Preface

In real life, a set of data invariably contains missing data. The problem then is to reconstitute the most probable values through processes such as interpolation and extrapolation before using that set.

Methods for resolving the problem of missing data have been extensively explored in statistical texts (Abdella, 2005; Little & Rubin, 1987). The initial work on compensating for missing data was focused on improving survey data. In this book, missing data interpolation is called *imputation* to distinguish it from the statistical approach. Imputation is viewed as an alternative approach to deal with missing data. There are two ways to deal with missing data: these are either to estimate the missing data or to delete any vector (data set) with missing value(s). This book focuses on methods that estimate the missing values.

Of particular importance to the area of missing data interpolation is to analyze the nature of the missing data, and this is termed the *missing data mechanism*. Little and Rubin (1987) categorized three missing data mechanisms, namely: Missing At Random (MAR), Missing Completely At Random (MCAR) and a non-ignorable case also known as Missing Not At Random (MNAR).

In the first case, MAR occurs when the probability that variable X is missing depends on other variables, but not on X itself. An example of this is the case where two variables: the vibration level of a machine and its temperature, X are measured. If a very high vibration level causes the temperature sensor to fall off and thus high and subsequently low values of X become missing because of the other variable *vibration level,* this is termed MAR.

MCAR occurs when the probability that variable X is missing is unrelated to the value of X itself or to any other variable in the data set. This refers to data sets where the absence of data does not depend on the variable of interest or of any other variable in the data set (Rubin, 1978).

MNAR occurs when the probability of variable X missing is related to the value of X itself even if the other variables are controlled in the analysis (Allison, 2000). An example of this is when in a survey of weights of candidates, a person omits mentioning his or her weight because its value is very high. In analyzing survey data, these mechanisms are very powerful and useful. Knowing these mechanisms assists one in choosing which missing data imputation method is best to use.

However, in many engineering problems, where on-line decision support tools are becoming widely used, these mechanisms are proving to be insignificant (Marwala & Hunt, 1999). For example, if an aircraft is flying over the Atlantic Ocean and one of its critical sensors fails, there is simply no time to investigate why that particular sensor has failed and, thereby, indentify its missing value mechanism. What ought to be done in this situation is to quickly estimate the sensor's value, so that an on-line auto-pilot system can continue to operate.

In using decision support tools, if data become missing, it is extremely important, particularly for critical applications, that the missing data estimation technique is accurate. The methods introduced in this book are *computational intelligence methods* and have proven to be very successful in modeling complex problems such as speech recognition (Nelwamondo, Mahola, & Marwala, 2006). In this book, many methods are considered. These include:

- The multi-layer perceptron model (Marwala, 2000),
- Radial basis functions (Bishop, 1995),
- Gaussian mixture models (Chen, Chen, & Hou, 2004),
- Rough sets (Wu, Mi, & Zhang, 2003),
- Support vector machines (Drezet & Harrison, 2001),
- Decision trees (Ssali, & Marwala, 2008),
- Fuzzy ARTMAP (Carpenter et al*.,* 1992) and extension neural networks (Mohamed, Tettey, & Marwala, 2006).

Descriptions and implementations for using these missing data estimation process follow (Bishop, 1995; Marwala, 2007). These methods are implemented in both the Bayesian and maximum-likelihood framework (Marwala, 2001).

It is still very difficult to know beforehand which of these computational intelligence methods are ideal for missing data imputation. For this reason, hybrid methods are also introduced and implemented in this book for missing data imputation. In particular, the ensemble methods that use more than one learning algorithm are considered (Perrone & Cooper, 1993). Some of these methods are computationally intensive, and as a result, the book introduces methods that are computationally efficient, such as the principal component analysis (Adams et al*.,* 2002) and dynamic programming method (Bellman, 1957; Bertsekas, 2000).

In this book, many optimization methods are used. For example, to train multi-layer perceptrons, a scaled conjugate gradient optimization method (Møller, 1993) is used. Other optimization methods used are:

- The expectation maximization algorithm (Dempster, Laird, & Rubin, 1977)
- Genetic algorithms (Goldberg, 1989)
- Particle swarm optimization (Poli, Langdon, & Holland, 2005)
- Hill climbing (Tanaka, Toumiya, & Suzuki, 1997)
- Simulated annealing (Tavakkoli-Moghaddam, Safaei, & Gholipour, 2006)

It is difficult to know in advance which optimization method to use for missing data estimation process and, therefore, this book also explores various hybrid optimization techniques. Some of the hybrid optimization techniques that are considered in this book include the hybrid of genetic algorithms and particle swarm optimization.

