Influence of Data Distribution in Missing Data Imputation

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

[1] Department of Informatics Engineering, Faculty of Sciences and Technology, CISUC, University of Coimbra, Coimbra, Portugal
{miriams,jastinps}@student.dei.uc.pt, pha@dei.uc.pt

[2] Department of Electrical and Computer Engineering, Faculty of Sciences and Technology, ISR, 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. Dealing with missing data is a crucial step in the preprocessing stage of most data mining projects. Especially in healthcare contexts, addressing this issue is fundamental, since it may result in keeping or loosing critical patient information that can help physicians in their daily clinical practice. Over the years, many researchers have addressed this problem, basing their approach on the implementation of a set of imputation techniques and evaluating their performance in classification tasks. These classic approaches, however, do not consider some intrinsic data information that could be related to the performance of those algorithms, such as features' distribution. Establishing a correspondence between data distribution and the most proper imputation method avoids the need of repeatedly testing a large set of methods, since it provides a heuristic on the best choice for each feature in the study. The goal of this work is to understand the relationship between data distribution and the performance of well-known imputation techniques, such as Mean, Decision Trees, k-Nearest Neighbours, Self-Organizing Maps and Support Vector Machines imputation. Several publicly available datasets, all complete, were selected attending to several characteristics such as number of distributions, features and instances. Missing values were artificially generated at different percentages and the 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, although some factors must be taken into account, such as the number of features per distribution and the missing rate at state.

Keywords: Missing data · Machine learning imputation · Data distribution · Healthcare contexts

A. ten Teije et al. (Eds.): AIME 2017, LNAI 10259, pp. 285–294, 2017.
DOI: 10.1007/978-3-319-59758-4_33

1 Introduction

In healthcare classification scenarios, the main goal is to provide strong classification results, whereas imputation is considered a necessary pre-processing step to achieve such goal [4]. Therefore, imputation is often evaluated using the classification error (CE): the method that minimizes the CE is considered the best. The use of CE is however controversial, in the sense that the imputation method that minimizes the classification error might produce biased estimates and affect the original data distribution, especially if the same method is used for all different types of features' distributions [11]. Furthermore, using the same method for all features raises two main issues: first, all techniques must be implemented for all features, which increases the number of necessary simulations and consequently computational cost; secondly, imputation is performed based on the assumption that the same technique should perform well for the great majority of features, which could be an over assumption, since different features may benefit the most from different imputation techniques, particularly if different missing rates are taken into account. Studying the influence of data distribution in imputation provides a heuristic on the most appropriate imputation strategy for each feature in the study, avoiding the need of testing a large set of methods.

In this work, we aim to assess which imputation techniques can efficiently reproduce the true, original values in data, without causing a distortion in their distribution, which can be evaluated by Predictive Accuracy (PAC) and Distributional Accuracy (DAC) metrics, respectively. Furthermore, we intend to investigate whether there is a relationship between the imputation methods and a particular distribution. Our study focuses on the best techniques for data imputation across several different distributions, in terms PAC and DAC, rather than CE. To achieve this goal, we have selected several complete healthcare datasets comprising features with different data distributions, and artificially generated missing data in all of them at several rates (5, 10, 15, 20 and 25%). Then the missing values are imputed with the methods most commonly used in related works: Mean imputation, Decision Trees (DT), k-Nearest Neighbours (KNN), Self-Organizing Maps (SOM) and Support Vector Machines (SVM) imputation. Our experiments show that the imputation methods are in fact influenced by data distribution, with the exception of SVM, that does not seem to be affected. Aside for SVM, that achieves the best PAC and DAC results for all distributions, SOM is overall winner in both metrics. However, the choice of the best imputation method depends also on the number of features per distribution and the missing rate at state.

The remainder of the manuscript is organized as follows: Sect. 2 presents some works that studied imputation for classification purposes. Sections 3 and 4 describe the setup used in this work and report on the experimental results, while Sect. 5 presents the conclusions and suggests some directions for future work.

