Pattern Recognition Letters 136 (2020) 111–119

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

How distance metrics influence missing data imputation with k-nearest neighbours☆

Miriam Seoane Santos[a,c,*], Pedro Henriques Abreu[a], Szymon Wilk[b], João Santos[c,d]

[a] CISUC, Department of Informatics Engineering, University of Coimbra, Coimbra 3030-790, Portugal
[b] Institute of Computing Science, Poznan University of Technology, Poznan 60-965, Poland
[c] Medical Physics, Radiobiology and Radiation Protection Group, IPO Porto Research Centre (CI-IPOP), Porto 4200-072, Portugal
[d] Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto, Porto 4050-313, Portugal

ARTICLE INFO

Article history:
Received 19 March 2019
Revised 8 May 2020
Accepted 27 May 2020
Available online 28 May 2020

MSC:
41A05
41A10
65D05
65D17

Keywords:
Missing Data
Data Imputation
k-nearest neighbours
Distance Functions
Heterogeneous Data
Imbalanced Data

ABSTRACT

In missing data contexts, k-nearest neighbours imputation has proven beneficial since it takes advantage of the similarity between patterns to replace missing values. When dealing with heterogeneous data, researchers traditionally apply the HEOM distance, that handles continuous, nominal and missing data. Although other heterogeneous distances have been proposed, they have not yet been investigated and compared for k-nearest neighbours imputation. In this work, we study the effect of several heterogeneous distances on k-nearest neighbours imputation on a large benchmark of publicly-available datasets.

1. Introduction

Dealing with Missing Data (MD) is crucial since absent observations compromise the reliability of models. Among several approaches, data imputation stands out as a way of preserving the existing information while replacing the absent with plausible values. k-nearest neighbours (KNN) imputation is a popular data imputation technique due to several properties: it takes advantage of the similarity between patterns to produce accurate estimates [1], it is a non-parametric method, which does not require any assumptions on the data or missing mechanism [2], it imputes absent information with already-occurring values in data ($k = 1$), and has proven to preserve the data distribution [3]. Although KNN has proven to be an efficient approach across different domains, it is largely dependent on its underlying distance function. By default, this technique uses the Euclidean distance, although it only applies when data is continuous.

In the literature, there are other distance functions that handle both continuous and nominal data (heterogeneous distance functions) and, in addition, deal with missing observations. Traditional distance functions include HEOM and HVDM [1] which are frequently applied in heterogeneous data contexts, such as healthcare domains [4–6]. Although other heterogeneous distances have been proposed, they have not yet been investigated for imputation purposes.

In this work, we perform a comprehensive search for heterogeneous distance functions that handle missing data and couple them with KNN for imputation. We study their behaviour for increasing percentages of missing data (5%, 10%, 20%, and 30%), aiming to answer two main research questions:

- *Do distance metrics significantly affect KNN imputation?*
- *Is there a distance more beneficial for some datasets?*

Other studies have investigated the influence of KNN or KNN-based algorithms on imputation and classification problems [7–9], although considering only the Euclidean distance or continuous features, poorly handling nominal features (e.g., applying the same distance as for continuous features) or neglecting missing data (us-

☆ Handle by Associate Editor: Shiguang Shan.

* Corresponding author.
E-mail address: miriams@dei.uc.pt (M.S. Santos).

https://doi.org/10.1016/j.patrec.2020.05.032
0167-8655/

ing complete datasets). However, as increasing research has come to acknowledge, handling heterogeneous data poses a more complex challenge for machine learning algorithms and should be addressed adequately [10]. Therefore, the main difference with our study is that we address all problems simultaneously, handling heterogeneous data with missing values with appropriate heterogeneous distance functions that account for missing observations. To the authors knowledge, no study has yet investigated the impact of different distance functions on the imputation of heterogeneous data and its effect on classification results, which constitutes the novelty and contribution of this work. Furthermore, we explore distance measures never before studied for imputation purposes, such as SIMDIST and MDE, extend MDE to handle nominal data and explore further HEOM and HVDM redefinitions, which constitute additional contributions. Additionally, this study may also provide interesting insights for general applications of heterogeneous distances with missing data such as classification algorithms operating with distances among patterns (e.g., instance-based learning, neural networks with radial basis functions, self-organising maps [11]), graph mining and clustering applications [12,13], and other fields in Pattern Recognition (e.g., Imbalanced Data research), where a plethora of oversampling approaches rely on the computation of distances to generate new synthetic data.

2. Heterogeneous distance functions for missing data

All distance functions considered in this work measure the distance between two patterns $\mathbf{x}_A$ and $\mathbf{x}_B$ through a sum of their individual distances in each j-th feature, $d_j(x_{Aj}, x_{Bj})$ as $D(\mathbf{x}_A, \mathbf{x}_B) = \sqrt{\sum_{j=1}^{p} d_j(x_{Aj}, x_{Bj})^2}$. However, they differ on the computation of individual d_j distances, as explained in what follows.

2.1. HEOM: Heterogeneous Euclidean-Overlap Metric

The Heterogeneous Euclidean-Overlap Metric (HEOM) distance [1], considers the normalised euclidean distance for continuous features, d_N (Eq. (2)), and an overlap metric for nominal features, d_O (Eq. (3)). However, d_O and d_N are only computed if both input values, x_{Aj} and x_{Bj} are available; otherwise, if either of them is missing, $d_j(x_{Aj}, x_{Bj})$ is defined as 1.

$$d_j(x_{Aj}, x_{Bj}) = \begin{cases} 1, & \text{if } j \text{ is missing in } x_{Aj} \text{ or } x_{Bj}, \\ d_O, & \text{if } j \text{ is a nominal feature}, \\ d_N, & \text{if } j \text{ is a continuous feature} \end{cases} \tag{1}$$

$$d_N(x_{Aj}, x_{Bj}) = \frac{\left|x_{Aj} - x_{Bj}\right|}{max(x_j) - min(x_j)} \tag{2}$$

$$d_O(x_{Aj}, x_{Bj}) = \begin{cases} 0, & \text{if } x_{Aj} = x_{Bj} \\ 1, & otherwise \end{cases} \tag{3}$$

2.2. HVDM: Heterogeneous Value Difference Metric

Similarly to HEOM, the Heterogeneous Value Difference Metric (HVDM) [1], defines d_j computation depending on the type of j: if both x_{Aj} and x_{Bj} are observed (Eq. (4)), d_{vdm} is used for nominal features (Eq. (5)) while d_{diff} is used for continuous features (Eq. (6)).

$$d_j(x_{Aj}, x_{Bj}) = \begin{cases} 1, & \text{if } j \text{ is missing in } x_{Aj} \text{ or } x_{Bj}, \\ d_{vdm}, & \text{if } j \text{ is a nominal feature}, \\ d_{diff}, & \text{if } j \text{ is a continuous feature} \end{cases} \tag{4}$$

