Estimating Missing Data Using Neural Network Techniques, Principal Component Analysis and Genetic Algorithms.

Abdul K. Mohamed, Fulufhelo V. Nelwamondo and Tshilidzi Marwala

School of Electrical and information Engineering, University of the Witwatersrand, Johannesburg, South Africa
abdul.mohamed@students.wits.ac.za f.nelwamondo@ee.wits.ac.za t.marwala@ee.wits.ac.za

Abstract

The common problem of missing data in databases is being dealt with, in recent years, through estimation methods. Auto-associative neural networks combined with genetic algorithms have proved to be a successful approach to missing data imputation. Similarly, two new auto-associative models are developed to be used along with the Genetic Algorithm to estimate missing data and these approaches are compared to a regular auto-associative neural network and Genetic algorithm approach. One method combines three neural networks to form a hybrid auto-associative network, while the other merges Principle Component Analysis and neural networks. The hybrid network and Genetic Algorithm approach proves most accurate, when estimating one missing value, while the PCA and neural network version is more consistent and captures patterns in the data most efficiently, in the chosen application.

1 Introduction

The presence of missing or inaccurate data in databases, such as those that house medical and specifically HIV data, is a common problem [1]. Missing data may result from inefficiencies in the data acquisition or data storage processes [1]. Non-response to various fields in a questionnaire, the incorrect insertion of data into a database, a break in the transmission line and failure of hardware are common causes of missing data [1, 2]. Many knowledge-discovery and data analysis techniques for databases depend heavily on complete data [1, 3]; hence an effective method of dealing with the missing data is required [1, 2, 3].

In the past, cases with missing data were simply deleted; however this approach may result in biased or erroneous analysis results [4, 5]. If a sensor fails, the value for that sensor will need to be estimated quickly and accurately based on the values of the other sensors [2], and as a result, case deletion will not be appropriate. This scenario illustrates the idea of data imputation, where missing data are predicted based on existing data [5].

Recent years have shown an increased interest in dealing with missing data by estimation or imputation [1, 2, 4]. An Auto-associative Neural Network (AANN) coupled with the Genetic Algorithm (GA) has been shown by Abdella and Marwala [2] to be a successful approach to missing data estimation [1, 2]. The efficient estimation of missing data relies on the extraction and storage of the relationships or correlations between the variables that make up the dataset [1]. AANNs allow this to be done [1, 6], but other techniques such as Principal Component Analysis (PCA) [6, 7] can also be used.

In this paper, a hybrid auto-associative network is developed and its performance in conjunction with the GA is compared to that of an ordinary AANN. A PCA and neural network missing data estimation system is also developed and compared to the other two systems. A description of the methodology is presented followed by the experimental implementation using HIV data from the department of health in South Africa. Results from this implementation are presented and conclusions are drawn.

2 Background

2.1 Missing Data

In order to deal with missing data effectively, it is important to understand how data goes missing so as to identify a possible cause to or pattern in the missing data [1]. Causes for missing data are commonly described as falling into three categories [5].

Firstly, 'Missing Completely at Random' (MCAR) describes the case when the probability of a value of a variable missing does not depend on itself or on any of the other variables in the dataset [2]. Effectively the cases with missing entries are the same as the complete cases [2, 5]. Secondly, 'Missing at Random' (MAR) describes the case where missing data can be described fully by the remaining variables in the dataset [5]. The missing datum depends on the remaining data, but not on itself or on any other missing data [1, 2]. The final case, 'Missing Not at Random' (MNAR), occurs when the missing datum depends on itself or on other missing data. [1]. Thus it cannot be deduced from the existing data and is hence termed the non-ignorable case [1, 2, 5].

Based on the mechanism through which the data is determined to have gone missing, a suitable approach for dealing with missing data can be adopted [1, 2]. MAR and MCAR are referred to accessible mechanisms, where the cause of missingness can be accounted for [5]. For the case of MNAR, which is described as inaccessible due to lack of knowledge concerning the cause of missingness [5], there is no choice but to apply listwise deletion, where variables are deleted for cases with missing data [4].