Traditional missing data imputation methods have been largely based on static models. Even computational intelligence methods are traditionally constructed in a static manner. These methods are static in the sense that they are the same over time. For situations where the concepts are drifting and, therefore, the data are non-stationary, these methods fail (Kubat, & Widmer, 1996). Many engineering problems have to model systems that are continuously changing because of aging. Therefore, for many engineering problems, data imputation methods are required that are immune or at best can handle these changes in the character of the systems. This problem, therefore, requires missing data models that evolve with the systems on which they are based. Evolutionary methods have been successful in designing learning machines that evolve with systems. These evolutionary methods include genetic algorithms (Goldberg, 2002), fuzzy maps (Carpenter et al*.,* 1992), particle swarm optimization (Kennedy & Eberhart, 1995) and are described in detail in this book.

Throughout this book, examples from the literature and case studies are used to illustrate the effectiveness of the presented missing data estimation methods. Some of the case studies used include the artificial taster, HIV and a mechanical system.

SUMMARY OF THE BOOK

In Chapter I, traditional missing data issues, such as missing data patterns and mechanisms, are described. Attention is paid to the best models to deal with particular missing data mechanisms. A review of traditional missing data imputation methods is conducted, and the methods reviewed include case deletion and prediction rules (Acork, 2005). The *case deletion* methods reviewed are list-wise and pair-wise deletion. The *prediction rule* imputation techniques reviewed are mean substitution, hot-deck, regression and decision trees. Two missing data examples are studied, namely, the Sudoku puzzle and a mechanical system.

Missing data estimation processes require mathematical models that capture interrelationships amongst the variables. In Chapter II, a method is presented that is aimed at approximating missing data and, thereby, capturing variables' interrelationships by combining genetic algorithms and autoassociative neural networks. The neural network architectures implemented are the multi-layer perceptron and the radial basis function neural networks (Russell & Norvig, 1995). The proposed procedures are tested and then compared for missing data imputation.

The ability to identify a model which captures the interrelationships between the variables is very important. Different models bring unique perspectives to the missing data problem and one way to maximize the performance of the missing data procedure is to hybridize different methods. In Chapter III, hybrid autoassociative neural networks models are developed and used in conjunction with genetic algorithms (Goldberg, 2002) to estimate missing data. One hybrid technique combines three neural networks to form a hybrid autoassociative network, while the other merges principal component analysis and neural networks. These procedures are compared to the Bayesian auto-associative neural network (Bishop, 1995) and the genetic algorithm approach.

In Chapter IV, two techniques, i.e., Gaussian mixture models trained using the Expectation Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) and the combined auto-associative neural networks and particle swarm optimization methods are implemented for missing data estimation and then compared. Of a particular interest is the nature of the data in the analysis that suits each of these methods.

Chapter V investigates an imputation technique based on rough sets computation (Wu, Mi, & Zhang, 2003). The characteristic relations are introduced to describe incompletely specified decision tables and then used for missing data estimation. Empirical results obtained using real data are given and insights into the problem of missing data are derived.

In Chapter VI, autoassociative neural networks, principal components analysis and support vector regression (Marivate, Nelwamondo, & Marwala, 2008) are all combined with genetic algorithms, and then used to impute missing variables. The impact of using the principal component analysis on the overall performance of the autoassociative network and support vector regression is then assessed.

In Chapter VII, a committee of networks is introduced for missing data estimation. This committee of networks consists of a multi-layer perceptron, support vector machines and radial basis functions. It is constructed through a weighted combination of the three networks. The networks committee is implemented collectively with a hybrid of the genetic algorithm and the particle-swarm optimization method for missing data estimation, and is then tested and assessed. Furthermore, evolutionary methods are used to evolve a committee of networks. The results of this committee are compared to the results from a traditional committee and stand-alone networks.

