Computational Statistics and Data Analysis 93 (2016) 373–387

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Bayesian network data imputation with application to survival tree analysis☆

Paola M.V. Rancoita[a,b,c,*], Marco Zaffalon[b], Emanuele Zucca[d], Francesco Bertoni[c,d], Cassio P. de Campos[e]

[a] University Centre for Statistics in the Biomedical Sciences, Vita-Salute San Raffaele University, Milan, Italy
[b] Dalle Molle Institute for Artificial Intelligence, Manno, Switzerland
[c] Institute of Oncology Research, Bellinzona, Switzerland
[d] Oncology Institute of Southern Switzerland, Bellinzona, Switzerland
[e] Queen's University Belfast, School of Electronics, Electrical Engineering and Computer Science, Belfast, UK

HIGHLIGHTS

- Retrospective clinical datasets have often small sample size and many missing data.
- We use Bayesian networks to impute missing data enhancing survival tree analysis.
- The Bayesian network is learned from incomplete data and used for the imputation.
- Our method generally achieved more accurate predictions than widely used approaches.

ARTICLE INFO

Article history:
Received 30 April 2014
Received in revised form 18 December 2014
Accepted 19 December 2014
Available online 3 January 2015

Keywords:
Bayesian networks
Data imputation
Missing data
Prognostic stratification
Survival tree

ABSTRACT

Retrospective clinical datasets are often characterized by a relatively small sample size and many missing data. In this case, a common way for handling the missingness consists in discarding from the analysis patients with missing covariates, further reducing the sample size. Alternatively, if the mechanism that generated the missing allows, incomplete data can be imputed on the basis of the observed data, avoiding the reduction of the sample size and allowing methods to deal with complete data later on. Moreover, methodologies for data imputation might depend on the particular purpose and might achieve better results by considering specific characteristics of the domain. The problem of missing data treatment is studied in the context of survival tree analysis for the estimation of a prognostic patient stratification. Survival tree methods usually address this problem by using surrogate splits, that is, splitting rules that use other variables yielding similar results to the original ones. Instead, our methodology consists in modeling the dependencies among the clinical variables with a Bayesian network, which is then used to perform data imputation, thus allowing the survival tree to be applied on the completed dataset. The Bayesian network is directly learned from the incomplete data using a structural expectation–maximization (EM) procedure in which the maximization step is performed with an exact anytime method, so that the only source of approximation is due to the EM formulation itself. On both simulated and real data, our proposed methodology usually outperformed several existing methods for data imputation and the imputation so obtained

☆ Software freely available at: http://code.google.com/p/csda-dataimputation/.

* Correspondence to: Vita-Salute San Raffaele University, University Centre for Statistics in the Biomedical Sciences, Faculty of Psychology, Via Olgettina 58, 20132 Milan, Italy. Tel.: +39 0226433844.
E-mail address: rancoita.paolamaria@unisr.it (P.M.V. Rancoita).

http://dx.doi.org/10.1016/j.csda.2014.12.008

improved the stratification estimated by the survival tree (especially with respect to using surrogate splits).

1. Introduction

Retrospective clinical data are often used to identify features that can help in classifying patients into groups of similar survival and in predicting the survival outcome of patients (i.e. a prognostic patient stratification). The identification of classes of patients with a different clinical course or response to a specific treatment allows the design of the most appropriate approach for the management of each individual patient. The survival tree is a state-of-the-art method to stratify patients for predicting survival on the basis of available clinical parameters (Ciampi and Thiffault, 1986). Although several algorithms exist for its estimation (Davis and Anderson, 1989; LeBlanc and Crowley, 1992, 1993; Segal, 1988; Keleş and Segal, 2002; Hothorn et al., 2006; Fana et al., 2009), the procedure always consists in finding, at each step, the best clinical variable able to divide the patients (with respect to the survival), so that the final stratification of the patients assumes a tree-like structure.

