Exploring the Effects of Data Distribution in Missing Data Imputation

Jastin Pompeu Soares[1], Miriam Seoane Santos[1], Pedro Henriques Abreu[1(✉)], Hélder Araújo[2], and João Santos[3]

[1] CISUC, Department of Informatics Engineering, University of Coimbra, Coimbra, Portugal
{jastinps,miriams}@student.dei.uc.pt, pha@dei.uc.pt
[2] ISR, Department of Electrical and Computer Engineering, University of Coimbra, Coimbra, Portugal
helder@isr.uc.pt
[3] IPO-Porto Research Centre (CI-IPOP), Porto, Portugal
joao.santos@ipoporto.min-saude.pt

Abstract. In data imputation problems, researchers typically use several techniques, individually or in combination, in order to find the one that presents the best performance over all the features comprised in the dataset. This strategy, however, neglects the nature of data (data distribution) and makes impractical the generalisation of the findings, since for new datasets, a huge number of new, time consuming experiments need to be performed. To overcome this issue, this work aims to understand the relationship between data distribution and the performance of standard imputation techniques, providing a heuristic on the choice of proper imputation methods and avoiding the needs to test a large set of methods. To this end, several datasets were selected considering different sample sizes, number of features, distributions and contexts and missing values were inserted at different percentages and scenarios. Then, different imputation methods were evaluated in terms of predictive and distributional accuracy. Our findings show that there is a relationship between features' distribution and algorithms' performance, and that their performance seems to be affected by the combination of missing rate and scenario at state and also other less obvious factors such as sample size, goodness-of-fit of features and the ratio between the number of features and the different distributions comprised in the dataset.

Keywords: Missing data · Data imputation · Data distribution

1 Introduction

Missing data imputation refers to the process of finding plausible values to replace those who are missing in a dataset and is a common data preprocessing technique applied in several fields [14]. Most often, imputation is performed using a brute force strategy, where a set of algorithms is used to impute all the features

W. Duivesteijn et al. (Eds.): IDA 2018, LNCS 11191, pp. 251–263, 2018.
https://doi.org/10.1007/978-3-030-01768-2_21

in a dataset. Then, the imputed datasets pass to the classification stage, where the imputation performance is evaluated through the classification error (CE) [1]. Although this is a standard approach to the missing data problem, it raises some important hitches: first, since all techniques must be implemented for all features, its computational cost is high; secondly, it assumes that the same technique should perform well for all or the great majority of features, which could be an over-assumption for features with different characteristics and finally, it uses the CE to evaluate the imputation quality, which for contexts other than classification, could be inappropriate. In general classification scenarios, the objective is to efficiently solve a classification problem, and therefore imputation is considered a required step to produce quality data. When imputation, rather than classification, is the focus, the use of CE is controversial. Some authors strongly defend that "imputation is not prediction" [22], and that the imputation method that minimises the classification error may produce biased estimates and affect the original data distribution.

Imputation methods should ideally be able to reproduce the true values in data – Predictive Accuracy (PAC) – and preserve the distribution of those true values – Distributional Accuracy (DAC) [6]. However, in the majority of imputation works, the nature of data (data distribution) is completely neglected and the above-mentioned properties are disregarded in favour of CE. Taking into account the distribution of data could be relevant to guide the choice of an appropriate imputation method: it considers the intrinsic characteristics of data and avoids the need to test a large set of methods for datasets where the features' distributions are known. Thus, studying the influence of data distribution in imputation presents a new challenge for missing data research and may provide a heuristic on the most appropriate imputation strategy for each feature in the study, allowing researchers to address missing data problems more easily and effectively.

This work follows from the initial research of Santos et al. [18], where authors showed that there was a relation between imputation methods and data distribution, when missing data is generated completely at random (MCAR mechanism). In this work, we extend their experimental set up to consider more datasets (15 datasets) and missing not at random (MNAR) mechanism, created in 6 different ways (scenarios T_1 to T_6, as will be explained in Sect. 3). Our experiments show that regardless of the missing data generation scenario, the imputation methods are in fact influenced by data distribution, with the exception of Support Vector Machine (SVM). Aside for SVM, that achieves the best PAC and DAC results for the great majority of distributions, Self-Organizing Map (SOM) is the overall winner in both metrics. However, the choice of the best imputation method depends also on the scenario and missing rate at state, besides other less obvious aspects as the Goodness of Fit (GoF), sample size and ratio of features per distribution.