Imputation methods can be applied to MAR and MCAR [5]. Older methods of imputation, such as mean substitution, regression-based methods and resemblance-based or 'hot deck imputation' may produce biased results and standard errors [4, 8]. Regression-based methods use a regression to predict an entry, while resemblance-based methods impute new values based on similar cases [4]. Two newer methods include multiple imputation and Expectation Maximisation [4] and are dealt with by [5] and [1] respectively. Local Global, Multilayer Perceptron (MLP) and Radial Basis Functions (RBF) AANNs are used with optimisation algorithms to successfully estimate missing data [1, 2, 9]. It is assumed that all data is MAR and that it is possible to deduce missing entries based on the remaining data [1].

2.2 Auto-associative Neural Networks

An auto-associative also referred to as auto-encoder neural network is a specific neural network, trained to recall its inputs [14]. Given a set of inputs, the network predicts these inputs as outputs and thus has the same number of output nodes as there are inputs [14]. However, the hidden layer is characterized by a bottleneck, with fewer hidden nodes than output nodes [1, 14]. This gives it a butterfly-like structure as shown in figure 1 [1]. The smaller hidden layer projects the inputs onto a smaller space, extracting linear and non-linear interrelationships, such as covariance and correlation, from the input space and also removes redundant information [1, 14]. This means that they can be used in applications to recall the inputs and in missing data estimation applications [1, 2, 9, 14].

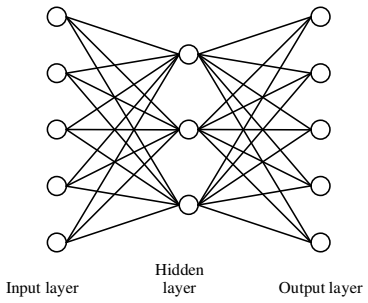

Figure 1: *Schematic of a five-input five-output auto associative neural network.*

2.3 Genetic Algorithm (GA)

The GA is a population-based, probabilistic technique that works to find a solution to a problem from a population of possible solutions [15]. It is based on Darwin's theory of evolution where members of the population compete to survive and reproduce while the weaker ones die out [16]. Each individual is assigned a fitness value according to how well it meets the objective of solving the problem [15]. New and more fitting individual solutions are produced during a cycle of generations, wherein selection and recombination operations, analogous to gene transfer [2] are applied to the current individuals [15]. This continues until a termination condition is met, after which the best individual thus far is considered the solution to the problem [15].

Unlike many optimisation algorithms, GA converges to a global optimal solution [1]. GA has also been proved to be very successful in many applications including the travelling salesman problem, adaptive control and database query optimisation [2]. Other optimisation techniques include particle swarm optimisation, simulated and quantum annealing as well as ant colony optimisation [1].

2.4 Principal Component Analysis

Principal Component Analysis (PCA) is a method of identifying patterns in data and displaying those patterns in such a way so as to highlight the similarities and differences amongst the data [7]. It is used in data analysis to identify and extract the main correlation variables amongst data [6]. This allows the dimensionality of the data to be reduced, without loss of essential information [6]. Hence the data is effectively compressed; consequently PCA finds an application in image compression [7].

PCA has been described as the optimum linear, information-preserving transformation [6, 17] and has been shown to facilitate many types of multivariate analysis, including fault detection and data validation [6, 18].

3 Design Methodology

For missing data estimation using an auto-associative model in conjunction with an optimisation technique, it is imperative that the auto-associative model be as accurate as possible [14]. Hence this paper attempts to find new and improved ways to capture and model the interrelationships between variables in a dataset and use the interrelationships along with an optimisation technique to predict missing entries.

The approach is summarised in figure 2 where X_k and X_u are the known and unknown variables respectively and constitute the input space [1].

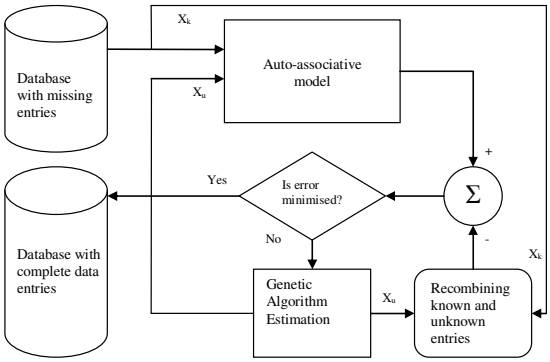

Figure 2. *Structure of missing data estimator using an auto-associative model and the Genetic Algorithm.*

