RatioRF: a novel measure for Random Forest clustering based on the Tversky's Ratio model

Manuele Bicego, *Member, IEEE,* Ferdinando Cicalese, and Antonella Mensi

Abstract—In this paper we propose $\mathrm{RatioRF}$, a novel Random Forest-based similarity measure for clustering. We build upon Tversky's ratio model definition of similarity [1] and specialize it to the Random Forest case. We study some properties of the proposed axiomatic similarity measure and present an extensive experimental clustering analysis involving different datasets and configurations. Results confirm that $\mathrm{RatioRF}$ represents a good alternative to other similar measures for clustering recently studied in the literature.

Index Terms—Random Forests, clustering, similarity measure, Tversky model, decision trees

1 Introduction

We start by introducing the two main ingredients of our study: Random Forests (RF) and Tversky axiomatic model of similarity.

A **Random Forest** (RF) [2], [3], [4] is a well known and widely exploited tool for Pattern Recognition and Machine Learning, in the class of ensemble methods [5], [6]. A RF realizes an ensemble of decision trees [7]. Each decision tree is a hierarchical partitioning of the object space, in which each split is determined by a threshold on a single feature. RFs achieve robustness by exploiting a randomization mechanism in the learning of the different trees, which are then aggregated to get the final model. In [2], Breiman shows that this aggregation exhibits different interesting theoretical properties: in particular, he derives an upper bound on the generalization error, in terms of the strengths of individual trees and their correlation. RFs have been extensively investigated for classification and regression, and shown to compete well with most effective approaches such as Support Vector Machines or Neural Networks. However, in other pattern recognition scenarios, such as clustering, RFs have received less attention, and their potential is far from being completely understood.

The second ingredient of our study is **Tversky ratio model of similarity**. Introduced in [1], this model is based on a set of axioms that are aimed at capturing the way humans choose levels of similarity[1]. Tversky assumes that objects from some universe U are represented by features from some space Φ. Precisely, Φ is assumed to be a "limited set of features which are relevant to perform the task of assessing pairwise similarity. Thus, the representation of an object as a collection of features is viewed as a product of *a prior process of extraction and compilation*." To compute a similarity, one should consider features that are in common to the two objects as well as features that characterize one object but not the other. More formally, in [1], it is argued that any similarity measure $s^{Tv}(\cdot,\cdot)$ that satisfies *the natural axioms*, and is scaled to have value between 0 and 1, should have the following structural definition: given objects $x, y \in U$, which are respectively *represented* by the sets of features X and Y taken from the feature space Φ, the similarity of x and y is given by

$$s^{Tv}(x,y) = \frac{f(X \cap Y)}{f(X \cap Y) + \alpha f(X - Y) + \beta f(X - Y)}, \quad (1)$$

with $\alpha, \beta \geq 0$ and f being a measure on the feature space.

Our aim is to exploit RF to perform—quoting [1]—"the *process of extraction and compilation*" needed in the ratio model, by selecting the "*limited set of features which are relevant to perform the task of assessing pairwise similarity*". In this respect, we consider a decision tree as a tool that selects the most significant features of the objects, i.e., those that can best describe the distinctive elements of the objects that are classified. We use Random Forests' ability to make such a selection more robust.

In order to properly asses the novelty of our approach let us start by commenting on previous uses of Tversky's model. There have been several re-elaborations of the ratio model together with proposals about how to use them in contexts where data are organized in tree structures [8], [9], [10], [11], [12]. However, most of these studies either explicitly assume that such a tree structure represents a taxonomy [8], [9], or implicitly rely on tests over disjoint paths of the tree to be distinct [12]. These assumptions are not generally valid in the case of decision trees (hence of RFs).