The remainder of this document is structured as follows: Sect. 2 discusses related works regarding missing data imputation in several contexts. Sections

3 and 4 describe the experimental setup used in this work and report on the achieved results, while Sect. 5 presents the conclusions and suggests some possibilities for future work.

2 Related Work

Nanni et al. [13] compared the performance of standard imputation techniques (including MMimp and KNNimp) and their proposed imputation method for classification purposes, by generating missing values on 5 health related datasets at different rates (10–50%). The researchers concluded that their imputation techniques, based on clustering and random sub-spaces, present better behaviour than all the others (in terms of CE), achieving a satisfactory performance for MR greater than 30%. Aisha et al. [2] studied the effects of data imputation (including MMimp, KNNimp and SVMimp) on the classification of an incomplete health dataset (MR of 48%). SVMimp, along with Local Least Squares, outperformed the remaining techniques (in terms of CE). Rahman and Davis [15], investigated the classification performance of several imputation methods (such as SVMimp, MMimp, DTimp) using CE metrics, on a real incomplete medical dataset with 0–30% MR per feature. The results showed that all imputation methods based on machine learning improved the sensibility (and in some cases accuracy) of the classification task, in relation to MMimp. García-Laencina et al. [7] studied the influence of imputation (including KNNimp and SOMimp) on classification accuracy, using synthetic and real datasets. In this work, the authors start by evaluating the imputation quality using PAC (Pearson's r) and DAC (Kolmogorov–Smirnov distance) metrics, but just applied KNNimp (with different k values) on the first feature of synthetics datasets (MR 5–40%). However, this approach was discarded in favour of CE metrics, since the main objective of the experiments was to solve a classification problem. Rahman and Islam [16] propose imputation techniques based on DT and compare them in terms of PAC - coefficient of determination (R^2), Mean Squared Error (MSE) and Mean Absolute Error (MAE). DAC metrics are, however, neglected. This work used 9 real datasets from different contexts, where missing values were generated (1–10%). The proposed imputation techniques outperformed the others. Amiri and Jensen [3] introduced three imputation methods based on Fuzzy Rough Sets and compared their performances with 11 standard techniques (including KNNimp and SVMimp), in terms of RMSE (PAC analysis). In this work, the authors used 27 complete and real datasets from different contexts and inserted missing values varying from 5 to 30%. The simulations showed that SVMimp, KNNimp and the three proposed techniques obtained the best results.

In the above-mentioned works, imputation techniques are frequently evaluated in terms of CE, and the effects they may have in data distribution are most often ignored. Moreover, in these approaches, the same technique is used to impute all features, without considering the possibility that different features may be more properly imputed with different techniques. This work conducts a study on the influence of data distribution in missing data imputation, aiming

to assess how different imputation techniques perform across different feature distributions and missing generation types, extending the work of Santos et al. [18].

3 Experimental Setup

Our experimental setup consisted in 4 main stages: Data Collection, Distribution Fitting and Missing Data Generation, Data Imputation and Evaluation (Fig. 1).

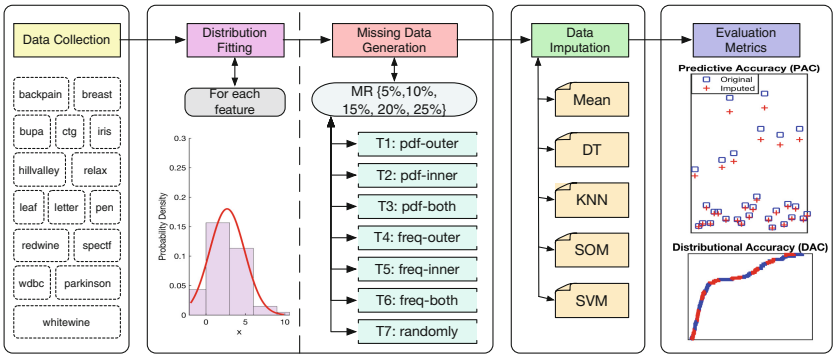


Fig. 1. Experimental Setup Architecture, comprising Data Collection, Distribution Fitting and Missing Data Generation, Data Imputation and Evaluation.

Data Collection comprised the selection of several publicly available datasets, from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml) and Kaggle Datasets (https://www.kaggle.com/datasets), attending to different contexts, sample sizes, number of features and number of different distributions (Table 1).