The Genetic Algorithm is used to estimate values for the unknown variables and these are fed into the auto-associative model along with the known inputs from the database [1]. The auto-associative model is trained, using complete data, to extract and store the correlations and interrelationships between the variables in the dataset and develop a mapping function F. An output $\vec{Y}$ can thus be formed as shown by (1).

$$\vec{Y} = F\left(\begin{bmatrix} \vec{X}_k \\ \vec{X}_u \end{bmatrix}\right) \tag{1}$$

Since the auto-associative model attempts to replicate the inputs at the outputs, the output is approximately equal to the input [2]. A small error exists between the input and output, the size of which depending on how accurate the auto-associative model is [1, 2]. This error is shown by (2)

$$e = \left(\begin{bmatrix} \vec{X}_k \\ \vec{X}_u \end{bmatrix} - F\left(\begin{bmatrix} \vec{X}_k \\ \vec{X}_u \end{bmatrix}\right)\right)^2 \tag{2}$$

The error will be at a minimum when the output comes closest to matching the input. This occurs only when the data inputted to

the auto-associative model carries the same correlations and interrelationships as those captured by the model during training [1]. Hence the minimisation of the error function implies the optimisation of the unknown variables such that the complete data vector $\vec{X}$ fits the pattern given by the complete data. Three designs for the auto-associative model are presented i.e. an ordinary auto-encoder neural network, a hybrid auto-encoder network and a combination of PCA and neural networks to form an auto-associative function.

3.1 Regular Auto-Encoder Neural Network

The regular auto-encoder network has a structure similar to that depicted in figure 1. MLP architecture is used due to its overall superior performance over RBF architecture in this application. The GA is used to optimise the number of hidden nodes and training cycles for the network so as to make the network as accurate as possible [14].

3.2 Hybrid Auto-Associative Network

Although MLP AANNs outperformed RBF AANNs, the latter showed superiority in predicting a few variables. The design of the Hybrid auto-associative network has three objectives:

- To combine the best from both the MLP and RBF architectures,
- To correct some of the distortion introduced by a single auto-encoder network and
- To capture complex non-linear relationships amongst variables more efficiently.

The structure of the Hybrid auto-associative network is shown in figure 3, where $\vec{X}$ is the set of inputs and $\vec{Y}$ is the predicted version of $\vec{X}$. The second layer MLP network is trained with a different part of the dataset to that used to train the MLP and RBF auto-encoder. This aids the corrective ability of the network. The number of hidden nodes in each network is optimised using the GA.

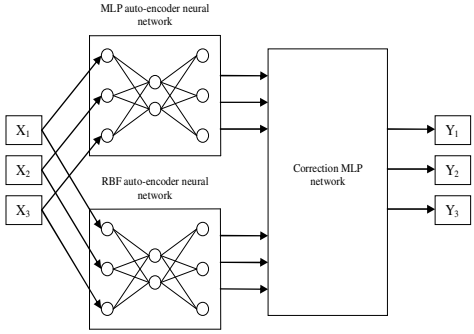

Figure 3: *Structure of a 3-input, 3-output hybrid auto-associative network*

3.3 PCA and Neural Network Approach

In order to make the auto-encoder network more efficient, when working with dimensionally complex and highly non-linearly related data, PCA is performed on the data before propagation through the auto-encoder. This performs much of the linear correlation extraction and reduces the dimensionality of the data, thus reducing the burden on the auto-encoder.

The auto-encoder is trained to recall the reduced data and the inverse PCA is then applied to restore the original data. MLP neural networks are trained to mimic the principal component extraction and original data reconstruction functions i.e. the PCA and inverse PCA functions. All architectures are optimised using the GA. Figure 4 depicts the arrangement of the networks, where *R* and *P* are the original and predicted dimensionally-reduced data respectively.