The main contribution of this paper is to derive and assess a more principled implementation of Tversky similarity, which is directly linked to the structure of a decision tree. A decision tree can be interpreted as an algorithm that, given a set of possible tests/features, characterizes each object x by selecting (adaptively) a specific subset of the tests/features (a root-to-leaf path of the tree). According to Tversky's model, when we compare two objects, we should restrict to a minimal set of features that identify the two objects, and quantify among these, the features that

- *M. Bicego, F. Cicalese and A. Mensi are with the Department of Computer Science, University of Verona, Italy*
  E-mail: {manuele.bicego,antonella.mensi,ferdinando.cicalese}@univr.it.

Manuscript received XXX; revised XXX

[1] On the basis of extensive experimental analysis, Tversky argues that any similarity measure should satisfy such axioms.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2021.3086147, IEEE Transactions on Knowledge and Data Engineering

distinguish the two objects from those that they share. Our implementation is to consider as set of features on which to compare x and y, *all and only* the tests encountered in the paths they follow in the tree. This contrasts to the approach used by most of the existing similarity measures that refer to Tversky's axiomatization and, like ours, are also based on decision trees. Such approaches basically identify features with the vertices of the tree, disregarding the semantic of such vertices, namely the actual test they represent. To be more specific, they consider as features shared by object x and y only the tests in the common prefix of their paths; and as differentiating features all the other tests in the two paths, ignoring that on many of these latter tests the two objects might actually agree.

We will elaborate on the structural differences between our approach and previous analogous definitions of ratio-model-based similarities in Section 3.2. In particular, we will formally compare to the information theoretic measure of [8], another axiomatic definition of similarity, and to the RF-based measures introduced in [12]. We then give experimental evidences of the efficacy of our novel model by employing it in several clustering scenarios. Our empirical evaluation involves 15 datasets, 4 clustering methods and different parametrizations of the forest, comparing $\mathrm{RatioRF}$ with the RF-clustering measures introduced in [12], the original one proposed in [13] and the two more recent RF-based measures [14] and [15]: the obtained results largely confirm the suitability of the proposed measure for RF-based clustering.

The rest of the paper is organized as follows: Section 2 discusses closely related work; Section 3 presents the proposed clustering measure; Section 4 contains the experimental evaluation, and Section 5 concludes the paper.

2 Related work

The basic idea of exploiting RFs to extract a meaningful similarity measure between objects—to be used in a classic distance-based clustering algorithm—has been proposed in [2], [13]: given a tree of the forest, we can consider two objects as similar if they end up in the same leaf, since they have answered in the same way to all tests in their path; a straightforward and natural similarity measure is thus represented by the number of times – over the whole set of trees – that two objects end up in the same leaf. Given the distance, in [13] the final clustering is then obtained using the PAM (Partitioning Around Medoids) algorithm. Despite its simplicity, this clustering approach has shown to be very useful in many different applications [16], [17], [18], [19]. Recently, [12] extended the approach of [2], [13] by improving the definition of the RF-similarity: the idea is to consider that objects which do not end up in the same leaf may be similar as well, since two objects that separate after t tests can be considered to be more similar than two objects that split after $t-1$ tests. Therefore, authors proposed a similarity which is proportional to the averaged length of the path that the two objects have in common in their traversal down to the leaves (in a second variant, the paths are weighted). Given the distances, the clustering is then obtained using spectral clustering. It is observed that this distance is based on another axiomatic definition of similarity, given in [8], which uses an information theoretic approach. In Section 3.2, we will discuss this approach and its relation with our proposal. We will show that as an implementation of the measure axiomatized in [8], the similarity proposed in [12] is somehow too restrictive and a bit myopic, and does not exploit all the information contained in the trees of a Random Forest.

Our method can be also somehow related to recent works on metric learning, not specifically designed for general purpose clustering, such as [14], [15], [20], [21], [22]. These methods, which exploit Random Forests in the metric learning process, are based on different ideas and approaches, often involving labels: in our experimental evaluation we consider [14], being strictly related to the weighted version of the RF-based distance proposed in [12], and [15], which can be considered as an extension of [14]. In particular, in [14] the defined similarity measure depends on the distribution of the data: the idea is that two objects are more similar if they are in a sparse region than if they are in a denser one. Based on this principle, the authors of [14] proposed a mass-based dissimilarity measure which exploits the probability mass of a region, computed using Isolation Forests [23] by considering the number of points falling in a particular node of the tree – this is analogous to the weighted version of the distance in [12]. In [15], authors introduce the m_0 distance as an extension of the m_p distances [24], a class of distances which can be also implemented with Random Forests: actually the distance in [14] represents the Random Forest implementation of m_1.