The computation of d_{vdm}, as shown in (5), requires information on the class targets of each pattern $\mathbf{x}_i$, herein referred to as c_i. Thus, d_{vdm} is computed as a sum over all C, the number of possible classes in the problem domain. $N_{x_{Aj},c}$ is the number of patterns in $\mathbf{X}$ that have value x_{Aj} in feature j and class target c, while $N_{x_{Aj}}$ is the number of patterns in $\mathbf{X}$ that have value x_{Aj} in feature j (the same is derived for x_{Bj}).

$$d_{vdm}(x_{Aj}, x_{Bj}) = \sqrt{\sum_{c=1}^{C} \left| \frac{N_{x_{Aj},c}}{N_{x_{Aj}}} - \frac{N_{x_{Bj},c}}{N_{x_{Bj}}} \right|^2} \tag{5}$$

In turn, similarly to HEOM, the continuous features are scaled by d_{diff}, considering 4 standard deviations (σ) of x_j.

$$d_{diff}(x_{Aj}, x_{Bj}) = \frac{\left|x_{Aj} - x_{Bj}\right|}{4\sigma_{x_j}} \tag{6}$$

2.3. HEOM-REDEF and HVDM-REDEF: redefinitions of HEOM and HVDM

Juhola and Laurikkala propose a redefinition for both HEOM and HVDM in what concerns the treatment of missing values [14]. Missing values are considered as "special values" and therefore HEOM and HVDM should be adjusted accordingly, following Eq. (7). In this work, we refer to these two redefinitions as HEOM-REDEF and HVDM-REDEF.

$$d_j(x_{Aj}, x_{Bj}) = \begin{cases} 1, & \text{if } j \text{ is missing only on } x_{Aj} \text{ or } x_{Bj}, \\ 0, & \text{if } j \text{ is missing in both } x_{Aj} \text{ and } x_{Bj} \end{cases} \tag{7}$$

2.4. HVDM-SPECIAL: an extension of HVDM

Inspired by the idea of considering a missing value as a "special value" [15], we consider another redefinition for HVDM, herein referred to HVDM-SPECIAL. Missing values are considered an "extra" nominal category and d_{vdm} is applied in the case that only x_{Aj} or only x_{Bj} are missing and j is nominal (Eq. (8)).

$$d_j(x_{Aj}, x_{Bj}) = \begin{cases} 0, & \text{if } x_{Aj} \text{ and } x_{Bj} \text{ are both missing}, \\ 1, & \text{if } x_{Aj} \text{ or } x_{Bj} \text{ are missing and } j \text{ is continuous}, \\ d_{vdm}, & \text{if } x_{Aj} \text{ or } x_{Bj} \text{ are missing and } j \text{ is nominal} \end{cases} \tag{8}$$

2.5. SIMDIST: similarity for heterogeneous data

Belanche and Hernandéz proposed an heterogeneous similarity function as a way to incorporate prior knowledge into a neural network [16], as defined by Eq. (9), where S_{ABj} represents the similarity between two patterns according to feature j.

$$S_{ABj} = \begin{cases} \frac{1}{2}, & \text{if either } x_{Aj} \text{ or } x_{Bj} \text{ are missing}, \\ z\left(\frac{s_{ABj}}{s_j}\right), & \text{if both } x_{Aj} \text{ and } x_{Bj} \text{ are observed} \end{cases} \tag{9}$$

s_{ABj} is an intermediate similarity distance between x_{Aj} and x_{Bj} and is determined according to the type of j; s_j represents the mean similarity among all patterns according to j and z is a normalisation function, $z(a) = \frac{a}{a+1}$. For nominal features, s_{ABj} is defined by Eq. (10), where P_{lj} in the fraction of patterns that takes value x_{lj} for feature j.

$$s_{ABj} = \begin{cases} 0, & \text{if } x_{Aj} \neq x_{Bj}, \\ 1 - P_{lj}, & \text{if } x_{Aj} = x_{Bj} \end{cases} \tag{10}$$

For continuous features, s_{ABj} is determined by Eq. (11), where $\max(x_j)$ and $\min(x_j)$ are the maximum and minimum values observed in j, respectively.

$$s_{ABj} = 1 - \frac{\left|\, x_{Aj} - x_{Bj} \,\right|}{max(x_j) - min(x_j)} \tag{11}$$

In Eq. (9), S_{ABj} is assumed to be $\frac{1}{2}$ when x_{Aj} or x_{Bj} are missing which is the equivalent of replacing the missing similarity between x_{Aj} or x_{Bj} by the mean similarities of all patterns according to j. Lastly, the individual similarities S_{ABj} are transformed to individual distances $D_{ABj} = 1 - S_{ABj}$ and aggregated to obtain $D(\mathbf{x}_A, \mathbf{x}_B)$.

2.6. Mean Euclidean Distance

Mean Euclidean Distance (MD_E) was proposed to handle incomplete data in clustering algorithms [17]. Three possibilities for comparing two values of a given attribute j are given. When both values are known, their distance is defined as the standard euclidean distance (Eq. (12)). When either x_{Aj} or x_{Bj} are missing, the MD_E is approximated as the mean distance of each value of x_j to the observed value, as defined by Eq. (13) (x_{Aj} is missing and x_{Bj} is observed). Finally, when both x_{Aj} and x_{Bj} are missing, the MD_E is approximated as the mean distance between all values of x_j (Eq. (14)).

$$MD_E(x_{Aj}, x_{Bj}) = (x_{Aj} - x_{Bj})^2 \tag{12}$$

$$\begin{aligned} MD_E(x_{Aj}, x_{Bj}) &= E\left((x - x_{Bj})^2\right) \\ &= \int p(x)(x - x_{Bj})^2 dx = (x_{Bj} - \mu_x)^2 + \sigma_x^2 \end{aligned} \tag{13}$$

$$\begin{aligned} MD_E(x_{Aj}, x_{Bj}) &= \iint p(x)p(y)(x - y)^2 dxdy \\ &= \left(E(x) - E(y)\right)^2 + \sigma_x^2 + \sigma_y^2 = 2\sigma_x^2 \end{aligned} \tag{14}$$

The original formulation of MD_E is only established for continuous features. To extend MD_E for nominal features, we shall consider the standard overlap distance, d_O (Eq. (3)) and define a nominal version of MD_E, which we will refer to as MD_O. When both values are known, MD_O is the same as d_O (Eq. (3)). When only one value is missing, MD_O is computed as the mean distance between all elements in x_j and the observed value, as shown in Eq. (15) (x_{Aj} is missing and x_{Bj} is observed). If both values are missing, MD_O is determined as the mean distance between all elements in x_j (Eq. (16)). Finally, $D(\mathbf{x}_A, \mathbf{x}_B)$ is obtained assuming d_j as MD_E or MD_O, depending on the feature type. Note that $MD_E(x_{Aj}, x_{Bj})$ already corresponds to $d_j(xA_j, xB_j)^2$ (Eqs. (12) to (14)). Therefore, only the $MD_O(x_{Aj}, x_{Bj})$ component should be squared when performing the aggregation.

$$\begin{aligned} MD_O(x_{Aj}, x_{Bj}) &= \sum_x p(x)\ d_O(x, x_{Bj}) \\ &= \sum_{x \neq x_{Bj}} p(x) = 1 - p(x_{Bj}) \end{aligned} \tag{15}$$

$$\begin{aligned} MD_O(x_{Aj}, x_{Bj}) &= \sum_x \sum_y p(x)p(y)\ d_O(x, y) \\ &= \sum_x \sum_{x \neq y} p(x)p(y) = 1 - \sum_x p^2(x) \end{aligned} \tag{16}$$

Fig. 1 presents the relationships of the distance functions considered in this work. According to the status of xA_j and xB_j values (either only one of them is missing, both are known or both are missing), and the type of j feature (either continuous or nominal), the schema depicts the d_j computation for each case. Each path (highlighted in different colours) presents the association between the presence/absence of values and feature type and aggregates the distances that perform the same computation of d_j.

