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

Holger Cevallos Valdiviezo
Department of Applied Mathematics, Computer Science and Statistics, Ghent University
Krijgslaan 281 S9, Ghent, Belgium
Holger.CevallosValdiviezo@ugent.be

Stefan Van Aelst
Department of Mathematics, Section of Statistics, KU Leuven
Celestijnenlaan 200B B-3001, Leuven, Belgium
Stefan.VanAelst@wis.kuleuven.be

Abstract

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. Therefore, it is in our interest that methods can deal with missingness in both training and test cases and are able to handle test cases that appear one-by-one, because this 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). This contribution is based on our recent work in [1]. We investigate which combinations of tree-based prediction method and missing data strategy yield the most satisfactory predictions when the available training data contains features with missing values. The methods are also able to generate predictions for future individual test cases with missing data. 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. 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. We consider as learning methods CART, conditional inference tree, random forest, conditional random forest, bagging and conditional bagging. To handle missing data we consider surrogates, single imputation by median/mode, proximity matrix or k-nearest neighbors (kNN), and multiple imputation by chained equations (MICE) or Multiple Imputation by Sequential Regression Trees (MIST). We also include an alternative to multiple imputation, as introduced in [2], that imputes B bootstrap samples. We adapted the procedure in [2] by using MIST to impute the bootstrap samples. In total we compare 26 techniques which are summarized in the table below.

In our study we perform an empirical and a theoretical comparison. In the empirical comparison we consider a simulated dataset with regression problem and real-life datasets with both classification and regression problems. The performance
Overview of the 26 techniques investigated in this study. Each mark ‘×’ corresponds to a technique. The second mark in the MIST + RF box corresponds to a special case of this technique that consists of imputing bootstrap samples by MIST + RF. N/I stands for “not implemented”.

<table>
<thead>
<tr>
<th>Strategy for miss. data</th>
<th>Imputation method</th>
<th>CART</th>
<th>CondTree</th>
<th>RF</th>
<th>CondRF</th>
<th>Bagg.</th>
<th>CondBagg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Surrogates</td>
<td>None</td>
<td>×</td>
<td>×</td>
<td>N/I</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Single Imp.</td>
<td>Median/mode</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>N/I</td>
<td>N/I</td>
</tr>
<tr>
<td></td>
<td>Prox. matrix</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>N/I</td>
<td>N/I</td>
</tr>
<tr>
<td></td>
<td>kNN</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>N/I</td>
<td>N/I</td>
</tr>
<tr>
<td>Multiple Imp.</td>
<td>MICE</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>N/I</td>
<td>N/I</td>
</tr>
<tr>
<td></td>
<td>MIST</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>×</td>
<td>N/I</td>
<td>N/I</td>
</tr>
</tbody>
</table>

is evaluated by the mean squared prediction error (MSPE) for regression or its equivalent misclassification error (MER) for classification. Moreover, missing values are generated according to missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR) mechanisms in all variables and in a randomly chosen third with various fractions of missing data: 10%, 20%, 30% and 40%. The procedure to generate datasets with missing values was repeated 1000 times for each mechanism, scheme and fraction of missing data. The mean root MSPE (RMSPE) or the mean MER across these 1000 iterations is reported as a final measure of predictive performance. A summary of mean RMSPE/MER values as the percentage of missing data increases is shown in Figures 1 and 2 for the real-life datasets analyzed in this study. Results for the simulated dataset (not shown) yield similar patterns. All methods with overall low performance are not plotted. The plots correspond to schemes with all variables containing missing values. The performance of all methods when a random third of the variables contains missing values is quite stable and resembles that of 10% missingness in all variables. The same point characters are used for techniques based on the same tree prediction method. Different line types and colors (gray scale) correspond to the different missing data treatments.

Overall, our results show that for smaller fractions of missing data an ensemble method combined with surrogates or single imputation suffices. In particular, conditional random forest with surrogates (dotted lines with triangle point-down symbols) performs as well as other conditional random forest combinations in three out of the four real-life datasets: Survival, Heart and Birthweight. Similarly, a single imputation + random forest strategy is sufficient to obtain competitive prediction results in the Fertility and Heart datasets. For instance, Prox. matrix + RF (long-dashed lines with triangle point-up symbols) performs very well for these datasets. For moderate to large fractions of missing values conditional random forest combined with multiple imputation by MICE or MIST is the safest option. MICE/MIST + conditional random forest are shown in Figures 1 and 2 with the triangle point-down joined with solid thick lines for MICE as opposed to solid thin lines for MIST. Both methods show good results in comparison to the other techniques in the Survival, Heart and Birthweight datasets, while in most instances of the Fertility dataset they are at least competitive to the other methods. Their satisfactory result is due to the mutual effort of the imputation strategy and prediction method to average out sampling variability and variability due to missing data. Results in the theoretical comparison confirm the potential better prediction performance of multiple imputation ensembles. Conditional bagging using surrogates (dotted lines with + symbol) also yielded quite competitive results for all datasets. Results of other experiments showed that Conditional bagging is also a good alternative for high-dimensional prediction problems, yielding a good compromise between performance and computation time. The new method of
Figure 1: Mean MER results for the Survival data (top row) and Heart disease data (bottom row). Results are shown for data MCAR (left panel), MAR (middle panel) and MNAR (right panel).

MIST imputed bootstrap samples when combined with random forest (in gray solid line in Figures 1 and 2) is not able to outperform multiple imputation.
Figure 2: Mean RMSPE results for the Fertility data (top row) and Birthweight data (bottom row). Results are shown for data MCAR (left panel), MAR (middle panel) and MNAR (right panel).
References