A line of research more loosely connected to our approach regards the use of decision trees and RFs (or RF-inspired mechanisms) to directly perform clustering. For early examples of clustering with decision trees, see, e.g., [25], [26], [27]. In [28] a clustering method is employed to build dictionaries for a Bag of Words classification of images; in particular, Extremely Randomized Trees [29] are used to build clustering forests in which each leaf represents a distinct visual word – in this sense, forests are interpreted as partitioners of the space. This approach has been extended in [30] along different directions: among others, by working directly on pixels (and not on visual descriptors), by considering the whole tree as a hierarchy of clusters, and by using the obtained forest also for classification. In [31], Perbet and colleagues propose a two-step clustering algorithm, in which they interpret the trees of the forest as multiple partitions of the input space, to be merged and refined using a graph-based algorithm. In [32], a K-means-style clustering algorithm has been proposed, in which every cluster is described using an Isolation Forest [23], a particular type of Random Forests designed for one-class classification. Finally, in [33], the authors introduce the "Cluster forests" algorithm, which employs a RF-like mechanism: the algorithm realizes an ensemble (similarly to RF) of clusterings by finding projections on which good local clusterings exist, aggregating these clusterings to get the final result.

3 The similarity measure RatioRF

Let U be the ground set of objects. Elements of U will be also referred to as points. A decision tree on U is a binary tree T

where: (i) each internal node ν is associated to a binary test θ_ν; (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_ν on an object from U. We use $r(T)$ to denote the root of T. Let ν be a node of T at level $h+1$ and $\theta_1, b_1, \theta_2, b_2, \dots, \theta_h, b_h$ be the sequence of nodes (tests) and edges (results), encountered on the unique path from $r(T)$ to ν. Then, we associate to ν the set of objects $S_\nu = \{x \in U \mid \theta_i(x) = b_i, i = 1, \dots, h\}$. In words, a node ν 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 ν. For each object x there is a single leaf containing it and we denote it by $\ell(x)$.

For each object $x \in U$ we let $P_T(x)$ be the set of pairs $(test, result)$ associated to x by the strategy/tree T

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

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

3.1 Tversky's Ratio model with decision trees

Recall the definition of Tversky ratio model in (1). We want to employ a decision tree T to select the set of features Φ relevant for the assessment of similarity between pairs of objects from the universe U. In particular, we define

$$\Phi = \{(\theta_\nu, b) \mid \nu \text{ is a node of } T,\ b \in \{Y, N\}\} \tag{2}$$

as the set of possible outcomes of the tests used by the decision tree.

For an object x we define its feature set $X = P_T(x)$ 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.

In Tversky's model, for comparing objects x and y, we have to distinguish in their set of representing features, those that account for more similarity (and will contribute to the term $f(X \cap Y)$) from those that account for more difference (and will contribute to the terms $f(X - Y), f(Y - X)$).

The crucial point in our implementation of Tversky's model of similarity, where features are selected using a decision tree, is recorded in the following postulates:

1) only features in $P_T(x)$ and $P_T(y)$ are relevant for the comparison of objects x and y;
2) a feature $(\theta, b) \in P_T(x) \cup P_T(y)$ is either an element of either commonality or an element of discrimination between x and y according to whether $\theta(x) = \theta(y)$ or $\theta(x) \neq \theta(y)$.