3. Experiments

We started by collecting 61 complete and binary-classification datasets from open-source repositories, as presented in Table 1. These consider different sample sizes, number of features, type of features (continuous and nominal) and imbalance ratios (IR). Then, missing data is generated at 4 different rates (5, 10, 20 and 30%) under a Missing Completely At Random (MCAR) mechanism [18]. Additionally, we guarantee that the same missing rate was inserted in both classes and 30 runs are performed for each dataset and missing rate (MR). The datasets with missing values are then *i)* directly classified with Classification and Regression Trees (CART) model (Baseline approach) or *ii)* first imputed with KNN ($k = 1$) and then classified with CART. For *ii)* 7 heterogeneous distance functions were considered, as detailed in Section 2. Finally, CART performance is evaluated using sensitivity, F-measure (F1) and G-mean, which are robust to the existing class imbalance of the collected datasets, as this is a common difficulty factor found in a huge variety of domains [19]. In sum, we have considered 61 datasets * 4 different missing rates * 30 runs * 7 distance functions for imputation + 61 datasets * 4 missing rates * 30 runs for the baseline. In what follows, experimental results are evaluated overall, presenting the average classification results, and also performing a rank analysis, both considering all results and a detailed analysis by group (Continuous, Nominal and Other Datasets).

4. Results and discussion

Here we study the effect of the distance function on the k-nearest neighbour imputation, aiming to address the research questions identified in the Introduction.

4.1. Do distance metrics significantly affect KNN imputation?

Table 2 reports on the sensitivity, F1 and G-mean of CART classifier for 8 different methods: Baseline (with missing values) and imputed with HEOM, HEOM-REDEF, HVDM, HVDM-REDEF, HVDM-SPECIAL, MDE and SIMDIST, for MRs of 5, 10, 20 and 30%. F1 and G-mean are presented on the left-side of the table, whereas sensitivity is used to perform the ranking of methods (on the right). Both Strategy 1 and 2 report on the sensitivity results, yet they differ on the computation of ranks. For Strategy 1, methods are ranked based on their average sensitivity: the results for all datasets are averaged by method, and then the ranking is computed. For Strategy 2, methods are first ranked for each dataset separately and the average rank is determined for each method.

Starting with the average sensitivity, F1 and G-mean, the first observation is that all imputation techniques are preferable to the classification with missing values, i.e., the datasets imputed with KNN (for any distance metric) outperform the Baseline results. Also, as the missing rate increases, so does the difference between the Baseline and the imputation methods: sensitivity, F1 and G-mean present average differences of 0.017, 0.005 and 0.008 for a MR of 5% and 0.119, 0.053 and 0.066 for 30%, respectively. Also, the differences between distance metrics is more noticeable with increasing missing rates: for a MR of 5% the performance results are similar between methods, with a difference from the best to worst method of 0.004, 0.003 and 0.002 (for sensitivity, F1 and G-mean). For a MR of 30%, those differences increase to 0.036, 0.037 and 0.029, respectively.

Overall, SIMDIST, MDE, HEOM and HVDM appear to be the best performing distances, although SIMDIST and MDE assume more prominent positions for higher missing rates (20% and 30%). In turn, HEOM-REDEF and HVDM-REDEF appear frequently at the bottom positions. As previously discussed in Section 3, the collected datasets are imbalanced and therefore we focus on sensitivity results in the following analyses (considering the classification performance on the positive/minority cases) [19]. Furthermore, we rely on Strategy 2 to analyse the sensitivity results more precisely for each dataset.

