Chapter II

Estimation of Missing Data Using Neural Networks and Genetic Algorithms

ABSTRACT

Missing data creates various problems in analyzing and processing data in databases. In this chapter, a method aimed at approximating missing data in a database that uses a combination of genetic algorithms and neural networks is introduced. The presented method uses genetic algorithms to minimize an error function derived from an auto-associative neural network. The Multi-Layer Perceptron (MLP) and Radial Basis Function (RBF) networks are employed to form an auto-associative network. An investigation is undertaken into using the method to predict missing data accurately as the number of missing cases within a single record increases. It is observed that there is no significant reduction in the accuracy of the results as the number of missing cases in a single record increases. It is also found that results obtained from using the MLP are better than from the RBF for the data used.

INTRODUCTION

Inferences made from available data for many applications depend on the completeness and the quality of the data being used in the analysis. Therefore, inferences made from complete data are most likely to be more accurate than those made from incomplete data are. However, there are time critical applications that necessitate estimation or approximation of the values of some missing variables, which have to be supplied with the values of other corresponding variables. Such situations may appear in a system that uses a number of instruments, where one or more of the sensors used in the system fail. In such a situation, the values from the missing sensor have to be estimated within a short time and with great precision by taking into account the values of other sensors in the system. In such situations, an approximation for the missing values involves estimating missing values while taking into account the inter-relationships that exists between the values of other variables.
Missing data in a database may arise for various reasons. They can arise from data entry errors, respondents’ non-response to some items in the data collection process, failure of instruments and other reasons. In Table 2.1 presents a database consisting of five variables, namely \(x_1, x_2, x_3, x_4,\) and \(x_5\), where the values of some variables are missing. If we assume that the observations from some variables in various records are not available, it becomes critical that a technique be formulated for estimating the values for the missing entries.

This is possible if there are techniques to approximate the missing data from the observed values that exploit the inter-relationships existing between the variables in the database. Therefore, the aim of this chapter is to use neural networks and genetic algorithms to approximate missing data in such situations. In this chapter, neural networks are used to identify the inter-relationships that exist amongst the variables while a genetic algorithm is used to identify the missing variables, given the inter-relationships that exist amongst the variables that have been identified by the neural network.

This chapter introduces the combination of neural networks and genetic algorithms for missing data estimation. An issue of the criticality of identifying the correct model that describes inter-relationships and rules that govern the data is solved by using the scaled conjugate gradient optimization method and model selection process through cross-validation (Bishop, 1995).

**BACKGROUND**

Missing data generates various problems in numerous applications that depend on access to accurate and complete data.

**Missing Data**

Methods that handle missing data have been areas of research in statistics, mathematics and other disciplines (Yuan, 2000; Allison, 2000; Rubin, 1978; Roth, 1994; Abdella, 2005; Abdella & Marwala, 2005; Nelwamondo & Marwala, 2007a; Nelwamondo & Marwala, 2008). The reasonable way to handle missing data depends upon how data points become missing.

As described earlier, according to Little and Rubin (1987), there are three types of missing data mechanisms. These are MCAR, MAR and the non-ignorable case, which is also called MNAR. MCAR arises if the probability of missing value for variable \(X\) is unrelated to the value \(X\) itself or to any other variable in the data set. This refers to data where the absence of data does not depend on the variable of interest or any other variable in the dataset (Rubin, 1978). MAR arises if the probability of missing

<table>
<thead>
<tr>
<th>(x_1)</th>
<th>(x_2)</th>
<th>(x_3)</th>
<th>(x_4)</th>
<th>(x_5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>3.5</td>
<td>?</td>
<td>5000</td>
<td>-3.5</td>
</tr>
<tr>
<td>?</td>
<td>6.9</td>
<td>5.6</td>
<td>?</td>
<td>0.5</td>
</tr>
<tr>
<td>45</td>
<td>3.6</td>
<td>9.5</td>
<td>1500</td>
<td>46.5</td>
</tr>
<tr>
<td>27</td>
<td>9.7</td>
<td>?</td>
<td>3000</td>
<td>?</td>
</tr>
</tbody>
</table>

*Table 2.1. Database with missing values*
Related Content

Graph and Neural Network-Based Intelligent Conversation System
www.igi-global.com/chapter/graph-and-neural-network-based-intelligent-conversation-system/213043?camid=4v1a

An Innovative Model for Detecting Brain Tumors and Glioblastoma Multiforme Disease Patterns
www.igi-global.com/article/an-innovative-model-for-detecting-brain-tumors-and-glioblastoma-multiforme-disease-patterns/197784?camid=4v1a

A Fast Two-objective Differential Evolutionary Algorithm based on Pareto-optimal Set
Xu Yu-long and Zhao Ling-dong (2016). International Journal of Software Science and Computational Intelligence (pp. 46-59).
www.igi-global.com/article/a-fast-two-objective-differential-evolutionary-algorithm-based-on-pareto-optimal-set/161712?camid=4v1a

Design and Implementation of a Cognitive User-Support System for Skin Progress Analysis Using a Smart Phone
www.igi-global.com/article/design-and-implementation-of-a-cognitive-user-support-system-for-skin-progress-analysis-using-a-smart-phone/88991?camid=4v1a