The Impact of Missing Data Imputation on HIV Classification

Nthabiseng Unathi Hlalele

A dissertation submitted to the Faculty of Engineering and the Built Environment, University of the Witwatersrand, fulfillment of the requirements for the degree of Master of Science in Engineering.

Johannesburg, 2008

Declaration

I declare that this dissertation is my own, unaided work, except where otherwise acknowledged. It is being submitted for the degree of Master of Science in Engineering in the University of the Witwatersrand, Johannesburg. It has not been submitted before for any degree or examination in any other university.

Signed this _____day of _____20__.

_____

Nthabiseng Unathi Hlalele

Abstract

Missing data are a part of research and data analysis that often cannot be ignored. Although a number of methods have been developed in handling and imputing missing data, this problem is, for the most part, still unsolved with many researchers still struggling with its existence. Due to the availability of software and the advancement of computational power, maximum likelihood and multiple imputations as well as neural networks and genetic algorithms (AANN-GA) have been introduced as solutions to the missing data problem. Although these methods have given considerable results in this domain, the impact that missing data and missing data imputation has on decision making has, until recently, not been assessed. This dissertation contributes to this knowledge by first introducing a new computational intelligent model that integrates Neuro-Fuzzy (N-F) modeling, Principal Component Analysis and the genetic algorithms to impute missing data. The performance of this model is then compared to that of the AANN-GA as well as the independent use of the N-F architecture. In order to determine if the data are predictable and also to assist in processing the data for training, an analysis on the HIV sero-prevalence data is performed.

Two classification decision making frameworks are then presented in order to assess the effect of missing data. These decision frameworks are trained to classify between two conditions when presented with a set of data variables. The first is the use of a Bayesian neural network which is statistical in nature and the second is based on the fuzzy ARTMAP (FAM) classifier which has incremental abilities. The two methods are used and compared in order to assess the degree in which missing data, and the imputation thereof, has on decision making. The effect of missing data differs for the two frameworks; while the Bayesian neural network fails in the presence of missing data, the FAM classifier attempts to classify with a decreased accuracy. This work has shown that although missing data and the imputation thereof has an effect on decision making, the degree of that effect is dependent on the decision making framework and on the model used for data imputation.

To my family and William

This work is dedicated to my mother Nomvuselelo Patricia Hlalele, my gogo Nandipha Jojwana, my sisters Philiswa, Dineo and Naledi and my precious nephews Katlego and Omphile.

Acknowledgements

I wish to thank my supervisors Prof. Tshilidzi Marwala and Dr. Fulufhelo Nelwamondo for their constant encouragement, advice and pressure throughout the course of this research. I especially thank Prof. Marwala for his insistence on excellence and the support he provided both intellectually and financially. Thank you for being such an inspiration. I would like to thank my mother, my grandma, my sisters and my nephews for all the support they have given me throughout my studies, ndiyabulela. A special thanks goes to William for all his encouragement and support during the course of this research, thank you for your unwavering belief in my abilities. Most importantly I would like to thank my fellow colleagues of the research unit: Vima, Mistry, and Lesedi for their support and for making this year an enjoyable journey. I would also like to thank Fulu and Rofhiwa for listening to my constant rambling and for assisting in proof reading this work. I would also like to especially thank Linda Mthembu and Cuthbert Nyamupangedengu for proof reading this work.

Lastly, I would like to acknowledge the financial assistance of the National Research Foundation (NRF) of South Africa towards this research. Opinions and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

Contents

List of Figures

List of Tables

Nomenclature

ANN Artificial Neural Networks

EM Expectation Maximisation

FAM Fuzzy ARTMAP

GA Genetic Algorithms

HIV Human Immunodeficiency Virus

MAR Missing at Random

MCAR Missing Completely at Random

ML Maximum Likelihood

MLP Multilayer Perceptron

MNAR Missing Not at Random

N-F Neuro Fuzzy

AANN-GA Auto-Associative Neural Network and Genetic Algorithm Combination

PCA Principal Component Analysis

Chapter 1

Introduction

1.1 Background and Motivation

Missing data have been an area of interest in the statistics community due to its inhibiting characteristics in data analysis (Little and Rubin, 2002). This has led to the development of models and methods to handle, and in some cases, impute missing data. When data are imputed, the missing value is substituted by an estimated value such that the dataset can be analyzed using standard techniques for complete data. Intuitively, it is expected that missing data should have an impact in data analysis and decision making; this impact, however, has not been evaluated in light of missing data imputation. This work adds to the knowledge of missing data by developing a new computational intelligence method to missing data imputation with the aim of evaluating the impact that missing data and the imputation thereof has on decision making.

1.2 Missing Data

Data mining and analysis techniques have been employed in many applications from the evaluation of a plant process in an engineering environment to the spread of a pandemic in a social community. Unfortunately, these techniques are prone to missing data that can lead to incorrect prediction and classification models. A number of methods have been investigated and implemented in order to deal with this problem, especially in large databases that require

computational analysis such as the case with some of the above mentioned applications (Little and Rubin, 2002). It has been observed that the most commonly accepted way of dealing with the problem of missing data is the imputation of the missing cases; computational intelligence techniques have also been employed to handle missing data with considerable success. Adding to this knowledge, a hybrid missing data imputation model is developed to impute missing data; this method is then improved to increase its accuracy.

1.3 Decision Making

An integral part of human interaction and intelligence is the ability to make a decision (French, 1986). It is because of this that intelligent systems are built to be able to make a decision based on the information or data that is given to them. Classification is a form of decision making that involves the assignment of objects to a class after some form of pattern recognition (Zhang, 2000) has been performed by a classifier. In the presence of missing data, the ability of these decision making systems (classifiers) to make a decision is affected. The extent to which the presence of missing data and the imputation thereof affects these frameworks is thus investigated in this work.

1.4 Outline of the Dissertation

As mentioned earlier, this dissertation adds to the missing data knowledge by first imputing missing data in a computational intelligent way and then evaluating the impact that this imputation has on a classification task. The dissertation is structured as follows:

- **Chapter 2** introduces the missing data problem and gives a background of the methods used in dealing with it.

- **Chapter 3** gives a background of decision making leading up to the use of statistical models in the development of the decision theory. The chapter also presents the use of classification models in decision making.

- **Chapter 4** introduces the HIV sero-prevalence dataset used in this work. The preprocessing and analysis of this data is also investigated.

- **Chapter 5** introduces computational intelligence methods in decision making followed by the introduction of the N-F model and a hybrid method for missing data imputation.

- **Chapter 6** presents the impact that the missing data and the imputation thereof has on decision making (classification) by evaluating the classification performance of two decision making frameworks, namely the Bayesian framework and the Fuzzy ARTMAP.

- **Chapter 7** concludes the work presented in this dissertation.

Although Chapters 2 and 3 of this dissertation are independent, it is suggested that the work be read in a sequential manner because of the interdependency of the rest of the chapters.

Chapter 2

Background on Missing Data Analysis

2.1 Background

Decision making processes often require comprehensive amount of information that is extracted from a dataset. Datasets are a collection of data commonly presented in tabular form with each column representing a variable and each row representing an observation. In most applications the analysis of the data is performed by statistical measures that analyze rectangular data sets. Missing data refers to a situation where no value is stored for a certain variable at the current observation resulting in a non-rectangular data matrix (Little and Rubin, 2002). In many cases, such as the nonresponse in a survey, the missing data matrix follows any one of the patterns in Figure 2.1 (Schafer and Graham, 2002).

The univariate pattern occurs when data are missing from one variable as shown by X_N in Figure 2.1 (a). : The monotone pattern occurs such that the missingness of one data variable results in the missingness of other variables as depicted in Figure 2.1 (b). Lastly, an arbitrary pattern occurs when the data that are missing follow some random pattern as shown in Figure 21 (c).

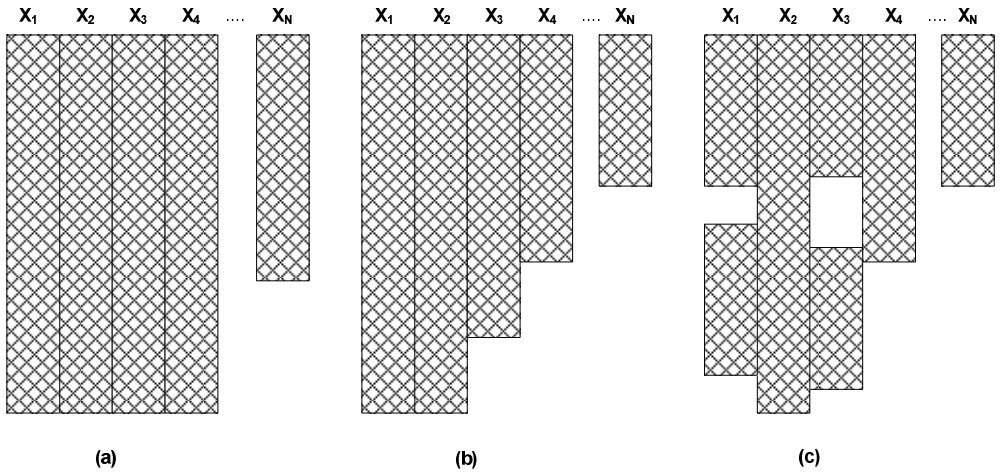

Figure 2.1: Patterns of nonresponse in rectangular data sets: (a) univariate pattern, (b) monotone pattern and (c) arbitrary pattern. In each case, rows correspond to observational units and columns correspond to variables.

Most data analysis techniques attempt to measure certain parameters in order to make a decision or a recommendation, for example, a survey might be carried out to determine the number of households that live below the poverty datum line in a certain area. In order to make an accurate inference from the data, it is essential that all individuals are represented in the data irrespective of whether there was non-response in their observations. An understanding of the relationship between the variables in the data set and missingness, that is the missing-data mechanism, is important for the treatment of data sets that contain missing data.

2.2 Missing Data Mechanisms

Consider a case where X defines the complete data set whilst X_{obs} denotes the observed entries of X and X_{miss} denotes the missing components. Little and Rubin (2002) distinguish between three mechanisms leading to missing data in which the following notion is used: In the presence of missing data, define the missing data indicator matrix, M, such that $m_{ij} = 1$ if

the datum, $x_{ij,}$ is missing, similarly, m_{ij} = 0 if the datum is observed. The missing data mechanism is then characterized by the conditional distribution of M given X, that is, $f(M|X,\phi)$ where ϕ denotes the unknown parameters. The three mechanisms are defined as 'Missing at Random', (MAR), 'Missing Completely at Random', (MCAR) and 'Missing Not at Random', (MNAR) and are described below.

2.2.1 Missing Completely at Random (MCAR)

This mechanism refers to a condition where the missingness is independent of the values of the missing or observed data X, that is (Little and Rubin, 1987),

$$f(M|X,\phi) = f(M|\phi) \text{ for all X and } \phi. \qquad (2.1)$$

This means any piece of data is just as likely to be missing as any other piece of data, that is, cases with complete data are identical to the cases with incomplete data.

2.2.2 Missing at Random (MAR)

This mechanism refers to a condition where the missingness is dependent only on the components of the data, X, that are observed (X_{obs}) and not on the components that are missing (X_{miss}), that is (Little and Rubin, 1987),

$$f(M|X,\phi) = f(M|X_{obs},\phi) \text{ for all } X_{miss} \text{ and } \phi. \qquad (2.2)$$

This means that missingness does not depend on the missing data entry even after controlling for another variable.

2.2.3 Missing Not at Random (MNAR)

This mechanism refers to a condition where the missingness is dependent only on the components of the data, X, that are missing (X_{miss}), that is, the missingness depends on the missing data entry even if the other variables are controlled; this means that the missing entry is dependent on its own value. An example of this is the case where people with low income are less likely to report their income on a data collection form.

2.3 Timeline of Missing Data Analysis

Advances in computer technology have made numerical analysis of data an easier task. With this development, missing data analysis has gained great popularity as new techniques are researched and implemented using computer software. The development of the Expectation Maximization (EM) algorithm, which enabled the computation of Maximum Likelihood (ML) estimates in missing data problems (Dempster, Laird and Rubin, 1977), created a new paradigm for dealing with missing data. This paradigm, fuelled by the progress in computer systems, gave rise to the formulation of Multiple Imputation (MI) due to the flaws of single imputation and case deletion (Little and Rubin, 1987). Since this formulation, methods such as Bayesian simulation (Schafer, 1997) and the use of computational intelligence (Nelwamondo, 2007; Abdella and Marwala, 2006) have been employed as solutions to the missing data problem. Recently, research has been conducted to model missingess including the assessment of the sensitivity of results to the distribution of missingness (Verbeke and Molenberghs, 2000), the modeling of MNAR missingness (Little, 1995) and the handling of missing values without the use of a full parametric model of the population (Robins, Rotnitzky

and Zhao, 1994). Other methods have also been employed to handle missing data without imputation, these include the use of Neural Networks (NN) in classification tasks in the presence of missing data without the actual imputation of the missing data (Wang, 2005) and the use of computational intelligence to preserve the dynamics of systems even in the presence of missing data (Qiao et al.,2005).

2.4 Treatment of Missing Data

There are a number of methods that are used to handle missing data. These methods depend on the data being analysed and the application the analysis is used for. This section briefly outlines these methods in an attempt to understand the build up of missing data approximation techniques in relation to the timeline of missing data analysis.

2.4.1 Listwise Data Deletion

This is the most common approach in handling missing data. This method simply removes cases or observations that contain missing data thus leaving the analysis to be performed on the data that remains. For example, if certain individuals have missing entries in one or more variables, the individuals are omitted from the analysis. This obviously results in a decreased sample size which is inefficient because energy is used to collect data that will not be used in the analysis. If we are certain that the data missingness is MCAR, or if we are certain that the deletion of missing cases will not significantly alter the precision and bias of the data (Little and Rubin, 2002) then listwise deletion is the simplest approach to use. Unfortunately, if missing data mechanism is not MCAR, then the deletion of cases results in bias and imprecise data.

2.4.2 Pairwise Data Deletion

Under this approach all the available data are used in the analysis phase. This means that an observation that is missing in one variable will only be used in the analysis that does not involve that variable. For example, if a participant neglects to mention his income but supplies his age in a survey, then he will be included in analyses involving the age of the participants in the survey but he will be excluded from the analyses involving the incomes of the participants. This approach leads to an analysis model that is based on different sets of data with different sample sizes and different standard errors. The other problem is that if the missing data mechanism is a function of the variables under the study, then comparability across variables can not be achieved (Little and Rubin, 2002).

2.4.3 Hot Deck Imputation and Mean Substitution

In the event of missing data, the hot deck imputation method uses the available data set to identify an observation that is most similar to the present case (that contains a missing value). When this observation has been identified, the missing value is then replaced by the corresponding value in the complete observation. This approach does not account for imputation uncertainty and standard errors as a result of the filled in data (Little and Rubin, 2002). Mean substitution refers to the replacement of a missing observation by the mean of the variable that is missing in that particular observation. The disadvantage of this method is that it adds no new information. The overall mean, with or without replacing the missing data, will be the same. Additionally, this method also leads to an underestimate of error.

2.4.4 Regression Techniques

This approach replaces the missing values by predicted values from a regression of the available variables in the present observation. This means that the missing value is treated as a dependent variable and the other variables in the observation are treated as the predictors. A regression equation is then developed in terms of the other variables. The disadvantage of this method is that by substituting a value that is perfectly predictable from other variables, there is no addition of new information, however, the sample size increases and the standard error is reduced. Methods that overcome this disadvantage include stochastic regression methods where the missing value is replaced by a value predicted by regression imputation and a residual to reflect uncertainty in the estimated value.

2.4.5 Maximum Likelihood and Expectation Maximization

Since its formulation, the maximum likelihood approach to estimate missing data has become widely accepted and used (Little and Rubin, 1987; Schafer and Olsen, 1998; Schafer, 1997) and is based on a statistical model of the data. The assumption of using ML in missing data analysis is that the data are MAR and that the objective is to maximize the likelihood L with respect to θ (Little and Rubin, 1987):

$$L\left(\theta \middle| X_{obs}\right) = \int f\left(X_{obs}, X_{miss} \middle| \theta\right) dX_{miss} . \quad (2.3)$$

where X_{obs} and X_{mis} represent the observed data and the missing data in the data set X respectively and θ is some control parameter of interest. The methods used to maximize this likelihood involve the calculation of the second derivatives of the loglikelihood. For complex missing data patterns, the entries into this matrix are complicated functions of θ. An alternative strategy to maximize the likelihood for missing data problems is the Expectation

Maximization (EM) algorithm because it does not require the calculation of second derivatives. The EM is an iterative algorithm for ML estimation involving two steps, Expectation (E) and Maximization (M). In the E step, the expected value of the unknown variables given the current estimated parameters is computed. The M step estimates the distribution parameters that maximize the likelihood of the data given the expected estimates of the unknown variables. The EM iterates through these two steps until convergence is reached, that is, until the change in parameter estimates from the one iteration to another is negligible (Little and Rubin, 2002).

2.4.6 Computational Intelligence

Research into the application of computational intelligence for handling missing data has advanced in recent years with the development of a paradigm that incorporates neural networks and genetic algorithms (Nelwamondo et al., 2007; Abdella and Marwala, 2006); this is covered in chapter 5. The effect of the imputation ability of this method, however, has not been extensively evaluated and there still exists some room for the development and use of other models that handle missing data imputation.

2.5 Conclusion

This chapter presented background material to the missing data problem and the various approaches that are used to handle it. Different missing data patterns and mechanisms were discussed with the intention of better understanding the missing data problem. Over the years, a number of methods have been employed for handling missing data with the use of the EM algorithm and ML becoming the most accepted method. Recently, the use of computational intelligence has also been explored as a candidate solution to the missing data problem. Despite all these developments in missing data, there exists some room for better models that impute missing data and methods of evaluating the impact that this imputation has on data handling models. This work adds to the missing data knowledge by investigating

and evaluating the use of other computational intelligent models, namely, the N-F model with genetic algorithms and PCA to impute missing data. The performances of these models are then evaluated using two classifiers.

Chapter 3

Decision Making and Classification

3.1 Introduction

The ability to choose and exercise free will is one of the characteristics that distinguish intelligent forms of life (French, 1986) from lower species. A decision, simply defined, is the selection of a course of action among several alternatives; therefore the decision making process always yields an action or an opinion of choice. This chapter is used as the basis of the classification evaluation techniques discussed in chapter 6 and presents an overview of human decision making and the developments of statistical analysis to formulate decision theories. The effect that technology has on the decision making process is also presented followed by the exploration of the use of computational intelligence in relation to decision making and, in particular, to classification tasks.

3.2 History of Decision Making

The ability of the human species to evolve and develop better civilizations is attributed to its intelligence, that is, the ability to learn. This translates to the ability to make choices given several alternatives that may or may not have already been learnt. In earlier times, humans have tried to model the environments around them in order to formulate the basis of decision making from the analysis of the positions of the stars to the development of statistical models. In the quest for better understanding which ultimately leads to better decision making, numbering systems, decision making philosophies as well as scientific innovations have been used to obtain analytical models of decision making. Great leaders in history are characterized

by their superior ability to make decisions that result in a growth in their influence and power (Kodish, 2006).

