Scalable Model-based Cascaded Imputation of Missing Data

Jacob Montiel[1] Jesse Read[2] Albert Bifet[1] Talel Abdessalem[1]

[1]LTCI, Télécom ParisTech, Université Paris-Saclay | [2]LIX, École Polytechnique, Université Paris-Saclay

Objectives

- A model-based method that casts the imputation process as a set of classification/regression tasks.
- Non-restrictive on the type of missing data to process, supports:
  - MAR and MCAR missing data mechanisms
  - Numerical and nominal data
  - Small to large data sets, including high dimensional data sets.
- A comprehensive evaluation of different imputation methods under multiple scenarios, identifying optimal operation and failure conditions.

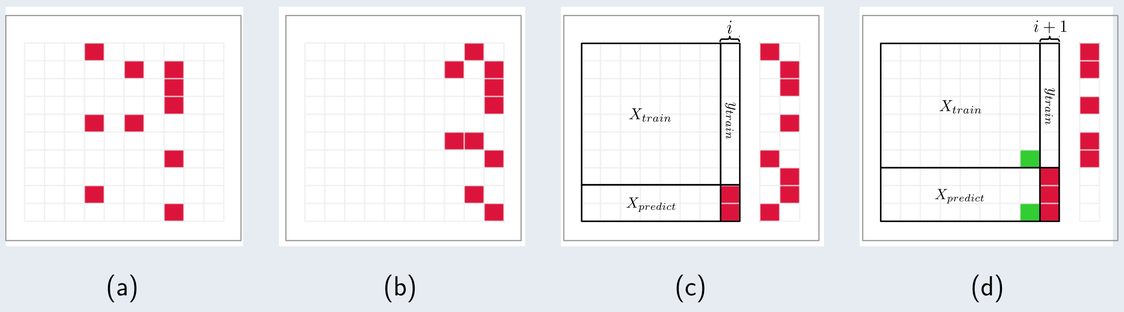
Missing data

- A common trait of real-world data.
- Can *negatively* impact the performance of supervised learning methods.
- On average, between 5% and 20% of data is missing [4].
- Missing data ratios between 5-15% require the usage of sophisticated methods while above 15% can compromise data interpretation [1].

Missing data mechanisms:

I. Missing Completely At Random (MCAR). Events behind missing values are independent of observable variables and the missing values themselves.
II. Missing At Random (MAR). Explained by observable variables.
III. Missing Not At Random (MNAR). Related to the value of the missing data itself.

Cascade Imputation (CIM)



(a) (b) (c) (d)

1. (a) Original data with missing values marked in red.
2. (b) After sorting attributes by count of missing values.
3. (c) Data in columns $0 \rightarrow i$ is re-arranged as a supervised learning problem. For each attribute i we train/apply a classification model if the attribute is nominal or a regression model if the attribute is numerical.
4. (c)-(d) Imputation iterations. Imputed values in green are used in following iterations. Repeated until the data set is complete.

Methodology

- Imputation Methods:
  - CONSTANT. Use a *constant* to fill missing values.
  - SIMPLE. Numerical values $\Leftarrow$ *mean* Nominal values $\Leftarrow$ *most-frequent*.
  - EMI (Expectation-Maximization Imputation) [3]. An iterative method with two steps. Expectation: values are imputed based on observed values. Maximization: imputed values are evaluated and updated if necessary according to the data distribution.
  - *k-Nearest Neighbor Imputation* (KNNI) [2] uses the neighborhood of a missing value to estimate the corresponding imputation value.
- Missing value ratios: 5%, 10%, 25%, 50
- Data set splits: 5%, 10%, 25%, 50%, 75% (scalability test)
- Missingness mechanisms: MAR, MCAR
- 10 files per *ratio-mechanism* and *split-ratio-mechanism* combinations
- 10-fold cross validation
- ML algorithms paired with CIM: Logistic Regression (CIM-LR), Random Forest (CIM-RF)
- Classifiers: Logistic Regression, Random Forest (Imputation-Classification Test), Extreme Gradient Boosting (Imputed vs Incomplete Test)

Data sets

Name	Domain	Instances	Attributes Num.	Attributes Nom.	Source
Adult	Demography	48,842	6	8	UCI Rep.
Census-IKDD	Demography	299,285	7	33	UCI Rep.
Music	Music	593	72	0	MEKA Rep.
Enron	Text	1,702	0	1,001	MEKA Rep.
Genbase	Biology	661	0	1,186	MULAN Rep.
Llog	Text	1,460	0	1004	MEKA Rep.
Medical	Text	978	0	1,449	MEKA Rep.
Scene	Image	2,407	294	0	MEKA Rep.
Yeast	Biology	2,417	103	0	MEKA Rep.
Covtype	Biology	581,012	10	44	UCI Rep.