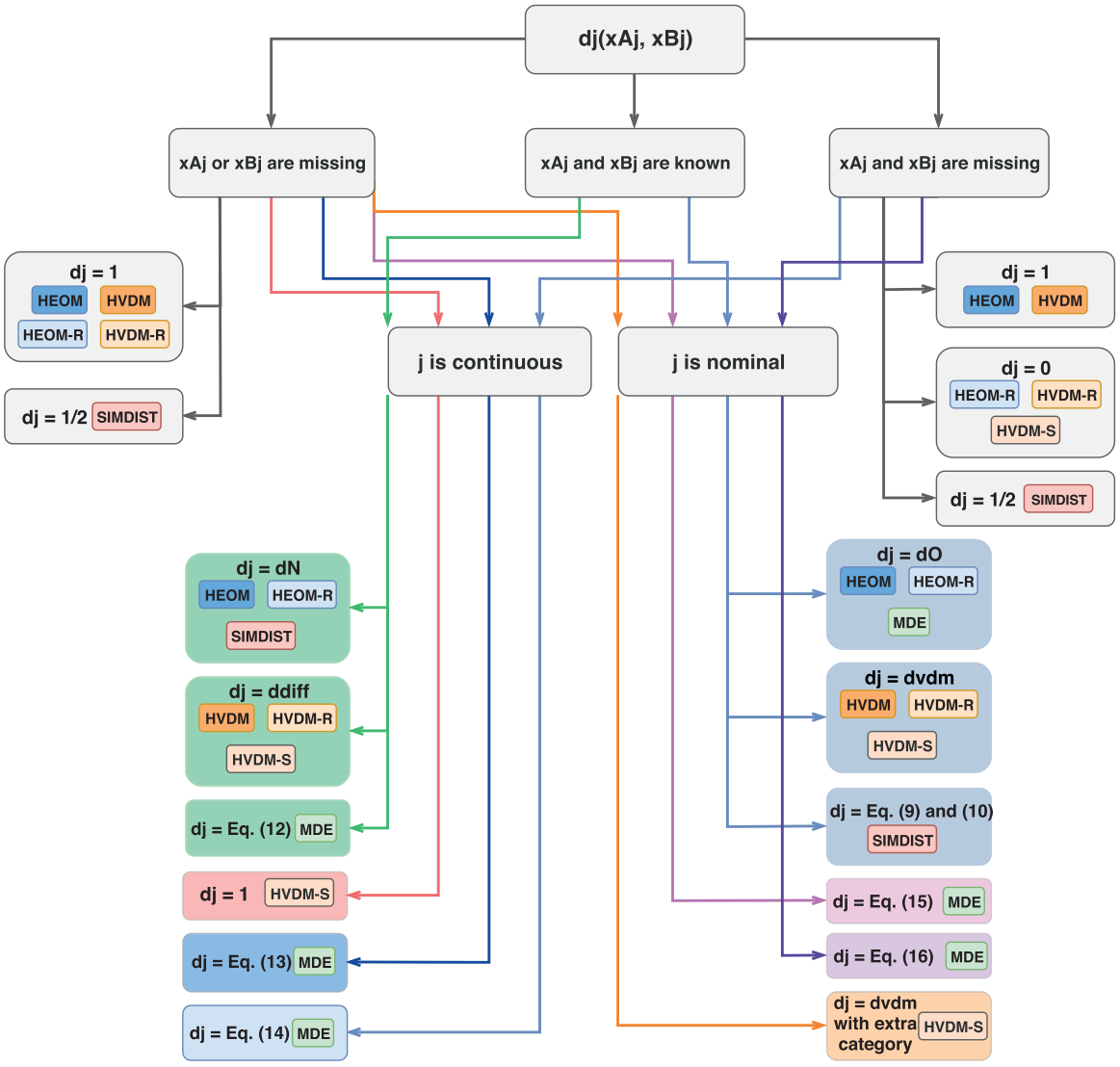

Fig. 1. Relationships between the distance functions considered in this work.

The ranks presented in Strategy 2 are consistent with our previous analysis, yet better illustrate the differences between methods. To determine whether there were significant differences between the methods, we compared them using the Friedman rank test. Under the null hypothesis, the different distances would assume equal ranks, i.e., the methods would be equivalent. We computed the F_F statistic [20] for all missing rates – $F_F = \{6.86, 8.73, 33.70, 35.22\}$ for 5, 10, 20 and 30% – and compared it with the established critical values for the F-distribution at a 5% significance level, $F(7, 420)_{0.05} = 2.03$. For all missing rates, the null hypothesis was rejected and therefore we proceeded to post-hoc testing (at a 5% significance level), computing the critical differences for Nemenyi test ($CD_n = 1.34$) so that all methods are compared with each other. For all missing rates, the difference between the rank of the Baseline and the ranks of remaining methods is higher than CD_n, which reveals that all imputation methods are significantly better than performing classification with incomplete data. Regarding the remaining methods, we further analyse the results by missing rate. For a MR of 5% and 10%, the post-hoc did not detect any significant differences between the methods (differences between the best and worst performing methods was lower than CD_n – 1.06 and 1.28 for 5% and 10%, respectively). In turn, for MRs of 20% and 30%, some methods proved to be significantly better than others. For 20% and 30% missing rates, the difference between distance ranks is reported in Table 3.

The values respect to the difference between the ranks of each method in the rows and the methods in the columns. Differences for 20% are presented in the upper part of table, whereas differences for 30% are shown in the lower part of the table. Significant differences (higher than CD_n, > 1.34) are marked in bold, except for the Baseline, since all methods proved to be significantly better. For both these missing rates, all distances were significantly better than HEOM-REDEF and HVDM-REDEF (except for HVDM-SPECIAL, that although achieving better results than both redefinitions, did not reach the CD_n value). Additionally, MDE and SIMDIST were significantly better than HVDM-SPECIAL. Considering the obtained results, it is interesting to observe that HEOM and HVDM are significantly better than their redefinitions, as previously discussed from Table 2. In turn, the experimental data was not sufficient to detect significant differences between HEOM/HVDM and HVDM-SPECIAL and although MDE and SIMDIST appear in the leading positions for MRs of 20% and 30%, the post-hoc was not enough to conclude on their superiority over HEOM or HVDM.

Table 1
Characteristics of collected datasets.

Dataset	Size	Features	C/N	IR	Dataset	Size	Features	C/N	IR
abalone	4174	8	(7/1)	1.89	ecoli_0_4_6_vs_5	203	6	(6/0)	9.15
acute-inflammations-nephritis	120	6	(1/5)	1.4	ecoli_0_6_7_vs_3_5	222	7	(7/0)	9.09
acute-inflammations-urinary	120	6	(1/5)	1.03	ecoli_0_6_7_vs_5	220	6	(6/0)	10
alzheimer-v1	317	9	(7/2)	1.5	ecoli_0_vs_1	220	7	(7/0)	1.86
arrhythmia	420	266	(205/61)	1.3	fertility-diagnosis	100	9	(2/7)	7.33
autism-adolescent	98	19	(1/18)	1.72	haberman	306	3	(3/0)	2.78
autism-adult	701	16	(1/15)	2.71	heart-statlog	270	13	(7/6)	1.25
bc-coimbra	116	9	(9/0)	1.23	immunotherapy	90	7	(5/2)	3.74
biomed	194	5	(5/0)	1.9	kala-azar	68	6	(5/1)	5.8
breast-tissue-2c	106	9	(9/0)	4.05	kidney	158	24	(11/13)	2.67
bupa	345	6	(5/1)	1.38	language-impairment-ENNI	377	61	(59/2)	3.9
cleveland_0_vs_4	173	13	(13/0)	12.31	language-impairment-conti	118	60	(59/1)	5.21
cryotherapy	90	6	(4/2)	1.14	language-impairment-gillam	667	61	(59/2)	2.92
ctg-2c	2126	21	(21/0)	11.08	lung-cancer-v1	27	56	(0/56)	2
dermatology-v2	182	34	(1/33)	1.56	lymphography-v1	142	18	(3/15)	1.33
dermatology_6	358	34	(34/0)	16.9	new-thyroid-N-vs-HH	215	5	(5/0)	2.31
diabetic-retinopathy	1151	19	(16/3)	1.13	newthyroid-v1	185	5	(5/0)	4.29
ecoli	336	7	(7/0)	8.6	newthyroid-v3	180	5	(5/0)	5
ecoli1	336	7	(7/0)	3.36	parkinson	195	22	(22/0)	3.06
ecoli2	336	7	(7/0)	5.46	pima	768	8	(8/0)	1.87
ecoli4	336	7	(7/0)	15.8	postoperative-SvsA	86	8	(1/7)	2.58
ecoli_0_1_4_6_vs_5	280	6	(6/0)	13	relax	182	12	(12/0)	2.5
ecoli_0_1_4_7_vs_2_3_5_6	336	7	(7/0)	10.59	saheart	462	9	(8/1)	1.89
ecoli_0_1_4_7_vs_5_6	332	6	(6/0)	12.28	spectf	267	44	(44/0)	3.85
ecoli_0_1_vs_2_3_5	244	7	(7/0)	9.17	thoracic	470	16	(3/13)	5.71
ecoli_0_1_vs_5	240	6	(6/0)	11	thyroid_3_vs_2	703	21	(21/0)	18
ecoli_0_2_3_4_vs_5	202	7	(7/0)	9.1	transfusion	748	4	(4/0)	3.2
ecoli_0_2_6_7_vs_3_5	224	7	(7/0)	9.18	vertebral-2c	310	6	(6/0)	2.1
ecoli_0_3_4_6_vs_5	205	7	(7/0)	9.25	wisconsin	683	9	(9/0)	1.86
ecoli_0_3_4_7_vs_5_6	257	7	(7/0)	9.28	wpbc	198	32	(32/0)	3.21
ecoli_0_3_4_vs_5	200	7	(7/0)	9					

C/N: Number of Continuous/Nominal features.

