Chapter V

Missing Data Estimation Using Rough Sets

ABSTRACT

A number of techniques for handling missing data have been presented and implemented. Most of these proposed techniques are unnecessarily complex and, therefore, difficult to use. This chapter investigates a hot-deck data imputation method, based on rough set computations. In this chapter, characteristic relations are introduced that describe incompletely specified decision tables and then these are used for missing data estimation. It has been shown that the basic rough set idea of lower and upper approximations for incompletely specified decision tables may be defined in a variety of different ways. Empirical results obtained using real data are given and they provide a valuable insight into the problem of missing data. Missing data are predicted with an accuracy of up to 99%.

INTRODUCTION

There are a number of general ways that have been used to approach the problem of missing data in databases (Little & Rubin, 1987; Rubin, 1976; Little & Rubin, 1989; Collins, Schafer Kam, 2001; Schafer & Graham, 2002). One of the simplest of these methods is the ‘list-wise deletion’, which simply deletes instances with missing values (Scheuren, 2005; King et al., 2001; Abdella, 2005). The major disadvantage of this method is the dramatic loss of information in data sets (King et al., 1988). Also, Enders and Peugh (2004) demonstrated that when there is a group of missing data, list-wise deletion gives biased parameters and standard errors. Another approach is ‘pair-wise deletion’ (Marsh, 1998).

Tsikritis (2005) observed that appropriately dealing with missing data has been underestimated by the operation management authors, unlike in other fields such as marketing, organizational behavior analysis, economics, statistics and psychometrics that have intensely attended to the issue. Tsikritis found from a review of 103 survey articles appearing in the Journal of Operations Management between 1993 and 2001 that list-wise deletion was the most widely used technique to deal with missing data. This is in spite of the fact that list-wise deletion is usually the least accurate method for handling missing
Kim and Curry (1997) found that when 2% of the features are missing and the complete observation is deleted then up to 18% of the total data may be lost. The second common technique imputes the data by finding estimates of the values and missing entries are then replaced with these estimates. Various estimates have been used and these estimates include zeros, means and other statistical calculations. These estimations are then used as if they were the observed values. Xia et al. (1999) estimated missing data in climatological time and investigated six methods for imputing missing climatological data including daily maximum temperature, minimum temperature, air temperature, water vapor pressure, wind speed and precipitation. These researchers used the multiple regression analysis with the five closest weather stations and the results obtained from the six methods showed similar estimates for the averaged precipitation amount.

Another common technique assumes some models for the prediction of the missing values and uses the maximum likelihood approach to estimate the missing values (Nelwamondo, Mohamed, & Marwala, 2007; Dempster, Laird, & Rubin., 1977; Abdella & Marwala, 2006). In Chapter IV, the hybrid auto-associative neural networks and simple genetic algorithms are compared to the Gaussian mixture models trained with the expectation maximization (GMM-EM) algorithm and tested using data sets from an industrial power plant, an industrial winding process and HIV sero-prevalence survey data. The results obtained show that both methods perform well. The GMM-EM method was found to perform well in cases where there was little or no inter-dependency between variables, whereas the hybrid auto-associative neural network and genetic algorithm was found to be suited to problems where there were some inherent nonlinear relationships between some of the given variables.

A great deal of research has been conducted to find new ways of approximating missing values. Among others, Abdella and Marwala (2006), Nelwamondo and Marwala (2007a), Ssali and Marwala (2007) as well as Mohamed and Marwala (2005) have used computational intelligence methods to approximate missing data. Qiao, Gao, and Harley (2005) have used neural networks and particle swarm optimization to keep track of the dynamics of the power plant in the presence of missing data. Nauck and Kruse (1999), Gabrys (2002) as well as Nelwamondo and Marwala (2007c) used fuzzy approaches to deal with missing data. A different approach was taken by Wang (2005) who replaced incomplete patterns with fuzzy patterns. The patterns without missing values were, along with fuzzy patterns, used to train the neural network. In this model, the neural network learned to classify without actually predicting the missing data and this approach is adopted in Chapter VIII. Similar work was also conducted by Arbuckle (1996)

Mitra, Pal, and Saddiqi (2003) described the incorporation of a minimal spanning tree, based on graph-theoretic technique, an expectation maximization algorithm as well as rough set initialization for non-convex clustering. The expectation maximization algorithm was found better able to handle uncertainties and the statistical model of the data. The rough set theory was found to assist in the increased speed of convergence and in avoiding local minima problem.

Hong, Tseng, and Wang (2002) solved the problem of effecting a set of certain and possible rules from incomplete data sets using rough set theory. They introduced a learning algorithm, which concurrently derived rules from incomplete data sets and estimated the missing values during the learning process. They first assumed unknown values as any possible values and then gradually refined these unknown values using the incomplete lower and upper approximations that were derived from the training examples. By so doing, the approximations and examples interacted with each other to give accurate rules.
Related Content

An Effective Hybrid Semi-Parametric Regression Strategy for Rainfall Forecasting Combining Linear and Nonlinear Regression
[www.igi-global.com/chapter/effective-hybrid-semi-parametric-regression/74935?camid=4v1a](www.igi-global.com/chapter/effective-hybrid-semi-parametric-regression/74935?camid=4v1a)

GBF Trained Neuro-Fuzzy Equalizer for Time Varying Channels
[www.igi-global.com/chapter/gbf-trained-neuro-fuzzy-equalizer/74929?camid=4v1a](www.igi-global.com/chapter/gbf-trained-neuro-fuzzy-equalizer/74929?camid=4v1a)

Relevant and Non-Redundant Amino Acid Sequence Selection for Protein Functional Site Identification
Chandra Das and Pradipta Maji (2012). *Breakthroughs in Software Science and Computational Intelligence* (pp. 139-164).
[www.igi-global.com/chapter/relevant-non-redundant-amino-acid/64607?camid=4v1a](www.igi-global.com/chapter/relevant-non-redundant-amino-acid/64607?camid=4v1a)

Learning Different Concept Hierarchies and the Relations Between them from Classified Data
[www.igi-global.com/chapter/learning-different-concept-hierarchies-relations/67441?camid=4v1a](www.igi-global.com/chapter/learning-different-concept-hierarchies-relations/67441?camid=4v1a)