Chapter X
Optimization Methods for Estimation of Missing Data

ABSTRACT

This chapter presents various optimization methods to optimize the missing data error equation, which is made out of the autoassociative neural networks with missing values as design variables. The four optimization techniques that are used are: genetic algorithm, particle swarm optimization, hill climbing and simulated annealing. These optimization methods are tested on two datasets, namely, the beer taster dataset and the fault identification dataset. The results that are obtained are then compared. For these datasets, the results indicate that genetic algorithm approach produced the highest accuracy when compared to simulated annealing and particle swarm optimization. However, the results of these four optimization methods are the same order of magnitude while hill climbing produces the lowest accuracy.

INTRODUCTION

Missing data problem creates a variety of problems in many disciplines which depend on good access to accurate data and because of this reason, techniques to deal with missing data have been the subject of research in statistics, mathematics and other disciplines (Yuan, 2000; Allison, 2000; Rubin, 1978). In the engineering domain the missing data problem occurs because of a number of reasons including sensors failures, inaccessibility of measuring surfaces and so on. If the data component from the missing data is needed for decision-making purpose then it becomes very important that the missing values be estimated.

A number of methods have been proposed to deal with missing data (Little & Rubin, 2002). These include the use of Expectation-Maximization methods (Dempster, Liard, & Rubin, 1977), autoassociative neural networks (Abdella & Marwala, 2006), rough sets approach (Nelwamondo & Marwala, 2007) and many more. Wang (2008) successfully dealt with the problem of the probability density estimation in the existence of covariates when data were missing at random whereas Shih, Quan, and Chang (2008) devised a method for estimating the mean of the data which included non-ignorable missing values. Keesing et al. (2007) introduced missing data estimation procedure for 3D spiral CT image reconstruction.

In this chapter, the methods that use autoassociative neural networks for missing data estimation are investigated. These techniques essentially are made up of two main components: the autoassociative neural network component, which is a network that maps the input onto itself (Hines & Garvey, 2006) and the optimization component (Snyman, 2005). This chapter investigates a number of optimization methods for missing data estimation and then compares these methods. Spiller and Marwala (2007) used genetic algorithms, particle swarm optimization and simulated annealing for warp control point placement and found that particle swarm optimization is ideally suited for this problem than the other methods.

The optimization methods that are investigated for missing data estimation, in this chapter, are genetic algorithm (Garcia-Martinez et al., 2008; Tavakkoli-Moghaddam, Safari, & Sassani, 2008), simulated annealing (Güngör & Ünler, 2007; Chen & Su, 2002), particle swarm optimization (van den Bergh & Engelbrecht, 2006; Marwala, 2007a) and hill climbing (Tanaka, Toumiya, & Suzuki, 1997; Johnson & Jacobson, 2002). In earlier chapters, it is assumed that the optimization method to be used for missing data estimation to minimize the missing data error equation must have global properties. In this chapter, three global optimization methods genetic algorithm, simulated annealing and particle swarm optimization are compared to the local optimization method, hill climbing, with a sole purpose of answering the crucial question that states that: Is the missing data estimation error equation, which is derived in detail in Chapter II, best solved by global optimization method or by local optimization method?

MISSING DATA ESTIMATION APPROACHES

As explained earlier, the logical approach to handle missing data depends upon how data points become missing. As indicated in earlier chapters, Little and Rubin (1987) as well as Rubin (1978) have demonstrated that there are three types of missing data mechanisms and these are: Missing Completely at Random, Missing at Random and Missing Not At Random. Depending on the mechanism of missing data, currently various methods are used to treat missing data. More information with detailed discussions on the various missing data estimation methods used to handle missing data can be found in Allison (2000); Rubin (1978); Little and Rubin (1987); Mohamed and Marwala (2005); Leke, Marwala, and Tettey (2006); and Nelwamondo (2008). In this chapter, as in earlier chapters, the method for estimating missing data is able to estimate missing data irrespective of the missing data mechanism as long as the rules that describe inter-relationships in the data are known.

The missing data estimation algorithm considered in this chapter involves a neural network which is trained to recall itself and is, therefore, called an autoassociative neural network. Successful deployment of autoassociative neural network include that by Pomi and Olivera (2006) who developed a context-sensitive autoassociative memories and applied this for medical diagnosis, in object recognition (Caldara & Abdi, 2006; Yokoi et al., 2004), in nuclear engineering (Marseguerra, Zio, & Marcucci, 2006), in mechanical engineering (Marwala & Chakraverty, 2006), in fault detection of gearboxes (Del Rincon et al., 2005) and in spotting consonants in speech (Gangashetty, Sekhar, & Yegnanarayana, 2004). As described in earlier chapters, the missing data estimation error equation can be written as follows: