Chapter 9

Estimating Semi-Parametric Missing Values with Iterative Imputation

Shichao Zhang

Zhejiang Normal University and Zhongshan University, China

ABSTRACT

In this paper, the author designs an efficient method for imputing iteratively missing target values with semi-parametric kernel regression imputation, known as the semi-parametric iterative imputation algorithm (SIIA). While there is little prior knowledge on the datasets, the proposed iterative imputation method, which impute each missing value several times until the algorithms converges in each model, utilize a substantially useful amount of information. Additionally, this information includes occurrences involving missing values as well as capturing the real dataset distribution easier than the parametric or nonparametric imputation techniques. Experimental results show that the author’s imputation methods outperform the existing methods in terms of imputation accuracy, in particular in the situation with high missing ratio.

INTRODUCTION

The real data usually are incomplete as some instances may have missing values. In fact, many reasons can result in missing values, for instance, malfunction of equipments, erroneous from human imputation, and so on. Missing values is an unavoidable problem in the real world, and various methods for dealing with such issues have been developed in data mining and in statistics. For example, case deletion, learning with no handling with missing data and missing values imputation.
In real application, the imputation method is a popular strategy comparing to other methods. Missing values imputation is to find an efficient way to “guess” the missing values (imputation) based on other information in datasets. One advantage of this approach is that missing values treatment is independent of the learning algorithm used. That allows users to select the most suitable imputation method for their applications.

Commonly used imputation methods for missing values include parametric regression imputation methods and non-parametric regression imputations. However, there are other relations within real world data, and both parametric imputation method and non-parametric imputation method are not adequate to capture the relations. That is, we know a part of relation between independent variables (condition attributes) and dependent variable (target attribute), e.g., we can regard this relation as parametric model, but we have no knowledge on the relation between other independent variables and dependent variable, e.g., we can take it as nonparametric model. However, combining these two parts, it is difficult for us to consider the compound relation with parametric model or nonparametric model. Moreover, the case is very general in real application. In this paper, we regard the relation containing two models as semi-parametric model or partial parametric model. In real application, semi-parametric model is natural than non-parametric model because users can always know some information but no all on the datasets, such some parameters in the datasets. To model this semi-parametric relation, in this paper, we design an efficient semi-parametric iterative imputation method (SIIA) that takes into account the advantages of parametric models and pure non-parametric models so as to overcome their certain shortcomings for each single model.

In the left parts, we will first review the existing literatures for dealing with missing values. And then we design the iterative imputation methods which can impute missing values with kernel method or even in the dataset with high missing ratio. After that, we will demonstrate our proposed methods with all kinds of experiments. Finally, we will conclusion our works and put forward our future work.

**RELATED WORK**

There are at least three different ways of dealing with missing data based on Little and Rubin (2002): single imputation, multiple imputation, and iterative procedure.

Single imputation strategies provide a single estimate for each missing data value. Many methods for imputing missing values are single imputation methods, such as, C4.5 algorithm, kNN method, and so on. We can partition single imputation methods into parametric methods and nonparametric ones. The parametric regression imputation methods are superior if a dataset can be adequately modeled parametrically, or if users can correctly specify the parametric forms for the dataset. Non-parametric imputation (Qin et al., 2007) offers a nice alternative if users have no idea on the actual distribution of a dataset because the method can provide superior fits by capturing structure in datasets. While much work focus on modeling data by parametric or nonparametric approaches, in Engle et al. (1986) have studied the semi-parametric model. They model the electricity demand \( y \) as the sum of a smooth function \( g \) of monthly temperature \( t \), and a linear function of \( x_1 \) and \( x_2 \), as well as 11 monthly dummy variables \( x_3, ..., x_{13} \), to build a semi-parametric model firstly. In fact, semi-parametric model is more ordinary in real application than nonparametric model or parametric model because we always contain a little but no all information on our datasets, however, there are a little literatures, such as, Nikulin (2008), focusing on this issue because of the analysis complexity, in this paper, we introduce SIIA algorithm to model the partial parametric model for filling up iteratively missing target values.
### Related Content

**An Efficient Pruning and Filtering Strategy to Mine Partial Periodic Patterns from a Sequence of Event Sets**  
Kung-Juian Yang, Tzung-Pei Hong, Yuh-Min Chen and Guo-Cheng Lan (2014). *International Journal of Data Warehousing and Mining* (pp. 18-38).  

**Web Mining: Creating Structure out of Chaos**  
[www.igi-global.com/chapter/web-mining-creating-structure-out/25769?camid=4v1a](www.igi-global.com/chapter/web-mining-creating-structure-out/25769?camid=4v1a)

**Constrained Cube Lattices for Multidimensional Database Mining**  
Alain Casali, Sébastien Nedjar, Rosine Cicchetti and Lotfi Lakhal (2010). *International Journal of Data Warehousing and Mining* (pp. 43-72).  
[www.igi-global.com/article/constrained-cube-lattices-multidimensional-database/44958?camid=4v1a](www.igi-global.com/article/constrained-cube-lattices-multidimensional-database/44958?camid=4v1a)

**Reliability Estimates for Regression Predictions: Performance Analysis**  
Zoran Bosnic and Igor Kononenko (2011). *Integrations of Data Warehousing, Data Mining and Database Technologies: Innovative Approaches* (pp. 320-338).  
[www.igi-global.com/chapter/reliability-estimates-regression-predictions/53082?camid=4v1a](www.igi-global.com/chapter/reliability-estimates-regression-predictions/53082?camid=4v1a)