3.3 The Decision Theory

Decision theory is an area of study that is concerned with identifying methods that aid in decision making. The vast literature studies statistical and mathematical techniques into modeling the complex structure of decision making (French, 1986). There are two types of theories concerned with decision making, the first deals with determining the optimal solution under the assumption that the decision maker is ideal and fully informed (normative). The other theory investigates what decision makers are likely to do, that is, a behavioral analysis is formed (descriptive). The two are closely linked because optimum, ideal solutions often create a hypothesis to test against actual behavior (French, 1986). Not all decisions require a theory, for example, an individual does not need a mathematical model to decide what to have for lunch, other decisions, however are better modeled in order to produce the optimum outcome. There are, however, complex decisions that require the use of a mathematical or statistical model to make a decision. The most modeled type of decision is the decision under uncertainty. This pertains to a decision that has 'risks' associated with it and is evaluated using the notion of expected utility. The idea of the expected utility theory is that, when faced with a number of alternatives, each of which could give rise to more than one possible outcome with different probabilities, the alternatives are each assigned a weighted average of its utility values under different states and the probabilities of these states are used as weights. The best decision will thus be the one that results in the highest total expected value (Raiffa, 1997).

Unlike the abovementioned type of decision, other decisions are more complex than this. Consider, for example that the utility values of different decisions might vary with time. This is often the case when people are faced with decisions that have different 'risks' associated with them depending on whether the decision is short or long term. These decisions are often dealt

with by taking human behavior into account leading to different models that represent the problem space. Other decisions are complex in the sense that the optimum solution is unknown especially if two variables that are inversely correlated need to be optimised. Decision theorists are mostly not concerned with this type of decision problem and base their assumption on the fact that, ideally, decision makers know what the optimum decision is whether it be winning a war or making profit in a business (French, 1986).

3.3.1 Bayesian Framework for Decision Making

According to Hansson (1994) the expected utility theory, with both subjective and objective probabilities, is known as Bayesian decision theory and is governed by the following principles (Hansson, 1994):

- The Bayesian subject has a coherent (compliant with the mathematical laws of probability) set of probabilistic beliefs.
- The Bayesian subject has a complete set of probabilistic beliefs, that is, a Bayesian subject has a degree of belief about everything.
- When exposed to new evidence, the Bayesian subject changes his or her beliefs in accordance with his or her conditional probabilities.
- Bayesianism states that the rational agent chooses the option with the highest expected utility.

The third bulletin point needs to be further explained. The Bayesian theory involves the collection of evidence and testing its consistency with a given hypothesis, as the evidence is gathered, the belief in the hypothesis also changes in the following way:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{3.1}$$

where A represents a specific hypothesis, P(A) is called the prior probability of A that was inferred before new evidence, B, became available. P(B| A) is called the conditional probability of seeing the evidence B if the hypothesis A happens to be true. P(B) is called the marginal probability of B: the priori probability of witnessing the new evidence B under all possible hypotheses. P(A|B) is called the posterior probability of A when B is given. The Bayesian theory has been used because of its ability to combine descriptive and normative points of view in decision making. The descriptive claim of the theorem is that actual decision-makers satisfy the four criteria above. The normative claim of the decision theory is that rational decision-makers also satisfy them. Bayesian analysis in decision making has been used in applications such as computational intelligence and expert systems. This is because the Bayesian inference (equation 3.1) can be applied to pattern recognition (Bishop, 2006).

3.3.2 Classification in Decision Making

The objective of classifiers is to assign an object or observation to a predefined class based on observation, this is achieved because the observed attributes form a pattern (Zhang, 2000). This assignment of objects can be viewed as a 'decision' because the analysis of patterns is performed prior to it and all possibilities (classes) are considered prior to the assignment of an observation to a specific class. In order to find the optimum decision, statistical tools that organize the evidence and evaluate the risk in decision making have been formulated. The risk associated with errors is one illustration of the statistical tools applied in decision making. In most applications the risk associated with misclassification is different depending on what the misclassification is. For example, in medical data, classifying a benign cell as malignant is less

harmful to the patient than the reverse case. In order to evaluate decision processes, the following confusion matrix shown in Figure 3.1 is often used:

Table 3.1: Confusion matrix depicting errors that often occur in decision making.

Decision Framework	**Actual Condition**	
	TRUE	FALSE
TRUE	TP (True Positive)	FP(False Positive)
FALSE	FN(False Negative)	TN (True Negative)

This matrix is used to specify the errors that exist in a model and provides a framework by which to build tradeoffs between the FP and FN errors. The accuracy of the decision model is measured by the following equation:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{3.2}$$

Measures such as the false positive rate and the false negative rate are also evaluated by equations 3.3 and 3.4 respectively (Dai et al., 2008).

$$TruePositiveRatio = \frac{TP}{TP + FN} \tag{3.3}$$

$$FalsePositiveRatio = \frac{FP}{FP + TN} \tag{3.4}$$

These equations, along with the confusion matrix, are used as the basis of the decision frameworks that are discussed in chapter 6. Alternative methods to probabilistic models have been explored (French, 1986). These include the use of fuzzy logic, agent based learning and other machine learning models that are non probabilistic in nature but have been employed for decision making (O' Hagan, 1987). Some approaches of incorporating the use of artificial

intelligence into descriptive decision modeling have also been looked at (Weber and Coskunoglu, 1990).

In recent years, much research has been done in the exploration of computational intelligence in the classification of HIV from demographic data (Leke et al., 2008, Tim, 2007). This work adds to this knowledge by applying two classification models to analyze the impact that missing data imputation has on a classification task. Because there are many different types of classifiers, it is not possible to give a generic analysis of the impact of data imputation on classification thus only two classifiers are used to determine and evaluate this impact.

3.4 Conclusion

Decision making is an integral part of intelligent life which is influenced and affected by a number of factors. This chapter presented an overview of what the decision making process entails in light of the development of decision making frameworks. The evolution of the human decision making process which has led to the development of statistically formulated frameworks, such as the Bayesian framework, has also been discussed. Finally, statistical aids in decision making and classification have been presented.

Chapter 4

Data Analysis

4.1 Introduction

Statistical data analysis provides a means to delve into the data in order to discover patterns that would not otherwise be obvious. This is achieved in a number of ways and assists in extracting information in the data that will be vital in the implementation stage. This chapter first introduces the South African sero-prevalence dataset used in this work followed by the statistical analyses that are used to discover the data patterns. Finally the results of the statistical analysis on this dataset are presented.

4.2 The Dataset

The implementation of the analysis described in chapter 5 is investigated using the South African HIV sero-prevalence data of 2001. The Human Immunodeficiency Virus (HIV) has been identified as the cause of AIDS (Acquired Immunodeficiency Syndrome) which has reached epidemic levels in South Africa where over 10.8 % of the population over 2 years old is HIV positive (South African Department of Health, 2005). It is with these numbers in mind that the South African Department of Health embarked on an annual survey of pregnant women in public clinics around the country. The dataset used in this document was obtained from the South African antenatal sero-prevalence survey of 2001 (South African Department of Health, 2001). The data for this survey were collected from questionnaires answered by pregnant

women visiting selected public clinics in South Africa and only women undertaking in the study for the first time were allowed to participate. This dataset has been used to investigate the effect that demographic information has on the HIV risk of an individual (Leke et al., 2008). This is especially helpful in countries such as South Africa, which have a high HIV infection rate as mentioned earlier.

The collection and analysis of this data may lead to a development of social and demographic structures that will curb the spread of HIV and the devastating effect that this disease has on developing countries. In order to analyse this data, models have to be accurate enough to firstly delve into the data and find patterns in the dataset and then to analyse the data such that there are no biases in the results. Although surveys such as this one are helpful in many ways, it should be noted that they are not perfect and carry with them inherent partialities for the HIV classification task of a complete population. The first of such in this dataset is that only women are surveyed which suggests that only a portion of the society is investigated. The second is that only pregnant women are surveyed, this again divides the society and gives biased results because women who have never been pregnant cannot be surveyed. As a result of these partialities, it should be noted that the analysis of this work is performed only on this antenatal data and does not necessarily infer to the rest of the population, that is, the rest of the population is not modeled.

This dataset consists of a total of 16608 data instances and the data attributes used in this work are the HIV status, Education level, Gravidity, Parity, the Age of the Mother (pregnant woman) and the Age of the Father responsible for the most recent pregnancy). The HIV status is represented in binary form, where 0 and 1 represent negative and positive, respectively. The education level indicates the highest grade successfully completed and ranges between 0 and 13 with 13 representing tertiary education. Gravidity is the number of pregnancies, successful or not, experienced by a female, and is represented by an integer between 0 and 11. Parity is the number of times the individual has given birth and multiple births (e.g. twin births) are considered as one birth event. It is observed from the dataset that the attributes with the most missing values are the age of the father (3972 missing values), the age of the mother (151 missing values) and the education level (3677 missing values) of the pregnant

woman. Imputing these data variables is helpful in educating people about HIV and the factors that render some individuals more risky than others. In situations where an attribute is missing in the questionnaire, it is almost impossible to retrieve this information from the woman who supplied it due to the anonymity of the study. It is for this reason that missing data imputation methods are employed.

4.3 Statistical Analysis

There a number of statistical methods that exists to determine patterns in data. The methods introduced here are those used in this document to extract features that will be used in the implementation of the imputation methods discussed in chapter 5.

4.3.1 Percentile Study of Data

The simplest method of statistical analysis is to look at the measures of dispersion of the data. This measures the spread of the data in the number line and includes measures such as the range, standard deviation and variance of the dataset. Although these methods are good in determining the patterns of the data, they are not robust enough to handle outliers that often plague databases in reality. Another option is to compute a number of the sample percentiles. This provides information about the shape of the data as well as its location and spread. Due to the graphical representation of this method, the data set can be fully represented to account for any outliers that might be observed. An example of this is now discussed following from data that are randomly permuted with 300 observations and one variable with values ranging from 1 to 7, the percentile study of the data is shown in Figure 4.1 (Tukey, 1977).

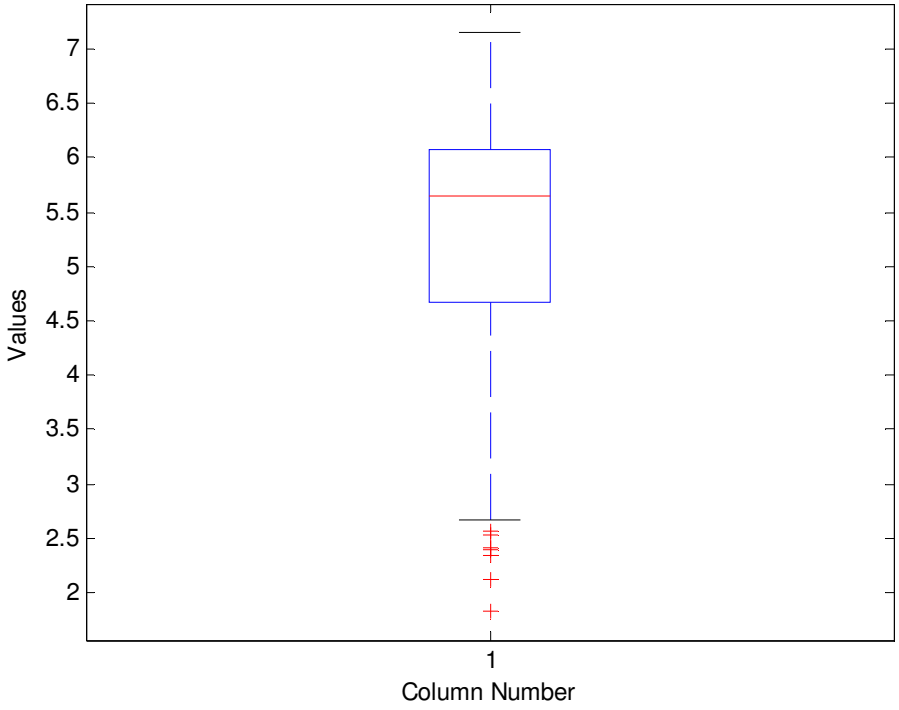

Figure 4.1: Graphical representation of the percentile analysis.

The graph shows an example of box plot based on statistical percentiles of the data. This plot has several graphic elements (Tukey, 1977):

- The lower and upper lines of the box are the 25^{th} and 75^{th} percentiles of the sample. The distance between the top and bottom of the box is the interquartile range. The box thus represents the middle 50 % of the data.

- The line in the middle of the box is the sample median. If the median is not centered in the box, as is the case here, it indicates skewness. Skewness of the data is a measure of the asymmetry of the data around the sample mean. If skewness is negative, the data are spread out more to the left of the mean than to the right. If

skewness is positive, the data are spread out more to the right. The skewness of the normal distribution (or any perfectly symmetric distribution) is zero.

- The "whiskers" are lines extending above and below the box. They show the extent of the rest of the sample (unless there are outliers). Assuming no outliers, the maximum of the sample is the top of the upper whisker. The minimum of the sample is the bottom of the lower whisker. By default, an outlier is a value that is more than 1.5 times the interquartile range away from the top or bottom of the box.

- The plus signs at the bottom of the plot are an indication of outliers in the data. These exist because the values that they represent fall outside of some statistical measure. These points may be the result of data entry error or poor measurements.

Percentile analysis, though valuable, deals with each column independently which may result in a reduced understanding of the interrelationships of the variables in the dataset. In order to investigate the intercorrelations in the data, Principal Component Analysis is employed.

4.3.2 Principal Component Analysis

Principal Component Analysis is a statistical technique used for data dimension reduction and pattern identification in high dimensional data (Jollife, 1986). The PCA orthogonalizes the components of the input vectors to eliminate redundancy in the input data thereby exploring correlations between samples or records. It then orders the resulting components such that the components with the largest variation come first. The compressed data (mapped into i dimensions) is presented by:

$$Y_{j\times i} = X_{j\times k} \times PCvector_{k\times i}. \qquad (4.1)$$

where the principal component vector, $PCvector$ is presented by the eigenvectors of the i largest eigenvalues of the covariance matrix of the input $X_{j\times k}$ with k dimensions and j set of

records $(i \leq k)$. The PCA is used in this work to orthogonalize the data, thereby revealing the internal structure of the dataset (Jollife, 1986) such that the missing data imputation model is better trained.

4.4 Results of the Data Analysis

When fitting a model in order to solve a problem, it is necessary to prepare the data such that the essence of the data is captured by the proposed model. First the data entries that contain logical errors are removed; these errors include ages of the mother that are greater than 60 and less than 12 (It is assumed that the reproductive health of a woman lasts from puberty to menopause) and instances where the parity is greater than the gravidity. Secondly, the data entries are normalized within the range [0 1] using min-max minimization:

$$X_{norm} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \qquad (4.2)$$

where X_{norm} is the normalized data set, X_{min} and X_{max} represent the minimum and maximum values for each variable in the dataset, respectively and X is the value to be normalized. This normalization is performed so that each of the inputs are of equal significance in the training of the methods implemented in chapter 5; this prevents the model from being biased towards a selected range of inputs. The data are then evaluated in order to see which attributes contribute the most outliers as shown in Figure 4.2.

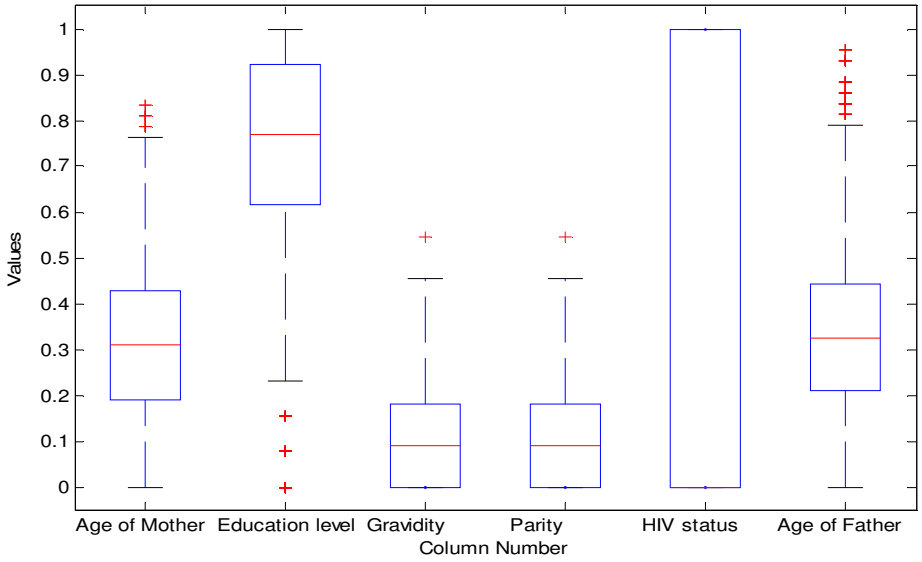

Figure 4.2: HIV dataset outliers per attribute.

It should be noted that the HIV status variable displays a different pattern from the other attributes because of its binary representation. All the other variables have medians that are in the middle of the box indicating no skewness in the data, that is, the data variables are (independently) symmetrically distributed. It can be seen from the percentile analysis that the age of the mother and the age of the father have a similar box graph which indicates correlation between the two variables; the same type of correlation can be seen for the gravidity and the parity attributes. In building certain models, such as the rule based model discussed in chapter 5, it might be necessary to assess each individual data attribute in order to assign characteristics to the system and this type of correlation might hinder the ability of the model to do so.

The crosses in the figure represent the outliers that are present as a result of each attribute and it is clear that the attributes that are more likely to be missing in the dataset i.e. the age of the mother, the education level and the age of the father, also produce the most outliers. It is important to note that these outliers are determined according to a statistical Gaussian measure (Tukey, 1977) and do not necessarily represent any false measurements in the

system. It can be deduced that the missingness of the data creates outliers which may hinder the ability of data analysis models to accurately analyse the data. It is thus important to remove outliers because not only do they often represent misplaced data points, they affect data analysis models resulting in longer training times and models that perform poorly.

In order to combat the effect that the correlation of data has on the missing data imputation model, the PCA is employed to orthogonalize the data ensuring that the model is better trained. Performing PCA on the input data results in orthogonal data that has variance percentages as illustrated by Figure 4.3 (the horizontal axis represent the principal components while the vertical axis indicates the percentage of the data represented in each principal component).

