Computing Random Forest-Distances in the Presence of Missing Data

MANUELE BICEGO, University of Verona, Verona, Italy
FERDINANDO CICALESE, University of Verona, Verona, Italy

In this article, we study the problem of computing Random Forest-distances in the presence of missing data. We present a general framework which avoids pre-imputation and uses in an agnostic way the information contained in the input points. We centre our investigation on RatioRF, an RF-based distance recently introduced in the context of clustering and shown to outperform most known RF-based distance measures. We also show that the same framework can be applied to several other state-of-the-art RF-based measures and provide their extensions to the missing data case. We provide significant empirical evidence of the effectiveness of the proposed framework, showing extensive experiments with RatioRF on 15 datasets. Finally, we also positively compare our method with many alternative literature distances, which can be computed with missing values.

CCS Concepts: • **Computing methodologies** → *Machine learning*; • **Information systems** → **Data cleaning;**

Additional Key Words and Phrases: Random forest distances, missing data, RatioRF measure

ACM Reference Format:
Manuele Bicego and Ferdinando Cicalese. 2024. Computing Random Forest-Distances in the Presence of Missing Data. *ACM Trans. Knowl. Discov. Data.* 18, 7, Article 180 (June 2024), 18 pages. https://doi.org/10.1145/3656345

1 INTRODUCTION

Random Forests (RFs) [9, 12] are ensembles of decision trees [27] which are successfully applied in Pattern Recognition and Machine Learning as models for regression, classification, and more recently, for clustering. RFs can also be employed to derive a (dis)similarity[1] measure, which can then be used in different types of applications. This is the case of several approaches in *distance-based RF-clustering* [5, 7, 9, 34, 38, 44], where the RF distance is used as input for a standard distance-based clustering method, such as Hierarchical clustering or Spectral clustering. Different definitions of such RF-based similarity measures have been proposed, ranging from the simplest and most employed one defined by Breiman [9, 34] up to more recent and complex similarities [5, 7, 38, 44]. In general, an RF-based similarity is computed in two phases: (i) a *learning phase*, where several

[1]In this article, we talk in an interchangeable way about similarity and dissimilarity/distance, the specific meaning will be clear from the context.

Authors' addresses: M. Bicego (Corresponding author) and F. Cicalese, University of Verona, Computer Science Department, Strada le Grazie, 15, 37135 – Verona, ITALY; e-mails: manuele.bicego@univr.it, ferdinando.cicalese@univr.it.

ACM 1556-4681/2024/06-ART180
https://doi.org/10.1145/3656345

decision trees are trained on random subsets of the input points; (ii) a *testing phase*, where the similarity, for a given pair of objects x, y, is first computed for each tree, by exploiting the test results when x, y traverse each tree T, and then aggregated at forest level by averaging over all trees.

An important issue when using RF-based approaches is how to deal with missing data [24], i.e., data where some variables do not have a value—a very common situations in many scenarios, like in the biomedical field [37]. Typical solutions to this problem are [41]: (i) to ignore objects with missing values, or, (ii) to attempt to complete the missing data with imputation methods [21, 24]. Both these approaches have drawbacks. Deleting or ignoring the data with missing values has the clear disadvantage of inducing a substantial decrease in the size of the dataset available, and passing from the original dataset to the reduced one may introduce bias. The introduction of spurious bias is also a problem with basic imputation methods, which replace a missing value with a new one, as well as with more elaborate ones, like k-NN [39] or multivariate imputation by chained equations [11], that depend on tuning parameters or the specification of a parametric model. In general, the common feeling is that pre-imputation should be avoided whenever it is possible, since it may introduce biases, especially if the underlying missing mechanism is unknown [4, 17, 19, 37].

The question is then: "Is it possible to perform a specific data analysis task on data with missing values without resorting to (and avoiding the pitfalls of) imputation"? This article attempts to answer such a question in the case of RF similarity, by studying approaches that make its computation robust to the presence of missing values by neither relying on imputing–and possibly biasing–the missing values nor completely ignoring them. This research is in line with some other analogous attempts at defining similarity measures that can deal with missing values without resorting to imputation, like the Bhattacharyya-based distance of [1], the Euclidean-like distances considered in [31, 32], the Feature Weighted Penalty based Dissimilarity (FWPD) introduced in [13, 14], or the Mahalanobis-based distance of [35] which is however specifically focused on time series.

The first issue that arises when computing an RF-similarity with missing values is what to do when a point x needs to be processed in a node of the tree where a feature/attribute is tested for which x's value is missing. For this, in principle one can rely on the approaches proposed in the context of standard classification decision trees (see, e.g., [10, 15, 25, 26, 28]). There are basically two options: to have x stop being processed when reaching a node u that tests one of x's missing values; or to follow on both children of u —we refer to these two basic alternative strategies as Nan-stop and Nan-both, respectively. However, the choice of the appropriate approach also depends on the fact that the result of processing a point in the single trees needs to be then aggregated at forest level.

One of our key observations, leading to the first and main methodological contribution, is that in general an RF-similarity can be formulated as the computation of a function over a family of sets. More precisely, each point gets associated a set of tests' results for each tree. Two points are then compared based on some specific function of the families of tests associated to them. If there are no missing values, the set associated with a point x for a tree T is the set of tests' results on the root to leaf path followed by x in T. This perspective is immediately evident for the RatioRF distance [7], an RF similarity measure which was recently introduced in the context of clustering, and whose original set based definition inspired our ideas. However, and we will show this in the article, many other RF-distances can be expressed from a set-based perspective. Given this set based perspective, it is very straightforward to derive an immediate extension of any RF-distance computation to the missing value scenario: when Nan-stop is employed, the set for a point x on the tree T contains only the tests encountered from the root to the node where x stops in T. When Nan-both is employed, the set contains the tests on the possibly several root-to-leaf paths followed by x in T.

The second contribution starts from the following observation: as a result of employing Nan-stop and Nan-both strategies, we have that trees are not any more equivalent in the amount of information that they provide for each given point. A point x in a tree containing tests on x's missing values might be associated to a very small number of tests (when Nan-stop is used and x stops being processed soon) or to very many tests (when Nan-both is used and x follows several root-to-leaf paths). In this article, we also investigate the possibility to replace the standard way of averaging distances computed at tree level by a more proper weighted average, where the weight assigned to each tree leverages the difference in the amount of tests' results the tree provides for the points (taken as a measure of the information content provided by the trees). The rationale at the basis of this variant is that it allows to differentiate the contribution of individual trees to the final similarity computation, hence it may be better suited to the missing values scenarios.

In this article, we will instantiate such ideas, presenting all the details together with an empirical evaluation, for the RatioRF mentioned described above, which was shown to outperform all other RF-based similarities [5, 9, 34, 38, 44] in the clustering scenario. Subsequently, we present a set-based reformulation of many other RF-distances, which permits to easily derive the definition of their robust-to-missing-data version—we also present the definition of such variants in the article. The experimental evaluation is based on 15 datasets, and is aimed at assessing the difference between the similarity measure computed on the complete dataset and the similarity measure computed by the proposed framework on the same dataset after injecting increasing percentages of missing values. In other words, we assessed a sort of absolute robustness of our proposed framework to the presence of missing values, independently of any task (e.g., classification, clustering) the distance might then be used for. Obtained results were promising, showing that the similarity measure computed by our approach (on the dataset with injected missing values) does not significantly change with respect to the similarity computed on the original complete dataset, also in the presence of a high rate of missing values. Moreover, results show that our approach compares very favourably to many alternative state-of-the-art distances which can be computed with missing data.

Summarizing, the main contributions of this article are the following:

— We present a general framework for dealing with missing data in the RF-based similarity computation, which starts from a set-based (re-)formulation of RF-distances;
— We instantiate such framework for the RatioRF case, empirically validating it with 15 different datasets;
— We show how the framework can be applied to other RF-based similarities, providing a set-based formulation of such distances together with their extension to the missing data case.