After the datasets were collected, the Distribution Fitting and Missing Data Generation follows. Each feature of each dataset is fitted against a comprehensive set of distributions (beta, birnbaum-saunders, exponential, extreme value, gamma, generalized extreme value, generalized pareto, inverse gaussian, logistic, loglogistic, lognormal, nakagami, normal, rayleigh, rician, t location-scale and weibull) and the Goodness of Fit (GoF) statistic is used to determine the distribution that best fits the data—GoF values vary from $-\infty$ (bad fit) to 1 (perfect fit). Then, based on the best fitting distribution, the probability density function (*pdf*) is determined and used to define several scenarios from which the missing values are introduced at different rates (5, 10, 15, 20 and 25%). Missing values are inserted following 7 distinct methods: the simplest method (T_7) consists on randomly selecting values to remove from each feature (MCAR mechanism); the remaining methods follow MNAR mechanism and are based on the probability density function (*pdf*-based methods: T_1 to T_3) and on the frequency distribution (*freq*-based methods: T_4 to T_6) of each feature. For each of these methods,

Table 1. Summary of datasets' characteristics.

Dataset	Context	Sample size	No. of features	No. of distributions (no. of features)	No. features/No. distributions^2
Backpain	Detect abnormal back pain	310	12	Beta(1), Gamma(2), Generalized Pareto(5) Normal(1), Nakagami(1), tLocationScale(2)	0.333
Breast	Identify breast carcinomas	106	9	Birnbaum-saunders(2), Generalized Extreme Value(4), Generalized Pareto(2), Lognormal(1)	0.563
Bupa	Detect alcoholism problems	345	6	Birnbaum-saunders(1), Exponential(1), Generalized Extreme Value(1), Inverse Gaussian(1), Loglogistic(2)	0.240
ctg	Detect pathologic fetal cardiotocograms	2126	21	Birnbaum-saunders(1), Gamma(4), Generalized Extreme Value(3) Generalized Pareto(2), Inverse Gaussian(1), Logistic(2) Normal(3), Nakagami(1), tLocationscale(2), Weibull(2)	0.210
Hillvalley	Detect hills and valleys	1212	100	Birnbaum-saunders(94), Generalized Extreme Value(6)	25
Iris	Distinguish between different types of iris plants	150	4	Extreme Value(1), Generalized Extreme Value(2), Inverse Guassian(1)	0.444
Leaf	Distinguish between different species of leafs	340	14	Beta(3), Birnbaum-saunders(1), Generalized Extreme Value(2) Generalized Pareto(5), Nakagami(1), Lognormal(1), Rayleigh(1)	0.286
Leaf	Identify the alphabet letters (A-Z)	5000	16	Exponential(1), Gamma(9), Generalized Pareto(2) Normal(2), Rayleigh(2)	0.640
Parkinson	Diagnose cases of parkinson's disease	195	22	Beta(1), Gamma(1), Generalized Extreme Value(14), Generalized Pareto(2), Inverse Gaussian(2), Loglogistic(1), Weibull(1)	0.449
Pen	Identify handwritten digits (0-9)	3498	16	Extreme Value(1), Gamma(2), Generalized Extreme Value(4) Generalized Pareto(1), Logistic(8)	0.640
Redwine	Classify red wine quality	1599	11	Birnbaum-saunders(2), Generalized Extreme Value(4) Logistic(1), Loglogistic(1), Nakagami(1), tLocationScale(2)	0.306
Relax	Distinguish between relaxed state and motor imagery state	182	12	Generalized Extreme Value(1), Logistic(3) Normal(1), tLocationScale(7)	0.750
Spectf	Detect abnormal SPECTF images	267	44	Extreme Value(30), Logistic(3), Weibull(11)	4.889
Wdbc	Diagnose breast cancer cases	569	30	Birnbaum-saunders(1), Gamma(5), Generalized Extreme Value(17), Generalized Pareto(1), Inverse Gaussian(1), Loglogistic(2), Lognormal(2), tLocationScale(1)	0.469
Whitewine	Classify white wine quality	4898	11	Generalized Extreme Value(4), Generalized Pareto(1) Loglogistic(3), Nakagami(2), tLocationScale(1)	0.440

the missing values are selected considering 3 different scenarios: removing from the inner areas, outer areas, or both. Inner and outer areas refer to high and low values of the *pdf* and *freq* histograms, respectively. Figure 2 depicts each of these methods and variations.