Imputation-Classification Test

Classifier	Mech.	CIM-LR	CIM-RF	CONSTANT	EMI	KNNI	SIMPLE
Logistic Regression	MAR	13	**17**	14	1	2	12
	MCAR	15	**21**	5	0	4	10
Random Forest	MAR	16	11	13	1	2	**18**
	MCAR	13	**18**	7	0	3	16

Performance ranking focusing on *overall behavior* of imputation methods over multiple test configurations (larger numbers are better). Top performer is CIM-RF, followed by CIM-LR and SIMPLE. Sensitivity of EMI and KNNI to data set size and ratio of missing values results on imputation failure for most tests.

Scalability Test

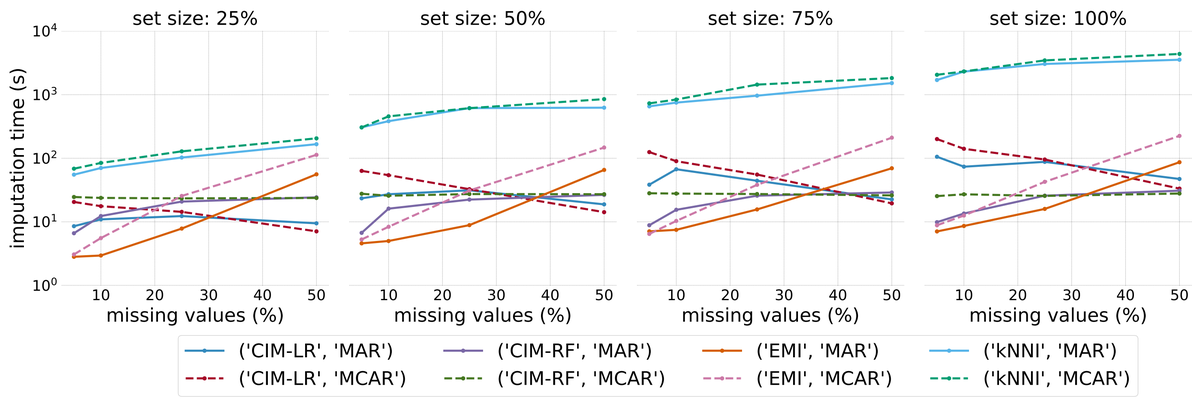
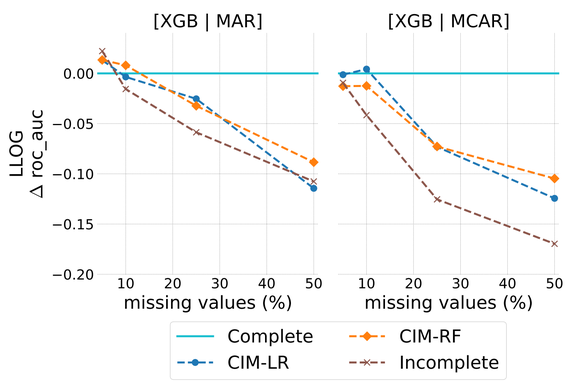


Figure 1: Scalability test results for CIM, EMI and KNNI on Adult data set.

CIM imputation time decreases with larger missing values ratios, while EMI and KNNI are weak against large data set sizes and ratio of missing values.

Imputed vs Incomplete Test



Algorithms that train on missing data are not guaranteed to create the best classifiers. Imputation methods shall be considered for ratios $> 5\%$.

Conclusion

CIM is an effective, robust and scalable imputation method. Test results show that CIM performs well over a wide range of test scenarios, overcoming limitations of popular imputation methods such as EMI and KNNI.

Implementing an imputation - classification pipeline is straightforward given that CIM does not require additional tools unrelated to the classification task.

[1] Edgar Acuña and Caroline Rodriguez. The Treatment of Missing Values and its Effect on Classifier Accuracy. *Classification, Clustering, and Data Mining Applications*, (1995):639–647, 2004.
[2] Gustavo E A P A Batista and Maria Carolina Monard. A study of k-nearest neighbour as an imputation method. *Frontiers in Artificial Intelligence and Applications*, 87:251–260, 2002.
[3] James Honaker, Gary King, and Matthew Blackwell. Amelia ii: A program for missing data. *Journal of Statistical Software*, 45(1):1–47, 2011. ISSN 1548-7660.
[4] Xiaoyuan Su, Russell Greiner, Taghi M Khoshgoftaar, and Amri Napolitano. Using Classifier-Based Nominal Imputation to Improve Machine Learning. *Advances in Knowledge Discovery and Data Mining, Pt I: 15th Pacific-Asia Conference, Pakdd 2011*, 6634(Mci):124–135, 2011.