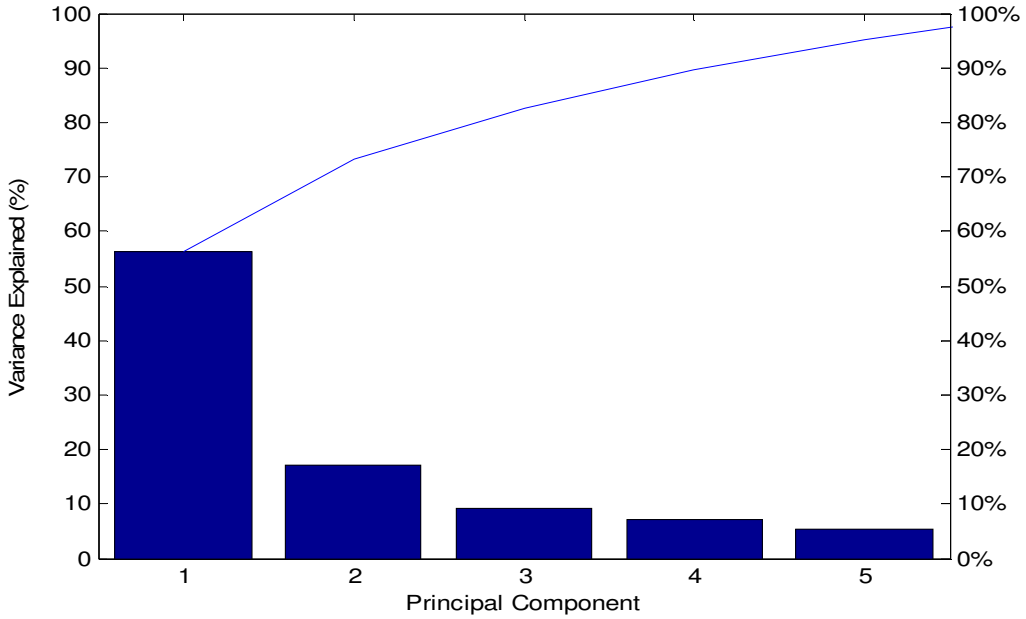

Figure 4.3: Variance of the training input data principal components.

It is clear that a greater percentage (75 %) of the variance in the principal components can be attributed to the first two principal components illustrating the orthogonality of the training data. The fact that the data are orthogonal gives the ability to build better models because

the data attributes can be independently modified (depending on the model) to determine their individual effects on the imputation model (Jollife, 1986).

4.5 Conclusion

In order to perform analyses on a dataset, it is important to understand the structure and statistical properties of the data. This chapter presented the statistical analysis that is used in this study to extract features of the South African HIV sero-prevelence data of 2001. This analysis includes the determination of the percentile characteristics of each data attribute in the dataset and the use of the PCA to orthogonalize the data in order to build an efficient model for missing data as discussed in Chapter 5.

Chapter 5

Computational Intelligence Approach to Missing Data

5.1 Introduction

In recent years, computational intelligence has been suggested as a candidate solution to the problem of missing data. The literature, however, endorses the use of ML with EM to impute missing data as discussed in chapter 2. Recently the two methods have been compared (Nelwamondo, Mohamed and Marwala, 2007) and it has been found that the EM algorithm is suitable in cases where there is little or no interdependence between the variables in the dataset; while AI is suitable when there are inherent non-linear relationships between the variables in the dataset. The AI technique used in recent literature has merged Auto Associative Neural Networks with Genetic Algorithms and has brought about significant results into the missing data problem. It should be noted that other computational intelligence methods have been applied in handling and imputing missing data. These methods include the use of decision trees and computational intelligence (Ssali and Marwala, 2007) and the use of ensemble based techniques for missing data handling (Nelwamondo and Marwala, 2007). These methods, however, are not exclusively discussed in this study because they are fundamentally based on the AI approach that forms the foundation for this chapter. In this chapter, the AI approach into missing data is investigated as the basis of the study; the uses of a pure neuro-fuzzy model and a hybrid version are also introduced and discussed as methods used to impute missing data.

5.2 Auto Associative Neural Networks and Genetic Algorithms

The AI approach incorporates the use of auto associative neural networks and genetic algorithms. These two models are discussed in this section.

5.2.1 Auto Associative Neural Networks

A neural network (NN) is composed of elements (neurons) inspired by biological nervous systems operating in parallel (Yoon and Peterson, 1990; Haykin, 1999). The NN is a machine that is designed to model the way in which the brain performs tasks or functions (Haykin, 1999). A typical neural network has a processing unit with an activation level, weighted interconnections between the processing units determining how one layer connects to another and a learning rule that determines how the weights are adjusted for a particular input-output pair. NNs have been used because of their ability to adapt to new environments and to derive meaning from complicated non-linear data. This has led to the NN being used in different applications such as pattern and speech recognition and financial modeling (Haykin, 1999). A simple example, and one that has been extensively used in literature is the multi-layer perceptron (MLP). These NNs are made up of multiple layers of computational units connected in a feed-forward architecture. Each neuron is connected to the neurons of the subsequent layer ensuring full interconnection between the layers. The MLP is made up of three layers, the input, output and hidden unit as shown in Figure 5.1 (Nabney, 2001).

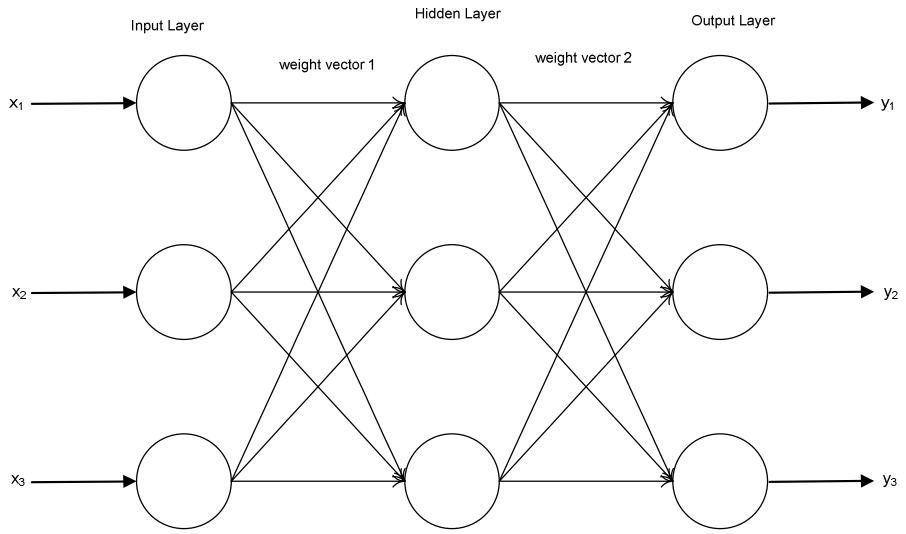

Figure 5.1: Structure of a MLP.

During training the weight vectors are adjusted by comparing the predicted output with the target output until the error between the two is minimized to a target value. When the error between the two is larger or lower than the targeted error, it is propagated back through the system and the weights are adjusted accordingly; this training is called backpropagation.

Auto-associative neural networks (AANN) are also called autoencoders and are trained to recall the input that is fed to them. The AANN is different from the MLP because not only does it map the input space, it also contains a bottleneck in the hidden layer in order to learn the interconnections of the variables in the input space such as covariance and correlation (Thompson et al., 2002). The structure of an autoencoder constructed using an MLP network is shown in Figure 5.2. The outputs produced in an autoencoder are such that they are the same as the inputs in order to map the input space.

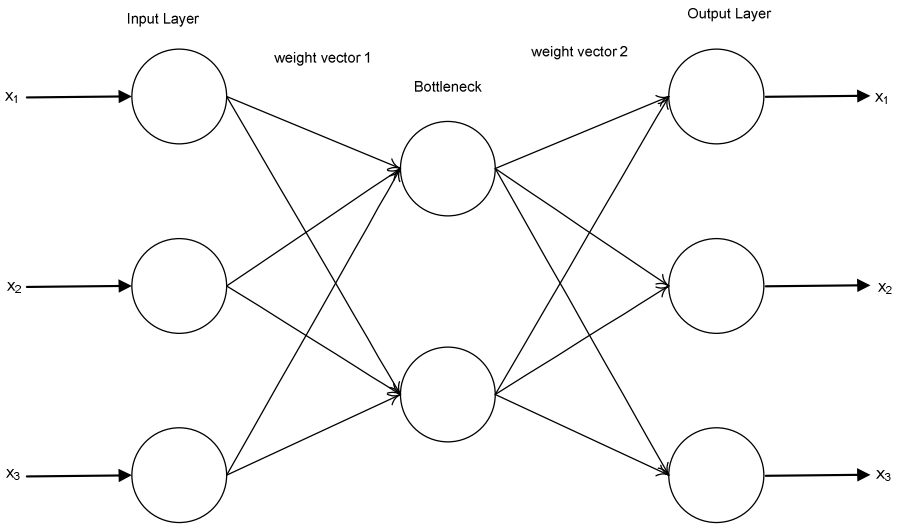

Figure 5.2: Structure of a three input three output autoencoder.

In the imputation of the missing data, the MLP neural network architecture is used and trained using backpropagation. Equation 5.1 is a representation of how the input vector, $\vec{x}$, is transformed using the weights W_{ij} and biases b_i associated with each hidden unit as follows:

$$a_j = \sum_{i=1}^{d} w_{ji} x_i + b_i \quad \text{j = 1,2.} \qquad (5.1)$$

where a_j represents a variable associated with the hidden unit. j represents the first and second layer, respectively. The input is further transformed using the activation function such as the hyperbolic tangent (tanh) in the hidden layer.

5.2.2 Genetic Algorithms

Genetic algorithms (GA) have been proven to be successful in optimization problems like scheduling, game playing and cognitive modeling. GAs use the concept of evolutionary biology, that is, the concept of survival of the fittest is applied over consecutive generations (Goldberg, 1989). Genetic algorithms view learning as a competition among a population of evolving contending problem solutions (Alfonseca, 1991). Through operations that are similar to gene transfer in biological evolution such as mutation, natural selection, inheritance and recombination, a new population of candidate solutions is formed (Banzhaf et al., 1998). A fitness function evaluates each solution and decides whether it will contribute to the next generation of solutions. Traditional optimization techniques are local in scope of their search and depend on well defined gradients in the search space. GAs are useful in this domain because of their ability to converge to global optimal solutions making them ideal for problem domains that have complex fitness landscapes. This is possible because, rather than focusing on a single candidate solution to an optimization problem, GAs operate on populations of solutions with the search process favoring the reproduction of individual (solutions) with

better fitness values than those of the previous generations. To optimize the operation of the GA, the following parameters need to be well chosen: population size, crossover rate, mutation rate, generation gap and mutation rate. Crossover is a simple process where two chromosomes and the crossover point are randomly selected and then letting the genes at the crossover point switch places. For example, if the first chromosome string is 10101010 and the second chromosome string is 00110011, if the crossover point is chosen to occur at the middle of the two strings, then the new string is 10100011. Mutation is a random process that seldom occurs where a gene (or bit) changes from 0 to 1 or vice versa. The procedure of the GA is given in the algorithm shown in Figure 5.3 as follows (Machalewicz, 1996).

```
BEGIN

   1. Randomly generate a population P of solutions at the
      initial generation g=0 from G generations.
   2. Evaluate the fitness of each of the population
      elements, P(g)
   3. While g < G
            begin
            Alter P(g) to form P(g+1) based on evolutionary
            parameters (crossover, mutation, inheritance)
            Evaluate the fitness of P(g+1)
            End
END
```

Figure 5.3: Structure of the Genetic Algorithm (Machalewicz, 1996).

5.2.3 Auto Associative Neural Networks and Genetic Algorithms (AANN-GA) for Missing Data imputation.

Nelwamondo, Mohamed and Marwala (2007) have compared the imputation accuracy of the EM algorithm and the AANN-GA combination. The AANN-GA method used in their analysis uses the set of the complete data to train an autoencoder to recall its input. The actual imputation occurs when the input in the trained autoencoder contains missing elements. When this occurs, the process depicted in Figure 5.4 is implemented.

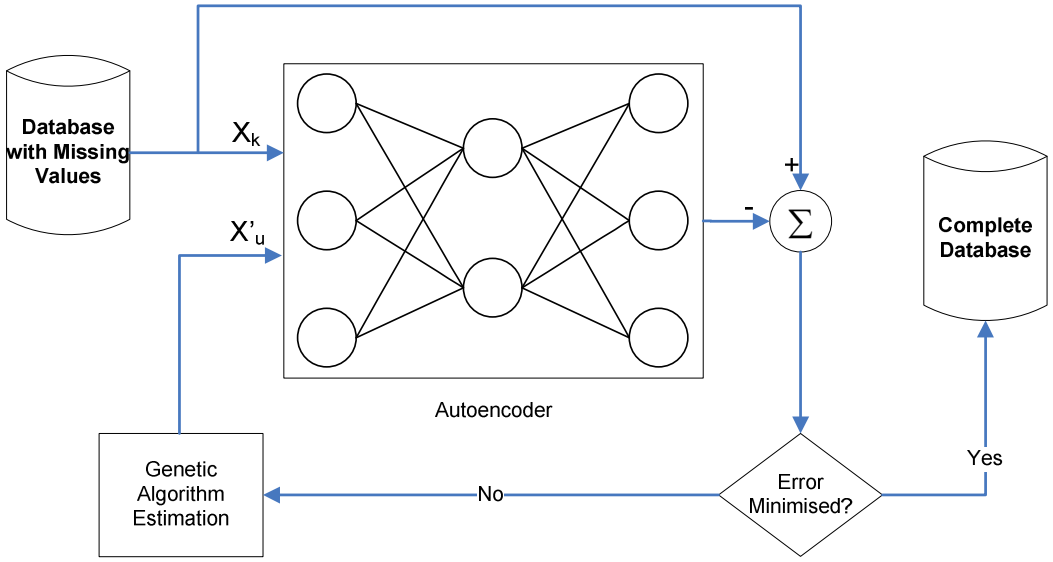

Figure 5.4: Autoencoder and GA based missing data estimator structure (Nelwamondo, Mohamed and Marwala, 2007).

The input into the model depicted in Figure 5.4 consists of the input data vector $\vec{x}$, that consists of known X_k and unknown X_u attributes. The autoencoder is trained such that the output in the system is the same as the input, that is (Abdella and Marwala, 2006)

$$f\left(\vec{W},\vec{x}\right)=\vec{x} \tag{5.2}$$

where $f\left(\vec{W},\vec{x}\right)$ is the output of the autoencoder represented by the weight vector $\vec{W}$ and the input vector $\vec{x}$. In reality, however, the dataset used is not similar to the problem space from which the autoencoder was trained, thereby, capturing intercorrelations in the dataset. Because of this, there exists an error between the target and actual outputs defined as follows (Abdella and Marwala, 2006):

$$e=\vec{x}-f\left(\vec{W},\vec{x}\right) \tag{5.3}$$

It is required that the error be minimal and nonnegative. To ensure that the error is positive, the square of the error function is used; and to ensure that the GA finds the minimum value (because the GA is designed to find an optimum maximum), the negative of the squared equation is supplied to the GA as a fitness function. Taking this into consideration and the fact that $\vec{x}$ is made up of known and unknown parameters, the fitness function, therefore, becomes:

$$e = -\left(\begin{Bmatrix} X_k \\ X_u \end{Bmatrix} - f\left(\vec{W}, \begin{Bmatrix} X_k \\ X_u \end{Bmatrix} \right) \right)^2 \tag{5.4}$$

Dhlamini et al. (2006) investigated other evolutionary computing methods such as Particle Swarm Optimisation (PSO) and Simulated Annealing (SA) and found that the GA has better performance in terms of the speed of convergence. The GA is thus used for the missing data imputation in this work.

5.2.4 Results of the AANN-GA Missing Data Imputation

The work done by Nelwamondo et al. (2007) analyzed the use of AANN-GA on missing data imputation using the HIV dataset. These results are discussed here to form a basis on the imputation accuracy that is expected of imputation models. In his results, he discusses the performance of the EM algorithm in comparison to the AANN-GA and found that both models impute the age of the mother with an 80 % accuracy within a 10 % tolerance. He also found that the EM performed better than the AANN-GA method for the prediction of the variables Education, Parity and Age gap because of its learning algorithm based on the ML method that can extract information even if there are no apparent interdependencies in the data. Although the EM algorithm outperforms the AANN-GA method of imputing missing data in a database, the prediction accuracy is still very low. Consider, for example, that if a person is 44 years old, their age can be imputed within 44±4.4 whilst a younger person who is 16 can have their age imputed within 16±1.6. This indicates that an older person is given a larger margin of error

than a younger one. These results indicate that an imputation model needs to be able to extract information in the dataset in order for it to impute missing data with greater accuracy.

5.3 Proposed Method

As discussed in chapter 2, there is still room for employing other computational intelligence models into the imputation of missing data. The results of the work done by Nelwamondo et al. (2007) indicate that, in the case of the HIV demographic dataset, the imputation model needs to have the ability to extract information on a dataset even without obvious interconnections of the data.

5.3.1 Neuro-Fuzzy Imputation

In order to improve the methods used for missing data imputation in social databases, it is important that the model used has the ability to interpret the interconnections in the database. The neuro-fuzzy (N-F) model is thus used for this purpose because of its ability to extract fuzzy rules from a dataset. The neuro-fuzzy architecture integrates the use of fuzzy models and intelligent processing in order to be able to learn and adapt through its environment (through the use of intelligent processing) and infer information and knowledge (as a result of the fuzzy model). A conventional fuzzy system uses expert knowledge to produce a linguistic rule base and reasoning mechanism for decision making. If artificial neural networks together with an optimization technique are incorporated into the fuzzy model to automatically tune the fuzzy parameters (antecedent membership functions and parametric consequent models), then the product is a neuro-fuzzy inference system (Jang et al., 1997; Bontempi and Bersini, 1997). There are a number of different fuzzy inference models including the Mamdani, Takagi-Sugeno (T-S) and the Tsukamoto fuzzy models (Jang et al., 1997). In this chapter, the T-S fuzzy inference system is used because of its ability to generate fuzzy rules from an input-output dataset which is especially useful in systems where the prior

knowledge of an expert is not available but a sample of input-output data is observed. A T-S neuro-fuzzy model is used in the implementation of the learning procedure depicted in Figure 5.5 (Bontempi and Bersini, 1997; Bontempi et al., 2001).