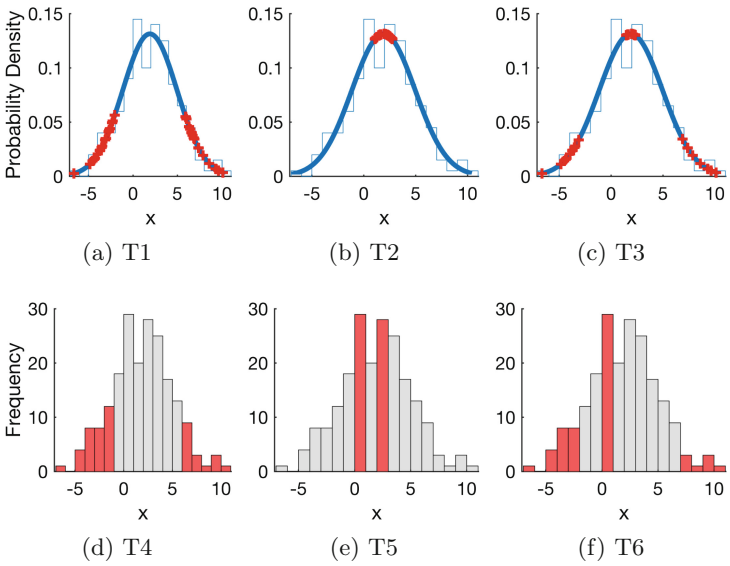


Fig. 2. Strategies for missing data generation: T_1 to T_3 are *pdf*-based methods while T_4 to T_6 are *freq*-based methods.

The Data Imputation stage considers the top five strategies used in recent works, attending also to different paradigms: statistical-based (Mean imputation - MMimp), tree-based models (Decision Trees - DTimp), neural networks-based (Self-Organizing Maps - SOMimp), similarity-based methods (k-Nearest Neighbours - KNNimp) and kernel-based methods (Support Vector Machines - SVMimp), which we briefly describe herein. MMimp is the most common and simple of imputation techniques: it imputes the missing values with the mean of the complete values on the respective features [8,13,19]. KNN imputes the incomplete patterns by finding its k nearest neighbours, found by minimising a similarity measure. Once those k neighbours are found, the missing values are imputed according to the type of feature [17]. The KNN implementation used in this work considers a weighted average of the k neighbours (1–20 neighbours) to determine the substitute value to impute. In DTimp, each incomplete feature is used as target, while the remaining features are used to fit the model: missing values are determined as if they were class labels [5]. SOMimp determines each incomplete pattern's Best Matching Unit (BMU) and imputes its missing values according to the BMU's weights on the incomplete features [11]. In this work, several network sizes were tested for SOMimp: 10–100 nodes. Support Vector

Machines can also be used for imputation (SVMimp), considering the feature to be imputed as the target. In this work, SVMimp was implemented considering both a linear (SVMlinear) and a gaussian (radial basis function, RBF) kernel (SVMrbf) [7]. For the linear kernel, we considered a value of $C = 1$, while for the gaussian kernel, different values of C and γ were tested ($1e^{-5}$ to $1e^{5}$, increasing by a factor of 10).

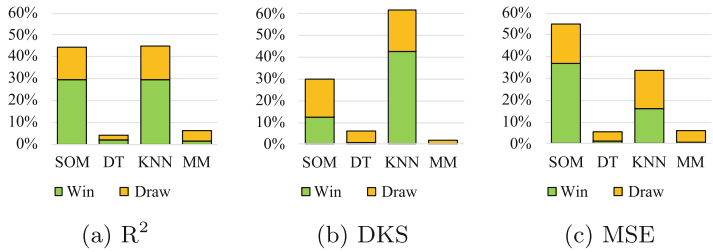
Finally, the quality of imputation is evaluated regarding two imputation properties proposed by Chambers [6]: Predictive Accuracy (PAC) and Distributional Accuracy (DAC). The former refers to a procedure's efficiency on retrieving the true values in data while the latter refers to its ability to preserve the original data distribution. PAC properties were assessed using the well-known coefficient of determination (R^2) and Mean Squared Error (MSE) [10] and DAC was assessed using the Kolmogorov–Smirnov distance (D_{KS}) [12]. R^2 provides a measure of the correlation between the original and imputed values (efficient imputations should have a value closer to 1), MSE measures the average squared deviation of the imputed values from the true values (values closer to 0 suggest more accurate imputations) and D_{KS} measures the distance between the cumulative distribution functions of the imputed values of a feature and its original values where better imputations are represented by smaller distance values.