2 Related Work

Addressing missing data to increase data quality for classification purposes is a standard procedure in a plethora of contexts, including healthcare. Nanni et al. [8] compared several imputation approaches (including Mean and KNN imputation) by randomly generating missing data at several rates, and used classification-related metrics such as accuracy (1-CE) and Area Under the ROC Curve (AUC) to evaluate the quality of imputation. Kang [7] also generated missing values in complete datasets, at several missing ratios. They compare their approach with other well-known imputation methods (also including Mean imputation and KNN), and the results were evaluated using accuracy. Aisha et al. [1] study the effects of several imputation techniques (including Mean, KNN and SVM imputation) on Bayesian Network classification of datasets with missing data, and evaluate the results also using accuracy. García-Laencina et al. [4] studied the influence of imputation (including KNN and SOM imputation) on the classification accuracy, using synthetic and real datasets. In this work, the authors start by measuring the quality of imputation using PAC (Pearson's coefficient and mean squared error) and DAC (Kolmogorov-Smirnov distance) metrics. However, this analysis in only performed for KNN imputation, and immediately discarded in favor of CE metrics, since the main objective is to solve a classification problem. Rahman and Islam [10] present two imputation techniques based on DT and compare them in terms of their predictive accuracy (PAC), using the Pearson's correlation coefficient, root mean squared error (RMSE) and mean absolute error (MAE) as performance indicators. DAC metrics are, however, completely disregarded. In what concerns healthcare contexts in particular, García-Laencina et al. [3] also compared the performance of standard imputation algorithms (including Mean and KNN imputation) on the survival prediction of breast cancer patients. The results were evaluated in terms of sensitivity, specificity, accuracy and AUC. Rahman and Davis [9] studied the influence of Mean, DT, KNN and SVM imputation on the survival prediction of cardiovascular patients, evaluating the quality of imputation also classification-related metrics (sensitivity, specificity and accuracy). Jerez et al. [6] use imputation (including Mean, KNN and SOM) to predict breast cancer recurrence in a real incomplete dataset, evaluating the results in terms of AUC. In all the previously mentioned works, imputation techniques are frequently evaluated in terms of CE, and the effects they may have in data distribution are ignored. Furthermore, all features are imputed with the same technique, without considering the possibility that some techniques may perform differently for different features. We herein conduct a study on the influence of data distribution in missing data imputation, aiming to assess how different imputation techniques perform across different feature distributions, which to the extent of our knowledge, as never been performed.

3 Methodology

This works comprised four main stages: Data Collection, Missing Data Generation, Data Imputation and Evaluation Metrics.

3.1 Data Collection

The first stage of this work consisted in choosing several publicly available datasets, all without missing values: Bupa Liver Disorders Dataset (*bupa*), Breast Tissue Dataset (*breast*), Cardiotocography Dataset (*ctg*), Haberman's Survival Dataset (*hsd*), Wisconsin Diagnostic Breast Cancer Dataset (*wdbc*), Parkinsons Dataset (*parkinson*) and Lower Back Pain Symptoms Dataset (*backpain*). All datasets were collected from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml), except for the latter, retrieved from Kaggle Datasets (https://www.kaggle.com/datasets). We have chosen only complete datasets composed exclusively of continuous features so that both the influence of different data distributions and missing rates could more efficiently studied. Table 1 summarizes the datasets' characteristics in what concerns their context, sample size, number of features and number of different distributions comprised in the data. In terms of data distributions, these datasets are somewhat heterogeneous, with the most common distributions being generalized extreme value (all 7 datasets), generalized pareto (6 datasets) and gamma distributions (4 datasets). We have also included the ratio of variables per distribution for each dataset (Ratio). Ratio is estimated as $\frac{\text{No. of features}}{\text{No. of distributions}^2}$, so that a greater weight is given to the number of distributions comprised in the dataset.

3.2 Missing Data Generation

Before generating missing values, each dataset's features were fitted against several standard continuous distributions and the distribution of each feature is saved for posterior analysis when assessing the imputation results (Table 1). Missing data was randomly inserted at several rates (5, 10, 15, 20 and 25%) for each feature in the dataset. Therefore, for each of the datasets, 5 different versions exist, one for each considered missing percentage.

3.3 Data Imputation

In this section, each imputation technique is briefly explained, with particular emphasis on the implementation details. **Mean imputation** is the most common of imputation techniques [5]. For continuous data, the missing values are replaced with the mean of the observed cases on each respective feature. In **k-Nearest Neighbours** (KNN), the incomplete patterns are imputed according to the values of their k closest neighbours on the missing features: mode for discrete data and the mean or a weighted average for continuous data [6], which is used in this work. Our implementation considers a range of k from 1 to 20

Table 1. Summary of datasets' characteristics.