Table 2
CART performance results without imputation (Baseline) and with KNN imputation using several distances.

MR	Distance	F1	GMEAN	Rank	Strategy 1	Strategy 2
5%	BASELINE	0.653 ± 0.240	0.724 ± 0.220	*1ST*	SIMDIST (0.6603 ± 0.2367)	SIMDIST (3.61 ± 1.96)
	HEOM	0.659 ± 0.237	0.731 ± 0.217	*2ND*	HVDM (0.6589 ± 0.2379)	HVDM (3.81 ± 1.95))
	HEOM-REDEF	0.658 ± 0.235	**0.732** ± 0.215	*3RD*	HVDM-SPECIAL (0.6584 ± 0.2341)	HVDM-SPECIAL (4.24 ± 1.94)
	HVDM	0.658 ± 0.238	0.731 ± 0.219	*4TH*	HEOM (0.6584 ± 0.2370)	HEOM (4.32 ± 2.25)
	HVDM-REDEF	0.657 ± 0.237	0.730 ± 0.218	*5TH*	HEOM-REDEF (0.6576 ± 0.2341)	HEOM-REDEF (4.54 ± 2.09)
	HVDM-SPECIAL	**0.660** ± 0.235	**0.732** ± 0.214	*6TH*	MDE (0.6570 ± 0.2320)	HVDM-REDEF (4.60 ± 2.01)
	MDE	0.659 ± 0.233	0.731 ± 0.213	*7TH*	HVDM-REDEF (0.6565 ± 0.2365)	MDE (4.67 ± 2.42)
	SIMDIST	**0.660** ± 0.237	**0.732** ± 0.218	*8TH*	BASELINE (0.6413 ± 0.2385)	BASELINE (6.20 ± 2.65)
10%	BASELINE	0.627 ± 0.236	0.699 ± 0.217	*1ST*	SIMDIST (0.6430 ± 0.2357)	SIMDIST (3.53 ± 2.15)
	HEOM	0.641 ± 0.235	0.716 ± 0.216	*2ND*	HVDM (0.6403 ± 0.2341)	HVDM (3.75 ± 1.99)
	HEOM-REDEF	0.638 ± 0.228	0.715 ± 0.209	*3RD*	HEOM (0.6385 ± 0.2338)	HEOM (3.94 ± 2.02)
	HVDM	0.643 ± 0.235	0.718 ± 0.215	*4TH*	MDE (0.6378 ± 0.2301)	MDE (4.40 ± 2.49)
	HVDM-REDEF	0.636 ± 0.233	0.713 ± 0.215	*5TH*	HVDM-SPECIAL (0.6366 ± 0.2307)	HVDM-SPECIAL (4.58 ± 1.72)
	HVDM-SPECIAL	0.638 ± 0.231	0.715 ± 0.210	*6TH*	HEOM-REDEF (0.6363 ± 0.2276)	HEOM-REDEF (4.63 ± 2.38)
	MDE	0.640 ± 0.233	0.716 ± 0.214	*7TH*	HVDM-REDEF (0.6335 ± 0.2320)	HVDM-REDEF (4.81 ± 1.69)
	SIMDIST	**0.645** ± 0.236	**0.720** ± 0.217	*8TH*	BASELINE (0.6027 ± 0.2329)	BASELINE (6.34 ± 2.53)
20%	BASELINE	0.563 ± 0.218	0.638 ± 0.204	*1ST*	SIMDIST (0.6082 ± 0.2314)	MDE (3.16 ± 1.83)
	HEOM	0.607 ± 0.228	0.689 ± 0.211	*2ND*	HEOM (0.6048 ± 0.2279)	SIMDIST (3.28 ± 2.18)
	HEOM-REDEF	0.587 ± 0.224	0.673 ± 0.206	*3RD*	MDE (0.6047 ± 0.2238)	HEOM (3.37 ± 1.91)
	HVDM	0.607 ± 0.230	0.688 ± 0.211	*4TH*	HVDM (0.6047 ± 0.2284)	HVDM (3.42 ± 1.86)
	HVDM-REDEF	0.591 ± 0.220	0.677 ± 0.202	*5TH*	HVDM-SPECIAL (0.5933 ± 0.2239)	HVDM-SPECIAL (4.63 ± 1.67)
	HVDM-SPECIAL	0.597 ± 0.224	0.681 ± 0.204	*6TH*	HVDM-REDEF (0.5880 ± 0.2189)	HVDM-REDEF (5.19 ± 1.53)
	MDE	0.608 ± 0.225	0.690 ± 0.207	*7TH*	HEOM-REDEF (0.5837 ± 0.2222)	HEOM-REDEF (5.76 ± 1.71)
	SIMDIST	**0.612** ± 0.232	**0.693** ± 0.213	*8TH*	BASELINE (0.5101 ± 0.2034)	BASELINE (7.20 ± 1.92)
30%	BASELINE	0.503 ± 0.204	0.580 ± 0.197	*1ST*	MDE (0.5694 ± 0.2201)	MDE (2.97 ± 1.97)
	HEOM	0.559 ± 0.224	0.650 ± 0.207	*2ND*	SIMDIST (0.5682 ± 0.2286)	SIMDIST (3.02 ± 1.99)
	HEOM-REDEF	0.537 ± 0.213	0.631 ± 0.197	*3RD*	HVDM (0.5571 ± 0.2252)	HVDM (3.49 ± 1.93)
	HVDM	0.561 ± 0.226	0.649 ± 0.210	*4TH*	HEOM (0.5563 ± 0.2228)	HEOM (3.79 ± 1.81)
	HVDM-REDEF	0.541 ± 0.214	0.634 ± 0.198	*5TH*	HVDM-SPECIAL (0.5456 ± 0.2188)	HVDM-SPECIAL (4.64 ± 1.76)
	HVDM-SPECIAL	0.547 ± 0.217	0.640 ± 0.198	*6TH*	HVDM-REDEF (0.5375 ± 0.2142)	HVDM-REDEF (5.18 ± 1.61)
	MDE	**0.574** ± 0.221	**0.660** ± 0.207	*7TH*	HEOM-REDEF (0.5334 ± 0.2122)	HEOM-REDEF (5.63 ± 1.64)
	SIMDIST	0.573 ± 0.229	0.659 ± 0.211	*8TH*	BASELINE (0.4331 ± 0.1805)	BASELINE (7.28 ± 1.82)

Table 3
Differences between ranks for each comparison of distance metrics for 20% and 30%. Significant differences are marked in bold.

