Chapter XXXIII

Fuzzy Imputation Method for Database Systems

J. I. Peláez
University of Málaga, Spain

J. M. Doña
University of Málaga, Spain

D. La Red
University of National of the Northeast, Argentina

ABSTRACT

Missing data is often an actual problem in real data sets, and different imputation techniques are normally used to alleviate this problem. Imputation is a method to fill in missing data with plausible values to produce a complete data set. In this chapter, we analyze the performance of the different traditional data imputation methods. A new fuzzy imputation approach is proposed using ordered weighted average operators and the majority concept. In order to form the majority concept, we propose the use of neat OWA operators and linguistic quantifiers with two fusion strategies for aggregation operators.

INTRODUCTION

The missing data and nonresponse problem is a usual difficulty of particular concern in medical and social science databases. Dealing with nonresponse can be a difficult matter and it is important to apply adequate missing data methods to obtain valid inference.

Missing data is a very common problem in real data sets, and different methods to solve this problem have been developed. A simple and common strategy is to ignore missing values, thus reducing the size of the useful data set. The experience in databases has demonstrated the dangers of simply removing cases (listwise deletion) from the original data set, and deletion can introduce
substantial biases in the study, especially when missing data are distributed in a nonrandom way. Missing data values may be frequent in data collection efforts, such as social surveys or scientific experiments, as well as in system data archives. This can be attributed to numerous factors, which include nonresponse from the sample of the study or malfunction of data collection devices.

In this sense, the imputation of missing data is an area of statistics that has attracted much attention in the last decades, and many different strategies have been developed to handle this problem.

Undoubtedly, imputation should be applied cautiously, and the analysts of the completed data set should be fully warned of the potential dangers created by the imputation. It is very important to reduce the impact of imputed data over the whole of database. In this way, the majority concept plays a main role in the imputation processes that complete the nonresponse or missing data, using values close to the real majority opinion.

This chapter reviews various imputation methods used within the social sciences to compensate for the item nonresponse bias and missing data, and provides guidance on how to use such methods in practice. Also, a new technique based on the fuzzy approach is presented. This new method uses the concept of fuzzy majority to analyze the data set, providing consistent imputation values.

The chapter introduces the principles and methods in the imputation literature to the novice. It also provides the reader who is familiar with imputation methods with new developments in this field. It is emphasised that the choice of an appropriate imputation method may strongly depend on the available data, and the application and purpose of the analysis, as illustrated from the social sciences.

The chapter is structured as follows. In the following subsection, different approaches to handle missing data are presented briefly; definitions and basic assumptions about the missing data mechanisms are also introduced. In the section “Imputation Methods,” several techniques are reviewed as ways of handling nonresponse problems. In “Fuzzy Majority Imputation,” OWA (ordered weighted averaging) operators and the majority concept are introduced; in addition, quantification methods are briefly defined to model a fuzzy majority, and their use for recovery data is proposed. Finally, the conclusions are exposed.

**Missing-Data Approaches**

By nonresponse, we mean that the required data are not obtained for all elements that are selected for observation. Generally, a distinction is made among nonresponse units, that is, the failure of a sample member to respond to a nonresponse item, the failure to obtain some required information from individual sample members, or missing data from traditional databases or fuzzy databases (Galindo, Urrutia, & Piattini, 2006). The first unit nonresponse occurs if it is not possible to interview certain sample members or if sample members did not want to take part in the survey. On the other hand, a nonresponse item occurs if the interviewer fails to ask a question or does not record the answer, or the sample member refuses to answer a question or does not know the answer. Missing data can appear in several cases, for example, as a result of hardware problems. There are several ways of dealing with nonresponse and missing-data problems. For nonresponse units, weighting methods are usually applied. There is a range of missing-data methods to compensate for the nonresponse item, such as the available case method, imputation methods, weighting methods, and model-based procedures such as maximum likelihood estimation. For missing data, imputation and weighting methods are used. Overviews of such methods are given in Little and Rubin (1990), Government Statistical Service (GSS, 1996), Schafer and Graham (2002), Durrant (2005), Raghunathan, Lepkowski, van Hoewyk, and Solenberger (2001), and Ibrahim, Chen, Lipsitz, and Herring (2005).

Some simple methods that are commonly used to handle nonresponse item are listwise or case deletion, pairwise deletion, and case analysis, which are only focused on observed cases (Allison, 2001;...
Related Content

An Interpreter Approach for Exporting Relational Data into XML Documents with Structured Export Markup Language
Joseph Fong and Herbert Shiu (2012). *Journal of Database Management* (pp. 49-77).
[www.igi-global.com/article/interpreter-approach-exporting-relational-data/62032?camid=4v1a](www.igi-global.com/article/interpreter-approach-exporting-relational-data/62032?camid=4v1a)

Challenges in Data Mining on Medical Databases
[www.igi-global.com/chapter/challenges-data-mining-medical-databases/7980?camid=4v1a](www.igi-global.com/chapter/challenges-data-mining-medical-databases/7980?camid=4v1a)

Dynamic Integration in Multidatabase Systems
[www.igi-global.com/article/dynamic-integration-multidatabase-systems/51160?camid=4v1a](www.igi-global.com/article/dynamic-integration-multidatabase-systems/51160?camid=4v1a)

Fuzzy Sequential Patterns for Quantitative Data Mining
[www.igi-global.com/chapter/fuzzy-sequential-patterns-quantitative-data/20375?camid=4v1a](www.igi-global.com/chapter/fuzzy-sequential-patterns-quantitative-data/20375?camid=4v1a)