The use of inferential sensors is common in on-line fault detection systems in various control applications. A problem arises when sensors fail while the system is designed to make a decision based on the data from those sensors. Various techniques to handle missing data are discussed in Chapter VIII. First, a novel algorithm that classifies and regresses in the presence of missing data is proposed. The algorithm is tested for both classification and regression problems. Second, an estimation algorithm that uses an ensemble of regressors within the context of the boosting mechanism is proposed. Hybrid genetic algorithms and fast simulated annealing are used to predict missing values and the results are compared.

In Chapter IX, a classifier method is presented that is based on a missing data estimation framework, and which uses auto-associative multi-layer perceptron neural networks and genetic algorithms. The method is tested and compared to conventional feed-forward neural network using classification accuracies and the area under the receiver operating characteristics curve.

In Chapter X, various optimization methods are compared with the aim of optimizing the missing data estimation equation, which is made out of the autoassociative neural networks with missing values as design variables. These optimization techniques are the genetic algorithm, particle swarm optimization, hill climbing and simulated annealing. They are tested and the results obtained are compared.

In implementing solutions to the missing data estimation problem, using optimization techniques, the definition of variable bounds is of critical importance. Chapter XI introduces a novel paradigm to impute missing data that combines decision trees with an auto-associative neural network and principal component analysis. This is designed to answer the crucial question on whether the optimization bounds actually matter in the estimation of missing data. In the model, a decision tree is used to predict search bounds for a hybrid simulated annealing and genetic algorithm that minimizes an error function derived from the respective models. The results obtained are compared.

Chapter XII presents a control mechanism to assess the effect of a demographic variable, *education level*, on the HIV risk of individuals. This is intended to assist for understanding the extent to which the spread of HIV can be controlled by using the variable *education level*. This control mechanism is based on missing data frameworks where the missing data are the set points for control. An inverse neural network model and a missing data approximation model, based on an auto-associative neural network and the genetic algorithm, are used for the control mechanism and the results obtained are then compared.

In Chapter XIII, a computational intelligence approach to predicting missing data in the presence of concept drift is presented, using an ensemble of multi-layered feed-forward neural networks. An algorithm that detects concept drift is presented. Six instances prior to the occurrence of missing data are used to approximate the missing values. The algorithm is applied to simulated time-series data set resembling the non-stationary data from a sensor. Second, an algorithm that uses dynamic programming and neural networks to solve the problem of missing data imputation is presented, tested and the results are assessed. Third, the impact of missing data estimation on fault classification in mechanical systems is studied. The missing data estimation method is based on auto-associative neural networks where the network is trained to recall the input data through some non-linear neural network mapping using genetic algorithm. The classification methods used are extension neural networks and Gaussian mixture models.

TARGET AUDIENCE OF THIS BOOK

This book is intended for researchers and practitioners who use data analysis to build decision support systems. In particular the target audience includes engineers, scientists and statisticians. The areas of engineering where decision support tools are becoming widely used (the target audience of this book) are aerospace, mechanical, civil, biomedical and electrical engineering. Furthermore, researchers in statistics and social science will also find the techniques introduced in this book to be highly applicable to their work. This book is carefully written to give a good balance between theory and application of various missing data estimation techniques. The applications selected reflect the target audience of this book and include examples from various branches of engineering.

Tshilidzi Marwala

REFERENCES

Abdella, M. (2005). *The use of genetic algorithms and neural networks to approximate missing data in database.* Unpublished master's thesis, University of the Witwatersrand, Johannesburg.

Acork, A. C. (2005). Working with missing values. *Journal of Marriage and Family, 67,* 1012–1028.

Adams, E., Walczak, B., Vervaet, C., Risha, P. G., & Massart, D. L. (2002). Principal component analysis of dissolution data with missing elements. *International Journal of Pharmaceutics*, *234* (1-2), 169-178.

Allison, P. (2000). Multiple imputation for missing data: A cautionary tale. *Sociological Methods and Research, 28*, 301-309.

Bellman, R. (1957). *Dynamic programming*. Princeton, NJ: Princeton University Press.

Bertsekas, D. P., (2000). *Dynamic programming and optimal control*. New York, NY: Athena Scientific.

Bishop, C. M. (1995). *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press.