Dataset	Context	Sample size	No. of features	Ratio	No. of distributions (no. of features)
bupa	Detect alcoholism problems	345	6	0.240	Generalized extreme value (1) Logistic (1), exponential (1) Loglogistic (2), lognormal (1)
breast	Identify breast carcinomas	106	9	0.360	Birnbaum-saunders (2) Generalized extreme value (1) Generalized pareto (2) Rayleigh (1), inverse gaussian (3)
ctg	Detect pathologic fetal cardiotocograms	2126	21	0.583	Generalized extreme value (5) Generalized pareto (10), gamma (1) Logistic (1), weibull (3), nakagami (1)
hsd	Predict 5-year survivability after breast cancer surgery	306	3	0.333	Generalized extreme value (1) Generalized pareto (1) Nakagami (1)
wdbc	Diagnose breast cancer cases	569	30	0.469	Generalized extreme value (16) Generalized pareto (2), gamma (1) Birnbaum-saunders (2), exponential (1) Inverse gaussian (1), loglogistic (1) Lognormal (4)
parkinson	Diagnose cases of parkinson disease	195	22	0.344	Generalized extreme value (9) Generalized pareto (5), gamma (1) Beta (1), inverse gaussian (1) Normal (1), weibull (1), lognormal (3)
backpain	Detect abnormal back pain	310	12	0.245	Generalized extreme value (2) Generalized pareto (4), gamma (2) Beta (1), birnbaum-saunders (1) Logistic (1), rayleigh (1)

closest neighbours and the Heterogeneous Euclidean-Overlap Metric (HEOM) as distance measure between patterns [12]. In **DT imputation**, each incomplete feature must be used as target: the remaining features are used as training data, to fit the model, and missing values are determined as if they were class labels. For this work, only regression trees are constructed, given the nature of all our features. In **Self-Organizing Maps** (SOM), each incomplete pattern is imputed according to its Best Matching Unit (BMU), its most similar unit in the SOM map. Several map configurations were tested: from 10 to 100 nodes. **Support Vector Machines** (SVM) are currently the state-of-the-art algorithms in pattern recognition, due to their good trade-off between the model's complexity, generalization and quality of fitting the training data, and have proven to perform well for missing data imputation [4]. In this work, only regression SVMs were used for imputation: in particular, we have implemented several Radial Basis Function (RBF) SVMs, with different values of C and γ (both from 1e^{-5} to 1e^{5}, increasing by a factor of 10).

3.4 Evaluation Metrics for Missing Data Imputation

The metrics used in this work concern mainly two aspects: Predictive Accuracy (PAC) and Distributional Accuracy (DAC) [2]. PAC relates to the efficiency of an imputation technique to retrieve the true values in data, while DAC represents the technique's ability to preserve the distribution of those true values. For PAC assessment, two measures were used: Pearson Correlation Coefficient (Pearson's r) and Mean-Squared Error (MSE). For DAC assessment, the Kolmogorov-Smirnov distance (D_{KS}) was implemented. Considering a complete feature x, and its imputed version $\hat{x}$, Pearson's r provides a measure of the correlation between the two, and is given by $r = \frac{\sum_{i=1}^{n}(x_i - \bar{x}_i)(\hat{x}_i - \bar{\hat{x}}_i)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x}_i)^2 \sum_{i=1}^{n}(\hat{x}_i - \bar{\hat{x}}_i)^2}}$, where an efficient imputation technique should have a value close to 1. MSE is traduced by $\frac{1}{n}\sum_{i=1}^{n}(x_i - \widehat{x_i})^2$ and measures the difference between the imputed and original values of a given feature j, the average square deviation of $\hat{x}_i$ from the true values x_i, for all n values of a feature j. In this case, values closer to 0 traduce a better imputation. Finally, D_{KS} is given by $\max(\|F_x - F_{\hat{x}}\|)$, where F_x and $F_{\hat{x}}$ are the empirical cumulative distribution functions of x and $\hat{x}$, respectively. Smaller distance values represent better imputations.