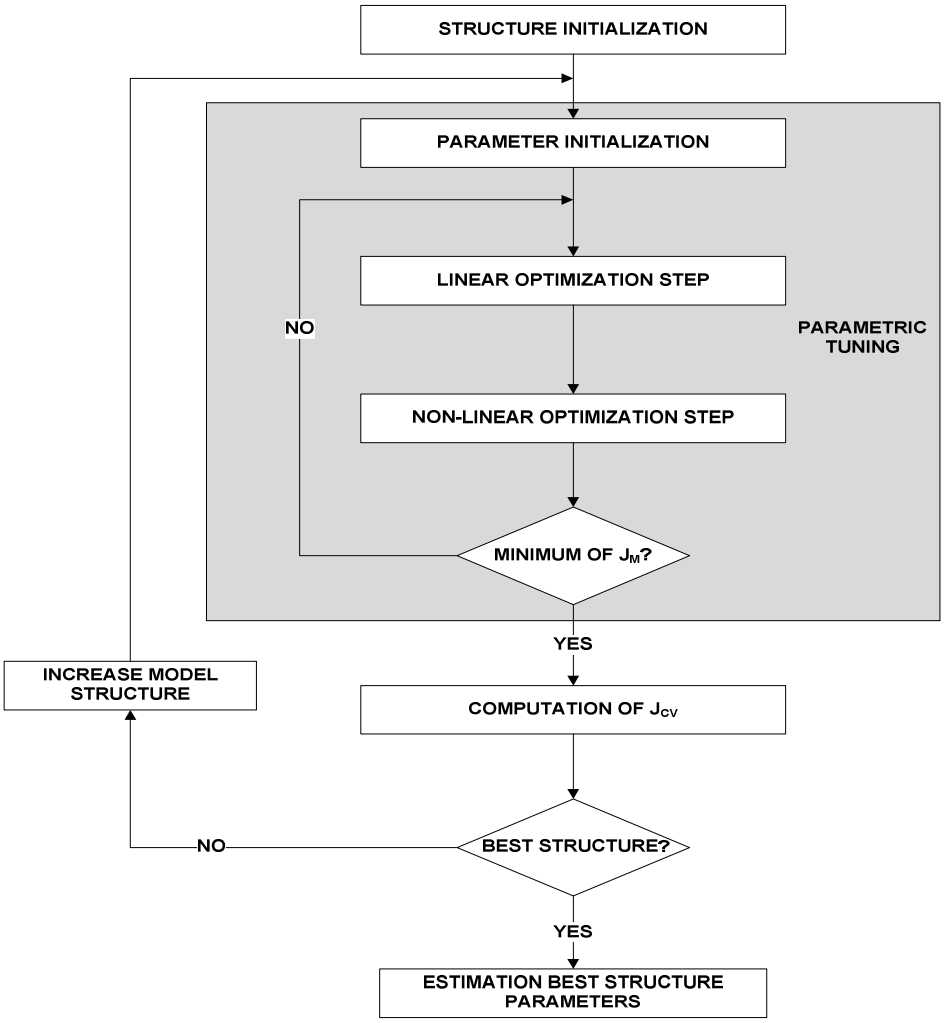

Figure 5.5: Structure of neuro-fuzzy learning procedure (Bontempi and Bersini, 1997; Bontempi et al., 2001).

The structural tuning (outer loop in Figure 5.6) is used to find the appropriate number of rules and partitioning of the input space. Once an optimum structure has been determined, the parametric tuning, which determines the optimum antecedent membership functions and consequent parameters is performed. The initialization of the architecture is performed by using a hyper-ellipsoid fuzzy clustering technique to cluster the input-output data. The parametric tuning (inner loop in Figure 5.5) searches for the best set of parameters by minimizing the sum of squares (J_M) cost function. This function is the error between the predicted and target values during training. Parametric tuning is dependent only on the training data. Prior to the training of the model to impute missing data, the optimum number of rules (outer loop) for the particular dataset has to be determined. It should be noted that for the HIV dataset, the training was initially performed with the least number of optimum rules in the architecture to minimize computational power and also to avoid over fitting. This is because evaluating the best structural ability of the input-output model is, as Figure 5.5 shows, computationally expensive because the parametric tuning (inner loop in Figure 5.5) needs to be evaluated for each number of rules (outer loop in Figure 5.5). Structural tuning aims to find the best number of rules and results in a graph depicting the computational accuracy against the number of rules. The initialization of the architecture is performed using hyper-ellipsoid clustering and the axes of the ellipsoids are used in initializing the consequent parameters. The cluster centers are projected into the input domain such that the antecedent membership functions are also initialized. From Figure 5.6 it is clear that the most optimum number of rules is between 4 and 8 because that gives the least sum squared error. Five rules were used for the training of the N-F model in this work.

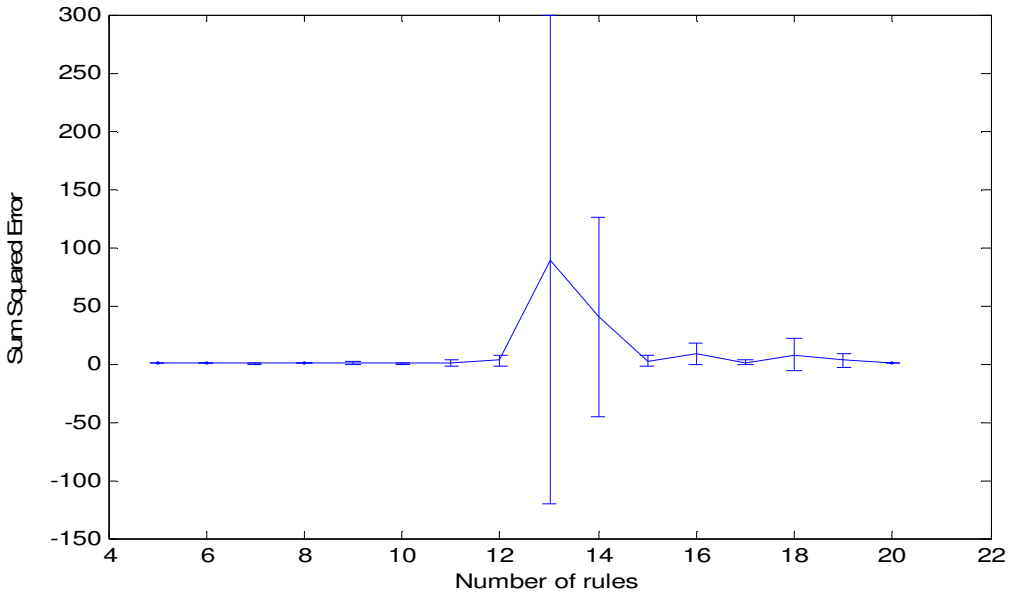

Figure 5.6: Cross validation vs. complexity.

The neuro-fuzzy architecture is trained using the Levenberg-Marquardt algorithm which provides a numerical solution to the problem of minimizing the sum squared error function by adjusting the fuzzy parameters (Bontempi and Bersini, 1997; Bontempi et al., 2001).

The neuro-fuzzy system was trained to recall the age of the father from a complete dataset consisting of 9745 instances; this variable is chosen because it is the variable with the most missing values. The results of this imputation are depicted in Figure 5.7.

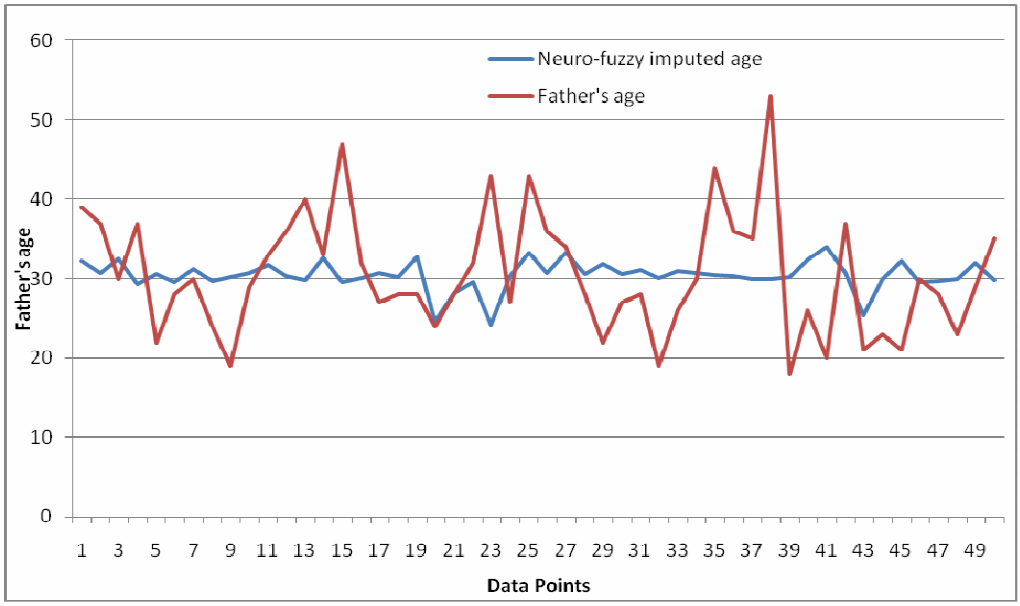

Figure 5.7: Neuro-fuzzy imputation of the father's age.

The N-F model has a 60 % accuracy in estimating the age of the father within 10 %. This is a lower accuracy than that of the NN-GA method and the EM algorithm which had better imputation accuracy. It can be seen from the figure that the imputation model tends to only impute the age of the father at around 30 years and could be indicative of the N-F system's inability to extract information and thus analyse the effect of each attribute on the system. It can be seen that the correlation between the age of the mother and the age of the father, as discussed in the percentile study of the data, creates a model that cannot extract features of each attribute from the dataset, therefore, the model simply imputes a value that is close to the mean of the data attribute.

In order to increase the accuracy of the N-F system, a hybrid method is proposed because hybrid systems attempt to capture complex natural intelligence by incorporating the use of complex connections in solving a problem. This hybrid method employs the use of the N-F, the PCA and the GA to impute missing data.

5.3.2 The Hybrid: Neuro-Fuzzy , Genetic Algorithms and PCA method

Because of the rule-base nature of the N-F model, it is important to orthogonalize the data prior to training because this eliminates the learning problem that exists as a result of the interdepencency of the data attributes. It is expected that the N-F model will perform better with orthogonal data because each component can be assigned its membership functions without the concern of interdependence with other datasets. Prior to the training of the N-F model for the HIV imputation, the PCA is employed to orthogonalize the data ensuring that the model is better trained. The orthogonal data are used to train three N-F models to impute the age of the father, the age of the mother and the education level of the mother, respectively, because these are the parameters with the most missing values. The flowchart of this method is depicted in Figure 5.8.

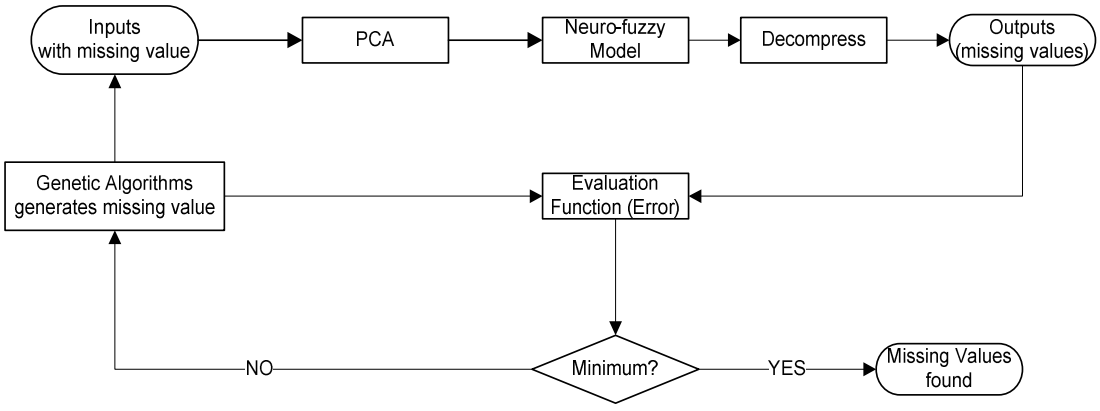

Figure 5.8: Flowchart of the proposed model for imputing missing data.

The model is trained with the same number of instances as above; the inputs of the model contain missing values that are randomly generated by the genetic algorithm. The input is then orthogonalized using the PCA and then fed into the N-F model that has been previously trained on complete data. The error between the imputed missing values and those generated

by the GA is used as the evaluation function that needs to be minimized. If the error is not minimal, the GA generates a new missing value that will minimize the error.

The proposed model is then implemented to impute the missing data yielding the results shown in figure 5.9 for the age of the father. There is obvious correlation between the imputed and actual age of the father.

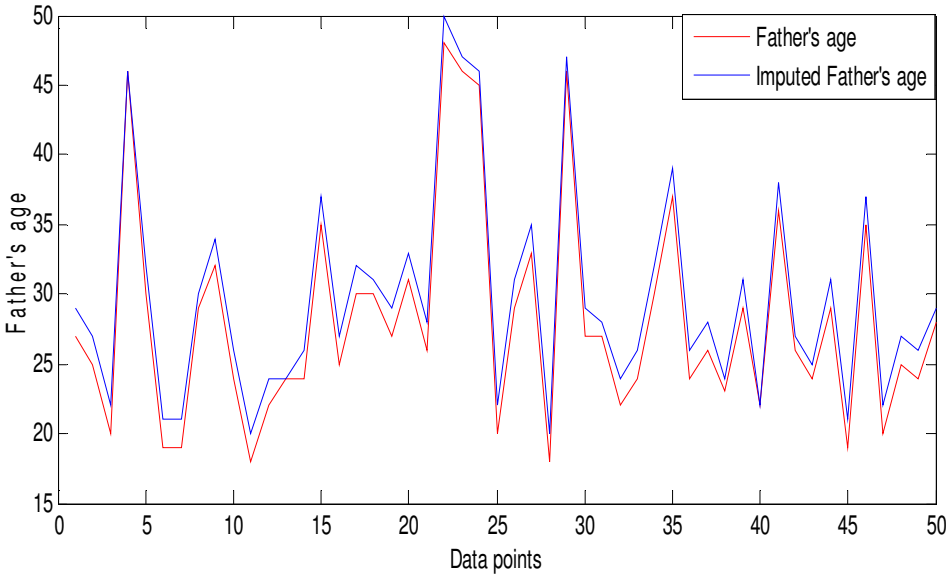

Figure 5.9: Proposed model's imputation of the father's age.

These results indicate an improved model that is able to extract information from the dataset and formulate rules that can model the system in terms of its inputs and the output. The same reason is considered for the results of the imputation of the age of the mother as can be seen from table 5.1 and figure 5.10. In the table, the accuracy of the mother or father's age is measured within 1 and 2 years and that of the education level is measured within 1, 2 and 5 grades.

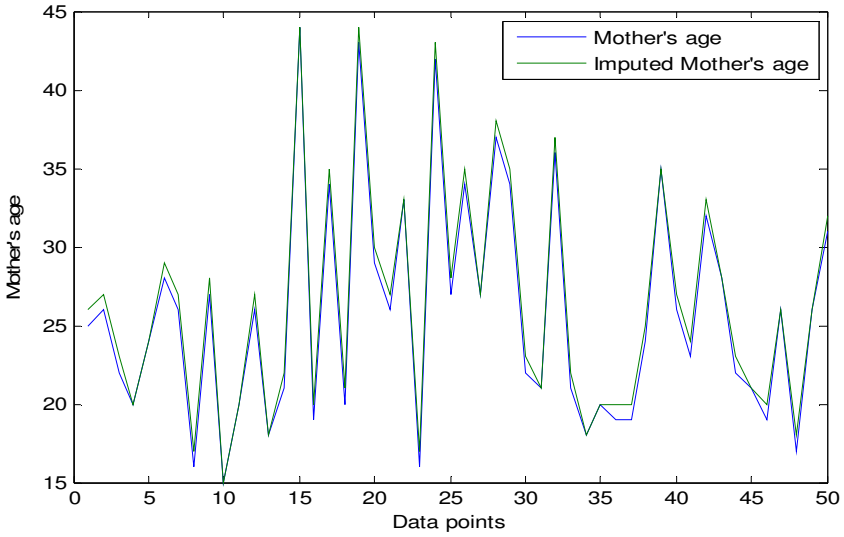

Figure 5.10: Proposed model's imputation of the mother's age.

The imputed education level of the mother has no correlation at all with the actual level indicating a low accuracy value measured within as indicated by table 5.1 and figure 5.11. It is observed, from the percentile analysis, that the education level had a differently shaped box plot with the outliers more spread out. The education level is also measured from 0 to 13 and gives a much smaller range than the age of the father and that of the mother. It is deduced, therefore, that the minimum accuracy in the imputation of the education level can be attributed to the lack of additional ranges of the data and the inability of the model to independently find the characteristic of this data attribute prior to training.

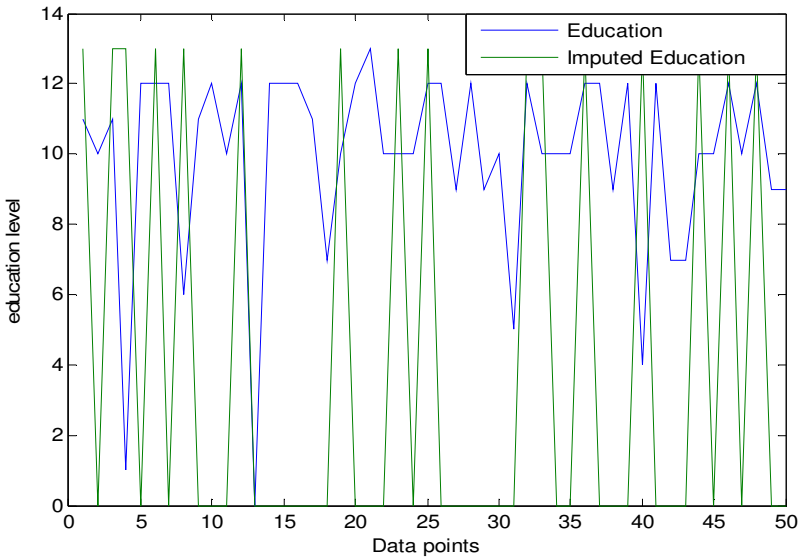

Figure 5.11: Proposed model's imputation of the mother's education level.

Table 5.1: Percentage of data that are correctly imputed.

Attribute	Exact Accuracy (within 0)	Accuracy within 1	Accuracy within 2	Accuracy within 5
Mother's age	43.638 %	98.99 %	100 %	100 %
Father's age	6 %	37.67 %	99.9 %	100 %
Education level	2 %	9 %	13 %	26.7 %

5.4 Conclusion

In this chapter, the use of computational intelligence methods as solutions to the missing data problem was presented. The results of a comparison between the AANN-GA method and the EM algorithm, as presented by Nelwamondo et al. (2007), to impute missing data in relation to

the HIV sero-prevalence data are also discussed as a basis of the expectation of a missing data imputation model. A N-F model was built for the imputation of missing data in the HIV sero-prevalence data and it was found that the lack of orthogonality between the variables inhibits the ability of the model to form accurate fuzzy rules. A novel computational intelligence hybrid method consisting of PCA, N-F and GA was also presented to impute missing data with an improvement in the accuracy of the system for variables that have less spread of the outliers.

Chapter 6

Impact of Missing Data Imputation

6.1 Introduction

This chapter analyses the impact that the presence of missing data and its imputation has on decision making. The 'decision' in this case is the prediction of the HIV status of an individual given their demographic factors. Two classifiers are used to assess this impact, namely the Bayesian classifier and the Fuzzy ARTMAP (FAM). Bayesian classification deals with the classification of unknown patterns based on statistical probabilities (Theodoridis and Koutroumbas, 2006). The Bayesian classifier used in this study employs the use of neural networks that are formulated in the Bayesian framework for HIV classification based on the sero-prevalence dataset. Fuzzy ARTMAP, introduced by Carpenter et. Al (1992), is a neural network based classifier that is capable of pattern classification. This supervised classifier has the ability of establishing an arbitrary mapping between an arbitrary analog input pattern and corresponding analog output patterns (Georgiopoulos et al. 1996). This network stems from the basic Adaptive Resonance Theory (ART) system that consists of a comparison field and a recognition field. The ART system is an unsupervised learning model that has been developed for pattern recognition and has led to its use in technological applications and biological analyses (Carpenter and Grossberg, 2003). Fuzzy ART implements fuzzy logic into ART pattern recognition thus enhancing generalization. Fuzzy ARTMAP combines two unsupervised modules to carry out supervised learning. The advantage of the fuzzy ARTMAP classifier over neural networks is its faster convergence and better efficiency. This system is thus also used to

determine the impact that the missing data imputation has on the decision (classification) problem.

6.2 Bayesian Classification of HIV

Appendix A presents the basis theory on the use of Bayesian classifiers. The classification results of using the abovementioned method are obtained for the classification of the HIV status on the dataset. Because of the nature and the biases that are inherent in the dataset (data are biased towards the HIV negative classification), it is not enough to simply present the accuracy of the classifier based on the number of data points that are correctly classified. This therefore means that a classifier that only predicts a negative HIV status will be deemed more accurate than a classifier that attempts to predict the status based on the data presented to it. To combat this problem, equations 3.2, 3.3 and 3.4 from chapter 3 are used to evaluate the performance of the classifier. A threshold is chosen for the classification of the HIV status using the Bayesian classifier and is chosen as 0.5, that is, all predictions less than 0.5 indicate a negative prediction and those greater than 0.5 indicate a positive prediction.

Because the dataset is skewed and biased towards the HIV negative cases, the classifier might bias the results towards the more common event (HIV negative). To combat this, training is performed on a balanced dataset of 12000 instances consisting of 6000 positive and 6000 negative cases and the testing is performed on an unbalanced dataset (2562 data points). The testing data consists of 612 positive cases and 1950 negative ones. When presented with the complete dataset (no missing entries) the classification results are depicted in the confusion matrix in table 6.1.

Table 6.1: Confusion matrix for the Bayesian classifier.

Bayesian classifier	Actual Condition	
	(+)	(-)
(+)	261	838
(-)	343	1120

As can be seen from the table and equations 3.1 to 3.3 in chapter 3, the true positive ratio is 43 % and the false positive ratio is 42 %. The overall accuracy of the classifier is 54 % which indicates that the classifier has a poor ability to efficiently classify the data based on its training model. However, the aim of this study is to assess the impact that the missing data imputation will have on the classifier performance and, therefore, the strength of the classifier is not considered to be the deterministic factor. In order to assess this impact, the Bayesian classifier is used to evaluate the classification results in the presence of missing data. Two scenarios are tested namely:

- The classification results when the missing field is imputed using the independent N-F system.
- The classification results when the missing field is imputed using the N-F, PCA, GA method.

The missing field chosen for all three cases is the age of the father because it is the field with the most missing values. In the presence of missing data, without its imputation, the model fails to classify. This is because the Bayesian classifier lacks the ability to perform when presented with 'new' information; 'new' information in this case refers to the missingness of a data attribute. Tables 6.2 to 6.3 depict the truth tables for the conditions mentioned above.

Table 6.2: Confusion matrix for the Bayesian classifier when the missing field is imputed using the N-F model independently.

Bayesian Classifier	Actual Condition	
	(+)	(-)
(+)	351	1037
(-)	253	921

From the table and equations 3.1 to 3.3 in chapter 3, the accuracy of the model when the age of the father is imputed using the N-F model is decreased to 50 %. The true positive ratio is 58 % and the false positive ratio is 53 % which indicates a significant depreciation of the capability of the classifier to correctly classify the HIV status of an individual as indicated by the increased false positive ratio and decreased accuracy.

Table 6.3: Confusion matrix for the Bayesian classifier when the missing field is imputed using the N-F, PCA, GA model.

Bayesian Classifier	Actual Condition	
	(+)	(-)
(+)	289	922
(-)	315	1036

Considering the accuracy of the N-F, PCA, GA model to impute missing data, it is anticipated that the accuracy of the classifier will be greater than the case when the N-F is used independently. The classifier gives a true positive ratio of 48 % and a false positive ratio is 47 % with an accuracy of 52 %. Although these results indicate a better 'decision' given the imputation model, the true positive ratio of the classifier given the use of the N-F model to impute missing data is higher than when the N-F, PCA, GA model is used. This means that the classifier is able to classify the positive HIV status better when it is presented with the mean of the missing attribute (the independent N-F system imputes values that are close to mean) and

can be due to the probabilistic nature of the Bayesian classifier (the mean is the most probable value).

6.3 Fuzzy ARTMAP Classification of HIV

The fuzzy ARTMAP (FAM) is used to evaluate the impact of missing data imputation on decision making. This is achieved by training the FAM using the complete dataset to map the inputs (Age of Mother, Age of Father, Gravidity, Parity, Education Level) to the corresponding output (the HIV status). The learning rate, β, is a factor by which the hyperboxes are adjusted with each training pattern during the training phase and in this problem, β = 1, which is known as fast learning. The vigilance parameter of ρ = 1, is used in the training of this architecture because of the need for detailed memories. This is because this problem consists of only two classes and finer categories are necessary for efficient classification. Because the dataset is skewed and biased towards the HIV negative cases the classifier might bias the results towards the more common event (HIV negative). To combat this, training is performed on a balanced dataset of 12000 inputs and the testing is performed on an unbalanced dataset (2652 data points). When presented with the complete dataset (no missing entries) the classification results are depicted in the confusion matrix in table 6.4.

Table 6.4: Confusion matrix for the FAM classifier.

Fuzzy ARTMAP	**Actual Condition**	
	(+)	(-)
(+)	571	305
(-)	41	1645

As can be seen from the table and equations 3.1 to 3.3, the true positive ratio is 93 % and the false positive ratio is 1.56 %. The accuracy for this model is 86 %. This indicates that the FAM has the ability to efficiently classify the data based on its training model.

In order to assess the impact of missing data imputation, the FAM is used to evaluate the classification results in the presence of missing data. Three scenarios are tested namely:

- The classification results when the missing field(s) is not replaced.
- The classification results when the missing field(s) is imputed using the independent N-F system.
- The classification results when the missing field(s) is imputed using the N-F, PCA, GA method.

As with the case of the Bayesian classifier, the missing field chosen for all three cases is the age of the father because it is the field with the most missing values. Tables 6.5 to 6.7 depict the truth tables for the conditions mentioned above.

Table 6.5: Confusion matrix for the FAM classifier in presence of missing data.

Fuzzy ARTMAP	**Actual Condition**	
	(+)	(-)
(+)	345	267
(-)	836	1114

As can be seen from table 6.5 and equation 3.1 in chapter 3, the true positive ratio is 56.4 % and the false positive ratio is 42.9 %. This indicates that in the presence of missing data, the FAM does not perform as well in classifying the given dataset. The presence of missing data also decreases the classification accuracy of the ARTMAP to 55 %.

Table 6.6: Confusion matrix for the FAM classifier when the missing field is imputed using the N-F method independently.

Fuzzy ARTMAP	Actual Condition	
	(+)	(-)
(+)	266	726
(-)	338	1232

Using equation 3.1 to 3.3 in chapter 3, the accuracy of the system is calculated to be 58 % with a true positive ratio of 44 % and a false positive ratio of 37 %. The false positive ratio indicates an improvement in the system as opposed to leaving the missing field as unknown. There is however only a 3 % increase in the accuracy of the classifier which indicates that the classifier handles missing data and the N-F imputation in the same way.

Table 6.7: Confusion matrix for the FAM classifier when the missing field is imputed using the N-F, PCA, GA method.

Fuzzy ARTMAP	Actual Condition	
	(+)	(-)
(+)	363	596
(-)	241	1362

When the missing field is imputed using the N-F, PCA, GA method, the FAM has an increased accuracy of 68 %. The true positive ratio increases to 60 % and the false positive ratio decreases to 30 %. This indicates an improved performance of the FAM to classify the HIV status when the proposed model is used to impute the missing data. Due to the fact that this classifier has the ability to learn and adapt when presented with 'new' information, it is able to classify in the presence of missing data. It should also be noted that the independent use of

the N-F in missing data imputation results in decreased accuracy for the FAM because this classifier is not probabilistically based and depends on the training model.

6.4 Discussions and Conclusions

Intuitively, missing data is expected to have an effect in decision making frameworks. In this chapter, two decision making frameworks are built, namely, a Bayesian classifier and the FAM for the HIV classification task. In both cases, the presence of missing data has detrimental effects with the neural network based Bayesian classifier failing to classify while the FAM's ability to classify reduces by 20 %. It is observed, however, that the imputation model has little effect on the accuracy of the Bayesian classifier while the effect on the FAM is fairly great. When considering the true positive ratio of the classification results, it is clear that the probabilistic nature of the Bayesian classifier enables it to perform better when the missing attribute is imputed using the independent N-F system (because of the system's imputation of the variable in chapter 5). The ability of the FAM to classify when the missing attribute is imputed using the independent N-F system depreciates (when considering the true positive ratio) because of the lack of knowledge of the imputed attribute from training. When the accuracy of classification is considered, it is clear in both cases that the classifier performs better when presented with the imputed results from the hybrid method. This is because these imputation results were considerably accurate as discussed in chapter 5. It is concluded, therefore, that the degree of the impact that missing data and the imputation thereof has on decision making is dependent on the both the imputation model and the decision (classification) framework.

It should be noted that the classification results obtained in this work should be subject to practical use in real situations in terms of classification of HIV within the sero-prevalance dataset. When presented with this demographic data, it might be useful for research institutions to analyse and classify the data so that it gives valuable and useful information. In this study, two forms of classifiers have been assessed in terms of their performance in the

presence of missing data. When the abovementioned classifiers are considered for the classification of HIV in the HIV sero-prevalence dataset, the following should be noted:

- If the dataset contains no missing values, then the FAM classifier should be used since it gives better classification accuracy than the Bayesian classifier.

- If the dataset contains missing data but no imputation method has been employed to deal with it, then it is suggested that the FAM be used to classify because, although its classification accuracy is reduced, it does not fail in the presence of missing data.

- If the N-F missing data imputation architecture is used in a dataset that contains missing elements, it is suggested that the Bayesian classifier be used in the classification of the data. This is because this classifier is probabilistically based giving less error in the classification results when considering the true positive ratio.

- If the N-F, PCA and GA missing data imputation method is used in the presence of missing data, it is noted that both classifiers retain a classification accuracy that is close to their abilities to classify HIV status with no missing values. In this case it is therefore suggested that the FAM be used because of its general high classification accuracy.

Chapter 7

Conclusions

7.1 Statistical Analysis of Data

In order to build meaningful models, the understanding of the data is necessary. From the percentile analysis of the data, it is clear that the statistical characteristics of the data attributes have an impact on the data analysis models. It was also observed that the missingness of a data attribute results in greater outliers that can often infringe the ability of the data analysis models. The PCA conducted proved useful in training the N-F model because of its ability to orthogonalize the data thereby enabling the formulation of efficient fuzzy rules.

7.2 Missing Data Imputation Models

The use of AANN-GA has been applied to missing data and because of the different computational intelligence methods available for pattern recognition, there is still room to investigate other models. In this study, it has been discovered that the use of a hybrid such as the PCA, N-F, GA model can improve the accuracy of an imputation model. This is expected because hybrid methods attempt to capture complex natural intelligence that is sometimes not captured by the use of less complex models. The orthogonalization of the data proved to be efficient in better training the N-F model resulting in an improvement of imputation accuracy from 60 % accuracy within a 10 % tolerance (when using the N-F model independently) to an accuracy of 99 % correct imputation within 2 years for the age of the father. It was therefore found that the imputation accuracy of a model not only depends on its

ability to extract information from data (Nelwamondo et al., 2007) but that the correlation, interdependency and the spread of the data outliers also play a pivotal role in the missing data imputation.

7.3 Impact of Missing Data Imputation

Because of the vast numbers of decision making frameworks, it is impossible to fully assess the impact of missing data on decision making. In order to combat this problem, two classifiers were assessed namely a Bayesian classifier and a FAM classifier. When the two methods are compared independently, the FAM outperforms the Bayesian classifier with an imputation accuracy greater than 80 %; this is due to its incremental ability. In the presence of missing data (without imputation) the Bayesian classifier breaks down because it lacks the ability to adapt in new situations; the FAM, however, does not break down but has a depreciation in classification accuracy that is greater than 20 %. When the N-F model is used independently to impute missing data, both classifiers exhibit a depreciation in accuracy and in the true positive ratio. However, the true positive ratio of the Bayesian classifier decreases when the hybrid imputation model is used; this is attributed to its probabilistic nature (because the independent N-F imputation gives results that are closer to the mean). The true positive ratio of the FAM when missing data is imputed using the hybrid method increases as can be expected when a model that has such close correlation to the original testing data is used. It is noted that the difference between the accuracy of the Bayesian classifier when in the presence of missing data and its accuracy when the missing variable is imputed using the hybrid method is 5 %. This is a small change in accuracy and indicates the ability of the classifier to be stable in the presence of missing data. The FAM depicts a different pattern with a significant decrease in accuracy in the presence of missing data (whether imputed or not). It can therefore be concluded that although missing data and its imputation has an impact in decision making, the extent of this impact is highly dependent on the decision making framework.

7.4 Future Work

It is clear that there is still room for future developments in the missing data field. Although this work has identified one of the gaps that still plague this field, that is, the impact of missing data on decision making, there still exist other problems that have not been dealt with here. Future work should investigate the individual role of each component in a hybrid computational intelligence model in order to assess the degree of impact each has on missing data imputation. Other decision making frameworks, such as regression models, should also be evaluated in order to fully assess the impact of missing data on decision making. It is also suggested, as future work, that other measures such as the Receiver Operating Characteristic (ROC) curve be used to measure the impact of missing data on classification accuracy.

References

Abdella, M. and Marwala, T.: 2006, The use of Genetic Algorithms and Neural Networks to Approximate Missing Data in Database. *Computing and Informatics* 24, 1001-1013.

Alfonseca, M.: 1991, Genetic Algorithms. *Proceedings of the International Conference on APL* 1-6. ACM Press.

Banzhaf, W., Nordin, P., Keller, R. and Francone, F.: 1998, *Genetic Programming-An Introduction: On the Automatic Evolution of the computer Programs and its Applications*. Morgan Kaufman, California, fifth edition.

Bishop C.: 2006, *Pattern Recognition and Machine Learning*. Springer, Singapore.

Bishop, C.M.:1996, Neural Networks for Pattern Recognition. *Oxford University Press*

Bontempi, G. and Bersini, H.: 1997, Now Comes the Time to Defuzzify Fuzzy Models. *Fuzzy Sets and Systems* 90(2), 161-170.

Bontempi, G., Bersini, H. and Birattan, M.: 2001, The Local Paradigm for Modeling and Control: From Neuro-Fuzzy to Lazy Learning. *Fuzzy Sets and Systems* 121(1), 59-72.

Carpenter, G.A and Grossberg, S.: 2003, Adaptive Resonance Theory. *Handbook of Brain Theory on Neural Networks*, MIT Press, Cambridge MA, 87-90. Second Edition.

Carpenter, G.A., Grossberg, S., Markuzon, N., Reynolds, J.H. and Rosen, D.B.: 1992, Fuzzy ARTMAP: A Neural Network Architecture for Incremental Supervised Learning of Analog Multidimensional Maps. *IEEE Transactions on Neural Networks (3) 693-713.*

Dai, Y., Hi, L., Liu, P., Liu, Y., Shi, Z., Sun, J., Pu, Y., Wang, C. and Yang, H.: 2008, False Positive and False Negative Predictive Value of HIV Antibody Test in Chinese Population. *Journal of Medical Screening* 15(2), 72-75.

Dempster, A.P., Laird N. M. and Rubin D. B.: 1977, Maximum Likelihood Estimation from Incomplete Data via EM Algorithm (with discussion). *Journal of the Royal Statistical Society*. Series B, 39. pp 1-38.

Dhlamini, S. M., Nelwamondo, F. V. and Marwala, T.: 2006, Condition monitoring of HV bushings in the presence of missing data using evolutionary computing, *WSEAS Transactions on Power Systems* 1(2), 280–287.

French, S.: 1986, *Decision Theory: An Introduction to the Mathematics of Rationality.* John Wiley and Sons, England.

Georgiopoulos, M., Fernulund, H., Berbis, G. and Heilima, G.L.: 1996, Order of Search in Fuzzy ART and Fuzzy ARTMAP: Effect of the Choice Parameter. *Neural Networks 9, (9)1541-1559.*

Golberg, D.E.: 1989, *Genetic Algotithms in Search, Optimization and Machine Learning*. Reading Mass. Addison-Wesley.

Hansson, S.O.: 1994, Decision Theory: A Brief Introduction. url: http://www.infra.kth.se/~sth/decisiontheoy.pdf, last accessed 01 October 2008.

Haykin, S.: 1999, *Neural Networks.* Prentice Hall, New Jersey, second edition.

Jang, T-S., Sun, C-T. and Mizutan, E.: 1997, *Neuro-Fuzzy and Soft Computing: A Computational Approach to Learning and Machine Intelligence*. Prentice Hall, New Jersey.

Jollife I.T.: 1986, *Principal Component Analysis.* Springler-Verlag, New York.

Kodish, S.: 2006, The Paradoxes of Leadership: The Contribution of Aristotle. *Leadership* 2, 451-468.

Leke, B., Marwala, T. and Manana, J.V.:2008, Computational Intelligence for HIV Modelling, *International Conference on Intelligent Engineering Systems*. 127-132.

Little, R.J. A and Rubin, D. B.: 2002, *Statistical Analysis with Missing Data*. John-Wiley and Sons, New York, Second Edition.

Little, R.J.A and Rubin, D. B (1987). *Statistical Analysis with Missing Data.* John-Wiley and Sons, New York. First Edition.

Little, R.J.A.: 1995, Modeling the Dropout Mechanism in Repeated Measures Studies. *Journal of the American Statistical Association* 90, 1112-1121.

Machalewicz, Z.: 1996, Genetic *Algorithms + Data Structures = Evolution Programs*. Springer-Verlag, New York, third edition.

MacKay, D.J.C.:1992, A practical Bayesian Framework for Backpropagation Networks. *Neural Computation (4), 448-472.*

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H. and Teller, E.: 1993, Equations of state calculations by fast computing machines. *Journal of Chemical Physics,* 21, 1087-1092.

Mohamed, S., Rubin, D. and Marwala, T.: 2006, Multiclass Protein Sequence Classification using Fuzzy ARTMAP. *IEEE Conference on Systems, Man and Cybernetics, 1676-1681.*

Nabney, I.: 2001, *Netlab: Algorithms for Pattern Recognition.* Springer-Verlag, United Kingdom.

