Chapter IX
Missing Data Approaches to Classification

ABSTRACT

In this chapter, a classifier technique that is based on a missing data estimation framework that uses autoassociative multi-layer perceptron neural networks and genetic algorithms is proposed. The proposed method is tested on a set of demographic properties of individuals obtained from the South African antenatal survey and compared to conventional feed-forward neural networks. The missing data approach based on the autoassociative network model proposed gives an accuracy of 92%, when compared to the accuracy of 84% obtained from the conventional feed-forward neural network models. The area under the receiver operating characteristics curve for the proposed autoassociative network model is 0.86 compared to 0.80 for the conventional feed-forward neural network model. The autoassociative network model proposed in this chapter, therefore, outperforms the conventional feed-forward neural network models and is an improved classifier. The reasons for this are: (1) the propagation of errors in the autoassociative network model is more distributed while for a conventional feed-forward network is more concentrated; and (2) there is no causality between the demographic properties and the HIV and, therefore, the HIV status does change the demographic properties and vice versa. Therefore, it is better to treat the problem as a missing data problem rather than a feed-forward problem.

INTRODUCTION

This chapter proposes missing data approaches to classification. Problems in classification have been dealt with by a number of researchers from different areas such as engineering, mathematics and statistics. Classification entails categorizing objects into different classes. As an example, human beings can be classified into two classes, and these are male and female. Diseases can also be classified into classes, and in the case for tumors these can be either malignant or benign. Classification of objects into two classes is called binary classification. In computational intelligence techniques such as neural networks classifying data into two classes involves constructing neural network architecture with an output which is either 0 or 1, with 0 being one class and 1 being another class.
There are many classification techniques that have been developed thus far and these include support vector machines and neural networks. Habtemariam, Marwala, and Lagazio (2005) as well as Habtemariam (2005) have successfully used support vector machines to classify interstate conflict where a 0 represented expected peace and a 1 represented expected conflict. Tettey and Marwala (2006) as well as Tettey (2006) extended this work by using neuro-fuzzy systems to classify inter-state conflict into either peace or conflict. Cai and Chou (1998) used artificial neural network model to construct a model that predicts human immunodeficiency virus (HIV) protease cleavage sites in protein whereas Chamjangali, Beglari, and Bagherian (2007) trained an artificial neural network model using the Levenberg–Marquardt optimization algorithm and used this model for predicting cytotoxicity data (CC50) of anti-HIV 5-pheny-l-phenylamino-IH-imidazole derivatives. Vilakazi and Marwala (2006a) used extension neural networks to classify bushings, which are critical components of electrical transformers, into two classes faulty or healthy while Marwala, Mahola, and Nelwamondo (2006) successfully used hidden Markov models and Gaussian mixture models for detecting faults in mechanical systems where faults could be classified as either present or absent. There are many classification problems that have been tackled using computational intelligence techniques including mechanical fault detection (Mohamed, Rubin, & Marwala, 2006a;b; Mohamed, Tettey, & Marwala, 2006; Nelwamondo & Marwala, 2006; Marwala, Mahola, & Nelwamondo, 2006); intrusion detection system (Vilakazi & Marwala, 2006b); and multi-class protein sequence (Mohamed, Rubin, & Marwala, 2006a; Mohamed, Rubin, & Marwala, 2007). This chapter proposes a novel classification technique that is based on missing data approaches that have been discussed in earlier chapters. Therefore, the classification results are viewed in this chapter as missing data amongst other variables that have been measured.

The missing data approach adopted in this chapter is composed of the combination of the multi-layer autoassociative neural network and genetic algorithm as implemented earlier (Abdella, 2005; Abdella & Marwala, 2005a;b; Mohamed & Marwala, 2005; Leke & Marwala, 2006). This classification technique is then compared to a conventional classification technique that is based on a standard multi-layer perceptron and formulated using the Bayesian approach and the posterior probability estimated using the Gaussian approximation. This technique was proposed by MacKay (1992) and was successfully implemented for classifying inter-state conflict by Lagazio and Marwala (2005). These two classification methods are compared using the classification of HIV problem that has been addressed in the past on numerous occasions (Leke, Marwala, & Tettey, 2006; Marwala, Tettey, & Chakraverty, 2006; Leke, Marwala, & Tettey, 2007).

Acquired immunodeficiency syndrome (AIDS) was first defined (Root-Bernstein, 1993) in 1982 to describe the first cases of unusual immune system failure that was identified in the previous year. HIV was later identified as the cause of AIDS. Risk factor epidemiology examines the individual demographic and social characteristics and attempts to establish factors that position an individual at risk of acquiring a life-threatening disease (Poundstone, Strathdee, & Celectano, 2004). In this chapter, the demographic and social characteristics of the individuals and their behavior are utilized to establish the risk of HIV infection; known as biomedical individualism (Fee & Krieger, 1993). By identifying the individual risk factors that result in HIV infection, it is possible to modify social conditions, which give rise to the disease, and consequently design effective HIV prevention policies. In this chapter, a model is created and used to classify the HIV status of individuals based on demographic properties.

An artificial neural network is an inter-connected structure of processing elements. The artificial neural network structure employed in this chapter consists of three major components (Bishop, 1995)