a training set for the RF, with the target being the *original* correct data. In this way, the RF is made to compensate for the error introduced by the AANN-GA. The experimental set is then used to test the RF.

5 Data Evaluation and Preprocessing

Preprocessing of data allows for the data to be of appropriate form for the machine learning paradigms. The data set used is based on a study performed in 2001 for national human immunodeficiency virus (HIV) and syphilis seroprevalence in women attending antenatal clinics in South Africa [26]. The variables contained in the data set are outlined in table 1. Note that the data ranges given are for once the variables have been processed according to the given rules. Gravidity refers to the number of times a woman has been pregnant, and parity the number of times the woman has given birth. Father's age refers to the age of the father responsible for the current pregnancy. Education is specified as 0 (no education); 1 - 12 (for grades 1 through to 12); or 13 (tertiary education). Province categorises a person in to one of the nine South African provinces, and race categorises a person in to one of six race categories [27].

Table 1. Outline of data set variables (adapted from [27])

Variable	Data Type	Range	Variable Type
Province (location)	Integer	1 - 9	Categorical
Age	Integer	12 - 50	Ordinal
Education	Integer	0 - 13	Ordinal
Gravidity	Integer	1 - 12	Ordinal
Parity	Integer	0 - 9	Ordinal
Father's age	Integer	12 - 90	Ordinal
HIV status	Binary	0/1	Categorical
Rapid Plasma Reagin (RPR) test status	Binary	0/1	Categorical
Race	Integer	0 - 5	Categorical

Since we are dealing with a real-world study involving missing data, the original data contain inherent errors. There were 16 743 pregnant women involved in the study, of which just under 12 000 instances are regarded as complete and/or valid according to the rules applied. A number of those regarded as outliers contained spurious data or missing data. The problem of missing data is immediately apparent from this statistic. In order to yield a complete set, all fields must be valid as specified according to the ranges indicated in table 1 and in accordance with the logical rules that data cannot be negative and that gravidity cannot be less than parity. The data are preprocessed to ensure this and for normalisation of the data. The categorical data of race and province are binary encoded, since these variables are not ordinal, and non-encoded variables may interfere with the performance of the learning paradigms [27].

6 Comparison and Results

Table 2 indicates the results on testing the missing data prediction ability of the various systems. The results are found by predicting missing data of the

indicated variables, and calculating the percentage of values accurately predicted within specified ranges. MAR and MCAR are not distinguished. Note that the C4.5 AANN-GA results, provided as benchmark results for comparison, may be biased due to the experiments being performed under different conditions, since they are obtained for the appropriate ranges from [1]. The ranges are indicated in the table (for example, age prediction is assessed for percentages of correct prediction within 1, 2, 4, 6 and 10 years).

Testing is performed to determine the best of the techniques specified in section 4. It is evident from the result that the RF and RF hybrids outperform the other methods of missing data prediction. There is significant improvement in the RF from the commonly used AANN-GA method, with an average percentage increase of 7.6 % for the indicated categories. The RF prediction of education increases by an average of 31.2 % when compared with the AANN-GA across the

Table 2. Results of percentage prediction accuracy for given methods within the specified ranges

Quantity	**Range** (Within)	**RF**	**RF-AANN-GA**	**AANN-GA RF**	**AANF-GA**	**AANN-GA**	**C4.5, AANN-GA** [1]
Age (years)	1	41.1	35.1	40.2	22.0	34.7	-
	2	62.3	55.9	60.8	36.7	54.7	52.3
	4	85.7	83.0	85.0	54.0	81.1	79.4
	6	95.0	92.8	93.9	68.7	90.5	89.6
	10	99.2	98.5	99.0	80.7	96.5	97.9
Education (grades)	0	16.7	19.7	15.4	6.5	5.5	-
	1	53.5	48.1	51.5	18.5	24.3	52.1
	2	76.9	69.4	75.7	34.5	35.8	69.5
	3	88.3	83.7	88.3	46.5	46.4	79.4
	5	93.1	90.8	93.2	70.0	54.8	91.8
Gravidity (Instances)	0	88.0	88.1	88.1	0	88.0	80.4
	1	98.3	98.2	98.2	13.7	98.2	97.1
	2	99.5	99.4	99.4	35.7	99.4	-
	3	99.8	99.7	99.7	67.0	99.8	99.6
	5	100.0	100.0	100.0	95.0	100.0	100.0
Parity (Instances)	0	89.4	87.6	89.5	0	87.9	60.8
	1	98.5	98.3	98.4	21.5	98.2	92.9
	2	99.6	99.5	99.6	52.0	99.4	-
	3	100.0	99.9	100.0	74.0	99.8	89.6
	5	100.0	100.0	100.0	94.0	100.0	97.9
Father's Age (years)	1	28.8	27.9	28.3	3.5	27.7	-
	2	45.7	45.6	46.2	11.5	45.9	41.7
	4	74.1	72.8	73.6	22.5	72.1	68.6
	6	86.3	86.2	86.7	32.0	86.1	82.7
	10	95.3	94.4	95.0	53.0	94.2	93.2

