MissForest - nonparametric missing value imputation for mixed-type data

Daniel J. Stekhoven, Peter Bühlmann

(Submitted on 4 May 2011 (v1), last revised 27 Sep 2011 (this version, v2))

Modern data acquisition based on high-throughput technology is often facing the problem of missing data. Algorithms commonly used in the analysis of such large-scale data often depend on a complete set. Missing value imputation offers a solution to this problem. However, the majority of available imputation methods are restricted to one type of variable only: continuous or categorical. For mixed-type data the different types are usually handled separately. Therefore, these methods ignore possible relations between variable types. We propose a nonparametric method which can cope with different types of variables simultaneously. We compare several state of the art methods for the imputation of missing values. We propose and evaluate an iterative imputation method (missForest) based on a random forest. By averaging over many unpruned classification or regression trees random forest intrinsically constitutes a multiple imputation scheme. Using the built-in out-of-bag error estimates of random forest we are able to estimate the imputation error without the need of a test set. Evaluation is performed on multiple data sets coming from a diverse selection of biological fields with artificially introduced missing values ranging from 10% to 30%. We show that missForest can successfully handle missing values, particularly in data sets including different types of variables. In our comparative study missForest outperforms other methods of imputation especially in data settings where complex interactions and nonlinear relations are suspected. The out-of-bag imputation error estimates of missForest prove to be adequate in all settings. Additionally, missForest exhibits attractive computational efficiency and can cope with high-dimensional data.

Comments:	Submitted to Oxford Journal's Bioinformatics on 3rd of May 2011
Subjects:	**Applications (stat.AP)**; Machine Learning (stat.ML)
Cite as:	**arXiv:1105.0828 [stat.AP]**
	(or **arXiv:1105.0828v2 [stat.AP]** for this version)

Submission history

From: Daniel Stekhoven [view email]

[v1] Wed, 4 May 2011 13:53:59 GMT (40kb,D)

Download:

- PDF
- Other formats

Current browse context:

stat.AP

< prev | next >

new | recent | 1105

Change to browse by:

stat

stat.ML

References & Citations

- NASA ADS

Bookmark(what is this?)

[v2] Tue, 27 Sep 2011 14:01:38 GMT (43kb,D)

Which authors of this paper are endorsers?

Link back to: arXiv, form interface, contact.