	BASELINE	HEOM	HEOM-R	HVDM	HVDM-R	HVDM-S	MDE	SIMDIST
BASELINE	–	3.83	1.43	3.78	2.01	2.57	4.04	3.92
HEOM	−3.49	–	**−2.39**	−0.05	**−1.82**	−1.26	0.21	0.09
HEOM-R	−1.65	**1.84**	–	**2.34**	0.57	1.13	**2.61**	**2.48**
HVDM	−3.79	−0.30	**−2.14**	–	**−1.77**	−1.21	0.26	0.14
HVDM-R	−2.10	**1.39**	−0.45	**1.69**	–	0.56	**2.03**	**1.91**
HVDM-S	−2.64	0.85	−0.99	1.15	−0.54	–	**1.48**	**1.35**
MDE	−4.31	−0.82	**−2.66**	−0.52	**−2.21**	**−1.67**	–	−0.12
SIMDIST	−4.25	−0.76	**−2.61**	−0.47	**−2.16**	**−1.61**	0.06	–

HEOM-R: HEOM-REDEF; **HVDM-R**: HVDM-REDEF; **HVDM-S**: HVDM-SPECIAL.

4.2. Is there a distance more beneficial for some datasets?

To tailor our analysis to the characteristics of each dataset, we divided the collected datasets into 3 groups on the basis of their types of features. Then, the ranks of each distance are evaluated for each group separately considering the highest percentage of missing data (30%), where differences between ranks were more noticeable. Table 4 reports on these results, where the 3 main groups are identified. *Continuous Datasets* consist entirely of datasets comprising continuous features while *Nominal Datasets* consist of datasets comprising predominantly nominal features. Among the collected datasets, only one had entirely nominal features (*lung-cancer-v1*), although several others were mainly composed of nominal features (comprising only 1, 2 or 3 continuous features). For this reason, we have decided to include them in this group. The remaining datasets were grouped in *Other Datasets*. This group contains heterogeneous datasets (with both continuous and nominal features) and comprises datasets that include a somewhat representative amount of each type of feature (*arrhythmia, heart-statlog, kidney*), although the majority is predominantly continuous (we have left them out of *Continuous Datasets* since there was a considerable amount of datasets exclusively continuous).

Similarly to the previous analysis, the F_F statistic was computed for all groups – $F_F = \{30.97, 9.01, 6.08\}$ for Group 1, 2 and 3, respectively – and compared to the F-distribution at a 5% significance level, $F(7, 252)_{0.05} = 2.05$, $F(7, 63)_{0.05} = 2.17$, $F(7, 91)_{0.05} = 2.11$. For all groups, the null hypothesis was rejected and Nemenyi test was performed. In what follows, we provide an analysis of each group, elaborating on the findings of the post-hoc and explaning some trends and hypothesis that were consistent with the experimental data.

4.2.1. Group 1: continuous datasets

The results are in agreement with the overall results presented in Table 2, with MDE and SIMDIST assuming the leading positions (2.78 and 2.81) and HEOM and HVDM falling just behind (3.30 and 3.31). All distances were significantly better than the Baseline, except HEOM-REDEF, although closer to the critical value, with a difference of 1.61 ($CD_n = 1.73$). Additionally, HEOM-REDEF, HVDM-REDEF and HVDM-SPECIAL proved to be significantly worse that the remaining distances.

Since these datasets comprise only continuous features, these distances can only differ on the way that continuous features are normalised and how missing values are treated. As discussed in Section 3, HEOM, SIMDIST and MDE perform min-max normalisation. HVDM scales features by $4\sigma_{x_j}$ and SIMDIST further uses a z normalisation function, yet overall the normalisation process is similar and therefore does not seem to be the reason behind the differences in performance. However, whereas HEOM and HVDM assume a distance of 1 if x_{Aj} and/or x_{Bj} are missing, SIMDIST and MDE apply more sensitive approaches: SIMDIST replaces the missing values by the mean similarity between all patterns according to j and MDE is more refined, further distinguishing situations where one value or both values are missing. Since two groups of similar ranks are identified among these top methods, {HEOM, HVDM} and {SIMDIST, MDE}, we hypothesise that the approach to handle missing values may be on the origin of differences found among these methods.

Also, if our hypothesis is true, it would explain why HEOM-REDEF and HVDM-REDEF are ranked lower than the remaining methods: although they treat missing values as "special values", the distance between two patterns on j is 1 if only x_{Aj} or x_{Bj} are missing, yet 0 if they are both missing (the same is valid for HVDM-SPECIAL, since for continuous datasets, is the same as HVDM-REDEF). This evaluation of distances between patterns with missing values seems extreme when compared to the top-performing distances (MDE and SIMDIST) and also HEOM and HVDM, which would explain their poor results.

4.2.2. Group 2: nominal datasets

When datasets are predominantly composed of nominal features, HVDM-SPECIAL stood out among all distances, obtaining a rank of 2.00. The post-hoc revealed that all distances were significantly better than the Baseline, except for HEOM-REDEF and HVDM-REDEF ($CD_n = 3.32$). Whereas overall HVDM-SPECIAL falls to the bottom positions (Table 2), for nominal datasets it seems to be the most beneficial (Table 4).

Nemenyi test also revealed that HVDM-SPECIAL was significantly better than HEOM-REDEF and HVDM-REDEF was near the critical value, with a difference between ranks of 3.4 and 3.2 respectively. No significant differences were found for HEOM, HVDM, MDE or SIMDIST, despite the considerable differences between ranks of HVDM-SPECIAL and each of the methods (1.5, 1.95, 2.3 and 1.75, respectively).

We hypothesise that the great advantage of HVDM-SPECIAL derives from the way it considers a missing value as an extra category and instead of simply applying a matching rule (as HVDM-REDEF), it applies d_{vdm}: when only x_{Aj} or x_{Bj} are missing, the distance computation is more refined, rather than being maximum (assigned a value of 1).

Another observation is that HEOM-REDEF and HVDM-REDEF are again ranked lower than HEOM and HVDM which, similarly to the previous group, indicates that differences rely on the treatment of missing values. In this case, assigning a a distance of 0 (minimum distance) between two missing values seems more prejudicial than assigning a distance of 1 (maximum distance). Nevertheless, the top-performing distance (HVDM-SPECIAL) also considers that the distance should be 0 between two missing values, although it uses a more refined approach (d_{vdm}) when only x_{Aj} or x_{Bj} are missing. This is consistent with the hypothesis that our proposed strategy of considering missing values as extra categories is a major advantage for nominal datasets.

In turn, MDE (which also considers a different distance assignment whether both values are missing or only one is missing) is

Table 4
Distance ranks for a 30% missing rate, divided by group.