Two final notes: (i) we stress that our goal is to build RF-similarities which are robust to missing data, and not to employ RFs to perform missing data imputation, as done in many other RF-based approaches to missing data such as [10, 18, 36]—some of them are also based on RF-distances, which are however computed on complete data and used to perform imputation. (ii) a preliminary version of the ideas presented in this article has been investigated in [29], where we studied some simple variants of RatioRF in the context of distance based RF-clustering in the presence of missing data. Here we extend such work and take a more structural approach, by (i) proposing a framework which can be applied to many RF-based similarity measures, and (ii) considering and focusing on the assessment of the similarity measure itself rather than its efficacy in a particular setting like clustering.

The rest of the article is organized as follows: in Section 2, we review the standard approaches to compute RF distances, fixing the notation, and introducing the basic concepts. The proposed approach is presented in Section 3, and evaluated in Section 4. Section 5 concludes the article.

Table 1. Table of Symbols

Symbol	Meaning
T	a tree
$r(T)$	the root of the tree T
U	set of objects
$\ell(x)$	the leaf the object $x \in U$ is reaching in the tree T
$L(T)$	the set of leaves of tree T
θ_v	a test in a node v
$\theta_v(x) \in \{yes, no\}$	answer of the object $x \in U$ to the test θ_v in the node v
v_Y (v_N)	the child of v connected to v via the edge associated with Y (N)
$P_T(x)$	set of tests/answers in the path x is taking in the tree T
$P_T^{MP}(x)$	set of tests/answers in the paths x is taking in the tree T, when the Nan-both policy is used
$\text{lca}(\ell(x), \ell(y))$	the *least common ancestor* of $\ell(x)$ and $\ell(y)$
$\text{depth}(v)$	the depth of node v in the tree T under consideration

2 COMPUTING DISTANCES WITH RFS

We start by defining some notation and recall the definition of several RF-based distance measures. A summary of the notation is reported in Table 1.

Let U be a set of objects/points and A (for attributes) a set of binary tests T defined over the whole set U. We assume that for each object $x \in U$ and each test $\theta \in T$ there is a unique value $\theta(x) \in \{yes, no\}$. A decision tree on the ground set of objects/points U and the test set A is a binary tree T where: (i) each internal node v is associated to a binary test $\theta_v \in A$; (ii) the two edges connecting the node to its children are associated with the two possible results—denoted Y for *yes* and N for *no*—of performing test θ_v on an object from U. Furthermore, v_Y (resp. v_N) denotes the child of v connected to v via the edge associated with Y (resp. N); $r(T)$ denotes the root of T. Let v be a node of T at level $h+1$ and $\theta_1, b_1, \theta_2, b_2, \ldots, \theta_h, b_h$ be the sequence of nodes (tests) and edges (results), encountered on the unique path from $r(T)$ to v. Then, it is possible to associate to v the set of objects $S_v = \{x \in U \mid \theta_i(x) = b_i, i = 1, \ldots, h\}$. In words, a node v is representative of (or it *contains*) all the objects that, when tested according to the adaptive strategy represented by the decision tree T, follow the path from the root to v. For each object x there is a single leaf containing it denoted as by $\ell(x)$. Let $P_T(x)$ be the set of pairs $(test, result)$ associated to x by the strategy/tree T

$$P_T(x) = \{(\theta, b_x^\theta) \mid \theta \text{ is a test on the path from the root } r(T) \text{ to the leaf } \ell(x) \text{ and } b_x^\theta = \theta(x)\}. \quad (1)$$

Let θ be a test and $b \in \{Y, N\}$. We say that x agrees with (θ, b) if $\theta(x) = b$. Similarly, objects x and y agree on test θ if $\theta(x) = \theta(y)$.

The decision tree T can be employed as a tool for selecting a set of features Φ relevant for assessing the similarity between pairs of objects from the universe U. In particular, in [7] the authors define

$$\Phi = \{(\theta_v, b) \mid v \text{ is a node of } T,\ b \in \{Y, N\}\}$$

as the set of possible outcomes of the tests used by the decision tree. For an object x its feature set $X = P_T(x)$ is defined as a set of test results on the path from $r(T)$ to the leaf $\ell(x)$ associated to x by the decision tree. These are the features from Φ that are most relevant for x, in the sense of being sufficient to identify x. Following the perspective of [7], the underlying principle of RF-based similarity is that tree T allows to compare objects x, y as represented by the set of features $P_T(x)$, and $P_T(y)$, respectively.

We now summarize the definition of some state-of-the-art representative measures in the field of RF-based similarities. The main idea, in all these distances, is that the similarity between two objects x and y can be measured by looking at the way they answer to the tests in the tree T, i.e.,

by exploiting $P_T(x)$ and $P_T(y)$. The set of measures we chose to include in this analysis is meant to cover the different approaches (leaf based, path based, probability mass based, etc.) used to derive similarity measures from RFs. For each measure μ we present here the computation of $\mu(x, y)$ on a single tree T of the forest, which we denote by $\mu_T(x, y)$. As discussed in the introduction, the final value $\mu(x, y)$ is in all cases obtained as the arithmetic average over the trees of the forest, i.e., $\mu(x, y) = \sum_T \mu_T(x, y)$.

Shi distance. Let $\text{Shi}_T(x, y)$ denote the distance proposed in [9, 34]—which was the first RF-based similarity measure—defining the similarity 1 if the paths $P_T(x)$ and $P_T(y)$ are identical (i.e., if the two objects end in the same leaf of the tree), and 0 otherwise. We have

$$\text{Shi}_T(x, y) = \begin{cases} 1 & \text{if} \ell(x) = \ell(y) \\ 0 & \text{if} \ell(x) \neq \ell(y) \end{cases} \tag{2}$$

Zhu distance. Let $\text{Zhu}_T(x, y)$ denote the distance proposed in [44]. In this case, the idea is that the larger the overlap between the two paths $P_T(x)$ and $P_T(y)$, the larger their similarity between x and y. We have[2]

$$\text{Zhu}_T(x, y) = \frac{\text{depth}(\text{lca}(\ell(x), \ell(y))}{\max\{\text{depth}(\ell(x)), \text{depth}(\ell(y))\}}, \tag{3}$$

where $\text{lca}(\ell(x), \ell(y))$ is the *least common ancestor* of $\ell(x)$ and $\ell(y)$.

Ting distance. Let $\text{Ting}_T(x, y)$ denote the distance proposed in [38], which exploits the concept of *probability mass* of the nodes in the sub-path in common between $P_T(x)$ and $P_T(y)$. We have

$$\text{Ting}_T(x, y) = \frac{|S_{\text{lca}(\ell(x), \ell(y))}|}{n}, \tag{4}$$

where $S_{\text{lca}(\ell(x), \ell(y))}$ is the set of points that when tested with decision tree T reach the least common ancestor of $\ell(x)$ and $\ell(y)$, and $n = |U|$ denotes the number of points in the dataset.

RatioRF. Let $\text{RatioRF}_T(x, y)$ denote the distance proposed in [7] following an axiomatic definition of similarity measures given by Tversky [40]. We have

$$\text{RatioRF}_T(x, y) = \frac{|P_T(x) \cap P_T(y)|}{|P_T(x) \cap P_T(y)| + |P_T(x) \dot{-} P_T(y)| + |P_T(y) \dot{-} P_T(x)|}, \tag{5}$$

where (i) $P_T(x) \cap P_T(y)$ is the set of tests on which x and y agree, among the features in $P_T(x) \cup P_T(y)$, and (ii) $P_T(x) \dot{-} P_T(y)$ (or, equivalently, $P_T(y) \dot{-} P_T(x)$) is the set of tests that are relevant for x (y) and on which y (x) disagrees, e.g.

$$P_T(x) \dot{-} P_T(y) = \{(\theta, b) \mid (\theta, b) \in X \text{ and } \theta(y) \neq b\}. \tag{6}$$

3 DEALING WITH MISSING VALUES

The above similarity measures are formalized assuming that: (i) all tests are defined on every object; (ii) there is a single root to leaf path associated to each object x, yielding the set $P_T(x)$ on which the measures are computed. The presence of data with missing values in a dataset means that for some object x and test θ the value $\theta(x)$ is not defined, a situation that is indicated by $\theta(x) = NAN$. We now detail the mechanisms in the learning and training phases which permit to deal with such situations.