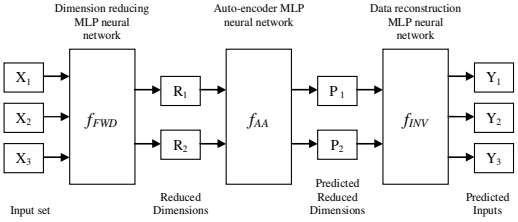

Figure 4: *Structure of a 3-input, 3-output PCA and neural network auto-associative model*

Equation (4) summarises the auto-associative function, where f_{FWD} is the PCA forward function, f_{AA} is the function of the auto-encoder neural network and f_{INV} is the inverse PCA function,

$$\vec{Y} = f_{INV}\left(\vec{W}_3, f_{AA}\left(\vec{W}_2, f_{FWD}\left(\vec{W}_1, \vec{X}\right)\right)\right) \quad (4)$$

where W_1, W_2 and W_3 are the weights of the dimensionally reducing, auto-encoder and reconstruction neural networks respectively.

4 Experimental Implementation

4.1 HIV Data Analysis

The dataset used was obtained from the South African antenatal sero-prevalence survey in 2001 and consists of information concerning pregnant women who have visited selected public clinics in South Africa [1, 14]. Only women participating for the first time in this national survey were eligible [1, 14].

As done in by Betechuoh *et al* [14], who achieved successful results with the above data, eleven attributes from the dataset are used. They are summarised in table 1. The HIV status is represented in binary form i.e. a 1 indicating a positive status, while a 0 implies a negative status [1, 14]. Gravidity refers to the number of combined complete and incomplete pregnancies that the woman has experienced, while parity refers to the number of occasions on which the woman has given birth [1, 14]. Age gap refers to the difference in age between the pregnant woman and the prospective father of the child [1, 14]. Rapid Plasma Reagin (RPR) refers to a screening test for syphilis for which HIV may cause a false positive test result [19]. Qualitative variables, stored as text, such as the race of the mother and her province and region of origin, are encoded as integers [14]. The education level of the pregnant woman is represented by an integer corresponding to the highest grade completed, with 13 corresponding to tertiary education [1].

The data consists of 16608 instances, however many of them contain missing variables and/or outliers. Such instances are removed, leaving behind a dataset of 10829 instances. This is normalised, randomised to improve NN performance and thereafter

split equally into training, validation and testing datasets. Hence 3609 training examples are used.

Table 1: *Summary of variables used from HIV data*

Variable number	Variable	Type	Range
1	Age	Integer	14 – 50
2	Age Gap	Integer	1 – 7
3	Education level	Integer	0 – 13
4	Gravidity	Integer	0 – 13
5	Parity	Integer	0 – 14
6	RPR	Integer	0 – 2
7	WTREV	Continuous	0.638 – 1.2743
8	Regions	Integer	1 – 77
9	Provinces	Integer	1 – 9
10	Races	Integer	1 – 5
11	HIV status	Binary	[0, 1]

The neural network architectures are optimised using the validation dataset, while their correct performances are verified and tested using the testing dataset. 100 instances are randomly extracted from the test dataset and variables therein removed. The resulting dataset, with missing values, is used to test the performance of the three missing data estimation schemes.

4.2 Neural Network Optimisation

MATLAB and the NETLAB toolbox [20] were used to implement all the neural networks. The ordinary auto-encoder network is optimised to an 11 input nodes, 9 hidden units and 11 output nodes (11-9-11) structure and trained with 180 cycles. The MLP auto-encoder of the Hybrid auto-associative network also has a similar structure while its RBF counterpart has one less hidden node in its architecture. The correction MLP is optimised to a 22-19-11 structure.

The optimum number of principal components is 9. These account for 99% of the correlation variances in the input space. Fewer principal components result in an ineffective reconstruction function. The dimensionality reduction, Principal Component auto-encoder and data reconstruction neural networks are optimised to 11-20-9, 9-8-9 and 9-13-11 node structures respectively.

4.3 GA Implementation

The Genetic Algorithm Optimisation Toolbox (GAOT) was used to implement the GA [21]. The initial population size was set to 100 and the process was run for 40 generations. Simple crossover, non-uniform mutation and normalised geometric selection were used as these were found to produce satisfactory results, which were not surpassed by using other combinations of GA parameters.