For 1) we note that for each object x, the decision tree identifies a set of tests and results that are sufficient to define x and "what is not x". This is the set of features $P_T(x)$. In Tversky's words, the pairs of tests and results on $P_T(x)$ are the features that are "selected and compiled" for x. "Being x" means to agree on each $(\theta, b) \in P_T(x)$; conversely, disagreeing on some $(\theta, b) \in P_T(x)$ means "not to be x". Therefore, features neither in $P_T(x)$ nor in $P_T(y)$ should not be considered relevant to compare the pair of objects x, y, since such features were not chosen as relevant for (the definition of) either object. This is, e.g., the case of a feature $(\theta, Y) \notin X \cup Y = P_T(x) \cup P_T(y)$ and such that $\theta(x) \neq \theta(y)$. Even though on such a test θ the two objects disagree, the test θ has not been selected by the decision tree to represent either of the two objects: it was not chosen to define what is not x (hence dissimilar from x) nor to define what is not y (hence dissimilar from y).

For 2), note that a feature (θ, b) that is in $P_T(x)$ and not in $P_T(y)$ is important to distinguish x from $U \setminus \{x\}$ and it is not necessary to distinguish y from $U \setminus \{y\}$. However, feature (θ, b) should be accounted to assess the dissimilarity between x and y *only if*, on the test it represents, the two objects disagree. Indeed, when this happens, we have that the result of this test provides a relevant feature of x, one of those on the basis of which we decide what is not x, and *in particular* this test says that y is different from x.

As a result of the above observations, we define

$$X \dot{-} Y = \{(\theta, b) \mid (\theta, b) \in X \text{ and } \theta(y) \neq b\} \tag{3}$$

to be the set of features that are relevant for x and on which y disagrees. Symmetrically the set of features relevant for y and on which x disagrees are given by the set

$$Y \dot{-} X = \{(\theta, b) \mid (\theta, b) \in Y \text{ and } \theta(x) \neq b\} \tag{4}$$

We also define

$$X \mathbin{\dot{\cap}} Y = \{(\theta, b) \in X \cup Y \mid \theta(x) = \theta(y)\} \tag{5}$$

to be the set of features on which x and y agree, among the features in $P_T(x) \cup P_T(y)$, which are those relevant for describing them (i.e., for identifying one or the other).

Remark 1. *$X \mathbin{\dot{\cap}} Y$ is the maximal set of features in $X \cup Y = P_T(x) \cup P_T(y)$ on which x and y agree.*

We have $X \cup Y = P_T(x) \cup P_T(y) = (X \dot{-} Y) \cup (Y \dot{-} X) \cup (X \mathbin{\dot{\cap}} Y)$ but the inclusions $X \mathbin{\dot{\cap}} Y \supseteq X \cap Y$, $X \dot{-} Y \subseteq X - Y$, $Y \dot{-} X \subseteq Y - X$, might be in general strict.

We conclude that, when features are selected using a decision tree, $X \mathbin{\dot{\cap}} Y, X \dot{-} Y, Y \dot{-} X$ are the correct terms to be used to represent the feature sets $X \cap Y, X - Y, Y - X$ that define Tversky's ratio model. Therefore, a general *decision tree-based* version of Tversky's ratio model is given by

$$s^{TvDT}(x, y) = \frac{f(X \mathbin{\dot{\cap}} Y)}{f(X \mathbin{\dot{\cap}} Y) + \alpha f(X \dot{-} Y) + \beta f(Y \dot{-} X)}, \tag{6}$$

with $\alpha, \beta > 0$ and f a non-negative and monotonically non decreasing function defined on subsets of features.

In particular, we define the *Ratio-DecisionTree* similarity measure $\mathrm{RatioDT}(\cdot, \cdot)$ from (6) by choosing $\alpha = \beta = 1$, and f to be the function returning the cardinality of its argument:

$$\mathrm{RatioDT}(x, y) = \frac{|X \mathbin{\dot{\cap}} Y|}{|X \mathbin{\dot{\cap}} Y| + |X \dot{-} Y| + |Y \dot{-} X|}, \tag{7}$$

These choices guarantee that the similarity measure is symmetric and the corresponding dissimilarity obtained as $\sqrt{1 - \mathrm{RatioDT}(x, y)}$ is a metric [34].