[2]Note that the max at the denominator is only meant to normalize the measure in the interval [0, 1].

3.1 Learning Phase: Training Trees with Missing Data

In the learning phase, we aim at building the decision trees of the forest based on a random training subset of the input objects. We select the test θ to associate with the root of the decision tree in order to split the training objects into two subsets according to the way they behave on such test. The two subsets are then used as training sets for recursively building the two subtrees rooted at the children of the node associated to test θ. In the presence of missing data, we need to decide how an object x used in such a procedure for building the tree is moved down (to the right or the left subtree) when processed at a node associated to a test θ for which x's value is missing/not known, i.e., $\theta(x) = NAN$. Here we considered two options, which are in line with our goal of avoiding pre-imputation and exploit all objects of the dataset. The first approach, referred to as NanBoth, is to have x passed to the training set of both subtrees. This scheme seems to be really suitable for our view; actually, we are trying to select tests that separate objects in the training test: since, unless we implicitly impute some value to x, the test θ does not provide information to separate x from the other objects, we delay this to the subtrees. The second option we consider is not to pass x to any of the training sets for the subtrees (NanStop). Also this alternative policy seems to be very suitable for our framework, since it keeps each split in the tree T to be decided only on the basis of the subset of the training set that reaches that split because of actual (non-missing) values.

As a final consideration, it is important to observe that the learning phase is aimed at creating a tree as an adaptive sequence of splits that separate the objects from one another. Clearly, when using the NanBoth approach, any missing value will induce an increase of the number of tests, as the object x will need to be separated both in the right and the left subtree. Conversely, with the NanStop approach, a tree built over a set of objects with many missing values may have very few splits, since after a split θ is chosen, all the objects x s.t. $\theta(x) = NAN$ will be ignored and no split will be needed to separate them from the others.

3.2 Computing the Distance with Missing Data

When computing the RF-distance, we have a pair of objects x, y traverse each decision tree T built in the learning phase, in order to select the sets/paths $P_T(x), P_T(y)$ on which the similarity is computed. There are two issues that need to be addressed if we want to extend the similarity measures defined in (2)–(5) to deal with missing values, and in particular with the situation when an object x reaches a node v such that $\theta_v(x) = NAN$: (i) Should the pair (θ_v, NAN) be part of the set of features $X = P_T(x)$ describing x? (ii) What is the next node/test to consider for x between v_Y and v_N, i.e., how should we complete the partial root-to-leaf path for x beyond v? How should we decide, considering that the test result $\theta_v(x)$ does not say which of the edges Y or N to follow?

Regarding (i), it appears to be most reasonable not to consider such a node as part of the set X, as this would imply *unfounded* dissimilarity of x with any other objects y for which the test θ_v is defined. By *unfounded* we mean that we do not know whether the missing value of x on test θ_v agrees or not with $\theta_v(y)$, hence it would not be fair to assume it is different. Dually, it can also be the case that $\theta_v(y) = NAN$, it would be unfounded to say that x and y have the same value for test θ.

Regarding (ii), we use again the NanBoth approach and proceed by extending the path in both directions. This means that x will be tested on both v_Y and v_N and from each of them it will continue to a single child node, or to both children nodes, if the test is again on a missing value of x. Please note that in the learning phase both choices NanBoth and NanStop do not affect the construction of the remaining part of the tree. On the other hand, in the computation of the distance, the use of NanStop might result in a significant difference in the number of tests performed on a point.

Therefore, a tree would end up giving too little information on a point even if it contains several test that can help assessing the distance of x from y. Clearly, as a result of employing NanBoth, instead of the single path $P_T(x)$, the object x gets associated to a subtree of T that starts at the root of T, equivalently, a collection of root-to-leaf paths parting from one another at some node associated to a test where x is not defined.

In order to make explicit the relationship and the difference with respect to the analogous structure $P_T(x)$ defined for data without missing values in (1), let us denote by $P_T^{MP}(x)$ this tree or set of features (pairs of test and result) selected by the application of the NanBoth approach.

It is important to note that, for each x, we have that $P_T^{MP}(x)$ is a subtree of T that stems from the root of T. The RF similarities of the previous section, which are defined on *single paths* $P_T(x)$ and $P_T(y)$, need to be extended to the subtrees case in order to work in the above defined missing data framework. We investigate here a possible strategy to achieve this: if the RF-distance is reformulated so that it is the result of set-based operations on the *set* of tests in $P_T(x)$ and $P_T(y)$, then the extension to the missing case is straightforward, since we can consider also $P_T^{MP}(x)$ and $P_T^{MP}(y)$ as sets. We will show the detailed instantiation of this principle in the RatioRF case in Section 3.3; in Section 3.4 we also provide the set-based reformulation of some other RF-similarities, thus suggesting that the proposed framework represents a general strategy to easily include mechanisms of missingness management in classic RF-similarities.

3.3 MissRatioRF—an Extension of RatioRF for Missing Values

In order to apply the proposed framework to the RatioRF measure, we have to formulate it as a result of operations on sets. Let $\Phi_T(XY) = P_T(x) \cup P_T(y)$ denote the set of features restricted to those employed by the tree to describe x and y. Then let X_T (resp. Y_T) be the element of $\Phi_T(XY)$ on which x (resp. y) agrees. Then, the RatioRF distance defined in Equation (5) can be rewritten as follows:

$$\text{RatioRF}_T(x, y) = \frac{|X_T \cap Y_T|}{|X_T \cup Y_T|} \tag{7}$$

i.e., the Jaccard distance computed on the restricted set of features, that the tree selected for x and y.

$\text{RatioRF}_T(x, y)$ is defined in terms of union and intersection of sets of tests from $P_T(x), P_T(y)$. It is then natural to consider a generalization of $\text{RatioRF}_T(x, y)$ where the role of $P_T(x), P_T(y)$ is taken by their generalizations $P_T^{MP}(x), P_T^{MP}(y)$. More, precisely, in the presence of missing value we take as the set of feature extracted by T for comparing x and y the tests in the subtree $\Phi_T(XY) = P_T^{MP}(x) \cup P_T^{MP}(y)$. Let X_T^{MP} (resp. Y_T^{MP}) be the element of $P_T^{MP}(x) \cup P_T^{MP}(y)$ on which x (resp. y) agrees. Then, we can compute the similarity of x and y with respect to a given decision tree T like in (7) by substituting X_T, Y_T, in the computation of $\text{RatioRF}_T(x, y)$, with X_T^{MP}, Y_T^{MP}. The resulting distance, denoted by MissRatioRF_T is defined as follows:

$$\text{MissRatioRF}_T(x, y) = \frac{|X_T^{MP} \cap Y_T^{MP}|}{|X_T^{MP} \cup Y_T^{MP}|} \tag{8}$$

From trees to forest. As usual in RF approaches, the final RF-similarity is obtained by aggregating tree-level information. We propose here two ways to perform such aggregation, leading to two variants of the MissRatioRF measure: the first represents the classical way of defining an RF-similarity, i.e., by averaging the tree-level measures on different trees, whereas the second is specifically designed for the missing data case. More formally, given a trained RF whose trees are $T_1, \ldots, T_m$, and a pair of points $x, y \in U$, let $\text{MissRatioRF}_t(x, y)$ be the similarity computed according to (8) from the decision tree T_t.

