

Imputation techniques on missing values in breast cancer treatment and fertility data

Xuetong Wu, Hadi Akbarzadeh Khorshidi, Uwe Aickelin, Zobaida Edib, Michelle Peate

University of Melbourne



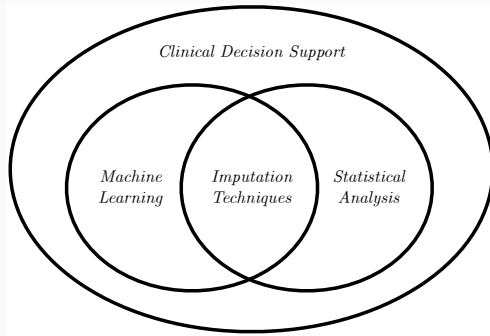
Table of contents

1. Introduction
2. Related Work
3. Experiments and Comparisons
4. Conclusion

Introduction

Introduction

- ▷ Clinical decision support using data mining techniques
- ▷ Find relationship between breast cancer (BC) treatment and chemotherapy-related amenorrhoea
- ▷ High missingness, negative impacts if handled improperly
- ▷ Imputation techniques → statistical analysis & machine learning methods



- Design a fertility calculation tool for BC patients
- Prepared dataset contains large amount of missing values
- How imputation improve the performance of classification?
- Imputation v.s. Classification

Research Question

1. Will imputation improve the classification performance?
2. Will imputation perform better by leveraging different datasets than using the single dataset?

- The work explores the impact of some notable imputation techniques using statistical and machine learning methods, and further evaluates the performance regarding the classification tasks on prediction of amenorrhoea after 12 months of breast cancer treatments.
- We examine whether the imputation across different datasets achieves significant improvements, even if the data has a large amount of missing values.

Related Work

Amenorrhea Status Prediction

- Chemotherapy-related amenorrhoea(CRA) can be caused by breast cancer treatment. The involvements to maintain the fertility options should be accessed before the treatment regularly and young women should be informed of the possibility of amenorrhoea or recovery of menstruation and contraceptive choices [7].
- Prediction of CRA involves complex factors such as age, lifestyle factors, knowledge, previous pregnancies, ovulation, history of previous medical and gynaecological diseases [4–7].
- The key challenge with fertility prediction is that the data usually contains substantial missing elements which adversely impact the prediction results. Imputation methods can help to accommodate this issue.

Missing Values Mechanism

- MCAR Missing Completely At Random* (MCAR) cases occur when the probability that an element missing is independent of the variable itself or any other related influences.
- MAR Missing At Random* (MAR) is the case such that the missing value is independent of the missing attribute itself but can be predicted from the observed responses.
- MNAR Missing Not At Random* (MNAR) is related to the missing feature itself and missing data can not be predicted only from the observed and missing entries from the database.

Imputation Techniques

- Deletion: Delete the cases that contains missing values
- Random: Imputation with the random values from observed ones
- Mean/Mode: Impute missing value with the mean/mode of the feature
- MICE: Multiple imputation using chained equations [1]
- EM: Expectation maximization algorithm [3]
- k NNI: Imputation with k nearest neighbour [9]
- RFI: Imputation with random forest for mixed type data [8]

Classifiers

SVM: Support vector machine classifier

RF: Random forest classifier

MLP: Multi-layer Perceptron classifier (DNN)

LR: Logistic Regression classifier

GNB: Gaussian Naive Bayesian classifier

Tree: Decision Tree

k NN: k nearest neighbour classifier

Experiments and Comparisons

Table 1: Data description for FoRECAST dataset, the dataset is split into six subsets regarding the sources. Observed feature here implies that the feature has at least one observation within the dataset, and missingness hinges on the observed features.

Data track	Instances	Observed Features	Categorical	Numerical	Missingness	Label
Track 0	725	19	19	0	8.6%	'Amen_ST12'
Track 1	280	36	36	0	9.1%	
Track 2	209	34	29	5	10.5%	
Track 3	154	20	19	1	22.2%	
Track 4	101	42	40	2	23.6%	
Track 5	96	47	43	4	18.3%	
Total	1565	87	76	11	72.5%	

Single Imputation

The imputation is implemented within the single dataset and missingness of each set is relatively low, missing values are under the assumption of MCAR or MAR.