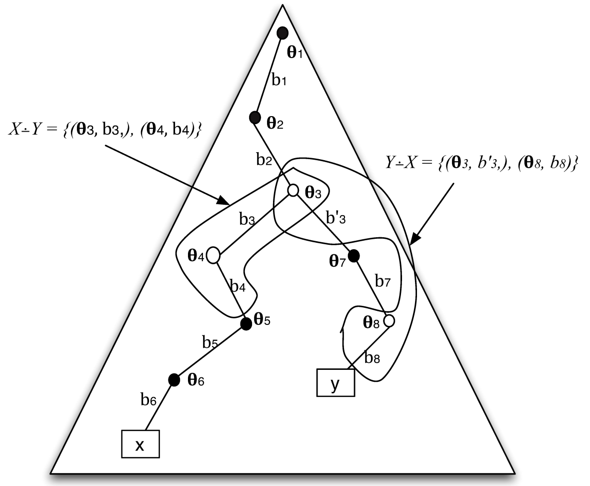

Fig. 1. The black nodes represent features shared by x and y, i.e., tests on which they agree. White nodes are tests on which x and y disagree.

Figure 1 gives a pictorial example of the terms in the computation of RatioDT. Assume that: (i) all tests $\theta_1, \ldots, \theta_8$ are different; (ii) $\theta_i(x) = \theta_i(y)$ for each $i \in \{1, 2, 5, 6, 7\}$. In the figure, tests on which x and y agree are shown as black dots and tests on which x and y disagree as white dots. Then, we have $|P_T(x) \cap P_T(y)| = |\{(\theta_1, b_1), (\theta_2, b_2)\}| = 2$. This accounts for the features on the intersection of the two paths. However, the example is meant to show a case where $P_T(x) \cap P_T(y)$ are not the only features that the two objects share: in the remaining 6 features used to define x and y, 3 more are features on which they agree (namely $(\theta_5, b_5), (\theta_6, b_6), (\theta_7, b_7)$). Then, according to our definitions, we have $|X \Cap Y| = 5$, $|X \dot{-} Y| = 2$, $|Y \dot{-} X| = 2$. Hence, $\text{RatioDT} = \frac{5}{5+2+2} = \frac{5}{9}$. Later we will refer again to this same example while considering other measures in order to clarify the difference from RatioDT.

3.2 Comparison with other measures

In this section we qualitatively compare our proposed measure with two other measures: i) the one introduced in [8], a general-purpose similarity based on information theory and ii) one of the measures introduced in [12], a set of recent and high performing Random Forest-based clustering measures inspired by [8].

In [8] Lin derives a similarity measure based on an information theoretic perspective which exploits the concepts of *commonality* and *description*. These terms can be seen as an information theoretic re-elaboration of the terms in Tversky ratio model. More in detail, given two objects x, y, [8] defines the commonality between x and y as measured by $Inf[common(x, y)]$ where $common(x, y)$ is a proposition that states the common features between x and y and $Inf(p)$ is the amount of information contained in proposition p. Analogously, $description(x, y)$ is defined as a proposition that describes what x and y are. Then, in [8] the similarity between x and y is defined by:

$$s^{Lin}(x, y) = \frac{Inf[common(x, y)]}{Inf[description(x, y)]}. \tag{8}$$

Let $A \subseteq \mathcal{F}$ be a set of features, as defined in our setting. Let $x(A) = \{x \in U \mid \forall(\theta, b) \in A, \theta(x) = b\}$ be the set of objects that agree on all features (θ, b) in A. Let

$$f(A) = -\log \frac{|x(A)|}{|U|} = -\log \mathcal{P}_U[A] = Inf(A) \tag{9}$$

be the information content of the set of features A with respect to the uniform distribution $\mathcal{P}_U$ on the set of objects. Setting $\alpha = \beta = 1$, Tversky's similarity in (6) coincides with $s^{Lin}(x, y)$ of [8], where

$$common_T(x, y) = X \Cap Y$$

and

$$description_T(x, y) = P_T(x) \cup P_T(y),$$

according to our definition of common and discriminating features between x and y.