Then, we define the first variant of RF similarity measure $\text{MissRatioRF}_1(x, y)$ by averaging over all decision trees, i.e.,

$$\begin{aligned} \text{MissRatioRF}_1(x, y) &= \frac{1}{m}\sum_{t=1}^{m} \text{MissRatioRF}_t(x, y) \\ &= \frac{1}{m}\sum_{t=1}^{m} \frac{|X_{T_t}^{MP} \cap Y_{T_t}^{MP}|}{|X_{T_t}^{MP} \cup Y_{T_t}^{MP}|} \end{aligned} \tag{9}$$

The second version starts from the following observation: with the scheme defined in (9) each tree contributes *with the same weight* $(1/m)$ to the final RF-similarity measure. However, the presence of missing values can make distinct trees provide a very different amount of information content. This is a consequence of two main aspects: (i) the use of the NanBoth training approach, by requiring that an object x must be separated by other objects on several paths, possibly induces many more eligible tests; (ii) the use of the NanBoth in the testing phase, which clearly increases the eligible tests considered for x. It is then natural to consider the weighted average of the similarity computed by the single trees weighted with respect to the ratio of the information content of the tree with respect to the information content of the whole forest. More formally, we can define the information content $IC(T, x, y)$ of the tree t with respect to the comparison of two objects x, y as the number of bits provided by the tests from the tree T that contribute to the computation of the similarity between x and y, i.e.

$$IC(T, x, y) = |X_T^{MP} \cup Y_T^{MP}| = |P_T^{MP}(x) \cup P_T^{MP}(y)|. \tag{10}$$

Therefore, the total information content of the forest $T_1, \ldots, T_m$ with respect to the comparison of two objects x, y is given by

$$IC(x, y) = \sum_{t=1}^{m} |X_{T_t}^{MP} \cup Y_{T_t}^{MP}|. \tag{11}$$

If we then take the weighted average (w.r.t. IC) of the tree-level MissRatioRF similarities, we have our second variant:

$$\begin{aligned} \text{MissRatioRF}_2(x, y) &= \sum_{t=1}^{m} \frac{IC(T_t, x, y)}{IC(x, y)} \frac{|X_{T_t}^{MP} \cap Y_{T_t}^{MP}|}{|X_{T_t}^{MP} \cup Y_{T_t}^{MP}|} \\ &= \sum_{t=1}^{m} \frac{|X_{T_t}^{MP} \cup Y_{T_t}^{MP}|}{\left(\sum_{t=1}^{m} |X_{T_t}^{MP} \cup Y_{T_t}^{MP}|\right)} \cdot \frac{|X_{T_t}^{MP} \cap Y_{T_t}^{MP}|}{|X_{T_t}^{MP} \cup Y_{T_t}^{MP}|} \\ &= \frac{\sum_{t=1}^{m} |X_{T_t}^{MP} \cap Y_{T_t}^{MP}|}{\sum_{t=1}^{m} |X_{T_t}^{MP} \cup Y_{T_t}^{MP}|}, \end{aligned} \tag{12}$$

where the third line follows from the second by noticing that the enumerator of the first fraction is equal to the denominator of the second fraction, and the denominator of the first fraction is the same for all terms of the outer summation, hence we can factor it out.

Let us say that a test is eligible for x and y if it appears on some root to leaf path followed by x or y in some tree of the forest. Then, the last expression in (12) says that $\text{MissRatioRF}_2(x, y)$ is equal to the ratio between the number of eligible tests for x and y (in the whole forest) on which they agree, and the total number of eligible tests for x and y.

3.4 Missing-Data Extensions of Other RF-Similarities

The MissRatioRF distance introduced in the previous section exploits the set based formulation of the RatioRF distance to derive a missing data variant. Our claim is that the approach is general enough to apply whenever an RF-based similarity measure $\mu_T(x, y)$ can be expressed in terms of intersections and union of the sets of tests in the paths $P_T(x), P_T(y)$ of the underlying decision tree T. Actually, in such cases, we can proceed as in the case of RatioRF to obtain the missing-value variant Miss$\mu(x, y)$ by simply replacing $P_T(x), P_T(y)$ with $P_T^{MP}(x), P_T^{MP}(x)$, which can be computed with the NanBoth approach in the testing phase. In this section, we provide the missing data formulation for all the measures presented in Section 2: these measures were not presented, in their original version, in a set-based formulation; in this section we show how they can be reformulated in such perspective, leading to their missing data extension.

In the following, we are going to use $P_T(x)$ to denote also the subtree (path) of T associated to its sequence of tests. Moreover, for a subtree T' of T , we denote the set of its leaves by $L(T')$.

Shi distance for missing values. By noticing that when $P_T(x)$ is a path, $\ell(x) = L(P_T(x))$ we can rewrite (2) as

$$\text{Shi}_T(x, y) = |L(P_T(x)) \cap L(P_T(y))|/|L(P_T(x)) \cup L(P_T(y))|, \tag{13}$$

which leads to the following generalization to the case of missing values

$$\text{MissShi}_T(x, y) = \frac{|L(P_T^{MP}(x)) \cap L(P_T^{MP}(y))|}{|L(P_T^{MP}(x)) \cup L(P_T^{MP}(y))|}. \tag{14}$$

Zhu distance for missing values. We start by observing that $\text{depth}(\text{lca}(\ell(x), \ell(y)) = |P_T(x) \cap P_T(y)|$, and $\text{depth}(\ell(x)) = |P_T(x)|$. Moreover, considering that the denominator $\max\{|P_T(x)|, |P_T(y)|\}$ in Equation (3) simply represents a normalization factor, and that the similarity is mainly defined by the numerator, if as a normalization factor we use $|P_T(x) \cup P_T(y)|$ we can rewrite (3) as

$$\text{Zhu}_T(x, y) = |P_T(x) \cap P_T(y)|/|P_T(x) \cup P_T(y)|, \tag{15}$$

which leads to the following generalization to the case of missing values

$$\text{MissZhu}_T(x, y) = \frac{|P_T^{MP}(x) \cap P_T^{MP}(y)|}{|P_T^{MP}(x) \cup P_T^{MP}(y)|}. \tag{16}$$

Ting distance for missing values. Using $\text{lca}(\ell(x), \ell(y)) = P_T(x) \cap P_T(y)$, and noticing that in the absence of missing values, in T we have one leaf per object, i.e., $n = \sum_{\ell \in L(T)} |S_\ell|$, we can rewrite (4) as

$$\text{Ting}_T(x, y) = \frac{\sum_{\ell \in P_T(x) \cap P_T(y)} |S_\ell|}{\sum_{\ell \in L(T)} |S_\ell|}, \tag{17}$$

which leads to the following generalization to the case of missing values

$$\text{MissTing}_T(x, y) = \frac{\sum_{\ell \in L(P_T^{MP}(x) \cap P_T^{MP}(y))} |S_\ell|}{\sum_{\ell \in L(T)} |S_\ell|}, \tag{18}$$

where we recall that S_v is the set of points that when tested with the decision tree will arrive at the node v.

As a concluding remark, we observe that, given the set-based formulation, the weighted average proposed for the RatioRF scheme might be also employed for all the other RF-based distance measures considered in this section.

Table 2. Datasets

Problem	#Obj	#Feat	#DifVal
Small size			
Iris	150	4	78
BreastTissue	106	9	925
Lung	32	54	4
Ecoli	336	7	92
Fertility	100	9	36
Seeds	210	7	1202
Cryotherapy	90	6	77
StoneFlakes	70	8	324
WBC	683	9	10
Moderate size			
Energy	1500	24	11847
Volcano	1078	65	70070
Flickr	1000	87	47158
Gas	1641	128	209127
UAV-1	1100	54	43859
UAV-2	1100	18	12887

In the table, "#Obj" represents the number of objects, "#Feat" the number of features, and "#DifVal" the number of different values in the whole dataset.

4 EXPERIMENTAL EVALUATION

In this section, we present the experimental evaluation of the proposed approach applied to the RatioRF framework. In particular, we present the empirical details, we evaluate the main difference between the different variants of the proposed scheme, and we conclude with a comparison with alternative distances which avoid pre-imputation of data.