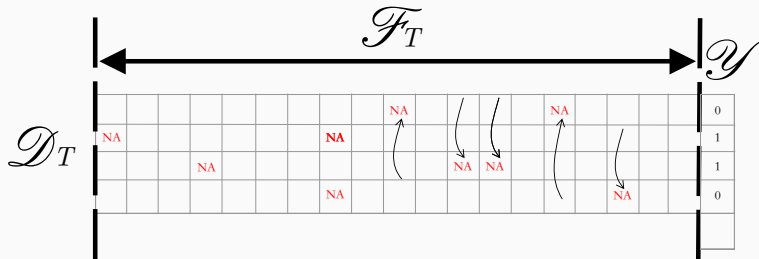
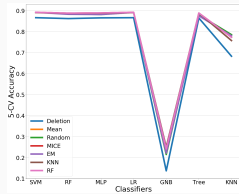
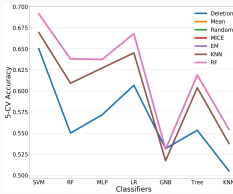


Figure 1: Single imputation

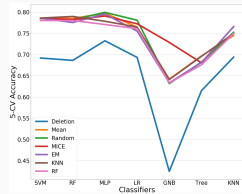
Single Imputation Results



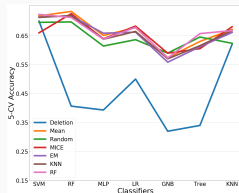
(a) Track 1



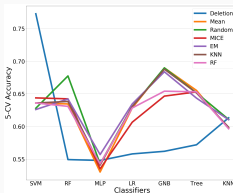
(b) Track 2



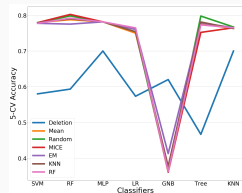
(c) Track 3



(d) Track 4



(e) Track 5



(f) Track 6

Figure 2: Results of prediction (Y axis) by seven classifiers(X axis) in terms of six sub-datasets, the imputation is performed on single sub-dataset.

Single Imputation Results

Table 2: Classification accuracy and its standard deviation in different imputed datasets (upper) and classifiers (lower), the results are averaged by seven classifiers (upper) and six datasets (lower), respectively, using 5-fold cross-validation. Highest accuracy and lowest standard deviation are highlighted in bold in each row.

Accuracy averaged by the results from different classifiers (single dataset)							
Data track	Deletion	Mean	Random	MICE	EM	KNN	RF
Track 1	0.736±0.253	0.781±0.0223	0.775±0.224	0.781±0.224	0.777±0.230	0.779±0.221	0.781±0.223
Track 2	0.567±0.045	0.621±0.055	0.620±0.055	0.620±0.055	0.620±0.055	0.601±0.055	0.620±0.055
Track 3	0.648±0.095	0.745±0.055	0.740±0.063	0.686±0.155	0.747±0.055	0.743±0.054	0.741±0.056
Track 4	0.469±0.134	0.665±0.045	0.656±0.055	0.669±0.053	0.650±0.045	0.655±0.045	0.665±0.045
Track 5	0.596±0.077	0.624±0.045	0.614±0.045	0.615±0.032	0.627±0.045	0.626±0.045	0.616±0.033
Track 6	0.605±0.077	0.716±0.141	0.719±0.145	0.714±0.141	0.719±0.145	0.718±0.137	0.716±0.146
Accuracy averaged by the results from different datasets (single dataset)							
Classifier	Deletion	Mean	Random	MICE	EM	kNNI	RFI
SVM	0.710±0.091	0.750±0.081	0.745±0.084	0.741±0.087	0.745±0.084	0.746±0.084	0.751±0.081
RF	0.608±0.141	0.745±0.089	0.741±0.092	0.747±0.090	0.738±0.088	0.739±0.095	0.742±0.089
MLP	0.635±0.151	0.714±0.120	0.708±0.121	0.711±0.117	0.701±0.129	0.709±0.117	0.711±0.124
LR	0.633±0.120	0.730±0.087	0.727±0.091	0.731±0.089	0.726±0.088	0.725±0.091	0.731±0.088
GNB	0.433±0.164	0.509±0.155	0.497±0.154	0.519±0.169	0.501±0.155	0.507±0.151	0.501±0.152
Tree	0.569±0.160	0.711±0.095	0.705±0.098	0.710±0.095	0.709±0.092	0.706±0.100	0.712±0.094
kNN	0.636±0.068	0.684±0.084	0.687±0.089	0.691±0.085	0.683±0.093	0.679±0.087	0.687±0.086

Discussion

- SVM and RF prediction models achieve the highest accuracy (74%), almost 24% percentage higher than Naïve Bayesian method (50%).
- The results show a commonly observed pattern that most imputation methods help improve the classification performance, except a few conditions.
- Elementary mean imputation can improve the performance to some extent and produce competitive results compared to other more complicated approaches.
- RF and MICE achieved the best performance and RF showed more robustness in prediction.
- The appropriate combination of imputation and classifier can lead to flexible and outstanding solutions for missing data.