There is another interesting relationship between our RatioDT and Lin's information theoretic similarity. For a decision tree T, define the dyadic probability distribution $\mathcal{P}_T$, by setting

$$\mathcal{P}_T(x) = 2^{-h(\ell(x))}, \tag{10}$$

where, for each node ν of T we denote by $h(\nu)$ the number of edges on the path from the root to ν, i.e., the depth of ν, and, $\ell(x)$ is the leaf containing x. Then, for each $x \in U$, each test on the path leading to $\ell(x)$ provides exactly one bit of information. Under this assumption, and our interpretation of $common_T(x, y) = X \Cap Y$ and $description_T(x, y) = P_T(x) \cup P_T(y)$ we have that $s^{Lin}(x, y)$ basically coincides[2] with $\text{RatioDT}(x, y)$.

Inspired by [8], Zhu et al. define an alternative way to extract a similarity measure from a decision tree, proposing three novel RF-based similarity measures for clustering [12]. Let us concentrate on the second variant introduced in [12], called *ClustRF-Strct-Unfm*. Let T be a decision tree defined over the set of objects U, and, as above, let $h(\nu)$ denote, for each node ν of T, the number of edges on the path from the root to ν, i.e., the depth of ν. Given objects x and y let

$$\lambda(x, y) = h(\texttt{lca}(\ell(x), \ell(y)) \tag{11}$$

where $\texttt{lca}(\ell(x), \ell(y))$ denotes the lowest common ancestor of the leaves containing x and y. The *ClustRF-Strct-Unfm* similarity measure, which we denote here by s^{Zhu2} (i.e. second variant) is defined by

$$s^{Zhu2}(x, y) = \frac{\lambda(x, y)}{\max\{h(\ell(x)), h(\ell(y))}. \tag{12}$$

The motivation for such a definition—quoting from [12]—is that "a larger value of $\lambda(x, y)$ signifies more split tests both x and y have gone through together, implying higher similarity shared between them. A lower value in $\lambda(x, y)$ suggests subtle and weak similarity [...]". The denominator is meant to scale the similarity measured by $\lambda(x, y)$ to a value between 0 and 1. In the third variant (called *ClustRF-Strct-Adpt*) each node in $\lambda(x, y)$ is weighted by the inverse of its hierarchical neighborhood (number of points reaching that node [35]). However, the reasoning remains the same: the similarity between two objects is proportional to the length of the common prefix between

[2] The test where the two paths part is counted twice in RatioDT.

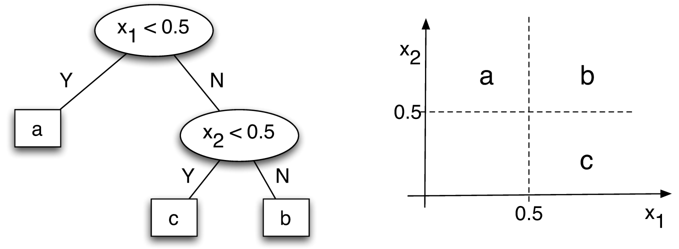

Fig. 2. Points $a, b,$ and c are separated by using tests on two different coordinates. If we ignore test $x_2 < 0.5$ when comparing a and b and only consider the fact that they disagree on the root test, then we cannot recognize that higher similarity between a and b with respect to the similarity between a and c.

$\ell(x)$ and $\ell(y)$—we will see the empirical behaviour of all variants of [12] in the experimental section.