4 Experimental Results and Discussion

Considering all five imputation methods (Mean, DT, KNN, SOM and SVM), the results clearly show that SVM is the winning method for all distributions (see Total and Total SVM in Table 2). For all metrics, SVM outperforms the remaining methods, with a maximum total mean MSE, Pearson's r and D_{KS} of 0.014, 0.993 and 0.01, respectively, versus the 0.039, 0.98 and 0.13 achieved by the remaining methods. Moreover, SVM does not seem to be affected by data distribution, with good performance indicators across all distributions. However, a preliminary analysis of our simulation results suggested that this was not the case for the remaining methods, which lead us to investigate them more closely, and further divide our analysis in particular ranges of missing data. Therefore, Table 2 also presents the winning methods with respective means and standard-deviations in several missing rate scenarios (5/10, 15/20 and 25%), and summarizes the number of victories and draws of each method. Note that for 25% missing rate, some methods do not show a mean/standard deviation, which happens in distributions included in only two datasets and where the methods tie (each wins in one dataset, and the presented value refers to the result achieved for that dataset). In what concerns PAC results, although DT and KNN may outperform or match SOM's results for low percentages of missing data (5–10%), SOM is generally the best approach for percentages above 10%. In terms of DAC, KNN and SOM have similar results for missing percentages between 5% and 20%. Nevertheless, for percentages higher that 20%, SOM is the method that better preserves the original data distribution. Due to space constraints, it is not possible to show the results for each dataset and distribution, but we provide a more detailed discussion for certain distributions in

Table 2. Simulation results by distribution: means and standard-deviations are shown for the winning methods regarding each distribution, metric and missing percentage (n.a - not applicable).