4 Experimental Results and Discussion

Considering all imputation methods, our experiments have shown that SVMimp is the winning method for the great majority of distributions, with an overall ratio of victories over 80%, regarding both PAC and DAC metrics. Considering all distributions, SVMimp obtains the highest mean value for R^2 – 0.765 versus 0.723 obtained with the remaining methods – and the lowest mean values for MSE and D_{KS} – 0.015 and 0.106 versus 0.019 and 0.136 of the remaining methods, for the respective measures, showing that it is not affected by data distribution and surpassing the remaining methods. However, a preliminary analysis of the results indicated that, if SVMimp was not considered, the remaining methods performed differently across different distributions, metrics, scenarios and missing rates. Therefore, we have investigated how the remaining methods behave in different configurations.

Overall, KNNimp and SOMimp are responsible for the highest performance results, with a percentage of wins of 46.8% and 43.2%, respectively. Regarding each individual metric, this tendency is maintained for R^2 (Fig. 3a), although it is slightly different for D_{KS} and MSE: KNNimp is more appropriate to keep the data distribution (Fig. 3b), while SOMimp is responsible for the best MSE values (Fig. 3c).

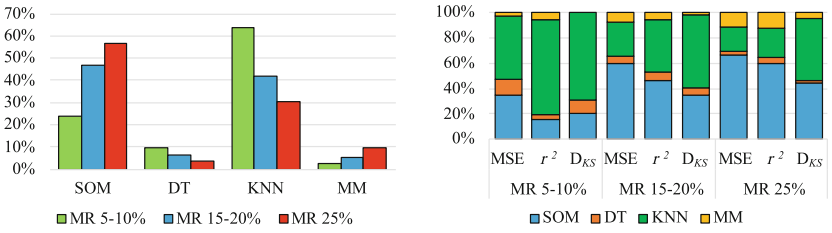
Figure 4a shows the victories and draws, altogether, for each range of considered missing rates (5–10, 15–20 and 25%). SOMimp and MMimp show a similar behaviour, where they surpass the other methods for increasing percentages of missing data. Contrariwise, DTimp and KNNimp tend to perform worse as the MRs increase. To further study this behaviour, Fig. 4b shows the overall victories and draws of each method, considering each specific metric (R^2, D_{KS} and



(a) R^2 (b) DKS (c) MSE

Fig. 3. Overall results (divided by wins and draws) for each metric: R^2, D_{KS} and MSE.

MSE). For low MRs (5–10%), KNN outperforms all other methods in terms of both PAC and DAC, being considered the most frequent winner in all metrics (50%, 75.2% and 68.6% for MSE, R^2 and D_{KS}, respectively). When the MR increases (15–20%), KNNimp loses its podium to SOMimp in terms of PAC (R^2 and MSE), though not DAC, where KNN appears as winner in 57.2% of times. When the missing rate increases to 25%, the previous behaviour is respected, although the differences between SOM and KNN are more accentuated. In terms of PAC, SOMimp's superiority becomes clearer (66.9% and 59.6% of wins for MSE and R^2), while KNNimp's dominance in terms of D_{KS} decreases to 49%.



(a) Divided by MRs. (b) Divided by MRs and metrics.

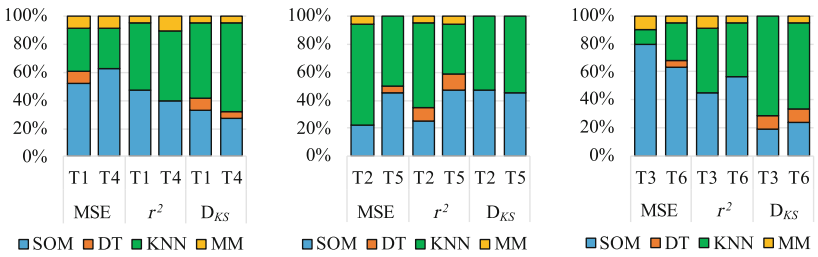
Fig. 4. Overall results (wins and draws altogether) for each imputation method, divided by MRs (a) and further specified by each metric (b).

