Information Sciences 311 (2015) 163–181

Contents lists available at ScienceDirect

Information Sciences

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

Tree-based prediction on incomplete data using imputation or surrogate decisions

H. Cevallos Valdiviezo [a], S. Van Aelst [a,b,*]

[a] Ghent University, Department of Applied Mathematics, Computer Science and Statistics, Krijgslaan 281 S9, Gent, Belgium
[b] KU Leuven, Department of Mathematics, Section of Statistics, Celestijnenlaan 200B B-3001, Leuven, Belgium

ARTICLE INFO

Article history:
Received 13 June 2014
Received in revised form 6 March 2015
Accepted 10 March 2015
Available online 23 March 2015

Keywords:
Prediction
Missing data
Surrogate decision
Multiple imputation
Conditional inference tree

ABSTRACT

The goal is to investigate the prediction performance of tree-based techniques when the available training data contains features with missing values. Also the future test cases may contain missing values and thus the methods should be able to generate predictions for such test cases. The missing values are handled either by using surrogate decisions within the trees or by the combination of an imputation method with a tree-based method. Missing values generated according to missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) mechanisms are considered with various fractions of missing data. Imputation models are built in the learning phase and do not make use of the response variable, so that the resulting procedures allow to predict individual incomplete test cases. In the empirical comparison, both classification and regression problems are considered using a simulated and real-life datasets. The performance is evaluated by misclassification rate of predictions and mean squared prediction error, respectively. Overall, our results show that for smaller fractions of missing data an ensemble method combined with surrogates or single imputation suffices. For moderate to large fractions of missing values ensemble methods based on conditional inference trees combined with multiple imputation show the best performance, while conditional bagging using surrogates is a good alternative for high-dimensional prediction problems. Theoretical results confirm the potential better prediction performance of multiple imputation ensembles.

1. Introduction

Many real datasets with predictive applications face the problem of missing values on useful features. Evidently, this complicates the predictive modeling process since predictive power may depend heavily on the way missing values are treated. In principle, missing data can occur in the training data only, in the individual test cases only, or in both the training data and test cases. In practice, however, missing data appear most often in both training and test set. Consider for instance customer data that is used to predict important outcomes such as buying preferences for individual costumers (based on their past actions). This type of data frequently contains missing values in both the training data and test cases, because the same amount of information is not available for all customers.

* Corresponding author at: KU Leuven, Department of Mathematics, Section of Statistics, Celestijnenlaan 200B B-3001, Leuven, Belgium. Tel.: +32 16 372383.
E-mail addresses: Holger.CevallosValdiviezo@ugent.be (H. Cevallos Valdiviezo), Stefan.VanAelst@wis.kuleuven.be (S. Van Aelst).

http://dx.doi.org/10.1016/j.ins.2015.03.018
0020-0255/

Most of the research work so far has addressed the problem of missing values in the training data (see e.g. [2,9,13,16,34,37]). On the other hand, [36] is one of the only contributions in which the prediction accuracy of classification techniques is compared when only test cases contain missing values. Tree-based classifiers have been investigated for test cases with data missing completely at random (MCAR), i.e. test cases with missingness which does not depend on any value of the data. The performance of prediction methods for different missing data strategies when missing data occur in both the training and test set has been assessed in [15,22,32]. However, in [32] k-nearest neighbors (kNN) imputation was applied separately on the training and test samples. This is a potential weakness for practical purposes because the kNN imputation is impossible for test cases that appear on a case-by-case basis. Similarly, in [15,22] imputation models were applied separately to the training and test cases. Moreover, the response variable was used in the imputation model for the training data so that the same imputation scheme cannot be applied to test cases arriving one-by-one. In this study, we are interested in methods that can deal with missingness in both training and test cases. Moreover, the methods should be able to handle test cases that appear one-by-one, because this case is often encountered in practical applications. Think for example of new potential patients for which a prediction needs to be made as soon as possible on a case-by-case basis, using the available information of the patient (such as clinical test results).

In this work we compare several strategies to handle missing data when using tree-based prediction methods. We focus on trees because they have several advantages and few limitations compared to other prediction techniques. Firstly, trees allow to handle data of different type (categorical, discrete, continuous). Other features that make trees highly popular among practitioners are their ability to capture important dependencies and interactions. Moreover, tree-based ensembles such as random forests can easily handle high dimensional problems and often show good performance without the need to fine-tune parameters. Trees also include a built-in methodology to process observations with missing data, called surrogate splits [6].

Evidently, if the missing data issue is not addressed correctly, misleading predictions may be obtained. Thus, one aims for prediction rules that have low bias (accurate enough) and low variability (stable enough) and at the same time take into account the additional uncertainty caused by missing values. Among the strategies to handle the missing values are:

1. Discard observations with any missing values in the training data.
2. Rely on the learning algorithm to deal with missing values in the training phase.
3. Impute all missing values before training the prediction method.

Approach 1 encompasses ad hoc procedures like complete case and available case analysis. They have been shown to work for relatively small amounts of missing data and under certain restrictive conditions [44,48]. However, this approach is not applicable when missing values are present in test cases. Tree methods with surrogate splits are an example of the second approach. An advantage of strategy 2 is that incomplete data need not be treated prior to model fitting. For most learning techniques, the third approach is necessary to handle incomplete values or it simply helps to improve predictive capability. Many imputation methods have been developed to address the missing data issue in general. Imputation methods have been studied extensively with regard to inference: unbiasedness of estimates, efficiency, coverage and length of confidence intervals or power of tests (see e.g. [8,11,26,38]). Other works study the performance of imputation methods when estimating the true values of the missing data, without considering the subsequent statistical analysis (see e.g. [24,39]). However, there is much less known about the properties of imputation methods in the context of prediction. An advantage of Approach 3 is that it completely separates the missing data problem from the prediction problem. This strategy thus gives freedom to (third party) analysts to apply any appropriate data mining method to the imputed data.

A few comparisons of approach 2 and 3 have already been considered in the literature. For instance in [13] CART using surrogates was compared to CART preceded by single or multiple imputation. Two classification problems were considered. Multiple imputation performed clearly better than both single imputation and surrogates. Single imputation outperformed surrogates for a fraction of missingness above 10%. No ensemble methods were considered.

The predictive performance of conditional random forests [20] with missing data was investigated in [32]. Conditional random forests (CondRF) combined with surrogates was compared to CondRF with prior kNN imputation. Both classification and regression problems were considered. No difference in performance was found between handling missing values by surrogates or with prior kNN imputation. Recently, [15] compared the predictive performance of CART, conditional inference tree (CondTree) and CondRF in combination with surrogates or Multiple Imputation by Chained Equations (MICE) to handle the missing data. Real datasets with and without missing cells were used. The complete data were used for a simulation study in which missing values were introduced completely at random. For the real data with missing values MICE did not show a convincing improvement compared to surrogates, while in their simulation study MICE was beneficial for large amounts of missing data introduced in many variables. However, the authors argue that their simulation results may lack generalizability due to restrictive and artificial simulation patterns. Therefore, it is suggested to extend their simulations to a wider range of patterns.

So far, there is no clear conclusion in the literature about which combinations of tree-based prediction method and missing data strategy yield the most satisfactory predictions. It seems that an answer to this question may depend on the structure of the predictors, the type of relationship between predictors and response variable, and the pattern and fraction of missing data.