In general, one can interpret $s^{Zhu2}(x, y)$ (and the third variant as well) as a variant of Lin's similarity $s^{Lin}(x, y)$ (resp. Tversky's ratio model) where $Inf[common(x, y)]$ (resp. $f(X \cap Y)$) is defined as the length of *only* the common prefix between $\ell(x)$ and $\ell(y)$, i.e., $\lambda(x, y)$. We believe that such a choice is both too restrictive and somehow myopic since it completely ignores to analyze the elements of commonality and discrimination in the sub-paths of $P_T(x), P_T(y)$ that extend below the test where they fork. In particular, all the tests not in the sub-path common to $P_T(x)$ and $P_T(y)$ are considered to account for dissimilarity between x and y; this directly contrasts with our postulate 2). For instance, on the example in Figure 1, we have $s^{Zhu2}(x, y) = 1/3$. More in general, it should be noted that most previous implementations of Tversky's models [8], [11], like Zhu [12], take only the common path followed by two objects as a proxy for the features they share. For an extreme example of the difference in the similarity computation between such approaches and our model, consider two objects that test different on the root vertex of the decision tree, and then test equally on all tests encountered in the remaining parts of the two distinct paths they follow. If h is the height of the tree, our measure consider the similarity of the two objects to be $\sim 1 - 1/h$, while the other approaches would assign similarity $\sim 1/h$. It is clear that in such cases, our measure takes better account of the possible correlation between tests that are significant for the objects and used in different parts of the tree, resulting in a more precise extraction of the features of similarity between the two objects.

As another example motivating our definition, let us consider the simple case in Figure 2. It is natural to assume that a is more similar to b than to c and also a is as similar to b as b is to c. However, if we consider the decision tree in the figure, we have $s^{Zhu2}(a, b) = s^{Zhu2}(a, c) = 0$ and $s^{Zhu2}(b, c) = \frac{1}{2}$, while, with our similarity measure, we have $\mathrm{RatioDT}(a, b) = \mathrm{RatioDT}(b, c) = \frac{1}{3}$, $\mathrm{RatioDT}(a, c) = 0$. The fact that the measure of similarity obtained using RatioDT is closer to the intuition is a direct consequence of the choice of considering as common features not only those in the common part of the paths, as measured by $\lambda(x, y)$. This is clearly a too restrictive measure of commonality, since, in general, we have

$$\lambda(x, y) \le |P_T(x) \cap P_T(y)| \le |X \Cap Y|$$

Moreover, it has other possible drawbacks when the similarity is meant to be used for clustering. In particular, in the construction of the decision trees, there may be situations when the first cuts splits an existing cluster (especially for not axis-aligned clusters); therefore, for several pairs of points in the same cluster it is very likely that the similarity extracted with the criterion of [12] will be zero or small; our measure, instead, can balance the loss of the first intercluster splits by considering also the later splits on which the points agree.

3.3 A RF-based clustering approach

The RatioDT similarity measure can be straightforwardly generalized to Random Forests by averaging the decision tree distance in eq. (7) over all the trees in the forest. More precisely, given a trained RF whose trees are $T_1, \ldots, T_m$, fix a pair of points $x, y \in U$ and let $\mathrm{RatioDT_t}(x, y)$ be the similarity computed according to (7) from the decision tree T_t. Then, we define the Random Forest similarity measure $\mathrm{RatioRF}(x, y)$ by averaging over all decision trees, i.e.

$$\mathrm{RatioRF}(x, y) = \frac{1}{m} \sum_{t=1}^{m} \mathrm{RatioDT_t}(x, y). \tag{13}$$

Given the measure, any distance-based clustering algorithm can be used to get the final clustering. If the clustering algorithm needs in input a dissimilarity, we transform our similarity into a dissimilarity using $\sqrt{1 - \mathrm{RatioRF}(x, y)}$, as done in [13].

In order to derive the measure we should have a trained Random Forest, which has to be learned without labels (clustering); in the literature, there are many strategies to derive it: the most common one consists in training a standard classification forest which, for clustering tasks, discriminates between the original data and a synthetically generated negative class [12], [13]. Typically, the negative class is obtained by sampling points from the product of empirical marginal distributions of the observed data: in this way the dependency structure of the original data is removed. Other options are based on Extremely Randomized Trees [29] (trees which are based on random splits – thus no need of labels), possibly also exploiting some extra information, if available [28], [30]. Please note that here we are not interested in optimizing this step, since we focus our attention on the derivation of the similarity measure.