The observed results are in agreement with the characteristics of the considered algorithms. Although MMimp is a rapid and simple solution to impute missing data, it is known to ignore the relations between the features, disturbing the original data variance [9]. As such, MMimp tends to have a poor performance compared to the other methods, in terms of DAC. Regarding KNNimp, previous works have shown that it has a robust behaviour even for large amounts of missing data [4,21]. The fact that it uses the information of the most similar cases rather than all the cases makes it superior to MMimp, being stronger in maintaining the distribution of data (DAC). DTimp is resilient to outliers and has the ability to cope with skewed distributions; however, the higher the amount of

missing data, the more difficult is to have a good decision tree to estimate the missing values [22]. SOM imputation somehow approximates a clustering solution, in the sense that the imputations are made in clusters, activation groups constituted by the k-closest BMUs of a given incomplete pattern. This type of mapping allows SOM to preserve the data topology, which is one of the factors that may contribute to its robust behaviour [20].

Out of these four methods, MMimp serves as a baseline, and behaved as expected, deteriorating the data distribution. DTimp does not seem to be a general good approach for imputation in terms of PAC and DAC: it estimates missing values based on the information of the remaining features and therefore it produces good estimates when the correlation is high. However, for low correlations between features it can lead to poor performances, which could be on the origin of its discouraging behaviour. Finally, imputation algorithms that approached a clustering-based solution (KNNimp and SOMimp) seem to be generally appropriate to keep the PAC and DAC properties of data: this fact could be related to the fact that both these methods properly address the similarity between patterns, using only resembling data points to impute the missing values.

Figure 5 specifies the overall victories and draws of each imputation by metric (MSE, R^2 and D_{KS}), for each scenario. It is clear that KNNimp achieves the best results for DAC, regarding all generation types. In terms of PAC, SOMimp seems to be the preferable approach for all scenarios except T_2, where the supremacy of KNN is noticeable both in terms of MSE and R^2.



(a) T1 versus T4. (b) T2 versus T5. (c) T3 versus T6.

Fig. 5. Comparison between analogous pairs: *freq*-based versus *pdf*-based generations types.

From Fig. 5 it is also possible to compare the analogous pairs of *freq*-based and *pdf*-based generation types. There are not relevant differences to point out, except for the imbalance between SOM's and KNN's results for PAC metrics in T_2 versus T_5 pairs. T_2 generation is most often better imputed with KNN for all metrics, with KNN gaining a clear advantage over SOM; in T_5, this gap is not so clear.

Since this work considers an extensive set of configurations (distributions, missing data rates, scenarios and metrics), summarising the conclusions to provide a clear heuristic is not a trivial process. Thus, we have decided to build a dataset including each existing variable from each studied dataset to analyse all the available information. Specifically, the produced dataset includes information on the name of distribution, missing rates, metrics, generation type, feature ratio, number of features, number of features with the same distribution included in the dataset, sample size, goodness-of-fit of the feature and the best imputation method, as the target class. An excerpt of such dataset is shown on Listing 1.1.

```
@relation LowLevelInfoT1T2T3T4T5T6T7

@attribute Distribution_class {Beta,BirnbaumSaunders,Exponential,
    ExtremeValue,Gamma,GeneralizedExtremeValue,GeneralizedPareto,
    InverseGaussian,Logistic,Loglogistic,Lognormal,Nakagami,Normal,
    Rayleigh,Weibull,tLocationScale}
@attribute MissingRate {5,10,15,20,25}
@attribute Metric_class {ksdistance,mse,pearson}
@attribute GenType_class {T1,T2,T3,T4,T5,T6,T7}
@attribute FeatureRatio numeric
@attribute FeatureNo numeric
@attribute SameFeature numeric
@attribute SampleSize numeric
@attribute GoF numeric
@attribute bestMethod_class {DT,KNN,Mean/Mode,SOM}

@data
Gamma,5,mse,T1,0.33333,12,2,310,0.91288,SOM
Gamma,5,pearson,T1,0.33333,12,2,310,0.91288,SOM
Gamma,5,ksdistance,T1,0.33333,12,2,310,0.91288,DT
```

Listing 1.1. Produced dataset regarding all the available information.

With the help of Waikato Environment for Knowledge Analysis (WEKA) software, we then started by analysing the simplest rules (ZeroR and OneR) that allowed a general classification of the data. ZeroR suggested classifying all instances as SOM (AUC of 0.5) and OneR used GoF to produce a larger set of rules for classification (AUC of 0.608). These results show that SOM is generally the overall winner for the great majority of configurations and suggest that GoF has a high discriminative power. Motivated by these results, we performed an attribute selection based on Information Gain, which revealed that GoF (0.229), Sample Size (0.165) and Feature Ratio (0.158) are the top three most discriminative features. We also ran a sequential forward selection to determine the subset of features that more accurately traduced the best imputation method for each input variable. This search returned a subset including the missing generation scenario (Generation Type), Sample Size and GoF, for which a 10-fold cross-validation of a C4.5 decision tree returned an average AUC of 0.725, decreasing just by 0.027 relatively to the AUC results including all information (0.752).