4.4 Performance Evaluation

The effectiveness of the missing data estimation system is evaluated using the Standard Error (*SE*), the Correlation Coefficient (*r*) and the relative prediction accuracy (*A*) [1, 2]. The Standard Error measures the error between the actual values and the predicted values and gives an indication of capability of prediction [2]. It is given by (5), where x_i is the actual value, $\hat{x}_i$ is the predicted value and n is the number of missing values [2].

$$SE = \sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \hat{x}_i\right)^2}{n}} \tag{5}$$

The Correlation Coefficient measures the linear similarity between the predicted and actual data [1, 2]. r ranges between -1 and 1, where its absolute value relates the strength of the relationship while the sign of r indicates the direction of relationship [2]. Hence a value close to 1 indicates a strong predictive capability [2]. The formula is given by (6), where $\bar{x}$ is the mean of the data [2].

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}_i\right)\left(\hat{x}_i - \bar{\hat{x}}_i\right)}{\left[\sum_{i=1}^{n}\left(x_i - \bar{x}_i\right)^2 \sum_{i=1}^{n}\left(\hat{x}_i - \bar{\hat{x}}\right)^2\right]^{1/2}} \tag{6}$$

The relative prediction accuracy is a measure of how many predictions are made within a certain tolerance, where the tolerance can be set based on the sensitivity demanded by the application [1]. When applied to HIV data, the tolerance is set to 10% as done by Nelwamondo *et al.* [1] since it seems suitable to this application. The accuracy is given by (7), where n_τ is the number of predications within tolerance [1].

$$A = \frac{n_\tau}{n} \times 100 \tag{7}$$

Using the above mentioned performance parameters, the performance of the three missing data estimation systems are evaluated by estimating each of the 11 attributes individually. Their abilities to estimate two, three and four missing attributes simultaneously are examined through the estimation of the Age, Age Gap, Parity and Gravidity variables.

5 Experimental Results and Discussion

The comparisons of the Standard Errors, Correlation Coefficients and relative prediction accuracies for the three systems, when estimating a single missing value, are shown in figures 5, 6 and 7 respectively. Table 2 summarises the mean performance of the three systems when estimating a single variable.

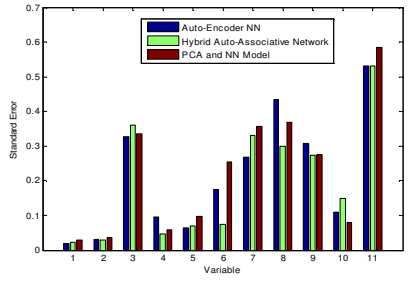

Figure 5: *Standard errors for all variables when one missing value is estimated by the three estimator systems*

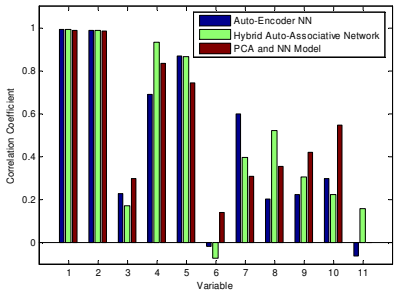

Figure 6: *Correlation Coefficients for all variables when one missing value is estimated by the three estimator systems*

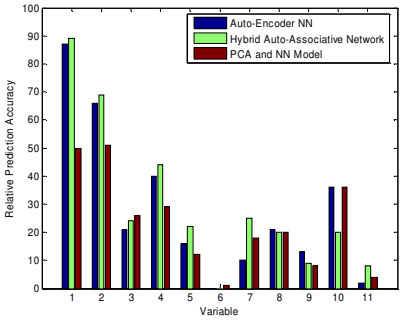

Figure 7: *Relative Predicative Accuracies for all variables when one missing value is estimated by the three estimator systems.*

Table 2: *Overall performance summary for 1 missing value*

	Standard Error	Correlation Coefficient	Relative Prediction Accuracy
MLP auto-encoder Neural network	0.2148	0.4560	28.3636
Hybrid Auto-associative network	0.1988	0.4989	30
PCA and Neural Network Model	0.2247	0.5113	24.3636

Figure 8, 9, 10 and 11 depict the respective estimations of a few instances of the age attribute, when one, two, three and four values are missing from an instance.