Neal, R. M.: 1994, Bayesian Learning for Neural Networks. *PhD Thesis.* Department of Computer Science, University of Toronto, Canada.

Nelwamondo, F., Mohamed, M. and Marwala, T.: 2007, Missing data: A Comparison of Neural Networks and Expectation Maximization. *Current Science* 93(12), 1514-1520.

Nelwamondo, F.:2007, Computational Intelligence Techniques for Missing Data Imputation, *PhD Thesis.* Faculty of Engineering and the Built Environment, University of the Witwatersrand, South Africa.

Nelwamondo, F.V. and Marwala, T.: 2007, Fuzzy ARTMAP and Neural Network Approach to online Processing of Inputs with Missing Values. *SAIEE Africa Research Journal. 98(3), 45-51.*

O'Hagan, M.: 1987, Fuzzy Decision Aids. *Proceedings of the 21st Conference on Signals, Systems and Computers* 2, 264-628.

Qiao, W., Gao, Z. and Harley R.G.: 2005, Continuous Online Identification of Nonlinear Plants in power Systems with Missing Sensor Measurements. *IEEE International Conference on Neural Networks*, 1729-1734.

Raiffa, H.: 1997, Introductory Lectures on Choices under Uncertainty. Reading, McGraw Hill.

Robins, J.M., Rotnitzky, A. and Zhao, L.P.: 1994, Estimation of Regression Coefficients when some Regressors are not Always Observed. *Journal of the American Statistical Association* 89, 846-866.

Rubin, D. B.: 1976, Inference and Missing Data. *Biometrica* 63, 581-592.

Schafer, J. L. and Graham, J. W.: 2002, Missing Data: Our View of the State of the Art. *Psychological Methods* 7(2), 147-117.

Schafer, J.L. and Olsen, M.K.: 1998, Multiple Imputation for Multivariate Missing-Data Problems: A Data Analyst Perspective. *Multivariate Behavioral Research* 33(4), 545-571.

Schafer, J.L.: 1997, *Analysis of Incomplete Multivariate Data*. Chapman and Hall, London.

South African Department of Health.: 2005, South African National HIV Prevalence, HIV Incidence, Behaviour and Communication Survey. url: http://www.hsrcpress.ac.za/product.php?productid=2134, last accessed, 01 October 2008.

South African Department of Health.; 2001, HIV and Syphilis Sero-Prevalence Survey of Women Attending Public Antenatal Clinics in South Africa. url: http://www.info.gov.za/view/DownloadFileAction?id=70347,last accessed, 01 October 2008.

Ssali, G. and Marwala, T.: 2008, Estimation of Missing Data Using Computational Intelligence and Decision Trees. *Proceedings of the IEEE International Joint Conference on Neural Networks*, 201-207.

Theodoridis, S. and Koutroumbas, K.: 2006, Pattern Recognition. *Academic Press, Amsterdam* , Third edition.

Thompson, B., Marks, R. and Choi, J.: 2002, Implicit Learning in Autoencoder Novelty Assessment. *IEEE International Joint Conference on Neural Networks* 3, 2878-2883.

Tim, Taryn.: 2007, Predicting HIV Status Using Neural Networks and Demographic Factors. *MSc Dissertation,* Faculty of Engineering and the Built Environment, University of the Witwatersrand, South Africa.

Tukey, J. W.: 1977, Exploratory Data Analysis. *Addison-Wesley*, Reading, MA.

Verbeke, G. and Molenberghs, G.: 2000, *Linear Mixed Models for Longitudinal Data*. Springler-Verlag, New York.

Wang, S.: 2005, Classification with Incomplete Survey Data: A Hopfield Neural Network Approach. *Computers and Operations Research* 24, 53-62.

Weber, E.U. and Coskunoglu, O.: 1990, Descriptive and Prescriptive Models of Decision Making for the Developments of Decision Aids. *IEEE Transactions on Systems, Man and Cybernetics* 20(2), 310-317.

Yoon, Y. and Peterson, L.: 1990, Artificial Neural Networks: An Emerging New Technique. *Proceedings of the 1990 ACM SIGBOP Conference on Trends and Directions in Expert Systems* 417-422.

Zhang, G. P: 2000, Neural Networks for Classification: A Survey. *IEEE Transactions on Systems, Man and Cybernetics-Part C: Applications and Reviews, 30 , 451–462.*

Appendix A

A. Bayesian Neural Networks for Classification Tasks

Neural networks are extensively used in pattern recognition because of their ability to model complex input-output relationships. Considering the use of neural networks in classifying the HIV status based on the demographic features, the five demographic variables are used as the inputs (x) to the MLP and the HIV status (y) as the output to be mapped to the input as depicted in Figure A.1.

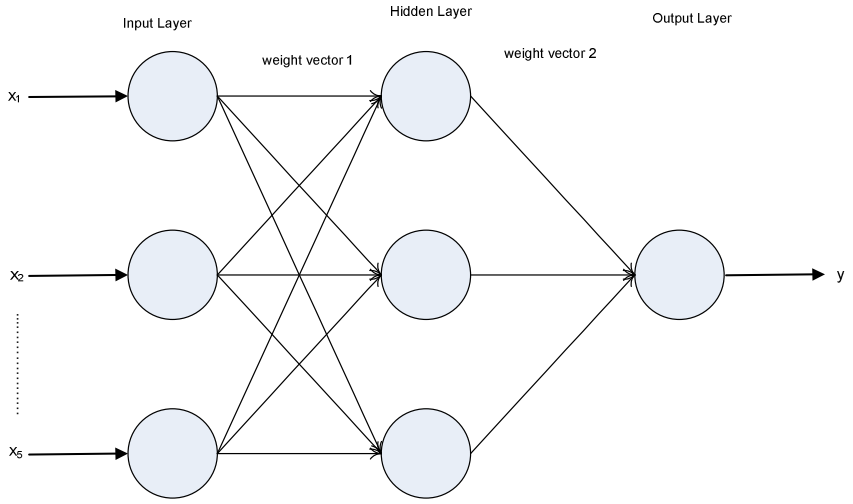

Figure A.1: Neural network structure for classification problems.

The relationship between the k^{th} HIV status to be predicted, y_k, and the demographic variables x may be written as follows (Bishop, 1996):

$$y_k = f_{outer}\left(\sum_{j=1}^{M} w_{kj}^{(2)} f_{inner}\left(\sum_{i=1}^{d} w_{ji}^{(1)} x_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)}\right) \quad \text{(A.1)}$$

where $w_{ji}^{(1)}$ and $w_{kj}^{(2)}$ indicate the weights in the first and second layers respectively going from input *i* to hidden unit *j*, *M* is the number of hidden units, *d* is the number of output units while $w_{j0}^{(1)}$ and $w_{k0}^{(2)}$ indicate the bias of the hidden unit *j* and the bias of the output unit *k*. The selection of the appropriate network architecture is necessary in forming the model. A MLP architecture trained using the scaled conjugate gradient method is used for this analysis. The dataset is divided into training, validation and testing data. The validation data are used to find the optimum number of training cycles that avoid overfitting of the model to the training data. A threshold is set for the testing data where all the predicted cases that are below that threshold are classified as HIV negative and those above the threshold are classified as HIV positive.

The calculation of the weights and biases of the neural network may be written as follows (Bishop. 1996):

$$P(w|D) = \frac{P(D|w)P(w)}{P(D)} \quad \text{(A.2)}$$

where $P(w)$ is the probability distribution function of the weight space in the absence of any data and $D \equiv (y_1, ..., y_N)$ is a matrix containing the HIV data. The quantity $P(w|D)$ is the posterior probability distribution function after the data has been seen. $P(D|w)$ is the likelihood function and $P(D)$ is the normalization function also known as the evidence. For the MLP, equation A.1 can be expanded using the cross-entropy error function to give equation A.3 (Bishop, 1996). The cross-entropy function is used because of its classification advantages with the weight decay for the prior distribution assumed because its penalization of weights of higher magnitudes (Bishop, 1996).

$$P(w|D) = \frac{1}{Z_s} \exp\left(\beta \sum_{n}^{N} \sum_{k}^{K} \{ t_{nk} \ln(y_{nk}) + \ln(1 - y_{nk}) \} - \sum_{j}^{w} \frac{\alpha_j}{2} w_j^2 \right) \qquad \text{(A.3)}$$

where:

$$Z_s(\alpha, \beta) = \left(\frac{2\pi}{\beta} \right)^{N/2} + \left(\frac{2\pi}{\alpha} \right)^{w/2} \qquad \text{(A.4)}$$

In equation A.3, n is the number index in the training pattern, hyperparameter β is the data contribution to the error, *k* is the index for the output units, t_{nk} is the target output corresponding to the n^{th} training pattern and k^{th} output unit and y_{nk} is the corresponding predicted output. The parameter α_j is a hyperparameter which determines the relative contribution of the regularization term of the training error. Equation A.3 can be solved using the Taylor expansion and approximating it by a Gaussian distribution and applying the evidence framework (MacKay, 1992). This equation can also be solved by numerically sampling the posterior probability using the Hybrid Monte Carlo (HMC) method (Neal, 1994).

a. Hybrid Monte Carlo Sampling

The main idea of this method is to solve a problem by calculating the mean of a function, *f(w)*, sampled from the posterior probability of the weights. The HMC method uses the gradient of the neural network error to ensure that the simulation samples throughout the regions of higher probabilities thus avoiding the random walk that is often associated with traditional Monte Carlo methods (Bishop, 1996). The gradient is calculated using the backpropagation method. According to Neal (1994), sampling using the HMC is performed by taking a series of trajectories and then either accepting or rejecting a resulting state at the end of each

trajectory. Each state is represented by the network weights and its associated momentum ρ_i. Each trajectory is achieved by following a series of leapfrog steps where for each leapfrog step size ε_0 and the number of leapfrog steps, *L*, the transition between two states of the HMC procedure is performed as follows (Neal, 1994):

- Randomly choose the direction of the trajectory, λ, to be either -1 for the backward trajectory and +1 for the forward trajectory,
- Starting from the initial state, (w, ρ), perform *L* leapfrog steps resulting in a new state, (w_{new}, ρ_{new}).
- Reject or accept (w_{new}, ρ_{new}) using the Metropolis criterion.

The Metropolis criterion (Metropolis et al, 1993) accepts the new sample if the current posterior probability given the weights and the data is higher than the previous posterior probability; otherwise it accepts it with a probability of $\exp -\left(dE/T\right)$ where dE is the change in error between the current and the previous samples.

Appendix B

B. Fuzzy ARTMAP

The fuzzy ARTMAP is a supervised learning system made up of two unsupervised ART networks ARTa and ARTb, as well an inter-ART module as depicted in Figure A.1. During training, ARTa receives an input, a, and ARTb receives the corresponding output, **O**. The inter-ART module includes a mapping field which determines whether the correct mapping has been achieved from inputs to outputs. Data pre-processing is required prior to the classification task of the fuzzy ARTMAP. Firstly, min-max normalization is performed in order to produce a vector, **a**, whose values lie in the interval [0 1]. The second pre-processing phase at field F_0^a performs complement coding by accepting a and producing a vector **I** such that

$$I = (a, a^c) \tag{B.1}$$

Complement coding is also performed for the class labels applied at the ART_b module producing the output pattern **O**.

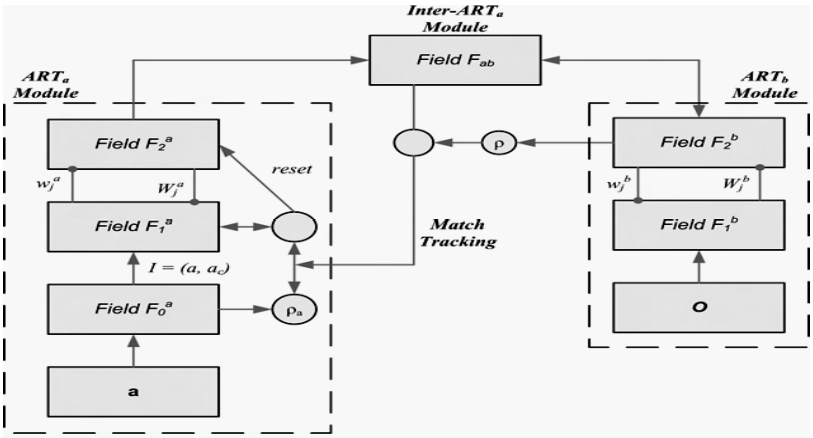

Figure B.1: Fuzzy ARTMAP architecture (Carpenter et. Al 1992).

During training, the ARTMAP architecture uses a mini-max learning rule which minimises the error and maximises the code compression (the number of patterns stored in the hidden units). Training stops when the weights can no longer be adjusted irrespective of the number of times the input-output patterns are presented to ART_a and ART_b respectively (Carpenter et. Al 1992). The performance of this classifier is dependent on a number of parameters such as the vigilance parameter, ρ, the learning rate, β, and to some extent the order in which the data are presented to the fuzzy ARTMAP architecture. The vigilance parameter controls the input data clustering providing a trade-off between the incremental learning ability of the fuzzy ARTMAP and the classification accuracy. Small values of ρ result in coarser clustering of the input space and larger values allow for fine categories and abstract memories. The fuzzy ARTMAP divides the input space into a number of hyperboxes and these are adjusted by the learning rate during training. When β =1, this is regarded as fast learning where hyperboxes are increased to include the points represented by the input vectors (Mohamed et al. 2006). Depending on the order in which the input data are presented to the ARTMAP, the classification performance will be different. As a result, the performance of the ARTMAP is averaged for different orders of the input data. The fuzzy ARTMAP has gained popularity over the years because of its incremental learning capabilities. This it achieves by assigning a flag when presented with a previously unseen class which ensures that a new class is detected instead of being classified incorrectly, thus the architecture has also found applications in the online domain (Nelwamondo and Marwala, 2007). The FAM can also be applied in offline applications as cited by Carpenter et al. (1992).

Appendix C

C. Publications

This is a list of published papers and papers that has been accepted to be published that has been derived from this work

- Hlalele, N.U., Nelwamondo, F.V. and Marwala,T.: 2008, Estimation of Missing Data using a Neuro-Fuzzy Architecure, *International Association of Science and Technology for Development: Modeling and Simulation (IASTED)*, 31-36.
- Hlalele, N.U., Nelwamondo, F.V. and Marwala,T.: 2008, "Imputation of Missing Data using PCA, N-F and Genetic Algorithm", *15th International Conference on Neural Information Processing of the Asia-Pacific Neural Network Assembly*, Springer (accepted).

ESTIMATION OF MISSING DATA USING A NEURO-FUZZY ARCHITECTURE

Nthabiseng U. Hlalele, Fulufhelo V. Nelwamondo, Tshilidzi Marwala
School of Electrical and Information Engineering, University of the Witwatersrand
Private Bag 3, Wits, 2050
South Africa
nthabiseng.hlalele@students.wits.ac.za, f.nelwamondo@ee.wits.ac.za, t.marwala@ee.wits.ac.za

ABSTRACT
In this paper, the analysis of a neuro-fuzzy inference architecture for missing data imputation is presented. The background of the missing data problem is sketched along with the methods currently used to impute missing data in databases. Three datasets, namely, the HIV South African seroprevelance data, Puma 560 Robot Arm data and the letter image recognition data are used to investigate the ability of the inference system to estimate the missing data. The inference is found to have an accuracy of 60 % when imputing the age of the father in the HIV dataset and an almost zero error when imputing the angular acceleration of the robot arm data. The accuracy of imputing the width of the letter attribute box proved to be satisfactory with a correlation coefficient of 0.9636 indicating close correlation.

KEY WORDS
Neuro-fuzzy, missing data, ANFIS, numerical methods

1. Introduction

The missing data problem has gained great popularity which has led to the development of a number of methods to handle and, in cases where the data have to be computationally analysed, impute the missing data [1], [2], [3]. In most cases, as is the case with large databases, computational analysis cannot be conducted without the availability of complete information. Traditionally, ad hoc methods have been used when dealing with missing data; these include mean substitution and the deletion of all data entries that contain missing variables. Although easy to implement, these methods often lead to loss of data resulting in a more biased database. This has led to the development of more advanced regression techniques and likelihood based approaches such as expectation maximization (EM). Recently, the use of autoencoders in conjunction with the genetic algorithm has also been employed for the missing data problem. This paper introduces the missing data problem and uses a neuro-fuzzy architecture to estimate missing data in three databases, namely, the South African HIV seroprevalance survey data, puma 560 Robot Arm data and letter (alphabet) image recognition data. The backgrounds of the missing data problem as well as neuro-fuzzy computing are presented. The experimentation and results are then presented followed by the conclusion section.

2. Background

The background of the missing data problem and its mechanisms as well as a brief background of the neuro-fuzzy architecture used in this paper are presented.

2.1 Missing data

The missing data problem is a widely researched topic [1], [2], [3], [4], [5], [6], [7] that has led to developments of many methods that analyse and impute the missing data with great accuracy. Missing data estimation depends, to a large extent, on the knowledge of how the data are missing. Three mechanisms of missing data have been documented [1], [2], [4], [8] which include missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR), also called the non-ignorable case. MCAR occurs when the probability of the missing data variable is independent of the variable value itself or on any other values in the database. In this case, when the missing records are small in comparison to that of complete records, list-wise or pair-wise deletion of cases may be chosen as methods of handling the missing data [9]. MAR occurs when the probability of the missing data variable is dependent on other variables in the database but not on the value of the variable itself. This means that there exists some complex relationship between the observed and missing data; i.e. the observed data can be used to approximate the missing data. When data are MNAR, the probability of the missing data variable is related to the value of the missing variable, this means that missing data variables cannot be predicted from the observed database. Dealing with this type of missing data is difficult and may involve imputing the missing variables based on external data not within the database [9]. There are a number of methods that are available for the imputation of missing data including the use of artificial intelligence methods. This paper investigates the use of a neuro-fuzzy architecture to impute the missing data. This is because of its ability to extract expect knowledge from the input-output dataset enabling the understanding of the system.

2.2 Neuro-fuzzy computing

The neuro-fuzzy architecture integrates the use of neural networks, which identify interrelationships and patterns in numerical datasets, and fuzzy systems, that incorporate expert knowledge and perform decision making [10]. This results in an inference system (as a result of fuzzy rules) that has the ability to learn and adapt through its environment (as a result of neural networks). A conventional fuzzy system uses expert knowledge to produce a linguistic rule base and reasoning mechanism for decision making. If artificial neural networks, together with an optimisation technique, are incorporated into the fuzzy model to automatically tune the fuzzy parameters (antecedent membership functions and parametric consequent models), then the product is a neuro-fuzzy inference system [10], [11]. There are a number of different fuzzy inference models including the Mamdani, Takagi-Sugeno (T-S) and the Tsukamoto fuzzy models [10]. In this paper, the T-S fuzzy inference system is used because of its ability to generate fuzzy rules from an input-output dataset which is especially useful in systems where the prior knowledge of an expert is not available but a sample of input-output data is observed. A T-S adaptive neuro-fuzzy inference system (ANFIS) architecture is used in the implementation of the learning procedure depicted in figure 1 [11], [12].