However, these features did not provide any insights regarding the different distributions. Therefore, we have tested several decision trees in order to obtain a model that included the most information possible, but without compromising the interpretability of the model: we looked for subsets of features that enabled a clear interpretation of a decision tree with a minimum performance drop, in order to produce meaningful rules. The subset of features that enable the most clear

decision tree is the distribution of the feature (Distribution), MR, the metric considered (Metric) and Generation Type, with a mean AUC of 0.675, showing a decrease of 0.077 relatively to the best AUC achieved (considering all features). Despite this drop in performance, this model allows the construction of general, heuristic rules that may be useful for researchers that know the distribution of data and want to select the best imputation method: an example branch of such a decision tree is shown in Fig. 6. From this heuristic, some imputation methods stood out for particular Distribution and Generation Types, e.g.: SOMimp for Birnbaum-saunders ($T_{1,2,3,4,5,6}$), Extreme Value ($T_{1,2,3,6}$) and Weibull ($T_{1,3,4,6}$); KNNimp for Logistic ($T_{1,2,3,4,5}$).

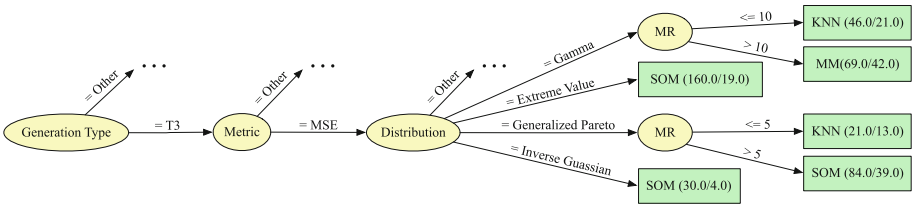


Fig. 6. Example of a branch of the decision tree generated from the considered subset of features. An example of a rule obtained by the presented model is: `Generation Type = T3 and Metric = MSE and Distribution = Gamma and Missing Rate <= 10: KNN(46,21)`

5 Conclusions and Future Work

This research follows from Santos et al. [18], where authors found a relation between data distribution and imputation quality, showing that the latter is influenced by the missing rate and the ratio of features per distribution, when missing data is generated completely at random. Herein, we extend the work of Santos et al. to more extreme setups, where missing values affect specific areas of features' frequency histograms and probability density functions. To this end, a set of comprehensive experiments were conducted in order to study the effect of several data distributions on well-known imputation algorithms. We collected several datasets with different characteristics, fitted the data to determine the best distribution that describes each feature and then inserted missing data in 7 different approaches (T_1 to T_7). After the insertion of missing values, five imputation methods were used to reproduce the original values and the results were evaluated in terms of PAC and DAC metrics.

From the results gathered we can summarise the following conclusions:

- SVMimp is the winning method for nearly all distributions in both PAC and DAC metrics, unaffected by data distribution;
- Overall, imputation algorithms that followed clustering-based solutions (KNNimp and SOMimp) seem to be generally appropriate to keep the PAC and DAC properties;

- KNNimp is more appropriate in terms of DAC and SOMimp seems preferable in terms of PAC;
- KNNimp outperforms all methods regarding both PAC and DAC metrics for MRs $< 15\%$, However, for MRs $\geq 15\%$ SOMimp is generally the best approach for PAC, though for DAC, KNNimp still maintains its superiority.

With more detail on the heuristic analysis we have the following conclusions:

- Overall, SOMimp is the most robust approach across several scenarios;
- GoF, Sample Size, Feature Ratio and Generation Type seem to be relevant features to determine appropriate imputation algorithms, although they do not provide insights regarding the different distributions;
- It was possible to obtain a clear decision tree model that allows the extraction of general rules comprising Generation Type, Metric, Distribution and MR;
- SOMimp is the most appropriate method for Birnbaum-saunders, Extreme Value and Weibull distributions. Logistic distributions tend to be better imputed with KNNimp.