		Distributions			
Metric	Scenario	**beta**	**bimbaumsaunders**	**gamma**	**generalized extreme value**
MSE	5%-10%	SOM[0.054 ± 0.049]	SOM[0.015±0.017]	SOM[0.023±0.012]	SOM[0.06±0.067]
	15%-20%	SOM[0.143 ± 0.011]	SOM[0.049±0.029]	DT[0.063±0.017]/SOM[0.087±0.027]	SOM[0.086±0.059]
	25%	KNN[0.022]/SOM[0.149]	SOM[0.091±0.053]	SOM[0.098±0.027]	SOM[0.151±0.106]
	Total	KNN[0.141±0.099]/SOM[0.099±0.059]	SOM[0.043±0.041]	SOM[0.069±0.04]	SOM[0.085±0.077]
	Total SVM	SVM [0.071±0.054]	SVM[0.029±0.044]	SVM[0.031±0.024]	SVM[0.05±0.0078]
Pearson	5%-10%	KNN[0.982±0.004]/SOM[0.97±0.035]	SOM[0.993±0.008]	SOM[0.988±0.006]	DT[0.989±0.008]/SOM[0.969±0.035]
	15%-20%	SOM[0.926±0.006]	SOM[0.976±0.015]	DT[0.968±0.009]/SOM[0.956±0.014]	SOM[0.955±0.044]
	25%	KNN[0.881]/SOM[0.923]	SOM[0.954±0.028]	SOM[0.949±0.014]	SOM[0.919±0.06]
	Total	KNN[0.939±0.051]/SOM[0.943±0.03]	SOM[0.978±0.021]	SOM[0.965±0.021]	SOM[0.956±0.042]
	Total SVM	SVM [0.964±0.028]	SVM[0.988±0.017]	SVM[0.984±0.012]	SVM[0.974±0.042]
D_{KS}	5%-10%	KNN[0.013±0.003]	SOM[0.012±0.005]	SOM[0.013±0.005]	KNN[0.015±0.009]
	15%-20%	KNN[0.038±0.005]	SOM[0.027±0.012]	DT[0.027±0.01]	KNN[0.029±0.013]
	25%	SOM[0.052±0.006]	SOM[0.037±0.013]	SOM[0.027±0.003]	KNN[0.028±0.004]
	Total	KNN[0.026±0.014]/SOM[0.039±0.023]	SOM[0.023±0.014]	SOM[0.019±0.007]	KNN[0.025±0.014]
	Total SVM	SVM[0.024±0.01]	SVM[0.015±0.01]	SVM[0.013±0.007]	SVM[0.016±0.009]
Metric	Scenario	**generalized pareto**	**logistic**	**rayleigh**	**inverse gaussian**
MSE	5%-10%	DT[0.027±0.023]/SOM[0.032±0.036]	DT[0.053±0.049], KNN[0.033±0.028]/SOM[0.031±0.0001]	SOM[0.025±0.018]	DT[0.008±0.006]/KNN[0.008±0.004]
	15%-20%	SOM[0.06±0.06]	SOM[0.083±0.052]	DT[0.068±0.022]	SOM[0.037±0.023]
	25%	SOM[0.075±0.059]	SOM[0.099±0.052]	SOM[0.132±0.085]	SOM[0.062±0.047]
	Total	SOM[0.064±0.065]	SOM[0.072±0.047]	SOM[0.060±0.068]	SOM[0.039±0.033]
	Total SVM	SVM[0.048±0.057]	SVM[0.061±0.063]	SVM[0.049±0.047]	SVM[0.014±0.026]
Pearson	5%-10%	DT[0.983±0.015]/SOM[0.984±0.018]	DT[0.973±0.025]/KNN[0.984±0.014]/SOM[0.985±0.001]	SOM[0.988±0.009]	DT[0.996±0.003]/KNN[0.996±0.002]
	15%-20%	SOM[0.969±0.031]	SOM[0.958±0.038]	DT[0.966±0.011]	SOM[0.982±0.011]
	25%	SOM[0.962±0.031]	SOM[0.949±0.027]	SOM[0.932±0.046]	SOM[0.968±0.025]
	Total	SOM[0.968±0.034]	SOM[0.963±0.024]	SOM[0.969±0.409]	SOM[0.98±0.017]
	Total SVM	SVM[0.977±0.029]	SVM[0.969±0.033]	SVM[0.976±0.024]	SVM[0.993±0.013]
D_{KS}	5%-10%	KNN[0.013±0.008]	KNN[0.009±0.004]	SOM[0.014±0.007]	DT[0.014±0.007]
	15%-20%	KNN[0.026±0.011]/SOM[0.024±0.007]	SOM[0.018±0.007]	KNN[0.032±0.013]	DT[0.018±0.004]/SOM[0.023±0.005]
	25%	KNN[0.038±0.013]	SOM[0.023±0.009]	DT[0.039]/SOM[0.047]	SOM[0.03±0.007]
	Total	KNN[0.02±0.012]/SOM[0.022±0.011]	SOM[0.018±0.006]	SOM[0.026±0.016]	DT[0.018±0.009]
	Total SVM	SVM[0.016±0.009]	SVM[0.012±0.006]	SVM[0.023±0.016]	SVM[0.015±0.009]
Metric	Scenario	**exponential**	**loglogistic**	**lognormal**	**nakagami**
MSE	5%-10%	SOM[0.049±0.027]	DT[0.035±0.011]	DT[0.015±0.012]/KNN[0.019±0.014]/SOM[0.052±0.04]	KNN[0.031±0.037]
	15%-20%	SOM[0.139±0.1]	KNN[0.064±0.001]/SOM[0.119±0.015]	SOM[0.091±0.06]	KNN[0.175±0.032]/SOM[0.041±0.013]
	25%	KNN[0.218]/SOM[0.038]	SOM[0.15±0.018]	SOM[0.152±0.074]	KNN[0.191] e SOM[0.062]
	Total	SOM[0.086±0.078]	SOM[0.101±0.056]	SOM[0.091±0.063]	KNN[0.121±0.086]/SOM[0.048±0.013]
	Total SVM	SVM [0.019±0.025]	SVM[0.06±0.044]	SVM[0.031±0.026]	SVM[0.072±0.081]
		Distributions			
Pearson	5%-10%	SOM[0.975±0.014]	DT[0.982±0.006]	SOM[0.974±0.018]	KNN[0.984±0.019]
	15%-20%	SOM[0.917±0.053]	KNN[0.968±0.0004]/SOM[0.939±0.008]	SOM[0.953±0.032]	KNN[0.908±0.018]/SOM[0.98±0.006]
	25%	SOM[0.933±0.067]	SOM[0.922±0.01]	SOM[0.92±0.04]	MEAN[0.899]/SOM[0.968]
	Total	SOM[0.947±0.046]	SOM[0.948±0.029]	SOM[0.953±0.034]	KNN[0.946±0.046]/SOM[0.976±0.006]
	Total SVM	SVM [0.99±0.012]	SVM[0.969±0.023]	SVM[0.984±0.013]	SVM[0.962±0.043]
D_{KS}	5%-10%	DT[0.006±0.001]/KNN[0.0145]/SOM[0.0116]	DT[0.006±0.0004]	KNN[0.015±0.007]/SOM[0.015±0.004]	DT[0.004]/KNN[0.007]/SOM[0.016±0.009]
	15%-20%	KNN[0.023±0.016]/SOM[0.018±0.004]	KNN[0.029±0.002]	SOM[0.025±0.011]	KNN[0.02±0.008]/SOM[0.028±0.002]
	25%	KNN[0.026]/SOM[0.028]	KNN[0.0319]/SOM[0.0378]	SOM[0.034±0.007]	SOM[0.035±0.006]
	Total	KNN[0.022±0.011]/SOM[0.019±0.007]	KNN[0.024±0.01]/SOM[0.026±0.016]	SOM[0.024±0.011]	KNN[0.016±0.01]/SOM[0.026±0.01]
	Total SVM	SVM [0.01±0.004]	SVM[0.016±0.011]	SVM[0.015±0.009]	SVM[0.011±0.012]
		Distributions		No. of vitories and draws per algorithm		
Metric	Scenario	**weibull**	**normal**	**SOM**	**KNN**	**DT**
MSE	5%-10%	DT[0.014±0.012]	KNN[0.0114] e SOM[0.036]	6/4	1/4	2/4
	15%-20%	SOM[0.041±0.013]	DT[0.0883]/SOM[0.163]	9/4	0/2	0/3
	25%	SOM[0.053±0.002]	SOM[0.124]	11/3	0/3	0
	Total	SOM[0.039±0.015]	SOM[0.108±0.065]	12/2	0/2	0
	Total SVM	SVM[0.024±0.023]	SVM[0.072±0.066]	n.a.	n.a.	n.a.
Pearson	5%-10%	DT[0.993±0.006]	KNN[0.994]/SOM[0.982]	5/5	1/4	2/4
	15%-20%	SOM[0.979±0.007]	DT[0.955]/SOM[0.915]	9/4	0/2	1/2
	25%	SOM[0.973±0.001]	SOM[0.935]	12/2	0/1	0
	Total	SOM[0.98±0.008]	SOM[0.944±0.034]	12/2	0/2	0
	Total SVM	SVM[0.988±0.012]	SVM[0.963±0.034]	n.a.	n.a.	n.a.
D_{KS}	5%-10%	KNN[0.005±0.002]/SOM[0.013±0.004]	KNN[0.01]/SOM[0.015]	3/5	4/5	2/2
	15%-20%	SOM[0.021±0.005]	DT[0.031]/KNN[0.031]/SOM[0.041]	4/5	4/4	1/2
	25%	KNN[0.026]/SOM[0.024]	SOM[0.026]	7/5	2/3	0/1
	Total	KNN[0.013±0.008]/SOM[0.018±0.006]	SOM[0.027±0.013]	6/5	2/5	1/0
	Total SVM	SVM[0.01±0.006]	SVM[0.019±0.007]	n.a.	n.a.	n.a.