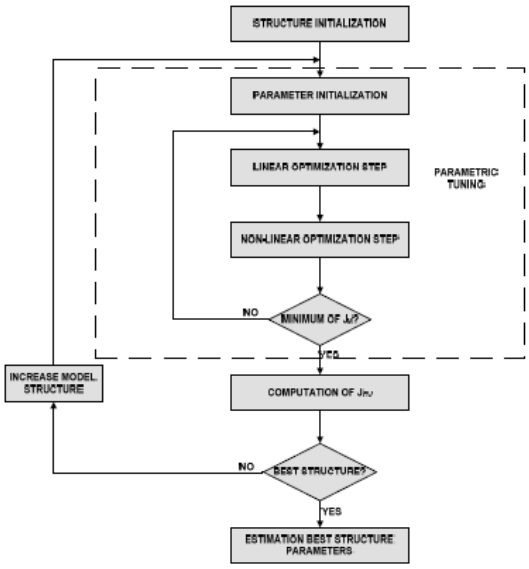

Figure 1: Flow chart of the ANFIS learning procedure [11], [12].

The structural tuning (outer loop in figure 1) is used to find the appropriate number of rules and partitioning of the input space. Once an optimum structure has been determined, the parametric tuning, which determines the optimum antecedent membership functions and consequent parameters is performed. The initialisation of the architecture is performed by using a hyper-ellipsoid fuzzy clustering technique to cluster the input-output data. The parametric tuning (inner loop in figure 1) searches for the best set of parameters by minimising the sum of squares (J_M) cost function. This parametric tuning is dependent only on the training data [11], [12].

3. Experimentation

The evaluation of the ANFIS architecture in imputing the missing variable in each dataset is dependent on the representation of the data as well as the structural selection of the architecture. The datasets as well as the structural selection for each dataset are presented on this section.

3.1 Datasets

Three datasets, viz. the HIV seroprevelence data, part of the Pumadyn datasets and the letter image recognition data, were used to measure the performance of the ANFIS architecture at imputing missing data. These datasets are explained further in this section. The missing variables in each dataset are the age of the father in the HIV database, the angular acceleration of the Pumadyn robotic arm and the width of the box containing the letter pixel in the letter recognition dataset. In order to incorporate the use of the ANFIS architecture to impute the missing variables in each case, the attributes in each dataset are scaled between the range [-1 1].

HIV dataset

This dataset was obtained from the South African antenatal seroprevalence survey of 2001. The data for this survey were collected from questionnaires answered by pregnant women visiting selected public clinics in South Africa and only women undertaking in the study for the first time were allowed to participate. This dataset has been used by analysts to investigate the effect that demographic information has on the HIV risk of an individual. This is especially helpful in countries such as South Africa, that have a high HIV infection rate. The data attributes used in this study are the HIV status, Education level, Gravidity, Parity, the Age of the Mother (pregnant woman) and the Age of the Father (responsible for the most recent pregnancy). The HIV status is represented in binary form, where 0 and 1 represent negative and positive respectively. The education level indicates the highest grade successfully completed and ranges between 0 and 13 with 13 representing tertiary education. Gravidity is the number of pregnancies, successful or not, experienced by a female, and is represented by an integer between 0 and 11. Parity is the number of times the individual has given birth and multiple births (e.g. twin births) are considered as one birth event. Both parity and gravidity are used in this dataset to indicate the reproductive health of the woman.

The age of the mother ranges from 14 to 50 years. It is observed from the dataset that the most missing variable is the age of the father. Imputing this data variable is helpful in educating people about HIV and the factors that render some individuals more risky than others. In situations where the age of the father is missing in the questionnaire, it is almost impossible to retrieve this information from the woman who supplied it due to the anonymity of the study. In order to capture the full understanding and contextualization of the data, it is necessary to have the full dataset which means that attributes that are missing, like the age of the father, should be imputed to capture the full dataset. Because of the nature of the data, many attributes were missing or not supplied. All the records with missing fields, outliers and with logical errors were removed for the training of the ANFIS system leaving a total of 12367 records from an initial 16743 records. These records were removed so as to use complete information when evaluating the missing data imputation ANFIS system. The remaining records were then randomly varied and divided to form training (9745 records) and testing data (2622 records). The testing data consists of previously unseen data with missing values of the father's age to be imputed by the ANFIS architecture. This data has been used in the evaluation of a number of computationally intelligent systems used to impute missing data [1], [2].

Puma 560 Robot arm dataset

This dataset forms part of the Pumadyn datasets which are a family of datasets that are generated from realistic simulations of the dynamics of a Puma 560 Robot arm [13]. The given dataset consists of 8 inputs and one output given by:

$$\ddot{\theta}' = f(\theta, \dot{\theta}, \tau) \qquad (1)$$

where the parameters $\theta, \dot{\theta}, \tau$ are vector elements of joint angles, velocities and torques respectively. Each data point consists of three elements of angles, three elements of velocities and two elements of torques that, together, give the targeted angular acceleration $\ddot{\theta}'$ i.e. [θ_1, θ_2, θ_3,τ_1, τ_2, τ_3, $\ddot{\theta}'$]. $\ddot{\theta}'$ is the angular acceleration consisting of $\theta', \dot{\theta}', \tau'$,which are the modified $\theta, \dot{\theta}, \tau$ terms. These are modified by adding zero mean Gaussian noise with standard deviation γ resulting in noisy inputs. This dataset thus has three levels of uncertainty; the noisy input, uncertainty of not observing some inputs (needed in the dataset) and the noisy output [13]. This data are used in predicting the angular acceleration of one of the links of the robotic arm. During the data capturing phase of this type of data, the angular acceleration might be measured along with the other variables in the dataset. If, however, the sensor used for capturing the angular acceleration malfunctions, the data cannot be recaptured. In this situation the imputation of the angular acceleration will be beneficial, especially in real time control situations. The ANFIS architecture is used to impute the angular acceleration assuming this real time control problem (of a malfunctioned sensor). These data were partitioned into 5460 training records and 2732 testing records to be used by the neuro-fuzzy system.

Letter image recognition dataset

This database is used to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The alphabet images were taken form 20 different fonts and each letter was randomly distorted to produce a dataset of 20 000 stimuli. Each stimulus was converted into 16 numerical attributes that include edge counts and statistical moments. These numerical attributes are scaled into integers ranging between 0 and 15 and these are shown in table 1 [14].

Table 1: Letter Image recognition data attributes [14]

Attribute number	Attribute	Description	Value type
1.	lettr	capital letter	(26 values from A to Z)
2.	x-box	horizontal position of box	(integer)
3.	y-box	vertical position of box	(integer)
4.	width	width of box	(integer)
5.	high	height of box	(integer)
6.	onpix	total # on pixels	(integer)
7.	x-bar	mean x of on pixels in box	(integer)
8.	y-bar	mean y of on pixels in box	(integer)
9.	x2bar	mean x variance	(integer)
10.	y2bar	mean y variance	(integer)
11.	xybar	mean x y correlation	(integer)
12.	x2ybr	mean of x * x * y	(integer)
13.	xy2br	mean of x * y * y	(integer)
14.	x-ege	mean edge count left to right	(integer)
15.	xegvy	correlation of x-ege with y	(integer)
16.	y-ege	mean edge count bottom to top	(integer)

17.	yegvx	correlation of y-ege with x	(integer)

If the width of the box containing each black-and-white pixel that represents each letter is unavailable or is invisible, then the width is a missing attribute that might distort the classification results if not imputed. However, assuming that the sensors used to measure the other attributes are available, this study desires to impute the box width using the neuro-fuzzy architecture. Out of the 20000 available records, 15000 were used for the inference training and the remaining 5000 were used for testing. Given the constraints of the dataset, no outliers were observed and the data was scaled in the range [-1 1] prior to the implementation of the neuro-fuzzy system. Since the ANFIS architecture only uses numerical data, the letter attribute is changed into numerical integer attributes ranging from 1 to 26 with 1 representing A and 26 representing Z.

3.2 Structural selection

The selection of the optimum number of antecedent fuzzy rules to estimate the missing attributes in each dataset is presented in this section. It should be noted that for each dataset, the training was initially performed with the least number of optimum rules in the architecture to minimise computational cost and also to avoid over fitting [11], [12]. This is because evaluating the best structural ability of the input-output model is, as figure 1 shows, computationally expensive as a result of the parametric tuning (inner loop in figure 1) that needs to be evaluated for each number of rules (outer loop in figure 1). Structural tuning aims to find the best number of rules and results in a graph depicting the computational accuracy against the number of rules shown in figure 2. The initialization of the architecture is performed using hyper-ellipsoid clustering and the axes of the ellipsoids are used in initialising the consequent parameters. The cluster centres are projected into the input domain such that the antecedent membership functions are also initialized [11], [12].

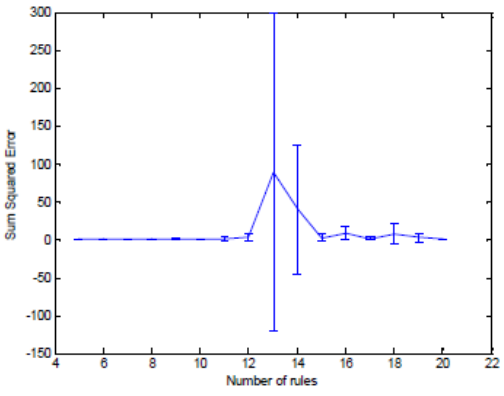

Figure 2: Cross validation vs. Complexity.

This figure is the cross-validation vs. complexity graph evaluated using the HIV database and it is clear that the least error is achieved with minimal number of rules (between 5 and 10). The same pattern is observed when evaluating the structural complexity of the Pumadyn and the letter image recognition data. Therefore, to minimise the computational cost and the effects of over fitting, the number of rules used for each database is chosen to be equal to the number of inputs for that database (5, 8 and 16 for the HIV, Pumadyn and letter recognition datasets respectively). In each case, the ANFIS is trained with triangular antecedent membership functions and a weighted architecture with no output bias [11], [12].

4. Performance analysis

The datasets mentioned are each used to evaluate the ability of the ANFIS architecture to estimate the missing data attributes. This section presents and discusses the results obtained in the evaluation of the inference.

HIV database

The results obtained in estimating the age of the father in the HIV database given the other data attributes are shown in figure 3 for the first 20 testing data points. The accuracy of this system is measured within a 10 % difference between the actual and predicted outputs i.e.

$$x_{predicted} \leq x_{actual} \pm 10\% \qquad (2)$$

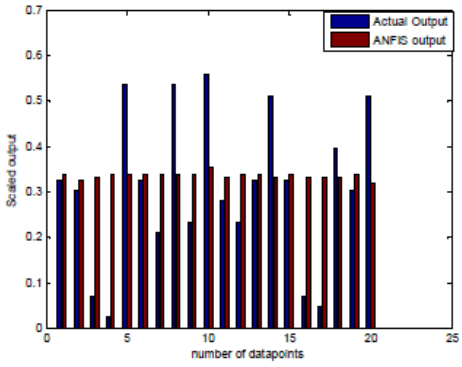

Figure 3: ANFIS and actual outputs for the HIV dataset

This error analysis method is used because, for the age attribute, imputing a value that is within a certain error would be considered acceptable because people are usually classified within an age range instead of using their specific ages (e.g. anyone between the ages of 13 and 19 years is considered a teenager etc.). A total of 1572 records were estimated within 10 % of the target output. This gives a 60 % accuracy of estimating the age of the father within a 10 % margin. This means that using

the ANFIS system, there is a 60 % chance of imputing the father's age within 10 %.

Pumadyn database

The accuracy of the ANFIS architecture for this dataset is tested using the standard error, which is the deviation between the target and predicted output for each dataset respectively [15]. For given data $x_1,x_2,...,x_n$ and corresponding approximated values x_1', x_2'...., x_n' the standard error (SE) is computed as:

$$SE = \sqrt{\frac{\sum_{i=1}^{n} x_i - x_i'}{n}} \qquad (3)$$

A low value of the SE indicates good accuracy for the model and vice versa. This error measure is used because of the continuous nature of the data variable to be imputed. The angular acceleration must be accurate for the robotic arm because of its possible real time control application and the standard error is used because it gives the best unbiased value of the deviation of the results. The results obtained in estimating the angular acceleration in the Pumadyn database given the other data attributes are shown in figure 4 for the first 50 testing data points.

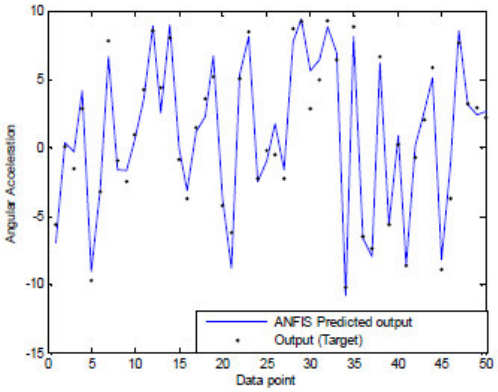

Figure 4: ANFIS and actual outputs for the Pumadyn dataset

The standard error, calculated from equation (3), for this approximation is 1.5777×10^{-5}. This error and figure 4 indicate that the ANFIS architecture employed for this estimation has satisfactory accuracy because not only is the SE low, but the results correlate quite well.

Letter image recognition database

The accuracy of the ANFIS architecture for this dataset is tested using equation 4. The standard error (3) and the correlation coefficient are used to measure the accuracy of the ANFIS architecture in estimating the letter image box width. The correlation coefficient is a measure of the linear relationship strength between two variables [15], namely, the ANFIS width and the original width. A correlation coefficient of 0 indicates no correlation and that of 1 indicates equivalent variables. The imputation results are depicted in figure 4 and indicate a good correlation between the letter box width and the values imputed by the ANFIS architecture.

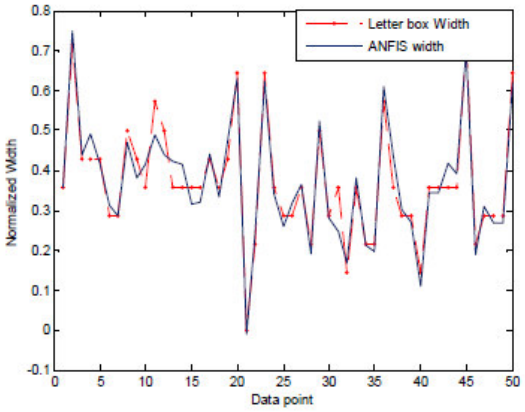

Figure 5: ANFIS and actual outputs for the width of the letter recognition dataset.

The standard error, calculated from equation (3), for this approximation is 0.0403 and the correlation coefficient is 0.9636, a value very close to 1. These results indicate that the ANFIS employed for this estimation has satisfactory accuracy because, not only is the SE low, but the results correlate quite well.

5. Conclusion

The neuro-fuzzy architecture has a strong ability of imputing continuous data such as the angular acceleration of the Puma robotic arm, which gave an almost zero error and the letter box width which gave a correlation coefficient close to 1. Estimating data from demographic attributes proved to be harder for the ANFIS architecture with only a 60 % chance of estimating the father's age within 10 %. This is due to the non-continuous nature of the data attribute that was estimated. It is recommended that this architecture be improved by using hybrid methods in order to give it the ability to impute any type of missing data within an accuracy measure of 75 % or more.

References

[1] G. Ssali & T. Marwala. Estimation of missing data using computational intelligence and decision trees. arXiv 0709.1640, 2007.

[2] F. Nelwamondo, S. Mohamed & T. Marwala. Missing data: A comparison of neural networks and Expectation Maximisation techniques". *Current Science, 93* (2), 1514-152.

[3] V. Tresp, R. Neuneir & S. Ahmad. Efficient methods of dealing with missing data in supervised learning". *Advances in Neural Information Processing Systems, 7* MIT Press.

[4] R. Little , D. Rubin. *Statistical analysis with missing data*. (New York: John Wiley, 1987).

[5] J. Schafer & M. Olsen. Multiple imputation for multivariate missing data problems: A data analyst perspective. *Multivariate Behavioural Research, 33*. 1997, 545-571.

[6] M. Abdella & T. Marwala. The use of genetic algorithms and neural networks to approximate missing data in database. *Computing and Informatics, .24*. 2005, 577-589.

[7] F. Nelwamondo & T. Marwala. Rough sets theory for the treatment of incomplete data. *IEEE International Conference on Fuzzy Systems*, 2007, 1-6.

[8] D. Rubin. Multiple imputations in sample surveys-A phenomenological Bayesian approach to nonresponse. *Proc. of the Survey Research Methods Section of the American Statistical Association*. 1978, 20-34.

[9] P. McKnight, K. McKnight, S. Sidani & A. Figuredo. *Missing data: A gentle introduction* (New York: Guilford Press, 2007).

[10] J-S. R. Jang, C-T. Sun & E. Mizutan. *Neuro-Fuzzy and soft computing; A computational approach to learning and machine intelligence* (Upper Saddle River, NJ: Prentice-Hall, 1997).

[11] G. Bontempi & H.Bersini. Now comes the time to defuzzify fuzzy models. *Fuzzy Sets and Systems,90*(2), 1997, 161-170.

[12] G. Bontempi, H. Bersini & M. Birattari. The local paradigm for modeling and control: From neuro-fuzzy to lazy learning. *Fuzzy Sets and Systems, 121*(1). 2001, 59-72.

[13] P. I. Corke. A robotics toolbox for MATLAB. *IEEE Robotics and Automation Magazine, 3* (1), 1996, 24-32.

[14] P. W. Frey & D. J. Slate. Letter recognition using Holland-style adaptive classifiers. *Machine Learning*, *6*(2), 1991, 161-182.

[15] N. Draper, H. Smith, *Applied regression analysis* (New York: J Wiley, 3rd edition, 1998).

Imputation of Missing Data using PCA, Neuro-Fuzzy and Genetic Algorithms

Nthabiseng Hlalele[1], Fulufhelo Nelwamondo[2], Tshilidzi. Marwala[1]

[1]School of Electrical and Information Engineering, University of the Witwatersrand
Private Bag 3, Wits, 2050.
2Graduate School of Arts Sciences, Harvard University, Hollyoke Center 350, Cambridge, Massachusetts, 023138
[1]nthabiseng.hlalele@students.wits.ac.za, [1]t.marwala@ee.wits.ac.za
[2]nelwamon@fas.harvard.edu

Abstract. This paper presents a method of imputing missing data that combines principal component analysis and neuro-fuzzy (PCA-NF) modeling in conjunction with genetic algorithms (GA). The ability of the model to impute missing data is tested using the South African HIV sero-prevalence dataset. The results indicate an average increase in accuracy from 60 % when using the neuro-fuzzy model independently to 99 % when the proposed model is used.