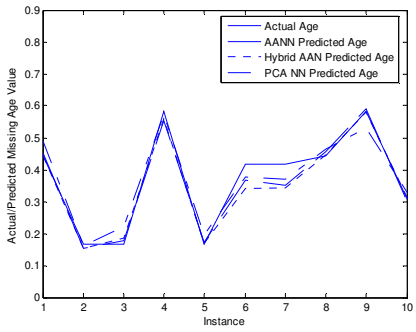

Figure 8: *Comparing the estimation of the normalised Age variable for 1 missing value.*

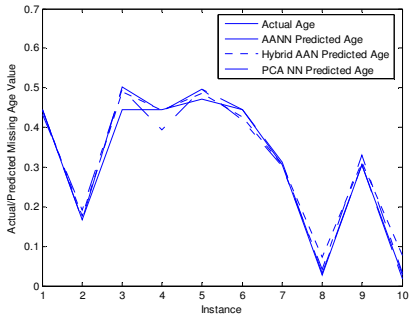

Figure 9: *Comparing the estimation of the normalised Age variable for 2 missing values.*

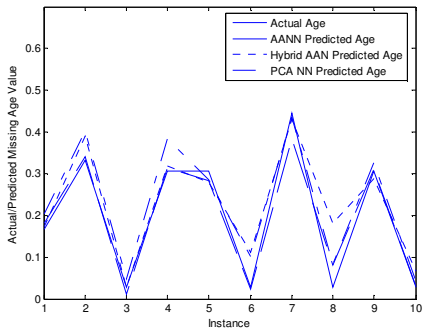

Figure 10: *Comparing the estimation of the normalised Age variable for 3 missing values.*

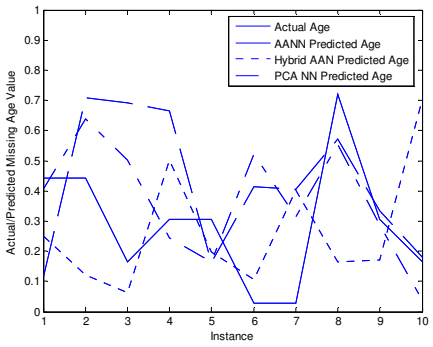

Figure 11: *Comparing the estimation of the normalised Age variable for 4 missing values.*

The average Standard Errors, Correlation Coefficients and relative prediction accuracies for the estimation of 1 – 4 simultaneous missing values are summarised in tables II – IV respectively. All the models show efficiency in the estimation of certain variables such as age, age gap, gravidity and parity. This is indicated by the high correlation coefficients, and low errors associated with these variables, as shown in figures 5 and 6. Conversely, all estimations of some variables such as HIV status and RPR are poor. This is clearly displayed in figures 5 – 7. This may be due to weak or highly non-linear interrelationships between these variables and the rest of the variables in the dataset.

Figures 5 – 6 show that the Hybrid auto-associative network, coupled with the GA is the most accurate when predicting a single missing value. This is supported by table 2, which shows that the hybrid auto-associative system produces the smallest mean Standard Error and has the highest overall correlation coefficient and relative prediction accuracy. The hybrid network performs better than a regular AANN overall.

The PCA and neural network auto-associative model demonstrates an increased ability to capture correlations within the data. This is shown by figure 6, where the model produces the highest correlations for most of the poorly estimated variables.

Tables 3 – 5 show that the PCA and neural network auto-associative model coupled with the GA performs best for multiple simultaneous estimations. The considerably higher correlation coefficients shown in table 4, for cases requiring multiple estimations, shows superiority over the other two models when it comes to capturing the patterns and correlations within the data. For this reason the PCA and neural network system would be best

suited for implementation in a real system since more than one missing value within an instance is likely, as is the case with the HIV data used.

Table 3: *Mean Standard Error for variables 1, 2, 4 and 5*

Number of missing values	1	2	3	4
MLP auto-encoder Neural network	0.0531	0.1443	0.1909	0.3154
Hybrid Auto-associative network	0.0422	0.1446	0.1874	0.2739
PCA and Neural Network Model	0.0542	0.1404	0.1286	0.2621

Table 4: *Mean Correlation Coefficient for variables 1, 2, 4 and 5*

Number of missing values	1	2	3	4
MLP auto-encoder Neural network	0.8855	0.7526	0.4465	0.0887
Hybrid Auto-associative network	0.9458	0.7925	0.4372	0.1191
PCA and Neural Network Model	0.8887	0.8219	0.7133	0.1815

Table 5: *Mean relative prediction accuracy for variables 1, 2, 4 and 5*

Number of missing values	1	2	3	4
MLP auto-encoder Neural network	52.5	28.5	19.83	6.5
Hybrid Auto-associative network	56	31.5	22	5
PCA and Neural Network Model	38.75	21.25	20.499	6.75

Figures 8 – 10 show that all the models are satisfactory for the estimation of up to 3 missing values and that accuracy decreases for increasing simultaneous estimations. However, figure 11 shows the poor estimation capabilities of all the models for the case of 4 missing values. This could be due to the data becoming MNAR when four values are missing i.e. one missing variable depends most one or more other variables that are also missing.

