UNIVERSITEIT
GENT

Experimental Comparison of Some Multiple Imputation Methods from the R Package mice

Wim De Mulder

University of Ghent
wim.demulder@ugent.be

UNIVERSITEIT
GENT

Short resume of the presenter

Wim De Mulder studied computer science and law. He has a research background in artificial intelligence and statistics, with positions at Ghent University, KU Leuven, the Norwegian University of Science and Technology, and the Technical University of Eindhoven. His previous research concerned diverse applications, such as molecular biology, predictive maintenance and engineering. Currently, his focus is on the use of artificial intelligence in legal practice.

Introduction

UNIVERSITEIT GENT

Background

- Missing data are common in real-world application
- Multiple imputation (MI) is a useful imputation methodology
  - For each missing data point multiple plausible values are drawn from an imputation model
  - Accounts for uncertainty in the imputed values
  - Allows to construct confidence intervals around the imputed values

Purpose of the paper

Experimental comparison of some MI methods from the R package mice on some real-world data sets

IARIA

Considered data sets

UNIVERSITEIT GENT

1. Financial ratios
   - 5 ratios describing financial information on companies based in Belgium and Luxemburg
   - The ratios are considered to be predictive of bankruptcy
   - Time series from 2010 till 2019 for about 1 million companies
   - Lots of missing values (cf. table on the next slide)
2. HTRU2 data set
   - Benchmark data set from the UCI Machine Learning Repository
   - Sample of pulsar candidates collected during the High Time Resolution Universe Survey
   - About 18 000 instances and 9 attributes

IARIA

UNIVERSITEIT
GENT

Table: Considered financial ratios

Ratio index	Description	% missing values
Ratio 1	Return on total assets	59%
Ratio 2	Interest cover	63%
Ratio 3	Solvency ratio	59%
Ratio 4	Liquidity ratio	61%
Ratio 5	Operating revenue per employee	97%

Considered imputation methods

UNIVERSITEIT GENT

1. `mean`: Imputes the arithmetic mean of the observed data
2. `norm`: Calculates imputations for missing data by Bayesian linear regression
3. `lasso.norm`: Imputes missing normal data using lasso linear regression with bootstrap
4. `lasso.select.norm`: Imputes missing data using Bayesian linear regression following a preprocessing lasso variable selection step
5. `rf`: Imputes missing data using random forests

Evaluation methodology

UNIVERSITEIT GENT

- Introducing missing values
  - Predefined percentage of non-missing values are randomly set to missing
  - Financial ratios: 2% of the non-missing values
  - HTRU2 data set: varying percentage of the non-missing values, ranging from 0.5% to 20%
- Evaluation measures:
  - Average relative difference (ARD): evaluates goodness-of-fit
  - Interval score (IS): evaluates the constructed confidence intervals

IARIA

Example results

UNIVERSITEIT GENT

Table: Mean of average relative difference (ARD): financial ratios

	`mean`	`norm`	`rf`	`l.norm`	`l.s.norm`
Ratio 1	**2.06**	38.86	8.87	39.01	38.96
Ratio 2	32.58	74.32	**13.00**	73.58	74.21
Ratio 3	13.49	20.85	**3.02**	20.87	20.83
Ratio 4	4.79	2.23	**1.80**	2.19	2.26
Ratio 5	11.73	48.11	**4.78**	44.21	48.17

IARIA

UNIVERSITEIT GENT

Table: Average interval score (AIS): financial ratios

	`mean`	`norm`	`rf`	`l.norm`	`l.s.norm`
Ratio 1	794.83	342.44	**239.08**	342.45	342.28
Ratio 2	1977.51	646.24	**458.91**	646.89	648.45
Ratio 3	158.03	47.85	**28.38**	47.82	47.82
Ratio 4	1206.81	116.77	**98.35**	117.24	117.05
Ratio 5	51076.13	25009.93	**11109.64**	23059.73	24990.42

UNIVERSITEIT GENT

Table: Mean of average relative difference (ARD): HTRU2 data set

	mean	norm	rf	l.norm	l.s.norm
0.5%	8.02	4.02	**1.10**	3.81	3.77
1%	8.00	4.19	**0.62**	3.72	3.71
5%	22.90	14.12	**2.32**	14.78	17.19
10%	8.06	4.51	**0.76**	4.42	4.32
15%	15.28	11.03	**1.69**	10.37	9.49
20%	11.59	5.59	**0.96**	5.62	5.62

UNIVERSITEIT GENT

Table: Average interval score (AIS): HTRU2 data set

	`mean`	`norm`	`rf`	`l.norm`	`l.s.norm`
0.5%	602.37	43.20	**12.29**	39.77	40.12
1%	539.44	42.33	**12.45**	41.10	42.55
5%	578.09	44.64	**14.00**	43.29	44.87
10%	587.01	49.05	**14.43**	49.22	49.08
15%	599.02	55.21	**16.81**	54.46	55.32
20%	593.21	60.14	**17.89**	61.69	60.29

IARIA

Conclusion

UNIVERSITEIT GENT

- The `rf` method, which relies on random forests, is superior to the other imputation methods
- The performance of imputation methods may vary significantly according to the specific data points that are missing.
- The relative number of missing values might not be a determining factor for the performance of imputation methods, except if that relative number is extremely high.
- In general, the absolute number of non-missing values is probably of more significance for the accuracy of the imputed values.
- Confidence intervals should always be computed and taken into account. Only in this way the multiple imputation methodology is fully exploited, by providing a measure of uncertainty about the accuracy of the imputed values.

IARIA