Keywords: Neuro-Fuzzy (NF), Principal Component Analysis (PCA), Genetic Algorithms (GA), Missing Data

1 Introduction

The missing data problem is a widely researched topic that has a huge impact on any field that requires the analysis of data in order to make a decision or reach a specific goal [1-7]. A number of methods have been investigated and implemented in order to deal with this problem, especially in large databases that require computational analysis [1], [2], [3]. Traditionally, ad hoc methods have been used when dealing with missing data; these include mean substitution and the deletion of all data entries that contain missing variables. Although easy to implement, these methods often lead to loss of data resulting in a more biased database. This has led to the development of more advanced regression techniques and likelihood based approaches such as expectation maximization (EM). Auto-associative neural networks (AANN) in conjunction with genetic algorithms have been employed and modified to improve the accuracy of computational methods in imputing missing data [1, 6]. This paper adds to this knowledge by employing principal component analysis, neuro-fuzzy modeling and genetic algorithms to impute missing data in the HIV sero-prevalence dataset. The backgrounds of the missing data problem, neuro-fuzzy computing and PCA are presented. The PCA-NF-GA method along with its testing data and measures are then presented followed by the results and discussions.

2 Background

The background of the missing data problem and its mechanisms is presented. Neuro-fuzzy networks, PCA and genetic algorithms are also briefly discussed.

2.1 Missing Data

Missing data estimation, like any other data analysis method, depends on the knowledge of how the data are missing. Three mechanisms of missing data have been documented [1-2], [4], [8-9] which include missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) also called the non-ignorable case. MCAR occurs when the probability of the missing data variable is independent of the variable value itself or on any other values in the database. MAR occurs when the probability of the missing data variable is dependent on other variables in the database but not on the value of the variable itself. This means that there exists some complex relationship between the observed and missing data; i.e. the observed data can be used to approximate the missing data. When data are MNAR, the probability of the missing data variable is related to the missing data, this means that missing data variables cannot be predicted from the observed database. Dealing with this type of missing data is difficult and may involve imputing the missing variables based on external data not within the database [9]. Because it is often difficult to determine what mechanism brought about the missing data, artificial intelligence methods have been investigated to solve the missing data problem irrespective of its missing mechanism [2, 7].

2.2 Neuro-Fuzzy Computing and Genetic Algorithm

The neuro-fuzzy architecture integrates the use of neural networks, which identify interrelationships and patterns in numerical datasets, and fuzzy systems, that incorporate expert knowledge and perform decision making [10]. This results in an inference system (as a result of fuzzy rules) that has the ability to learn and adapt through its environment (as a result of neural networks). A conventional fuzzy system uses expert knowledge to produce a linguistic rule base and reasoning mechanism for decision making. If artificial neural networks, together with an optimization technique, are incorporated into the fuzzy model to automatically tune the fuzzy parameters (antecedent membership functions and parametric consequent models), then the product is a neuro-fuzzy inference system [10-12]. In this paper, the Takagi-Sugeno fuzzy inference system is used because of its ability to generate fuzzy rules from an input-output dataset thus encapsulating expert knowledge that is otherwise lost when using traditional neural networks or autoencoders. Genetic algorithms (GA) are inspired by the theory of evolution involving genetic processes such as mutation, selection and cross over [13]. The Genetic Algorithms Optimization Toolbox (GAOT) is used to optimize the value of an evaluation function [14]. When AANN, which recall the input, are used for missing data imputation, the GA attempts to minimize the error between the output and the input data.

2.3 Principal Component Analysis

Principal Component Analysis is a statistical technique used for data dimension reduction and pattern identification in high dimensional data [15]. The PCA orthogonilizes the components of the input vectors to eliminate redundancy in the input data thereby exploring correlations between samples or records. It then orders the resulting components such that the components with the largest variation come first. The compressed data (mapped into i dimensions) is presented by:

$$Y_{j\times i} = X_{j\times k} \times PCvector_{k\times i}\,. \qquad (1)$$

where the principal component vector, $PCvector$ is presented by the eigenvectors of the i largest eigenvalues of the covariance matrix of the input $X_{j\times k}$ with k dimensions and j set of records $(i \le k)$.

3 Proposed Method

The method used to impute the missing data is presented. First the dataset as well as the data preprocessing are presented followed by the method used to impute the missing data.

3.1 HIV Sero-Prevalence Data

This dataset was obtained from the South African antenatal sero-prevalence survey of 2001 [16]. The data for this survey were collected from questionnaires answered by pregnant women visiting selected public clinics in South Africa and only women undertaking in the study for the first time were allowed to participate. This dataset has been used to investigate the effect that demographic information has on the HIV risk of an individual [17]. This is especially helpful in countries such as South Africa, which have a high HIV infection rate. The data attributes used in this study are the HIV status, Education level, Gravidity, Parity, the Age of the Mother (pregnant woman) and the Age of the Father (responsible for the most recent pregnancy). The HIV status is represented in binary form, where 0 and 1 represent negative and positive respectively. The education level indicates the highest grade successfully completed and ranges between 0 and 13 with 13 representing tertiary education. Gravidity is the number of pregnancies, successful or not, experienced by a female, and is represented by an integer

between 0 and 11. Parity is the number of times the individual has given birth and multiple births (e.g. twin births) are considered as one birth event. It is observed from the dataset that the attributes with the most missing values are the age of the father (3972 missing values), the age of the mother (151 missing values) and the education level (3677 missing values) of the pregnant woman. Imputing these data variables is helpful in educating people about HIV and the factors that render some individuals more risky than others. In situations where an attribute is missing in the questionnaire, it is almost impossible to retrieve this information from the woman who supplied it due to the anonymity of the study. It is for this reason that missing data imputation methods are employed.

3.2 Data Preprocessing

When fitting a model in order to solve a problem, it is necessary to prepare the data such that the essence of the data is captured by the proposed model. First the data entries are normalized within the range [0 1] in order to implement the neuro-fuzzy model and, secondly, the data entries that contain logical errors are removed. The data is then evaluated in order to see which attributes contribute the most outliers as shown in figure 1.

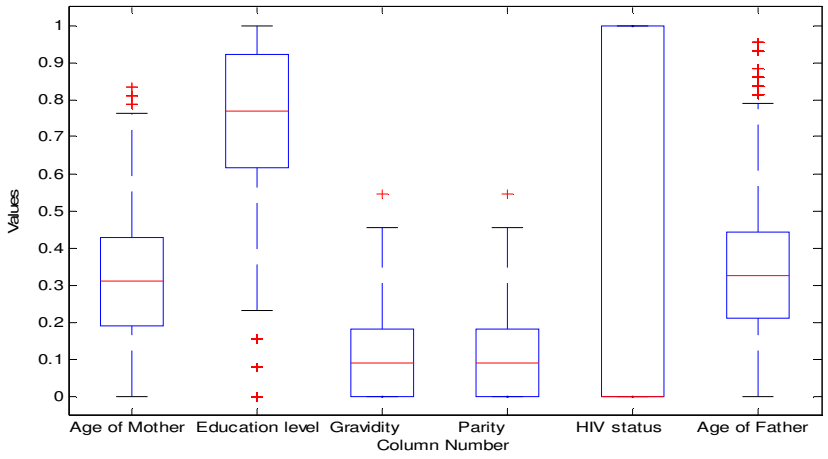

Fig. 1. HIV dataset outliers per attribute.

It is important to remove outliers because they often represent misplaced data points which result in longer training times and models that perform poorly. The crosses in the figure represent the outliers that are present as a result of each attribute, it is clear that the attributes that are more likely to be missing in the dataset i.e. the age of the mother, the education level

and the age of the father, also produce the most outliers. The data is then arbitrarily partitioned into 9745 datasets to train the model and 2462 testing datasets.

3.3 Proposed Method and Simulation

During training, PCA is employed to orthogonalize the data ensuring that the model is better trained. Performing PCA on the input data results in orthogonal data that has variance percentages as illustrated by figure 2.

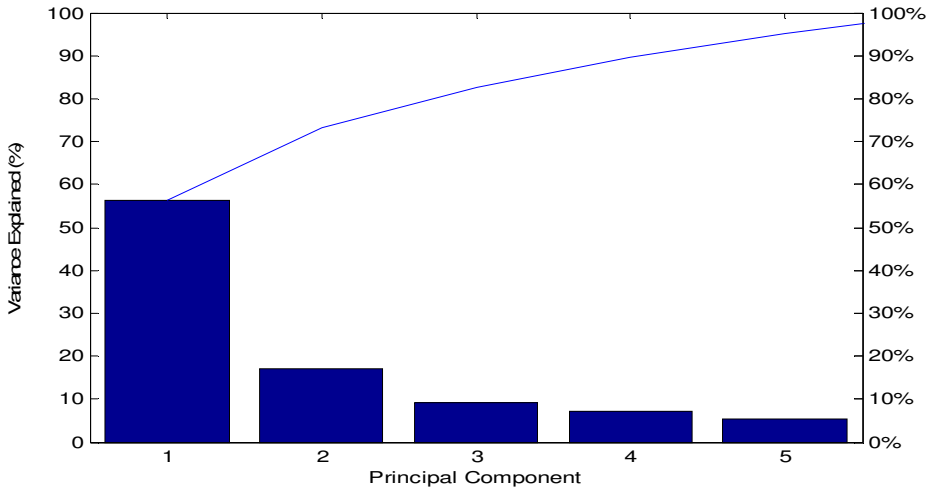

Fig. 2. Variance of the training input data principal components.

It is clear that a greater percentage (75 %) of the variance in the principal components can be attributed to the first two principal components illustrating the orthogonality of the training data. The PCA compressed data (using equation (1)) is then used during training to model all three of the data attributes that are likely to be missing. Figure 3 represents the proposed missing data imputation model.

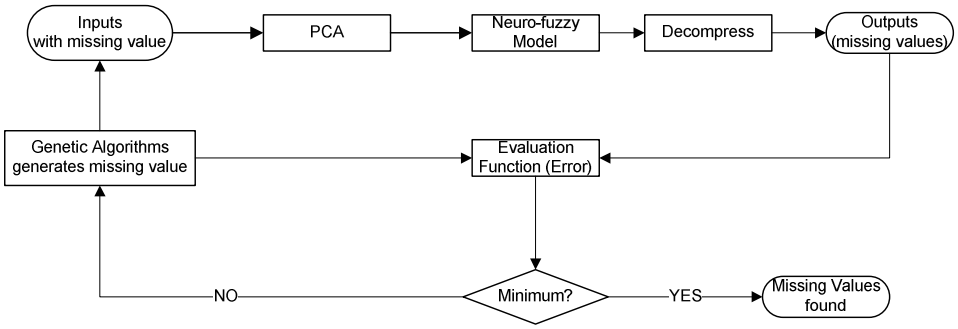

Fig. 3. Flowchart of the proposed model for imputing missing data

When an input matrix that contains a missing variable is fed into the proposed model, it is compressed and a neuro-fuzzy model (that has already been trained) is used to impute the variable. The sum squared error between the data that contains the imputed variable and the neuro-fuzzy output is then evaluated and minimized using the GA.

4 Results and Discussions

The results found when using the model are then discussed.

4.1 Results

A test sample is first evaluated using the neuro-fuzzy model on its own to impute the father's age without compressing the data. The father's age is chosen as the test experiment because it the field with the most missing values. The results of this test experiment are shown in Figure 4. These results indicate that the neuro-fuzzy model is unable to impute the missing data with great accuracy. The age of the father can only be imputed with an accuracy of 60 % within a 10 % margin. At first glance these results might seem satisfactory due to the fact that people are usually classified within a certain age group (e.g. anyone from the ages of 13 and 19 years is considered a teenager etc), however, the measure used indicates poor performance since an older person is given a larger margin of error than a younger person (10 % of 50 is 5 whereas 10 % of 16 is 1.6).

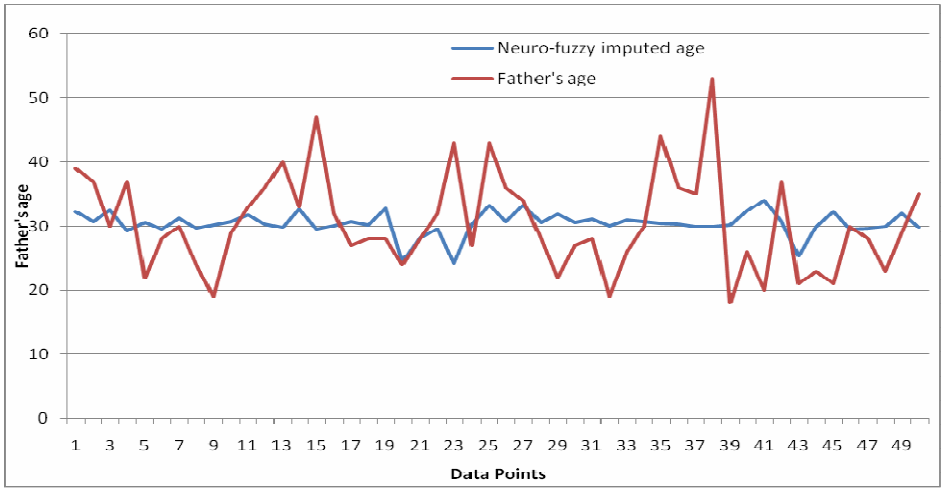

Fig. 4. Neuro-fuzzy imputation of the father's age

By employing a hybrid method such as the one proposed here, the accuracy of the imputation is expected to increase (because hybrid systems attempt to capture complex natural intelligence). Following the test experiment, the proposed model is then implemented to impute the missing data yielding the results shown in figure 5 for the age of the father. There is obvious correlation between the imputed and actual age of the father.

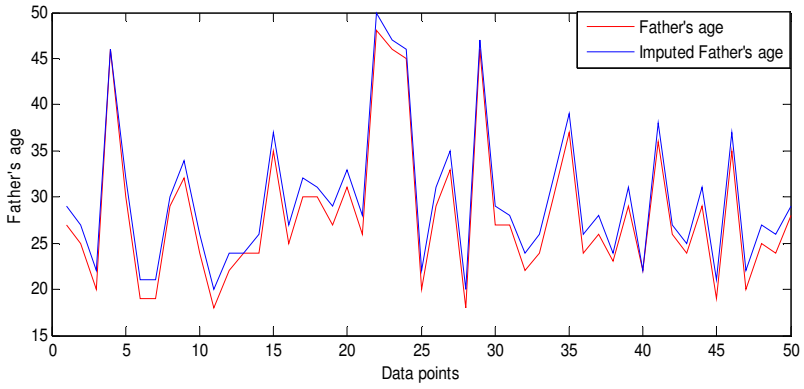

Fig. 5. Proposed model's imputation of the father's age.

The imputation results of the age of the mother also correlate quite well as indicated by table 1. The accuracy of the mother or father's age is measured within 1 and 2 years. The imputed education level of the mother has no correlation at all with the actual level indicating a low accuracy value measured

within 1, 2 and 5 grades. This, for example, means that the system has 98.99 % accuracy in estimating a woman's age within ± 1 year.

Table 8. Percentage of data that are correctly imputed.

Attribute	Exact Accuracy (within 0)	Accuracy within 1	Accuracy within 2	Accuracy within 5
Mother's age	43.638 %	98.99 %	100 %	100 %
Father's age	6 %	37.67 %	99.9 %	100 %
Education level	2 %	9 %	13 %	26.7 %

4.2 Discussion and Conclusion

From figure 1, it is deducible that the attributes with the most outliers are also the attributes that are likely to be missing. This is because both the missing data problem and the problem of outliers contain extreme values that provide erroneous information and modelling. This type of information is useful when building models that impute missing data. When the results in Figure 4 are compared with that of Figure 5 (imputation of the age of the father), it is clear that using the hybrid method provides better accuracy in imputing the age. When the ability of the system to impute the age of the mother is compared to Figure 1, it is deducible that the less varied the outliers of a variable (such as the case with the age of the mother), the higher the imputation accuracy of the model for that variable. The inverse can thus be the reason for the low accuracy of the system to impute the education level. This suggests that other methods be looked into for imputing variables that have such a varied outlier model as the education level of this database.

References

1. Ssali, G., Marwala, T.: Estimation of Missing Data using Computational Intelligence and Decision Trees. In: arXiv 0709.1640 (2007)
2. Nelwamondo, F., Mohamed, S., Marwala, T.: Missing Data: A comparison of Neural Networks and Expectation Maximisation Techniques. In: Current Science, vol. 93, pp. 1514-1521 (2007)
3. Tresp, V., Neuneir, R., Ahmad, S.: Efficient Methods for Dealing with Missing Data in Supervised Learning. In: Tesauro, G., Touretzky, D. S., Leen, T. K. (eds.) Advances in Neural Information Processing Systems 7. MIT Press, Campbridge (1995)
4 Little, R., Rubin, D.: Statistical Analysis with Missing Data. John Wiley, New York (1987).
5. Schafer, J., Olsen, M.: Multiple imputation for Multivariate Missing Data Problems: A Data Analyst Perspective. In: Multivariate Behavioural Research. 33, 545-571 (1997)
6. Abdella, M., Marwala, T.: The use of Genetic Algorithms and Neural Networks to Approximate Missing Data in Database. Computing and Informatics.24, 577-589 (2005)

7. Nelwamondo, F., Marwala, T.: Rough Sets Theory for the Treatment of Incomplete Data. Proceedings of the IEEE International Conference on Fuzzy Systems. 338-343 (2007)
8. Rubin, D.: Multiple Imputations in Sample Surveys-A Phenomenological Bayesian Approach to Nonresponse. Proceedings of the Survey Research Methods Section of the American Statistical Association. 20-34 (1978)
9 McKnight, P., McKnight, K., Sidani, S., Figuredo A.: Missing data: A gentle Introduction. Guilford Press, New York (2007)
10. Jang, J-S. R., Sun, C-T., Mizutan, E.: Neuro-Fuzzy and Soft Computing; A Computational Approach to Learning and Machine Intelligence. Prentice-Hall, Upper Saddle River, NJ, (1997)
11. Bontempi, G., Bersini, H.: Now Comes the Time to Defuzzify Fuzzy Models. In: Fuzzy Sets and Systems, vol. 90 (2), pp. 161-170 (1997)
12. Bontempi, G., Bersini, H., Birattari, M.: The Local Paradigm for Modeling and Control: From Neuro-Fuzzy to Lazy Learning. In: Fuzzy Sets and Systems, vol. 121(1), pp.59-72 (2001)
13. Goldberg, D.: Genetic Algorithms in Search Optimization and Machine Learning. Addison-Wesley, Boston, MA, USA (1989)
14. Houck, C., Joines, J., Kay, M.: A Genetic Algorithm for Function Optimization: A Matlab Implementation. North Carolina State University, Raleigh, NC, Tech. Rep. NCSUIE-TR-95-09, (1995)
15. Jollife, I. T.: Principal Component Analysis. Springer-Verlag, New York, USA (1986)
16. South African Department of Health: HIV and Syphilis Sero-Prevalence Survey of Women Attending Public Antenatal Clinics in South Africa, http://www.info.gov.za/view/DownloadFileAction?id=70247.
17. Betechuoh, B.L, Marwala, T., Manana, J.V.: Computational Intelligence for HIV Modeling. Proceedings of the International Conference on Intelligent Engineering Systems, 127-132 (2008)