Cross Imputation

We use a subset of dataset to impute missing values in other subsets and we call it as *cross imputation*.

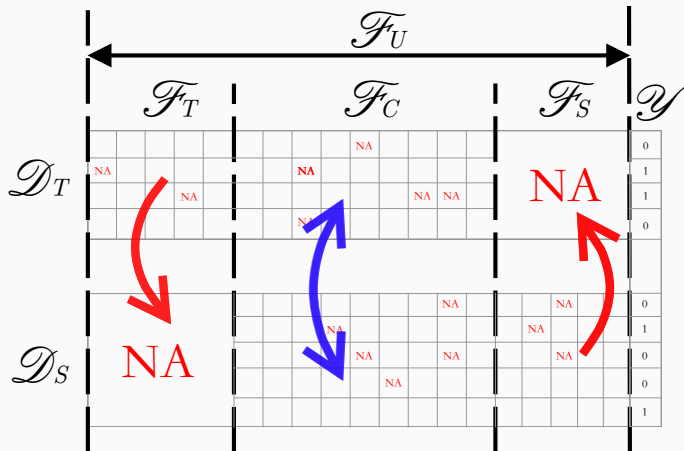
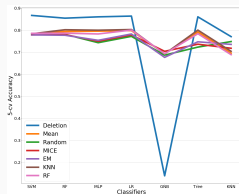
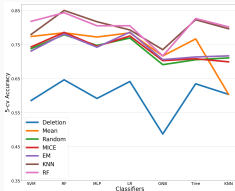


Figure 3: Cross imputation

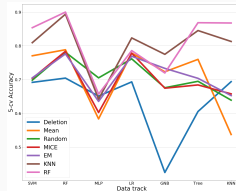
Cross Imputation Results



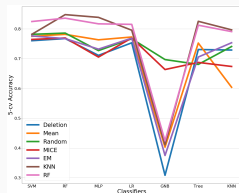
(a) Track 1



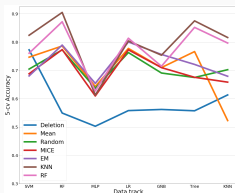
(b) Track 2



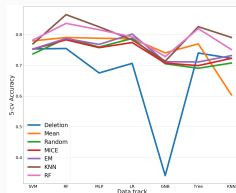
(c) Track 3



(d) Track 4



(e) Track 5



(f) Track 6

Figure 4: Results of prediction (Y axis) by seven classifiers(X axis) in terms of six sub-datasets, the imputation is performed on single sub-dataset.

Cross Imputation Results

Table 3: Classification accuracy and its standard deviation in imputed whole dataset utilizing cross imputation with different sub-datasets (upper) and classifiers (lower), the results are averaged by seven classifiers (upper) and six datasets (lower), respectively, using 5-fold cross-validation.

Accuracy averaged by the results from different classifiers (whole dataset)								
Data track	Deletion	Mean	Random	MICE	EM	KNN	RF	Missingness
Track 1	0.745±0.249	0.764±0.045	0.748±0.032	0.749±0.029	0.750±0.034	0.769±0.045	0.760±0.044	27.0%
Track 2	0.599±0.051	0.744±0.061	0.735±0.031	0.738±0.032	0.740±0.03	0.8±0.034	0.803±0.038	46.5%
Track 3	0.638±0.093	0.706±0.095	0.708±0.045	0.698±0.06	0.711±0.05	0.8±0.073	0.807±0.084	50.2%
Track 4	0.680±0.153	0.693±0.132	0.740±0.038	0.719±0.044	0.697±0.133	0.757±0.143	0.760±0.138	46.1%
Track 5	0.588±0.081	0.708±0.088	0.707±0.044	0.698±0.055	0.726±0.053	0.799±0.088	0.777±0.078	65.7%
Track 6	0.670±0.137	0.751±0.062	0.739±0.037	0.742±0.03	0.752±0.033	0.796±0.045	0.789±0.036	48.2%
Accuracy averaged by the results from different datasets (whole dataset)								
Classifier	Deletion	Mean	Random	MICE	EM	KNNI	RFI	
SVM	0.739±0.085	0.772±0.012	0.740±0.033	0.738±0.032	0.738±0.036	0.792±0.019	0.804±0.031	
RF	0.713±0.097	0.788±0.004	0.781±0.005	0.779±0.006	0.78±0.007	0.861±0.034	0.846±0.035	
MLP	0.665±0.110	0.724±0.081	0.720±0.041	0.695±0.065	0.714±0.051	0.756±0.092	0.751±0.078	
LR	0.703±0.094	0.783±0.008	0.770±0.009	0.775±0.003	0.785±0.013	0.800±0.012	0.802±0.011	
GNB	0.377±0.136	0.664±0.118	0.691±0.009	0.694±0.018	0.660±0.130	0.680±0.123	0.666±0.109	
Tree	0.688±0.101	0.768±0.012	0.695±0.016	0.699±0.02	0.717±0.015	0.832±0.023	0.827±0.027	
KNN	0.689±0.061	0.595±0.056	0.708±0.035	0.688±0.027	0.712±0.035	0.786±0.038	0.783±0.054	