what follows. For birnbaum-saunders datasets, SOM was always chosen as the best approach regarding all metrics. For datasets with a considerable number of features following the generalized extreme value distribution (*wdbc*: 16, *parkinson*: 9 and *ctg*: 5) and considering the range 5–10% of missing data, DT achieves the best results for in terms of PAC, although KNN achieves better results in terms of DAC. When the missing percentage increases, SOM is then considered the best approach in both metrics. Nevertheless, for datasets where only one variable of this type exists (*hsd* and *breast*), KNN outperforms or match SOM's results in both PAC and DAC metrics, for all missing rates. Dataset *bupa*, also with one variable of this type, seems to be an exception, with SOM achieving better results in all metrics, except when the missing rate increases (25%), where KNN is considered the best approach. Datasets *backpain*, *ctg* and *bupa* have one variable following the logistic distribution, where for small percentages of missing data, DT and KNN are feasible approaches. As the missing rate increases, only *bupa* includes KNN as best approach, while the remaining are better imputed with SOM. Dataset *bupa* seems to be a special case, where results are somewhat variable with increasing rates of missing values. This fact could be due to the ratio of features per distribution of this dataset (see Table 1). In fact, in a total of 6 features, *bupa* includes 5 different distributions, which causes it to have the lowest feature per distribution ratio (0.240). Intrigued by these results of *bupa*, we have further compared the overall MSE, Pearson's r and D_{KS} results for datasets with the lowest (*bupa* and *backpain*) and highest (*wdbc* and *ctg*) ratio of features per distribution, where a particular distribution is present in only one feature: exponential and logistic distributions (see Table 1). In the case of logistic distribution, PAC results of *backpain* and *bupa* differ from *ctg*: a mean MSE of 0.1/0.12 versus 0.04 and a mean Pearson's r of 0.95/0.94 versus the 0.98, respectively. Regarding DAC, all datasets are similar (maximum difference of 0.01). For the exponential distribution, the results follow the same trend: a mean MSE of 0.025/0.147 and Pearson's r of 0.99/0.92 for *wdbc*/*bupa*. DAC results are practically the same, with a difference of 0.005. This suggests that, when a particular distribution is present in only one feature, datasets with a low ratio of features per distribution (*backpain*: 0.245, *bupa*: 0.240), are more challenging than datasets with a higher ratio (*wdbc*: 0.469, *ctg*: 0.583), in what concerns retrieving the true values in data. However, imputation algorithms are able to considerably preserve the data distribution in both cases.

