Chapter 16

Combining kNN Imputation and Bootstrap Calibrated: Empirical Likelihood for Incomplete Data Analysis

Yongsong Qin
Guangxi Normal University, China

Shichao Zhang
Zhejiang Normal University, China & University of Technology, Australia

Chengqi Zhang
University of Technology, Australia

ABSTRACT

The k-nearest neighbor (kNN) imputation, as one of the most important research topics in incomplete data discovery, has been developed with great successes on industrial data. However, it is difficult to obtain a mathematical valid and simple procedure to construct confidence intervals for evaluating the imputed data. This chapter studies a new estimation for missing (or incomplete) data that is a combination of the kNN imputation and bootstrap calibrated EL (Empirical Likelihood). The combination not only releases the burden of seeking a mathematical valid asymptotic theory for the kNN imputation, but also inherits the advantages of the EL method compared to the normal approximation method. Simulation results demonstrate that the bootstrap calibrated EL method performs quite well in estimating confidence intervals for the imputed data with kNN imputation method.

INTRODUCTION

Item non-response is usually handled by some form of imputation to fill in the missing item values. Imputation not only permits the creation of a general purpose complete public-use data file that can be used for standard analyses, but also recovers part of lost information. For example, we may intend to get \((X,Y)\), \(1 \leq i \leq n\), from \(n\) individuals. If \(X\)-values are observed completely and some of \(Y\)-values may be missing, we could use some ways to fill in the missing values of \(Y\) by the relationship between the variables \(X\) and

DOI: 10.4018/978-1-61350-474-1.ch016
Combining kNN Imputation and Bootstrap Calibrated
Y as each pair \((X, Y)\) comes from one individual. It then would be better than to drop the \(i\)th pair \((X, Y)\) if \(Y\) alone is missing. The \(k\)-nearest neighbor (kNN) imputation is one of the most important hot deck methods used to compensate for nonresponse in incomplete data mining. It is a nonparametric imputation method, which is simple but effective for many applications. With the method, a case is imputed using values from its \(k\) nearest neighbors (points). A thorough description of the kNN technique was presented by McRoberts et al. (2002).

The nearest neighbor (NN) imputation, a special case of kNN when \(k = 1\), has widely been used in the data analysis applications, such as surveys conducted at Statistics Canada, the U.S. Bureau of Labor Statistics, and the U.S. Census Bureau (Chen & Shao, 2000). Although this happened, it is difficult to obtain a mathematical valid and simple procedure to construct confidence intervals for evaluating the imputed data. Instead of building a validation theory, the efficiency of kNN imputation is evaluated with few experiments in data mining and machine learning. This leaves open the confidence of the compensated values with the kNN imputation.

In this chapter, we propose a new estimation for missing (or incomplete) data, named \(k\)NN-BEL, that is a combination of the \(k\)NN imputation and bootstrap calibrated EL (Empirical Likelihood). The \(k\)NN-BEL first uses the Cross-validation method to choose \(k\). And then missing values are estimated with their \(k\) nearest neighbors. Finally, a bootstrap calibrated empirical likelihood (EL) method is applied to construct confidence intervals for the mean of the dependent variable, i.e., evaluating the confidence of the compensated values with the \(k\)NN imputation.

In the complete data setting, the original idea of empirical (or nonparametric) likelihood (EL) dates back to Hartley and Rao (1968) in the context of sample surveys, and Owen (2001) made a systematic study of the EL method. The EL confidence intervals are range preserving and transformation respecting, and the shape and orientation of EL intervals are determined entirely by the data, unlike the normal approximation based intervals.

In the incomplete data setting (Banek et al., 2008; Golfarelli & Rizzi, 2009; Gomez et al., 2009; Nikulin, 2008; Pighin & Ieronutti, 2008; Yu et al., 2009), the \(k\)NN imputation is used associated with the bootstrap calibrated EL. The use of the Cross-validation method can guarantee the smallest mean squared errors in the process of imputation, and the bootstrap calibrated EL method releases the burden of having a mathematical valid asymptotic theory for the estimators of parameters in interest, while this method also keeps the advantages of the EL method compared to the normal approximation method. Without loss of generality, we simulate the \(k\)NN-BEL and demonstrate that the combination of the \(k\)NN imputation and the bootstrap calibrated EL method perform quite well in data analysis with missing data.

The rest of the chapter is organized as follows. We first outline some commonly used imputation methods. And then, we propose the cross-validation method to choose \(k\) in \(k\)NN imputation, and propose a bootstrap calibrated EL method to construct confidence intervals. After that, some simulation results are reported on studying the performance of the proposed confidence intervals. Finally, we conclude this chapter.

PRELIMINARY

In this section, we first list all symbols used in this chapter. Second, we briefly describe the missing mechanisms. Third, we review some imputation methods used in the literatures. Finally, some related work is briefly introduced.

Symbols

In this chapter, all symbols are defined as follows.

- \(X\): A \(d\)-dimensional independent variable.
- \(Y\): An univariate dependent variable.