4.1 Experimental Details

The method, in all its variants, has been evaluated on 15 different datasets, briefly summarized in Table 2. We considered two main classes: small datasets and datasets of moderate size. The former class is representative for all those problems which involve few objects, as the classical medical domains which involve few patients, in which the missing data problem is highly impacting. The second class of problems is meant at investigating how the method scales when larger datasets are considered. Please note that, since the final output is the complete pairwise distance, scaling to very large datasets can be too computationally demanding, therefore we focussed on problems of moderate size. All the datasets of small size are taken from the standard UCI ML repository,[3] using them as provided (except Lung, for which we removed two features which already contain missing values, and StoneFlakes and WBC, for which we removed subjects which already contain missing values). The remaining are taken from real world larger size challenging problems. In particular, (i) Energy is a part of the ElectricityLoadDiagrams20112014 Data Set of the UCI ML repository, in which we used hourly profiles for the first 30 days of the first 50 subjects; (ii) Volcano is related to the analysis of volcano seismic events [23], using data provided by J.M. Londoño-Bonilla and the Observatorio Vulcanológico y Sismológico de Manizales, Colombia —we thank him for the data—,

[3] https://archive.ics.uci.edu/ml/index.php

and preprocessed as described in [23]; (iii) Flickr is part of the dataset "Favourite Images", a dataset used to study personal aesthetics, which can be downloaded from the author's website,[4] where we considered only the first five users; (iv) Gas is part of the dataset Gas Sensor Array Drift Data Set of the UCI ML repository, which contains only objects from the class 3; (v) UAV-1 and UAV-2 are parts of the Unmanned Aerial Vehicle (UAV) Intrusion Detection Datasets, available from the authors' website,[5] including the first 1100 signals of Dataset 1 and 4.

For all the experiments, we evaluated the proposed approach with a specific pipeline, particularly designed for the case at hand. Instead of focusing on validating the obtained RF-distance in a given application domain (like, for example, clustering), here we evaluate how *different* is the RF-distance computed on the incomplete data with respect to the RF-distance computed on the original data. In this way, we assess in a better way the impact of the missing data on the distance computation, independently of the task the distance is used for. The proposed pipeline works as follows. First, for the given dataset, we simulated missingness by artificially removing some values using the MCAR (Missing Completely At Random) protocol [30], which removes data completely at random, without considering relations among features. The only constraint we used was not to remove all the values of a given object, i.e., all objects should have at least one value. Then we applied the proposed MissRatioRF measure. Finally, we computed the original RatioRF similarity on the complete dataset and we compared MissRatioRF with it. In order to have a "baseline value", we also derived a second copy of the RatioRF similarity matrix on the complete dataset, comparing it with the reference RatioRF distance. This is meant to give an idea of the variability derived from the intrinsic randomness in the RFs method.

To quantitatively compare two similarity matrices, we can employ different strategies [33], based on statistical tests (e.g., the classic Mantel test [22]) or others ideas. Here we used the topological approach proposed in [45], which compares two similarity matrices by measuring the distance between the adjacency matrices of the two graphs derived from them. These graphs are derived here with the K-NN approach, i.e., the vertices are the objects and each object is adjacent to all and only its K nearest neighbours, as given by the similarity matrix. In our experiments we set $K = \sqrt{n}$, where n is the number of objects of the problems. The measure, which we call "TopDist" (Topological Distance), ranges between 0 and 1, the higher this value the more different the two similarity matrices.

In our experiments, we evaluated six levels of data missingness: 0.05, 0.1, 0.2, and so on up to 0.5. For each level and each dataset, we repeated the whole procedure 30 times. For what concerns the proposed approach, after a preliminary empirical evaluation (not shown here), we decided to use the following parametrization: 100 trees, building each of them with a random selection of 50% of the objects of the dataset. For datasets with more than 400 objects, we followed the suggestions in [7], and employed only a fixed amount of objects to learn each tree (in our experiments we set this number to 128). This would permit to reduce the computational burden introduced by large datasets, still keeping reasonable performances, as shown in [7]. We evaluated the two learning strategies described in previous section, namely NanStop and NanBoth. In the former case, each tree of the forest is grown until the maximum depth, whereas in the second each tree is grown only until the depth of $\log(n)$. We investigated the two versions of the distance MissRatioRF_1 and MissRatioRF_2, as given in Equations (9)–(12). For what concerns the original RatioRF distance, we used the same parametrization as that used with our approach: 100 trees, building each of them with a random selection of 50% of the objects of the dataset (or 128 for datasets with more than 400 objects). We employed Extremely Randomized Trees [16], which can be trained without any

[4] http://www.cristinasegalin.com/research/projects/phd/soft_biometrics/Dataset_IEEEForensics.zip
[5] http://mason.gmu.edu/~lzhao9/materials/data/UAV/

supervision: these methods have been shown to be surprisingly good not only for classification [16], but also in other unsupervised contexts such as anomaly detection [20], clustering [8], and, crucially, RF-distance computation [6].

4.2 Part 1: Analysis of MissRatioRF

In this section, we report some results aimed at evaluating some aspects of the proposed approach. One of the most interesting ones is the difference between the two variants MissRatioRF$_1$ and MissRatioRF$_2$, which represent two different strategies to aggregate the tree-level similarities. We report in Figure 1 the comparison between the two variants for all possible datasets. In each plot, we report the topological distance Topdist between the two MissRatioRF variants and the original distance on the complete data, for an increasing level of missingness. Each curve represents the average over the 30 different repetitions and the two learning strategies NanStop and NanBoth. The horizontal line at the bottom of the plot represents the topological distance between the reference RatioRF and the second copy of RatioRF. The first observation is that the topological distance, for all variants, is very low. Considering that the topological distance ranges between 0 and 1, we can say that, in general, our MissRatioRF is very close to the RatioRF distance computed on the complete data, also for drastic levels of missingness (0.5, i.e., when half of the values are missing). Please note that the topological distance between two copies of the RatioRF measure (on the complete data) is larger than zero: This represents the minimum displacement which is intrinsic in the RatioRF measure, due to the randomness in the process.

Concerning the difference between the two variants, we can observe that for low levels of missingness the two distances are more or less equivalent. When considering larger values of missingness, two different behaviours appear: On the datasets of small scale, there is not a predominance of one of the two variants, with 5 datasets on which MissRatioRF$_2$ is better than MissRatioRF$_1$, and 4 on which the opposite behaviour holds. On the contrary, on datasets of larger scale MissRatioRF$_2$ is always better than the counterpart. This seems to be reasonable: on these sets each tree is built with a small fraction of the objects (128), and this fact probably induces more diverse trees with respect to the case of small datasets, in which each tree is built with 50% of the dataset. MissRatioRF$_2$ is able to explicitly consider and quantify the diversity between the trees: considering that this diversity is drastically amplified by the mechanisms of missingness management we employed, it seems reasonable to note a more remarkable improvement with large levels of missingness.

To get a quantitative analysis of the difference, we performed a paired t-test between the performances of the two variants along the different repetitions, using as significance level 0.05. As a summary, we can say that in 74 cases over 90 (15 datasets times 6 missingness levels) there is a statistically significant difference: this is particularly true for large levels of missingness (i.e., 14 over 15 for level 0.5, 13 over 15 for level 0.4), and for datasets of moderate size (35 over 36, considering all levels of missingness).