In retrospective studies, clinical and survival data may contain many missing values for several reasons. If the study includes data over a long period, some clinical parameters might not have been measured for some patients, because they were not systematically collected at diagnosis, and data might be missing in individual patients due to technical issues. More importantly, data might be missing in some particular subset of patients, causing biases in the analysis: for example, patients with a very aggressive course might have died before performing a test, or a test might have been skipped in patients expected to have a very good clinical course. Therefore, a retrospective study usually contains many missing covariate data, and this can heavily affect the statistical analysis, especially when the sample size is small. This issue worsens if the dataset contains many censored survival data. In fact, the missingness of the covariates added to the censoring issue increases the hardness of identifying an accurate prognostic stratification of the patients. In this work, we denote by missing data only the incomplete information happening in clinical and biological variables that are available in the analysis, and not the incomplete lifetime information of patients (censoring). A naive, still very used, approach to handle this issue consists in discarding all patients with missing variables from the analysis, decreasing the power of any model, which is clearly undesirable. Instead, survival tree procedures decide the best splits to define the tree using only the observed data for each variable, and they resort to surrogate splitting in case of missing values, that is, they use a splitting rule based on another variable which most resembles the behavior of the original missing one (Breiman et al., 1984).

Another widely used approach to handle missing data is to impute the missing values, thus considering a complete dataset in further analyses (Little and Rubin, 1987). The data imputation problem regards completing the dataset in some particular manner such that the important characteristics of the dataset are preserved. This is mostly done by assuming that missing data are *missing completely at random* (that is, their missingness is independent of both unobserved and observed data), which implies that data imputation can be safely performed by analyzing each variable separately. Widely used methods, such as single expected value imputation and single mode imputation, are based on this assumption. However, missing data in clinical datasets can be more realistically considered as *missing at random* instead of missing completely at random (that is, their missingness, conditional on the observed data, is independent of the unobserved values). In fact, some of the examples discussed before in this introduction are missing at random, but not completely at random. In the literature, many statistical approaches that account for the dependencies among covariates have been used for data imputation. In case of categorical or discrete variables (which is often the case for clinical parameters), these methods are usually based on maximum likelihood (ML) estimation of the joint distribution of the covariates from the partially classified contingency table built using the observed data (Little and Rubin, 1987). Unfortunately, they suffer from the small sample size and tend to overfit, even with a small number of covariates, because they consider all dependencies among all variables.

We propose to use a methodology based on Bayesian networks as a way to impute accurately the missing data and improve the quality of the inference, especially in the application to survival tree analysis. For this application, our imputation method is employed only for imputing the covariates (without any knowledge about the survival data) and the survival tree is applied to the (supposedly accurate) completed dataset so obtained. A Bayesian network is a probabilistic graphical model that relies on a directed acyclic graph to encode the structured dependency among random variables and compactly represent a joint probability distribution. Learning and inference in these models benefit from fast and accurate procedures (Koller and Friedman, 2009). More specifically, learning a Bayesian network from data consists in searching for the structure of the network, as well as its parameters, such that some criterion of quality is maximized. The most common criterion for this purpose is the Bayesian Dirichlet Equivalent Uniform (Heckerman et al., 1995), which is based on maximizing the posterior probability of the structure given the data. Although this is a particularly challenging problem when data are incomplete, suitable algorithms do exist (Friedman, 1998; Meila and Jordan, 1998; Singh, 1998; Riggelsen and Feelders, 2005; Ramoni and Sebastiani, 1997; Riggelsen, 2006). These methods are mostly based on turning the incomplete data into a complete dataset (or even directly updating the sufficient statistics), and then recurring to particular methods for complete data. We adopt a meta-search composed of a few distinct methods (Jaakkola et al., 2010; de Campos and Ji, 2011; Cooper and Herskovits, 1992; Silander and Myllymaki, 2006) that selects the best procedure to run depending on the number of covariates