Table 3. Relative computation time taken for the various indicated methods for propagation through 5000 instances with missing data

Method	Training Time (s)	Propagation Time (s)
RF	0.04	0.5
AANF-GA	797.4	50964
AANN-GA	20.7	628.3

specified ranges. The improvement from the AANN-GA to the AANN-GA-RF is a significant one, indicating that the hybrid method of section 4 is working, but the results are comparable to the standalone RF. Furthermore, in the case of the RF-AANN-GA it is observable through experimentation that narrowing the search bounds of the GA improves the performance. Thus, introducing the AANN-GA with larger search bands starts to degrade the performance of the hybrid, indicating that this hybrid's results are suffering from problems within the AANN-GA and not within the RF. The AANN-GA and AANF-GA perform relatively badly in different aspects: age prediction (for the AANF-GA) and education prediction (for the AANN-GA). The RF performs well in all respects and does not suffer the drawbacks of the other paradigms in predicting age or education.

While the hybrid methods appear to show potential, the computational time trade-off for the use of these methods (due to the cascade with AANN-GAs) is not warranted for the performance improvement. This is especially so in lieu of the relative computation time taken, as indicated in table 3. Note that the study to obtain this table was performed in MATLAB and the programming is therefore not standardised. This result should thus be treated as a basic evaluation. That said, it is to be noted that RF is generally documented as being a relatively fast machine learning tool [10,13,14,24], and this is reflected by the table. It is evident that there are vast improvements in the computational efficiency of the RF algorithm.

The HIV status is predicted by an RF classifier and the results are presented in table 4. The other configurations were also used to predict HIV status. The AANN-GA obtains prediction accuracy of 64.2 % with an F-measure of 0.43. This is not, however, discussed further in this work. The classification results obtained for the RF are lower than those found in [9].

Table 4. HIV Status Prediction Confusion Matrix for RF Classifier run on experiment set

Confusion Matrix	Predicted Negative	Predicted Positive	Percentage Error (%)
Actual Negative	2902	1732	37.4
Actual Positive	449	875	33.9

7 Discussion and Recommendations for Future Work

This work is extended in a missing data impact assessment [28]. It is notable that the AANN-GA method is relatively computationally expensive and are shown to be outperformed when compared with RFs used for the purpose of estimating missing data for this data set [8,9]. In this paper, either regression or classification RFs are used. RF does, however, have built in functionality to estimate missing data through computing terminal node proximities [14]. This functionality was not tested within this work, but RFs implementing this method may show further improvement. The RF HIV classifier does not perform well when compared with classifiers using the same data [9]. The results obtained from the AANF-GA are relatively poor. Despite the fact that ANFIS struggles with high dimensionality data, training was possible through subtractive clustering. It was not feasible, however, to train the ANFIS system with grid partitioning unless variables were removed, and this impacts on the standardisation of the comparison with the other systems.

8 Conclusion

Missing data causes significant information loss in studies as information is wasted and insight cannot be gained into the underlying causes of the absence. Through the use of data resulting from an HIV seroprevalence survey, this paper investigates and compares five machine learning paradigms in order to impute missing data: RFs, AANN-GA, AANF-GA, RF-AANN-GA and AANN-GA-RF. It is evident from the presented results that the RF algorithm as a regression system to impute the missing data outperforms the other paradigms investigated for the studied data set. This is true for both computation time and computation accuracy, with RFs outperforming the other paradigms by up to 32 % on average for some categories.