For what concerns the two different learning strategies, we follow the same procedure described above, by comparing them considering the average over the two variants of the distance. We did not report the results of the comparison dataset per dataset, but we show in Figure 2 two aggregated results: the average over datasets of small size (Figure 2(a)) and the average over datasets of moderate size (Figure 2(b)), to analyse the difference between these two scenarios. From the plots we can observe that, for small size problems, the two learning strategies are rather equivalent, whereas, when considering larger size datasets, we can observe a clear trend: with a small level of missingness, then NanStop seems to be more adequate. On the other side, when the level of missingness is increasing, then NanBoth becomes the favourite choice. This is somehow expected: NanStop stops the propagation of objects for which we have a missing value for the given rule. With a large level of missingness, it is highly probable to stop many objects at the early depths

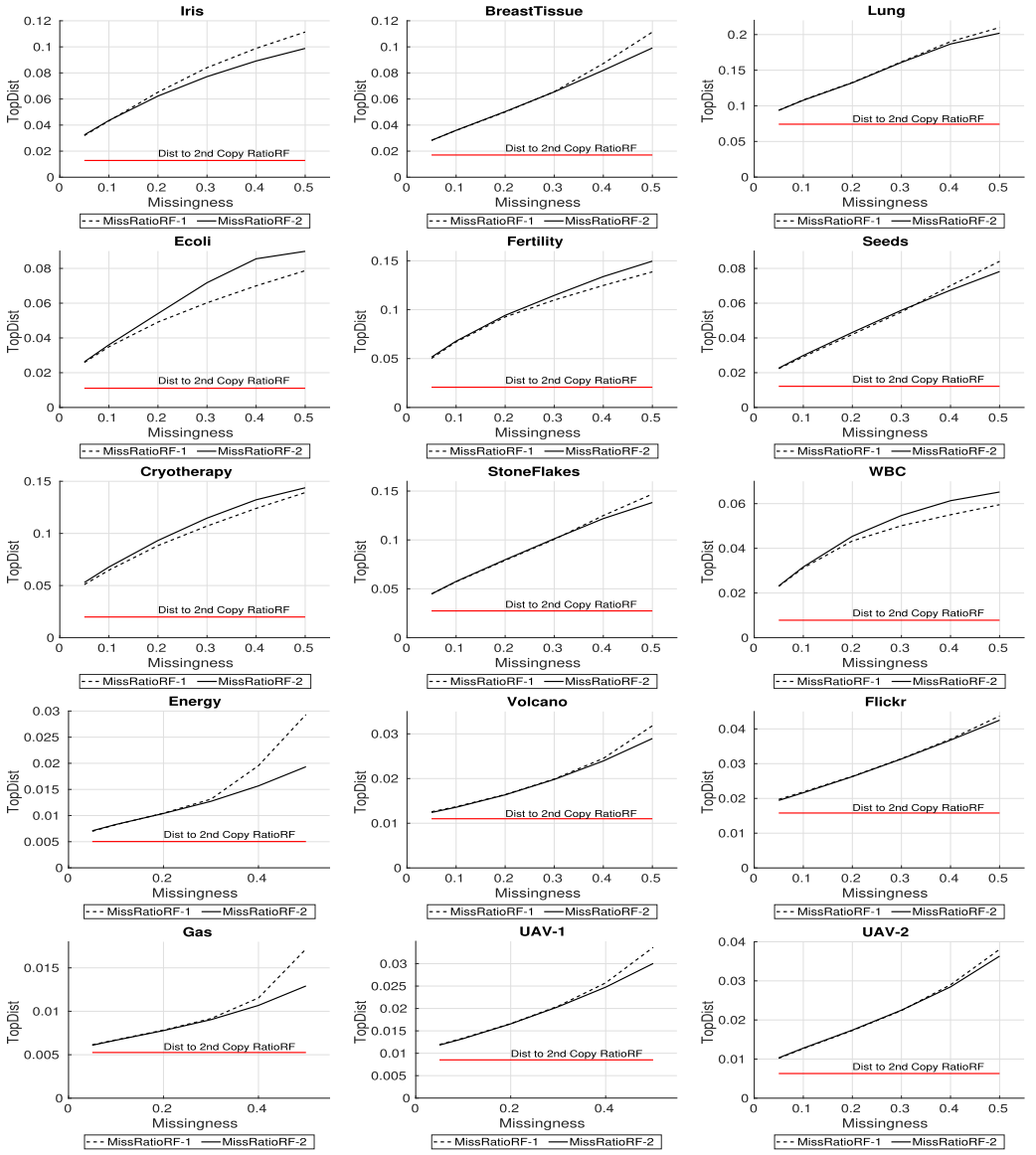

Fig. 1. Comparison between the two variants MissRatioRF_1 and MissRatioRF_2 on all datasets.

of the tree, resulting in worse descriptions. On the contrary, the NanBoth scheme propagates the objects with missing values along the two different branches, thus resulting in more descriptive trees for large levels of missingness. Also in this case, we performed a statistical paired t-test to assess the statistical significance of the curves, following the strategy described above. As a summary, we can see that the differences are statistically significant in 77 cases over 90, again with the majority for higher levels of missingness (15 over 15 for both levels 0.4 and 0.5) and for datasets of moderate size (36 over 36). If we perform the paired t-test by aggregating all the results for the datasets of small size and all the results for the datasets of moderate size (corresponding to the plots (a) and (b) in Figure 2), then the statistical significance is always present.

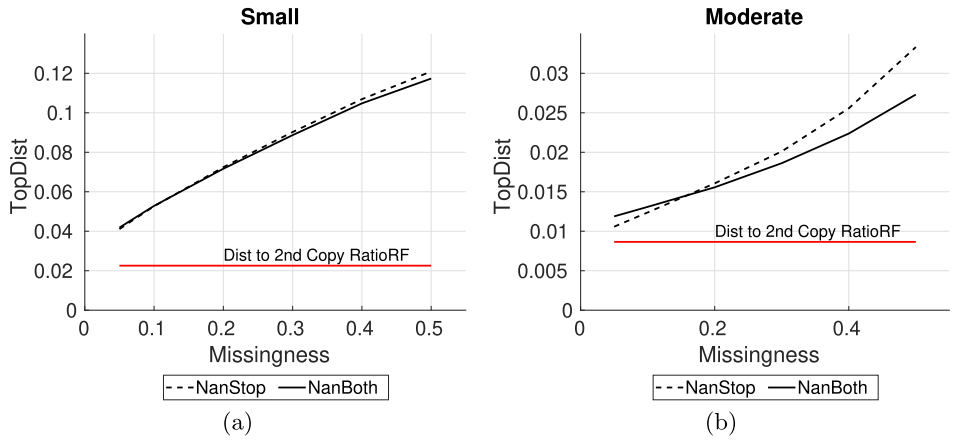

Fig. 2. Comparison between the two learning strategies NanBoth and NanStop. Averages over (a) datasets of small size; (b) datasets of moderate size.

4.3 Part 2: Comparison with Other Literature Distances

In this part, we compare the proposed approach with other distances which can be computed with missing values. In all cases, we have two variants of the distance: the original, which is computed on the complete set, and the robust-to-missing-data variant. In particular, we follow the same protocol as before, starting from a dataset with missing values, computing the robust-to-missing-data distance, comparing it in terms of topological distance with the reference measure computed on the complete data. Please note that, for having a fair comparison, the starting datasets with missing values were kept identical to those used with our method. We analysed the following measures:

- **HEOM, HVDM, HEOM-Redef, HVDM-Redef**: extensions of Euclidean distance able to deal with missing data—see [31, 32]. Even if these distances can also work with heterogeneous data, here we used their version for continuous numeric data. The reference distance, for the complete case, is the Euclidean distance.
- **Mean ED**: another extension of the Euclidean Distance, proposed in [2, 3] (also considered in [31, 32]). Also in this case, the reference measure is the Euclidean Distance.
- **Bhattacharyya**: an extension of the Mahalanobis distance for data with missing entries, based on the Bhattacharyya distance [1]. In this case, the reference distance is the Mahalanobis distance.
- **FWPD-Clas, FWPD-Clus, FWPD-Best**: three variants of the FWPD measure [13, 14]. In particular, here we select the parameter α in three different ways: (i) for FWPD-Clas we set $\alpha = 0.1$, which represents the best value when the distance is used for classification, according to [14]; (ii) for FWPD-Clus we set $\alpha = 0.25$, the best value when the distance is used for clustering [13]; (iii) finally, for FWPD-Best we select, in every experiment, the best value of α (i.e., the one leading to the lowest topological distance).

Summarizing results are reported in Figure 3. In particular, we displayed in part (a) the results averaged on all the datasets of small size and the different 30 repetitions, and in part (b) the results are averaged over the datasets of moderate scale. In all cases, for MissRatioRF we select the best version.