Discussion

- Missing information is no longer under MAR or MCAR mechanism.
- Cross imputation provides a possible solution for initialising the missing values under MNAR.
- Imputation improves classification performance in most cases.
- Mean imputation is no longer as effective as some machine learning methods such as KNN and RF, which gained significant advantages and robustness.
- Cross imputation can even outperform single imputation in some cases.
- For the datasets of mixed types, RFI can handle the missing values well in both single and cross imputation.

Conclusion

Conclusion

- Results from various classifier with single imputation were similar, despite that results from Naïve Bayesian is much less precise.
- Even simple mean imputation can achieve good results.
- In general, RF and MICE are likely to be the best approaches within a small scale database.
- Mean imputation is no longer as effective as some machine learning methods such as KNN and RF, which gained significant advantages and robustness.
- Machine learning based imputation such as RF and KNN, which are probably the best choices with the least standard deviation for larger scale datasets.
- Overall, most of the imputation techniques show strong robustness and high efficiency in cross-dataset imputation with regard to high missingness, even outperform than single imputation cases.

References

- [1] AZUR, M. J., STUART, E. A., FRANGAKIS, C., AND LEAF, P. J. Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research* 20, 1 (2011), 40–49.
- [2] BARAKAT, M. S., FIELD, M., GHOSE, A., STIRLING, D., HOLLOWAY, L., VINOD, S., DEKKER, A., AND THWAITES, D. The effect of imputing missing clinical attribute values on training lung cancer survival prediction model performance. *Health information science and systems* 5, 1 (2017), 16.

- [3] GOLD, M. S., AND BENTLER, P. M. Treatments of missing data: A monte carlo comparison of rbhdi, iterative stochastic regression imputation, and expectation-maximization. *Structural Equation Modeling* 7, 3 (2000), 319–355.
- [4] JOHNSON, N., BAGRIE, E., COOMARASAMY, A., BHATTACHARYA, S., SHELLING, A., JESSOP, S., FARQUHAR, C., AND KHAN, K. Ovarian reserve tests for predicting fertility outcomes for assisted reproductive technology: the international systematic collaboration of ovarian reserve evaluation protocol for a systematic review of ovarian reserve test accuracy. *BJOG: An International Journal of Obstetrics & Gynaecology* 113, 12 (2006), 1472–1480.

- [5] LEE, S., KIL, W. J., CHUN, M., JUNG, Y.-S., KANG, S. Y., KANG, S.-H., AND OH, Y.-T. Chemotherapy-related amenorrhea in premenopausal women with breast cancer. *Menopause* 16, 1 (2009), 98–103.
- [6] LIEM, G. S., MO, F. K., PANG, E., SUEN, J. J., TANG, N. L., LEE, K. M., YIP, C. H., TAM, W. H., NG, R., KOH, J., ET AL. Chemotherapy-related amenorrhea and menopause in young chinese breast cancer patients: analysis on incidence, risk factors and serum hormone profiles. *PloS one* 10, 10 (2015), e0140842.

- [7] PEATE, M., MEISER, B., FRIEDLANDER, M., ZORBAS, H., ROVELLI, S., SANSOM-DALY, U., SANGSTER, J., HADZI-PAVLOVIC, D., AND HICKEY, M. It's now or never: fertility-related knowledge, decision-making preferences, and treatment intentions in young women with breast cancer—an australian fertility decision aid collaborative group study. *J Clin Oncol* 29, 13 (2011), 1670–1677.
- [8] STEKHOVEN, D. J., AND BÜHLMANN, P. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (2011), 112–118.
- [9] ZHANG, S. Nearest neighbor selection for iteratively knn imputation. *Journal of Systems and Software* 85, 11 (2012), 2541–2552.

Thanks! Questions?