Carpenter, G. A., Grossberg, S., Markuzon, N., Reynolds, J. H., & Rosen, D. B. (1992). Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multidimensional maps. *IEEE Transactions on Neural Networks*, *3*, 698-713.

Chen, C. T. Chen, C., & Hou, C. (2004). Speaker identification using hybrid Karhunen–Loeve transform and Gaussian mixture model approach. *Pattern Recognition*, *37*(5)*,* 1073-1075.

Dempster, A. P, Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *B39*, 1-38.

Drezet, P. M. L., & Harrison, R. F. (2001). A new method for sparsity control in support vector classification and regression. *Pattern Recognition*, *34*(1)*,* 111-125.

Goldberg, D. E. (1989). *Genetic algorithms in search, optimization, and machine learning*. Reading, MA: Addison-Wesley.

Goldberg, D. E. (2002). *The design of innovation: Lessons from and for competent genetic algorithms*. Reading, MA: Addison-Wesley.

Kennedy, J. E, & Eberhart, R. C. (1995). Particle swarm optimization. In *Proceedings of the IEEE International Conference on Neural Networks*, (pp. 942-1948).

Kubat, M., & Widmer, G. (1996). Learning in the presence of concept drift and hidden contexts. *Machine Learning, 23,* 69–101.

Little, R., & Rubin, D. (1987). *Statistical analysis with missing data*. New York: John Wiley and Sons.

Marivate, V. N., Nelwamondo, V. F., & Marwala, T. (2008). Investigation into the use of autoencoder neural networks, principal component analysis and support vector regression in estimating missing HIV data. In *Proceedings of the 17th World Congress of the International Federation of Automatic Control* (pp. 682-689).

Marwala, T. (2000). On damage identification using a committee of neural networks. *American Society of Civil Engineers, Journal of Engineering Mechanics*, *126*, 43-50.

Marwala, T. (2001). Scaled conjugate gradient and Bayesian training of neural networks for fault identification in cylinders. *Computers and Structures 79/32*, 2793-2803.

Marwala, T. (2007). *Computational intelligence for modelling complex systems*. India, New Delhi: Research India Publications.

Marwala, T., & Hunt, H. E. M. (1999). Fault identification using finite element models and neural networks. *Mechanical Systems and Signal Processing, 13,* 475-490.

Mohamed, S., Tettey, T., & Marwala, T. (2006). An extension neural network and genetic algorithm for bearing fault classification. In *Proceedings of the IEEE International Joint Conference on Neural Networks* (pp. 7673-7679).

Møller, A. F. (1993). A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks*, *6*, 525-533.

Nelwamondo, F. V., Mahola, U., & Marwala, T. (2006). Multi-scale fractal dimension for speaker identification system. *Transactions on Systems 5*(5), 1152-1157.

Perrone, M. P., & Cooper, L. N. (1993). When networks disagree: Ensemble methods for hybrid neural networks. In R. J. Mammone (Ed.), *Artificial neural networks for speech and vision* (pp. 126-142). London: Chapman and Hall.

Poli, R., Langdon, W. B., & Holland, O. (2005). Extending particle swarm optimization via genetic programming. *Lecture Notes in Computer Science, 3447*, 291-300.

Rubin, D. B. (1978). Multiple imputations in sample surveys: A phenomenological Bayesian approach to non-response. In *Proceedings of the Survey Research Methods Section of the American Statistical Association* (pp. 20-34).

Russell, S., & Norvig, P. (1995). *Artificial intelligence: A modern approach*. Englewood Cliffs, NJ: Prentice Hall.

Ssali, G., & Marwala, T. (2008). Estimation of missing data using computational intelligence and decision trees. In *Proceedings of the IEEE International Joint Conference on Neural Networks* (pp. 201-207).

Tanaka, T., Toumiya, T., & Suzuki, T. (1997). Output control by hill-climbing method for a small scale wind power generating system. *Renewable Energy*, *12*(4), 387-400.

Tavakkoli-Moghaddam, R., Safaei, N., & Gholipour, Y. (2006). A hybrid simulated annealing for capacitated vehicle routing problems with the independent route length. *Applied Mathematics and Computation*, *176*(2), 445-454.

Wu, W., Mi, J., & Zhang, W. (2003). Generalized fuzzy rough sets. *Information Sciences*, *151*, 263-282.