By looking at the plots, we can observe that MissRatioRF is largely better than alternatives, especially for (i) large levels of missingness and (ii) datasets of moderate size. Please note that the

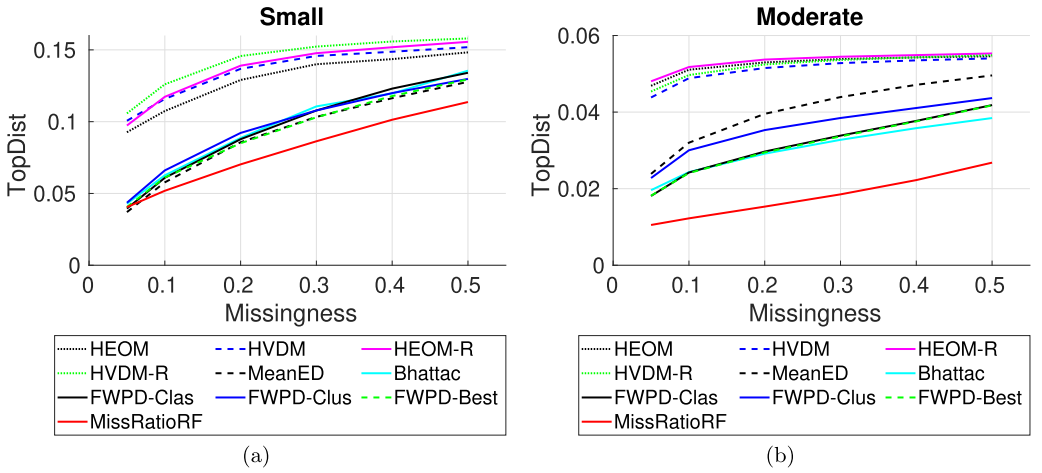

Fig. 3. Comparison with other distances: averages over different groups of datasets: (a) small size, (b) moderate size.

comparison is fair: the starting datasets were identical, and the TopDist measure used to compare the two versions of a given distance is, by construction, independent of the scale of the distance. Actually, the distance is used to derive an unweighted KNN graph (which only depends on the ranking, not on the scale of the input distance); then the TopDist measure is obtained by comparing the graph obtained with the original distance and that obtained on the robust-to-missing-data version.

Also for this set of results we made a set of statistical tests, comparing with a t-test MissRatioRF with competitors (significance level 0.05). After performing a Bonferroni correction for multiple tests, we have found that MissRatioRF is better than all alternatives, with a statistical significance and for all the levels of missingness, for all the large datasets. For the small datasets, the statistical significances are present in all cases when MissRatioRF is compared with HEOM, HVDM, HEOM-Redef, and HVDM-Redef, in 48 cases over 54 (9 datasets times 6 levels of missingness) when compared with MeanED, in 47 over 54 when compared with Bhattacharyya, FWPD-Clas, and FWPD-Clus, and on 49 over 54 when compared with FWPD-Best. This confirms that the proposed approach represents a valid alternative to classic as well as advanced distances robust to missing data.

4.4 An Analysis on Extremely High Levels of Missingness

We conclude our experimental section by presenting an analysis to understand the behaviour of the proposed scheme when an extremely high level of missingness is present. In particular we repeated the experiments reported in Section 4.2 for a level of missingness ranging from 0.6 to 0.9 (step 0.1). We performed the analysis only on those datasets for which the highest level of missingness could be computed, i.e., datasets for which at least one feature per object can be guaranteed: Lung, Energy, Volcano, Flickr, Gas, UAV-1, and UAV-2. Please note that we could compute all the 30 repetitions only for the NanStop configuration, whereas for the NanBoth configuration, we only compute one run. Actually, the NanBoth training, for such high levels of missingness, is very computationally demanding, at least in our implementation, since NanBoth propagates to both children all objects which are not able to provide an answer to the test (which, for a high missing level, represent the majority). Results are displayed in Figure 4, reporting the averages over the

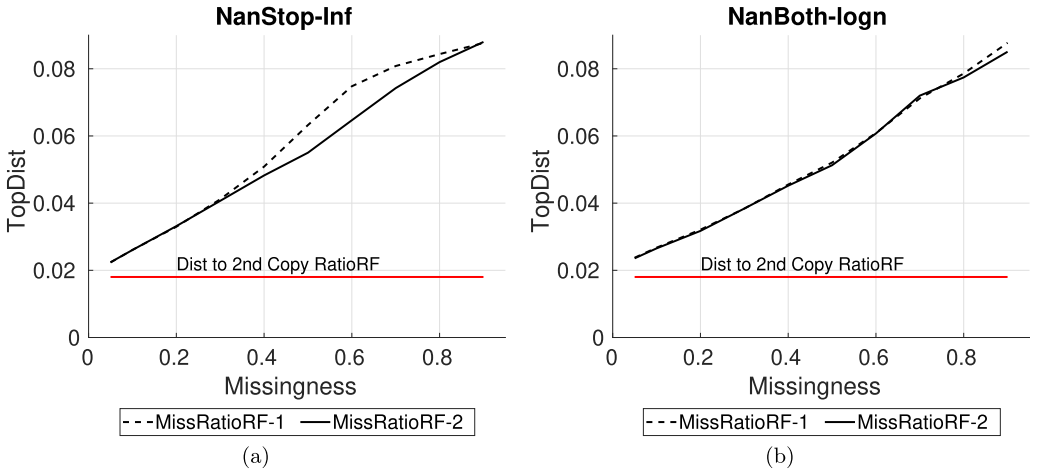

Fig. 4. Extreme levels of missingness: averages datasets Lung, Energy, Volcano, Flickr, Gas, UAV-1, and UAV-2: (a) configuration NanStop, (b) configuration NanBoth (due to high computational requirements, displayed results are for one run only).

different datasets. For the sake of readability, we also reported the same average for levels 0.05–0.5, in order to have an idea of the trend. It is evident how the proposed distance scales well, even in so large levels of missingness. The curves also appear to confirm the observation made before that the MissRatioRF$_2$ variant is better suited when trees are more diverse in the amount of information provided. Such diversity is expected to increase with higher values of missingness—explaining the diverging behaviour of the curves in the high middle range—up to a point where, for extremely high values of missing values, all the trees collapse to a set of very few tests so that the weighted average of MissRatioRF$_2$ does not make a difference anymore.

5 CONCLUSIONS

In this article, we presented a general framework for dealing with missing data in the RF-based similarity computation, which neither removes the data with missing values nor applies any pre-imputation step. We instantiated such framework in the case of RatioRF similarity measure, providing significant empirical evidence of its effectiveness with extensive experiments on 15 datasets. We show that the same framework can be applied to several other state-of-the-art RF-based similarities, providing their extensions to the missing data case.

As future directions, the immediate option would be to empirically evaluate also the other missing-data distances, to quantify the improvement provided by the introduced mechanisms. As a second direction, we consider it worth investigating more the weighting scheme introduced in the second variant of MissRatioRF. Our present approach has the advantage of not being task-dependent, while still trying to weight/balance the importance of the trees on the basis of the distribution of missing values. However, given a specific task (e.g., classification), a different promising approach could be to try and learn the weights for precise exploitation of the distance, in a context-dependent way. In this respect, the framework introduced in this papers opens the possibility of deriving advanced imputation-free approaches for clustering or classification, for example by combining them with recent interesting distance-based classification approaches (e.g., [43]): techniques like K-Nearest Neighbour still have many possibilities for improvements [42], and may drastically benefit from carefully designed distances robust to missing data.

REFERENCES

[1] L. AbdAllah and I. Shimshoni. 2013. A distance function for data with missing values and its application. *International Journal of Computer Science and Engineering* 7, 10 (2013).

[2] Loai AbdAllah and Ilan Shimshoni. 2014. Mean shift clustering algorithm for data with missing values. In *Proceedings of the 16th International Conference on Data Warehousing and Knowledge Discovery, DaWaK 2014*. Springer, 426–438.

[3] Loai AbdAllah and Ilan Shimshoni. 2016. k-means over incomplete datasets using mean euclidean distance. In *Proceedings of the 12th International Conference on Machine Learning and Data Mining in Pattern Recognition, MLDM 2016*. Springer, 113–127.