6 Conclusions

This paper investigates the estimation of missing data through novel techniques. The estimation system involves an auto-associative model to predict the input data, coupled with the genetic algorithm to approximate the missing data. Three auto associative models are investigated i.e. a regular auto-encoder neural network, a hybrid auto-associative network consisting of three neural networks, and the series combination of three neural networks to incorporate Principal Component Analysis into an auto-associative function. The performance of each model in conjunction with the GA is investigated. Results show that the hybrid network is most accurate for single missing value estimation, while the PCA and neural network model provides more consistency for multiple estimations. The latter also appears to perform better than the other two models when dealing with data exhibiting very little interdependencies.

References

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

[2] Abdella M, Marwala T. *Treatment of Missing Data Using Neural Networks and Genetic Algorithms*, Computing and Informatics, vol.24, 2005, pp.577-589.

[3] Fujikawa Y. *Efficient Algorithms for Dealing with Missing Values in Knowledge Discovery*, March 2001.

[4] Yuan K. H. and P.M. Bentler. *Three Liklihood-Based Methods for Mean and Covariance Structure Analysis with Non-Normal Missing Data*, Sociological Methodology, 2000, pp 165-200.

[5] Wayman J C. *Multiple Imputation For Missing Data: What Is It And How Can I Use It?*, Annual Meeting of the American Educational Research Association, 2003.

[6] Kramer F A. *Nonlinear Principal Component Analysis Using Autoassociative Neural Networks*, AIChE Journal, Vol. 3 No. 2, February 1991, pp 233- 243.

[7] Smith L I. *Tutorial on Principle Component Analysis*, February 2002, http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

[8] Gold M S, Bentler P M. *Treatments of missing data: A Monte Carlo comparison of RBHDI, iterative stochastic regression imputation, and expectation-maximization*. Structural Equation Modelling, 7, 2000, pp 319-355.

[9] Fariñas M, Pedreira C E. *Missing Data Interpolation By Using Local-Global Neural Networks*, ISSN 0969-1170, Vol. 10, No 2, 2002, pp. 85-92.

[10] M. Bosque, *Understanding 99% of Artificial Neural Networks*, pp 3, 4, Writers Club Press, New York , 2002.

[11] K. Gurner, *An Introduction to Neural Networks*. CRC Press, pp. 1, 6, 2003

[12] T Marwala, *Probabilistic Fault Identification using Vibration Data and Neural Networks*, PhD thesis, University of Cambridge, 2001.

[13] Hagan M T, Demuth H B, Beale M. Neural Network Design, PWS publishing company, Boston MA, 1996.

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

[15] J. Kubalík and J. Lazanský, *Genetic Algorithms and their Testing*, AIP Conference Proceedings, Volume 465, pp. 217-229, 1999.

[16] T. Marwala, *Bayesian training of Neural Networks using Genetic Programming*, Pattern recognition Letters, http://dx.doi.org/10.1016/j.patrec.2007.03.004, 2007.

[17] Fukunaga, **K,** Koontz W, *Application of Karhunen-Loeve Expansion to Feature Selection and Ordering*, IEEE Transactions. Comput**,** C-19, 1970.

[18] Wise B M, Ricker N L, *Upset and Sensor Failure Detection in Multivariate Processes*, paper 164b, AIChE Meeting, San Francisco, 1989.

[19] University of Pennsylvania Health System, *RPR*, http://pennhealth.com/ency/article/003533.htm, last accessed 28 August 2007.

[20] I. Nabney, *Netlab Neural Network Software*, http://www.ncrg.aston.ac.uk/netlab/, last accessed 13 May 2007.

[21] C. R. Houck, J. A. Joines and M. G. Kay, A Genetic Algorithm for Function Optimization: a Matlab Implementation, NCSU-IE TR 95-09, 1995.