5 Conclusions and Future Work

Our results show that SVM is the winning method for all distributions in both PAC and DAC metrics. Aside for SVM, SOM is generally the best approach in terms of PAC when the missing rates increases above 10%, although for DAC its superiority its only noticeable for percentages higher that 20%. Regarding particular distributions, SOM was the best approach for birnbaum-saunders distributions in all considered missing percentages. In datasets with a great number of features following a generalized extreme value distribution, DT and SOM are the best approaches in terms of PAC, in 5–10% and 15–20% ranges of missing data, respectively. Furthermore, PAC metrics seem to be affected by the ratio of features per distribution, when a particular distribution is present in only one feature. Lower ratios generally achieve worst PAC results, although the data distribution is not significantly affected (DAC results are similar for both cases). There are several directions for future work the authors would like to address. To the extent of authors' knowledge, this approach has never been applied in imputation studies for healthcare contexts in particular or other subjects in general. Therefore, its application for other contexts and other data distributions is yet to be addressed. The extension of this methodology for discrete features, fitting discrete distributions and investigating how the studied imputation techniques perform in this case, could also be a possibility for future work. An ongoing work is the evaluation of the proposed approach in more extreme setups, where missing values are not generated completely at random, but rather affecting specific areas of features' probability density functions. Finally, from a classification perspective, it would also be interesting to study whether the best imputation techniques regarding PAC and DAC metrics also achieve reasonable results in terms of classification error.

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. 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)
2. Chambers, R.: Evaluation Criteria for Statistical Editing and Imputation. National Statistics Methodological Series No. 28. University of Southampton, Southampton (2001)
3. García-Laencina, P.J., Abreu, P.H., Abreu, M.H., Afonso, N.: Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values. Comput. Biol. Med. **59**(2015), 125–133 (2015)
4. 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)
5. García-Laencina, P.J., Sancho-Gómez, J.L., Figueiras-Vidal, A.R.: Classifying patterns with missing values using multi-task learning perceptrons. Expert Syst. Appl. **40**(4), 1333–1341 (2013)
6. Jerez, J.M., Molina, I., García-Laencina, P.J., Alba, E., Ribelles, N.: Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif. Intell. Med. **50**(2), 105–115 (2010)
7. Kang, P.: Locally linear reconstruction based missing value imputation for supervised learning. Neurocomputing **118**, 65–78 (2013)
8. Nanni, L., Lumini, A., Brahnam, S.: A classifier ensemble approach for the missing feature problem. Artif. Intell. Med. **55**(1), 37–50 (2012)

9. 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, vol. 1, pp. 391–395 (2012)
10. Rahman, M.G., Islam, M.Z.: Missing value imputation using decision trees and decision forests by splitting and merging records: two novel techniques. Knowl.-Based Syst. **53**, 51–65 (2013)
11. Van Buuren, S.: Flexible Imputation of Missing Data. CRC Press, Boca Raton (2012)
12. Wilson, D.R., Martinez, T.R.: Improved heterogeneous distance functions. J. Artif. Intell. Res. **6**, 1–34 (1997)