[4] Deepak Adhikari, Wei Jiang, Jinyu Zhan, Zhiyuan He, Danda B. Rawat, Uwe Aickelin, and Hadi A. Khorshidi. 2022. A comprehensive survey on imputation of missing data in internet of things. *Computing Surveys* 55, 7 (2022), 1–38.

[5] S. Aryal, K.M. Ting, T. Washio, and G. Haffari. 2020. A comparative study of data-dependent approaches without learning in measuring similarities of data objects. *Data Mining and Knowledge Discovery* 34, 1 (2020), 124–162.

[6] Manuele Bicego and Ferdinando Cicalese. 2023. On the good behaviour of extremely randomized trees in random forest-distance computation. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD2023)*. Springer, 645–660.

[7] M. Bicego, F. Cicalese, and A. Mensi. 2023. RatioRF: A novel measure for random forest clustering based on the Tversky's ratio model. *IEEE Transactions on Knowledge and Data Engineering* 35, 1 (2023), 830–841.

[8] M. Bicego and F. Escolano. 2020. On learning random forests for random forest clustering. In *Proceedings of the 2020 International Conference on Pattern Recognition*. 3451–3458.

[9] L. Breiman. 2001. Random forests. *Machine Learning* 45 (2001), 5–32.

[10] L. Breiman, J. H. Friedman, R. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Wadsworth.

[11] S. Van Buuren and K. Oudshoorn. 1999. *Flexible Multivariate Imputation by MICE*. TNO, Leiden.

[12] A. Criminisi, J. Shotton, and E. Konukoglu. 2012. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision* 7, 2–3 (2012), 81–227.

[13] Shounak Datta, Supritam Bhattacharjee, and Swagatam Das. 2018. Clustering with missing features: A penalized dissimilarity measure based approach. *Machine Learning* 107 (2018), 1987–2025.

[14] Shounak Datta, Debaleena Misra, and Swagatam Das. 2016. A feature weighted penalty based dissimilarity measure for k-nearest neighbor classification with missing features. *Pattern Recognition Letters* 80 (2016), 231–237.

[15] Jerome H. Friedman. 1977. A recursive partitioning decision rule for nonparametric classification. *IEEE Transactions on Computers* 26, 4 (1977), 404–408.

[16] P. Geurts, D. Ernst, and L. Wehenkel. 2006. Extremely randomized trees. *Machine Learning* 63, 1 (2006), 3–42.

[17] Md Kamrul Hasan, Md Ashraful Alam, Shidhartho Roy, Aishwariya Dutta, Md Tasnim Jawad, and Sunanda Das. 2021. Missing value imputation affects the performance of machine learning: A review and analysis of the literature (2010–2021). *Informatics in Medicine Unlocked* 27 (2021), 100799.

[18] T. Ishioka. 2013. Imputation of missing values for unsupervised data using the proximity in random forests. In *Proceedings of the International Conference on Mobile, Hybrid, and On-Line Learning*. 30–6.

[19] Janus Christian Jakobsen, Christian Gluud, Jørn Wetterslev, and Per Winkel. 2017. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC Medical Research Methodology* 17, 1 (2017), 1–10.

[20] Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2012. Isolation-based anomaly detection. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 6, 1 (2012), 1–39.

[21] Qian Ma, Yu Gu, Wang-Chien Lee, Ge Yu, Hongbo Liu, and Xindong Wu. 2020. REMIAN: Real-time and error-tolerant missing value imputation. *ACM Transactions on Knowledge Discovery from Data (TKDD)* 14, 6 (2020), 1–38.

[22] N. Mantel. 1967. The detection of disease clustering and a generalized regression approach. *Cancer Research* 27, 2_Part_1 (1967), 209–220.

[23] M. Orozco-Alzate, P. A. Castro-Cabrera, M. Bicego, and J. M. Londoño-Bonilla. 2015. The DTW-based representation space for seismic pattern classification. *Computers & Geosciences* 85, Part B (2015), 86–95.

[24] T. D. Pigott. 2001. A review of methods for missing data. *Educational Research and Evaluation* 7, 4 (2001), 353–383.

[25] J. R. Quinlan. 1986. Induction decision trees. *Machine Learning* 1 (1986), 81–106.

[26] J. R. Quinlan. 1989. Unknown attribute values in induction. In *Proceedings of the 6th International Machine Learning Workshop*. 164–168.

[27] J. R. Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc.

[28] J. Ross Quinlan. 1987. Decision trees as probabilistic classifiers. In *Proceedings of the 4th International Workshop on Machine Learning*. Elsevier, 31–37.

[29] Matteo Raniero, Manuele Bicego, and Ferdinando Cicalese. 2022. Distance-based random forest clustering with missing data. In *International Conference on Image Analysis and Processing*, Cham: Springer International Publishing. 121–132.

[30] D. B. Rubin. 1976. Inference and missing data. *Biometrika* 63, 3 (1976), 581–592.
[31] M. S. Santos, P. H. Abreu, S. Wilk, and J. Santos. 2020. How distance metrics influence missing data imputation with k-nearest neighbours. *Pattern Recognition Letters* 136 (2020), 111–119.
[32] Miriam Seoane Santos, Pedro Henriques Abreu, Alberto Fernández, Julián Luengo, and João Santos. 2022. The impact of heterogeneous distance functions on missing data imputation and classification performance. *Engineering Applications of Artificial Intelligence* 111 (2022), 104791.
[33] J. W. Schneider and P. Borlund. 2007. Matrix comparison, Part 1: Motivation and important issues for measuring the resemblance between proximity measures or ordination results. *Journal of the American Society for Information Science and Technology* 58, 11 (2007), 1586–1595.
[34] T. Shi and S. Horvath. 2006. Unsupervised learning with random forest predictors. *Journal of Computational and Graphical Statistics* 15, 1 (2006), 118–138.
[35] D. Sitaram, A. Dalwani, A. Narang, M. Das, and P. Auradkar. 2015. A measure of similarity of time series containing missing data using the Mahalanobis distance. In *Proceedings of the 2015 International Conference on Advances in Computing and Communication Engineering*. IEEE, 622–627.
[36] D. J. Stekhoven and P. Buhlmann. 2011. MissForest: Non-parametric missing value imputation for mixed-type data. *Bioinformatics* 28, 1 (2011), 112–118.
[37] J. A. C Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter. 2009. Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. *BMJ* 338 (2009), b2393.
[38] K. M. Ting, Y. Zhu, M. Carman, Y. Zhu, and Z.-H. Zhou. 2016. Overcoming key weaknesses of distance-based neighbourhood methods using a data dependent dissimilarity measure. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*. 1205–1214.
[39] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17, 6 (2001), 520–525.
[40] A. Tversky. 1977. Features of similarity. *Psychological Review* 84, 4 (1977), 327.
[41] K. Wagstaff. 2004. Clustering with missing values: No imputation required. In *Classification, Clustering, and Data Mining Applications*, D. Banks, F. R. McMorris, P. Arabie, and W. Gaul (Eds.). Studies in Classification, Data Analysis, and Knowledge Organisation. Springer, Berlin, Heidelberg.
[42] Shichao Zhang. 2021. Challenges in KNN classification. *IEEE Transactions on Knowledge and Data Engineering* 34, 10 (2021), 4663–4675.
[43] Shichao Zhang, Xuelong Li, Ming Zong, Xiaofeng Zhu, and Ruili Wang. 2018. Efficient kNN classification with different numbers of nearest neighbors. *IEEE Transactions on Neural Networks and Learning Systems* 29, 5 (2018), 1774–1785.
[44] X. Zhu, C. C. Loy, and S. Gong. 2014. Constructing robust affinity graphs for spectral clustering. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition, CVPR 2014*. 1450–1457.
[45] D. A. Zighed, R. Abdesselam, and A. Hadgu. 2012. Topological comparisons of proximity measures. In *Proceedings of the Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*. Springer, 379–391.

Received 28 November 2023; revised 25 March 2024; accepted 29 March 2024