There are several directions for future work. One is the extension of this methodology for datasets comprising also discrete features, fitting discrete distributions and investigating how the studied imputation techniques perform in each scenario. Also, from a classification perspective, it would be interesting to assess whether the best imputation techniques regarding PAC and DAC metrics would also achieve good results in terms of classification error. An ongoing work is focused on a sensibility analysis of SVMimp, studying the best set of parameters that achieve high PAC and DAC results and looking for the absolute most missing data rate for which SVMimp is still able to maintain the original data values and distribution.

Acknowledgments. This article is a result of the project NORTE-01-0145-FEDER-000027, supported by Norte Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 Partnership Agreement, through the European Regional Development Fund (ERDF).

References

1. Abreu, P.H., Santos, M.S., Abreu, M.H., Andrade, B., Silva, D.C.: Predicting breast cancer recurrence using machine learning techniques: a systematic review. ACM Comput. Surv. (CSUR) **49**(3), 52 (2016)
2. Aisha, N., Adam, M.B., Shohaimi, S.: Effect of missing value methods on bayesian network classification of hepatitis data. Int. J. Comput. Sci. Telecommun. **4**(6), 8–12 (2013)
3. Amiri, M., Jensen, R.: Missing data imputation using fuzzy-rough methods. Neurocomputing **205**, 152–164 (2016)
4. Batista, G.E., Monard, M.C.: A study of k-nearest neighbour as an imputation method. HIS **87**(251–260), 48 (2002)
5. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A.: Classification and Rregression Trees. CRC Press, Boca Raton (1984)

6. Chambers, R.: Evaluation criteria for statistical editing and imputation, national statistics methodological series no. 28. University of Southampton (2001)
7. García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R.: Pattern classification with missing data: a review. Neural Comput. Appl. **19**(2), 263–282 (2010)
8. García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R.: Classifying patterns with missing values using multi-task learning perceptrons. Expert Syst. with Appl. **40**(4), 1333–1341 (2013)
9. Howell, D.C.: The treatment of missing data. The Sage Handbook of Social Science Methodology, pp. 208–224. Sage Publications, Thousand Oaks (2007)
10. Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., Kolehmainen, M.: Methods for imputation of missing values in air quality data sets. Atmos. Enviro. **38**(18), 2895–2907 (2004)
11. Kohonen, T.: Self-Organizing Maps. Springer, Berlin (1995)
12. Lopes, R.H.: Kolmogorov-smirnov test. International Encyclopedia of Statistical Science, pp. 718–720. Springer, New York (2011)
13. Nanni, L., Lumini, A., Brahnam, S.: A classifier ensemble approach for the missing feature problem. Artif. Intell. Med. **55**(1), 37–50 (2012)
14. Pigott, T.D.: A review of methods for missing data. Educ. Res. Eval. **7**(4), 353–383 (2001)
15. Rahman, M.M., Davis, D.N.: Fuzzy unordered rules induction algorithm used as missing value imputation methods for k-mean clustering on real cardiovascular data. In: Proceedings of the World Congress on Engineering I, pp. 391–394 (2012)
16. Rahman, M.G., Islam, M.Z.: Missing value imputation using decision trees and decision forests by splitting and merging records: two novel techniques. Knowledge-Based Syst. **53**, 51–65 (2013)
17. Santos, M.S., Abreu, P.H., García-Laencina, P.J., Simão, A., Carvalho, A.: A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients. J. Biomed. Inf. **58**, 49–59 (2015)
18. Santos, M.S., Soares, J.P., Henriques Abreu, P., Araújo, H., Santos, J.: Influence of data distribution in missing data imputation. In: Artificial Intelligence in Medicine, pp. 285–294. Springer International Publishing, Cham (2017)
19. Sivapriya, T., Kamal, A.N.B., Thavavel, V.: Imputation and classification of missing data using least square support vector machines-a new approach in dementia diagnosis. Int. J. Adv. Res. Artif. Intell. **1**(4), 29–33 (2012)
20. Sorjamaa, A., Corona, F., Miche, Y., Merlin, P., Maillet, B., Séverin, E., Lendasse, A.: Sparse linear combination of soms for data imputation: application to financial database. In: Príncipe, J.C., Miikkulainen, R. (eds.) WSOM 2009. LNCS, vol. 5629, pp. 290–297. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-02397-2_33
21. Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., Altman, R.B.: Missing value estimation methods for dna microarrays. Bioinformatics **17**(6), 520–525 (2001)
22. Van Buuren, S.: Flexible Imputation of Missing Data. CRC Press, Boca Raton (2012)