Group 1: Continuous Datasets	C/N	B	HEOM	HEOM-R	HVDM	HVDM-R	*HVDM-S	MDE	SIMDIST
bc-coimbra	(9/0)	8	3	**1**	6	4.5	4.5	2	7
biomed	(5/0)	8	3	7	**1**	5.5	5.5	4	2
breast-tissue-2c	(9/0)	8	2	7	3	5.5	5.5	4	**1**
cleveland_0_vs_4	(13/0)	7	3	4	8	5.5	5.5	2	**1**
ctg-2c	(21/0)	8	4	7	3	5.5	5.5	2	**1**
dermatology_6	(34/0)	8	4	7	3	5.5	5.5	2	**1**
ecoli	(7/0)	8	4	7	2	5.5	5.5	**1**	3
ecoli1	(7/0)	8	2	5	3	6.5	6.5	**1**	4
ecoli2	(7/0)	8	4	7	**1**	5.5	5.5	2	3
ecoli4	(7/0)	2.5	7	2.5	8	5.5	5.5	**1**	4
ecoli_0_1_4_6_vs_5	(6/0)	8	3	7	2	5.5	5.5	4	**1**
ecoli_0_1_4_7_vs_2_3_5_6	(7/0)	8	**1**	5	3	6.5	6.5	2	4
ecoli_0_1_4_7_vs_5_6	(6/0)	8	2	5	3	6.5	6.5	**1**	4.5
ecoli_0_1_vs_2_3_5	(7/0)	8	4	5	3	6.5	6.5	2	**1**
ecoli_0_1_vs_5	(6/0)	8	3	5	2	6.5	6.5	4	**1**
ecoli_0_2_3_4_vs_5	(7/0)	8	2	7	3	5.5	5.5	4	**1**
ecoli_0_2_6_7_vs_3_5	(7/0)	6	8	7	2	4.5	4.5	**1**	3
ecoli_0_3_4_6_vs_5	(7/0)	8	2	5	4	6.5	6.5	3	**1**
ecoli_0_3_4_7_vs_5_6	(7/0)	8	3	7	2	4.5	4.5	**1**	6
ecoli_0_3_4_vs_5	(7/0)	8	2	5	3	6.5	6.5	4	**1**
ecoli_0_4_6_vs_5	(6/0)	8	3	7	2	5.5	5.5	4	**1**
ecoli_0_6_7_vs_3_5	(7/0)	2.5	8	4	2.5	6.5	6.5	**1**	5
ecoli_0_6_7_vs_5	(6/0)	2	4	5	8	6.5	6.5	**1**	3
ecoli_0_vs_1	(7/0)	8	**1**	7	2	5.5	5.5	4	3
haberman	(3/0)	8	4	7	**1**	5.5	5.5	3	2
new-thyroid-N-vs-HH	(5/0)	8	**1**	6	3	4.5	4.5	7	2
newthyroid-v1	(5/0)	8	**1**	7	3	5.5	5.5	4	2
newthyroid-v3	(5/0)	8	3	6	**1**	3	3	7	5
parkinson	(22/0)	8	4	7	3	5.5	5.5	2	**1**
pima	(8/0)	8	3	5	2	6.5	6.5	**1**	4
relax	(12/0)	8	6	**1**	7	3.5	3.5	5	2
spectf	(44/0)	8	2	6	7	4	4	**1**	4
thyroid_3_vs_2	(21/0)	**1**	3	6	4	6	6	2	8
transfusion	(4/0)	8	7	4	6	2.5	2.5	**1**	5
vertebral-2c	(6/0)	8	**1**	7	2	4.5	4.5	6	3
wisconsin	(9/0)	8	**1**	5	2	6.5	6.5	4	3
wpbc	(32/0)	8	4	7	2	5.5	5.5	3	**1**
	Rank:	7.27	3.30	5.66	3.31	5.43	5.43	**2.78**	2.81

Group 2: Categorical Datasets	C/N	B	HEOM	HEOM-R	HVDM	HVDM-R	HVDM-S	MDE	SIMDIST
acute-inflammations-nephritis	(1/5)	8	5	6	1.5	7	4	3	1.5
acute-inflammations-urinary	(1/5)	8	3	4	2	7	5	6	**1**
autism-adolescent	(1/18)	8	3	5	2	7	**1**	4	6
autism-adult	(1/15)	8	4	6	5	7	2	3	**1**
dermatology-v2	(1/33)	8	4	6	5	**1**	2	3	7
fertility-diagnosis	(2/7)	8	3	6	2	5	**1**	7	4
lung-cancer-v1	(0/56)	8	4	7	5	2	**1**	6	3
lymphography-v1	(3/15)	8	3	5	6	7	2	**1**	4
thoracic	(3/13)	8	3	5	6	7	**1**	2	4
postoperative-SvsA	(1/7)	7	3	4	5	2	**1**	8	6
	Rank:	7.90	3.50	5.40	3.95	5.20	**2.00**	4.30	3.75

Group 3: Other Datasets	C/N	B	HEOM	HEOM-R	HVDM	HVDM-R	HVDM-S	MDE	SIMDIST
abalone	(7/2)	8	2	7	3	4	5	6	**1**
alzheimer-v1	(7/2)	8	5	3	2	7	6	4	**1**
arrhythmia	(205/61)	8	7	6	2	4	5	**1**	3
bupa	(5/1)	8	6	4	3	5	2	**1**	7
cryotherapy	(4/2)	8	6	7	5	2	4	**1**	3
diabetic-retinopathy	(16/3)	8	5	7	2	4	6	3	**1**
heart-statlog	(7/6)	8	6	5	2	7	4	3	**1**
immunotherapy	(5/2)	2	4	5	6	8	7	**1**	3
kala-azar	(5/1)	8	5	2	3	**1**	7	4	6
kidney	(11/13)	8	6	3	2	4	5	7	**1**
language-impairment-ENNI	(59/2)	4	7	8	6	5	3	**1**	2
language-impairment-conti	(59/1)	3	4	8	7	5	2	**1**	6
language-impairment-gillam	(59/2)	7	6	8	4	5	3	**1**	2
saheart	(8/1)	8	5	7	4	2	3	**1**	6
	Rank:	6.86	5.29	5.71	3.64	4.50	4.43	**2.50**	3.07

C/N: Number of Continuous/Nominal features.
B: BASELINE; **HEOM-R**: HEOM-REDEF; **HVDM-R**: HVDM-REDEF; **HVDM-S**: HVDM-SPECIAL.
*For continuous datasets, HVDM-S is equivalent to HVDM-R.

ranked lower than HEOM and HVDM. Further investigation on this effect is required, although we argue that computing the mean distance between patterns (Eqs. (15) and (16)) might not be adequate for nominal features (as it seems to be for continuous).

4.2.3. Group 3: other datasets

Only HVDM, MDE and SIMDIST proved to be significantly better than the Baseline ($CD_n = 2.91$). MDE stood out as the winning approach, followed by SIMDIST, whereas HEOM and HEOM-REDEF were at the bottom positions. The Nemenyi test also revealed that MDE proved to be significantly better than HEOM-REDEF (with a difference between ranks of 3.21) and HEOM was near the critical value (2.79). This is an interesting observation since, as stated in the Introduction, HEOM is traditionally used for handling heterogeneous data with missing values, although for this group of datasets it was frequently assigned the worst ranks.