References

1. Ssali, G., Marwala, T.: Computational Intelligence and Decision Trees for Missing Data Estimation. In: Proceedings of the International Joint Conference on Neural Networks, part of the IEEE World Congress on Computational Intelligence, WCCI 2008, IJCNN, pp. 201–207. IEEE, Los Alamitos (2008)
2. Horton, N.J., Kleinman, K.P.: Much Ado About Nothing: A Comparison of Missing Data Methods and Software to Fit Incomplete Data Regression Models. The American Statistician 61(1), 79–90 (2007)
3. Markey, M.K., Tourassi, G.D., Margolis, M., DeLong, D.M.: Impact of Missing Data in Evaluating Artificial Neural Networks Trained on Complete Data. In: Computers in Biology and Medicine, vol. 36, pp. 517–525. Elsevier, Amsterdam (2006)
4. Little, R.J., Rubin, D.B.: Statistical Analysis with Missing Data. John Wiley & Sons, Chichester (2002)
5. Ziegler, M.L.: Variable selection when confronted with missing data. PhD thesis, University of Pittsburgh (2006)

6. Fogarty, D.J.: Multiple Imputation as a Missing Data Approach to Reject Inference on Consumer Credit Scoring. Intersat 41(9) (2006)
7. Yuan, K.H., Bentler, P.M.: Three likelihood-based methods for mean and covariance structure analysis with non-normal missing data. Sociological Methodology, 165–200 (2000)
8. Nelwamondo, F.V., Mohamed, S., Marwala, T.: Missing data: a of neural network and expectation maximisation techniques. Current Science 93(11), 1514–1521 (2007)
9. Betechuoh, B.L., Marwala, T., Tettey, T.: Autoencoder Networks for HIV Classification. Current Science 91(11), 1467–1473 (2006)
10. Biau, G., Devroye, L., Lugosi, G.: Consistency of Random Forests and Other Averaging Classifiers. Journal of Machine Learning Research 9, 2015–2033 (2008)
11. Masisi, L., Nelwamondo, F.V., Marwala, T.: The Effect of Structural Diversity of an Ensemble of Classifiers on Classification Accuracy. In: IASTED International Conference on Modelling and Simulation (Africa-MS) (2008)
12. Qi, Y., Klein-Seetharaman, J., Bar-Joseph, Z.: Random Forest Similarity for Protein-Protein Interaction Prediction from Multiple Sources. In: Pacific Symposium on Biocomputing, vol. 10, pp. 531–542 (2005)
13. Ho, T.K.: Random Decision Forests. In: ICDAR 1995: Proceedings of the Third International Conference on Document Analysis and Recognition, vol. 1 (1995)
14. Breiman, L., Cutler, A.: Random Forests. Department of Statistics, University of California Berkeley (2004)
15. Brence, J.R., Brown, D.E.: Improving the Robust Random Forest Regression Algorithm (2006)
16. Engelbrecht, A.P.: Computation Intelligence, an Introduction. John Wiley & Sons, Ltd., Chichester (2002)
17. Haykin, S.: Neural Networks: A Comprehensive Foundation. Macmillan, New York (1994)
18. Jang, J.S.R., Gulley, N.: Fuzzy Logic Toolbox. The MathWorks Inc. (1997)
19. Abraham, A.: Neuro fuzzy systems: Sate-of-the-art modeling techniques. In: Mira, J., Prieto, A.G. (eds.) IWANN 2001. LNCS, vol. 2084, pp. 269–276. Springer, Heidelberg (2001)
20. Jang, J.S.R.: Neurofuzzy Modelling and Control. Proceedings of the IEEE 83 (1995)
21. Jang, J.S.R.: Input Selection for ANFIS Learning. In: Proceedings of IEEE International Conference on Fuzzy Systems (1998)
22. Wong, H.: Genetic Algorithms. Surprise 96 Journal. Imperial College of Science Technology and Medicine (1996)
23. Zalzala, A.M.S., Fleming, P.J.: Genetic Algorithms in Engineering Systems. IET (1997)
24. Breiman, L.: Random Forests. Machine Learning 45, 5–32 (2001)
25. Yuan, Y., Lorenzo, R., Andrea, C.: On Early Stopping in Gradient Descent Learning. Constructive Approximation 26(2), 289–315 (2007)
26. Ntsaluba, A.: Summary Report: National HIV and Syphilis Sero-prevalence Survey of Women Attending Public Antenatal Clinics in South Africa, 2001 Department of Health, South African Government (2001)
27. Mistry, J., Nelwamondo, F.V., Marwala, T.: Investigation of Autoencoder Neural Network Accuracy for Computational Intelligence Methods to Estimate Missing Data. In: IASTED International Conference on Modelling and Simulation (2008)
28. Pantanowitz, A., Marwala, T.: Evaluating the Impact of Missing Data Imputation. LNCS (LNAI), vol. 5678. Springer, Heidelberg (to appear, 2009)