Given the MCAR generation, there are no constraints on which type of features missing values were inserted, which is a question to be investigated on ongoing work. For instance, it was expected that MDE performed worse for datasets comprising nominal features, but the number of nominal features comprised in these datasets (plus the randomisation process associated with MCAR) does not allow a full characterisation of such effect. For *kidney* dataset, which contains an even distribution of continuous/nominal features (11/13), MDE ranks the lowest (7.00). However, *arrhythmia* includes a considerable number of both (205/61) and MDE achieves the best rank (1.00). In this case, the superiority of MDE may be explained by the fact that the continuous features constitute the vast majority, although this question should be further addressed in future work.

5. Conclusions and future work

We perform a comparison of several heterogeneous distances that handle missing values across a benchmark of 61 publicly-available datasets with different characteristics. From the results obtained with the experimental data, four main conclusions may be derived:

- Distance metrics significantly affect KNN imputation, especially for higher missing rates (20% and 30%). HEOM-REDEF and HVDM-REDEF performed the worst, occasionally achieving lower performance than classifying data with missing values;
- Differences in performance between distance functions mostly rely on their respective approaches to missing values. Overall, the distance assignment of 0 when two values are missing seems rigid and may be prejudicial for imputation;
- There seems to be an advantage in distinguish situations where only one value is missing from situations when both are missing. However, depending on the type of feature, these situations should be subjected to different approaches: e.g., considering the mean similarity of values for continuous features (similarly to MDE) and considering the missing value as an extra category for nominal features (similarly to HVDM-SPECIAL);
- Finally, although further investigation is required, we also argue that HEOM, widely used across several domains, may not be the go-to approach, as others have shown to be more beneficial (MDE and SIMDIST).

Regarding future work, there are three main ongoing directions:

- Collect more datasets to further investigate the trends found within this work;
- Focus on which features the missing values will be placed (continuous or nominal) for heterogeneous datasets, to perform a more thorough analysis;
- Investigate other values of k and determine whether strategies of weighting features differently (based on their mutual information or discriminative power) would be useful to improve the imputation results.
- Explore KNN sensitivity to both the chosen distance function and features considered for distance computation, either chosen randomly or according to their relevance to classification. A precursor work on the latter topic may be found in [21].

Authorship Confirmation

1. This manuscript, or a large part of it, has not been published, was not, and is not being submitted to any other journal.

2. If presented at or submitted to or published at a conference(s), the conference(s) is (are) identified and substantial justification for re-publication is presented below. A copy of conference paper(s) is(are) uploaded with the manuscript.

3. If the manuscript appears as a preprint anywhere on the web, e.g. arXiv, etc., it is identified below. The preprint should include a statement that the paper is under consideration at Pattern Recognition Letters.

4. All text and graphics, except for those marked with sources, are original works of the authors, and all necessary permissions for publication were secured prior to submission of the manuscript.

5. All authors each made a significant contribution to the research reported and have read and approved the submitted manuscript.

Declaration of Competing Interest

None.

Acknowledgments

This work was supported in part by the project NORTE-01-0145-FEDER-000027 (Norte Portugal Regional Operational Programme – Norte 2020) and in part by the FCT Research Grant SFRH/BD/138749/2018.

References

[1] D.R. Wilson, T.R. Martinez, Improved heterogeneous distance functions, J. Artif. Intell. Res. 6 (1997) 1–34.
[2] G. Tutz, S. Ramzan, Improved methods for the imputation of missing data by nearest neighbor methods, Comput. Stat. Data Anal. 90 (2015) 84–99.
[3] M.S. Santos, J.P. Soares, P.H. Abreu, H. Araújo, J. Santos, Influence of data distribution in missing data imputation, in: Conference on Artificial Intelligence in Medicine in Europe, Springer, 2017, pp. 285–294.
[4] P.J. García-Laencina, P.H. Abreu, M.H. Abreu, N. Afonoso, Missing data imputation on the 5-year survival prediction of breast cancer patients with unknown discrete values, Comput. Biol. Med. 59 (2015) 125–133.
[5] M.S. Santos, P.H. Abreu, P.J. García-Laencina, A. Simão, A. Carvalho, A new cluster-based oversampling method for improving survival prediction of hepatocellular carcinoma patients, J. Biomed. Inform. 58 (2015) 49–59.
[6] J.A. Sáez, B. Krawczyk, M. Woźniak, Handling class label noise in medical pattern classification systems, J. Med. Inform. Technol. 24 (2015).
[7] J. de Andrade Silva, E.R. Hruschka, An experimental study on the use of nearest neighbor-based imputation algorithms for classification tasks, Data Knowl. Eng. 84 (2013) 47–58.
[8] L. Beretta, A. Santaniello, Nearest neighbor imputation algorithms: a critical evaluation, BMC Med. Inform. Decis. Mak. 16 (3) (2016) 74.
[9] L.-Y. Hu, M.-W. Huang, S.-W. Ke, C.-F. Tsai, The distance function effect on k-nearest neighbor classification for medical datasets, SpringerPlus 5 (1) (2016) 1304.
[10] A. Nazabal, P.M. Olmos, Z. Ghahramani, I. Valera, Handling Incomplete Heterogeneous Data using VAEs, arXiv e-prints (2018) arXiv:1807.03653.
[11] B. Tang, H. He, S. Zhang, Mcenn: A variant of extended nearest neighbor method for pattern recognition, Pattern Recogn. Lett. 133 (2020) 116–122.
[12] Z. Abu-Aisheh, R. Raveaux, J.-Y. Ramel, Efficient k-nearest neighbors search in graph space, Pattern Recogn. Lett. (2018), doi:10.1016/j.patrec.2018.05.001.
[13] B. Shi, L. Han, H. Yan, Adaptive clustering algorithm based on knn and density, Pattern Recogn. Lett. 104 (2018) 37–44.
[14] M. Juhola, J. Laurikkala, On metricity of two heterogeneous measures in the presence of missing values, Artif. Intell. Rev. 28 (2) (2007) 163–178.

[15] P. Domingos, Rule induction and instance-based learning: a unified approach, in: International Joint Conference on Artificial Intelligence (IJCAI), 2, Springer, 1995, pp. 1226–1232.
[16] L.A. Belanche Muñoz, J. Hernández González, Similarity networks for heterogeneous data, in: ESANN 2012, 2012, pp. 215–220.
[17] L. AbdAllah, I. Shimshoni, k-means over incomplete datasets using mean euclidean distance, in: Machine Learning and Data Mining in Pattern Recognition, Springer, 2016, pp. 113–127.
[18] M. Santos, R. Pereira, A. Costa, S. J., J. Santos, P. Abreu, Generating synthetic missing data: a review by missing mechanism, IEEE Access 1 (1) (2019) 1–18.
[19] M.S. Santos, J.P. Soares, P.H. Abreu, H. Araujo, J. Santos, Cross-validation for imbalanced datasets: avoiding overoptimistic and overfitting approaches [research frontier], IEEE Comput. Intell. Mag. 13 (4) (2018) 59–76.
[20] J. Demšar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (Jan) (2006) 1–30.
[21] C. Domeniconi, B. Yan, Nearest neighbor ensemble, in: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., 1, IEEE, 2004, pp. 228–231.