4 Experimental Evaluation

In this section our proposed similarity measure RatioRF, defined in eq. (7), is evaluated and compared with 5 alternative measures, hereafter referred to as: d^{Shi}, s^{Zhu2}, s^{Zhu3}, d^{Ting}, and d^{Aryal}.

The measure d^{Shi}, defined in [2], [13], has been the first version of a Random Forest-based distance used for clustering: the similarity $s^{Shi}(x, y)$ between points x and y is defined as the number of trees where x and y fall in the same leaf, divided by the total number of trees; the dissimilarity d^{Shi} is then obtained as $\sqrt{1 - s^{Shi}(x, y)}$.

The two measures s^{Zhu2}, s^{Zhu3} were both introduced in [12] as refinements of s^{Shi}. More precisely, s^{Zhu2} represents the second variant introduced in [12] (there called *ClustRF-Strct-Unfm*), discussed in Section 3.2, whereas s^{Zhu3} is the third variant introduced in [12] (there called *ClustRF-Strct-Adp*). The latter extends s^{Zhu2} by weighing every node in the common path – the weight of a node is computed as the inverse of the number of points which reach such node[3].

The fourth measure we consider for comparison, d^{Ting}, was defined in [14]. It is another Random Forest-based similarity which shares some ideas with s^{Zhu3}. In particular, the authors of [14] define the distance between two points x and y as the ratio of points of the training set reaching the LCA (Lowest Common Ancestor) of x and y.

Finally, measure d^{Aryal} is an adaptation to the RF case of the very recent distance introduced in [15]. Here, the authors introduce the m_0 distance, an extension of the class of m_p distances [24], which can be also implemented with Random Forests – d^{Ting} is the Random Forest implementation of m_1.

The Matlab code used for the experiments is available at the first author's webpage[4].

4.1 Experimental details

As commonly done in clustering, the evaluation is performed using supervised datasets, removing labels, determining the clustering and comparing the obtained clusters with the original classes to assess clustering performances. We quantified the quality of the clustering results with two classical measures, the *purity index* and the *adjusted Rand index* (ARI) [36], [37].

To determine the purity, each cluster is assigned to the class label that is most frequent in that cluster. The purity index, ranging from 0 (worst) to 1 (best), is then the proportion of examples assigned to the correct label.

For the computation of the ARI, a contingency table is first built between the clustering and the true labeling. The classic Rand index is determined by assessing the agreement between the two partitions; in the ARI, such index is also corrected for the chance of the formation of the clusters. Also in this case, the higher the index value, the better the clustering.

For testing we used 15 different datasets, which are described in Table 1. The datasets are divided into two groups: "Group 1" contains those of small/moderate size (up to 1.500 objects), while "Group 2", contains larger datasets (up to ∼20.000 objects). Datasets in the first group are used to thoroughly test the framework with respect to the different parametrizations. The datasets in the second group are instead analysed with respect to a single parametrization in order to evaluate the performances of the proposed framework on larger scale problems.

Most of the datasets are available on the UCI ML Repository[5], except UAV, Volcano and Energy: UAV [38] was downloaded from the authors' web site[6], whereas the Volcano and Energy derive from two real world challenging non classical problems. Volcano deals with the classification of volcano seismic events [39][7], whereas Energy represents a peculiar behavioural biometrics scenario which exploits energy load profiles to identify users [40]. In the Volcano dataset the seismic signals have been collected at the Nevado del Ruiz volcano in Colombia, and preprocessed by the Observatorio Vulcanológico y Sismológico de Manizales, Colombia; each signal is represented using the averaged spectrogram (65 bands): there are five different seismic events, which represent the classes. In the Energy case, the goal is to recognize people by exploiting the way a person employs electrical energy in their house, from a behavioural biometrics perspective: here, we employed profiles from internet [41], using daily recordings (one registration every hour) of 50 users for one month; the signal is used as it is (i.e. a 24-dimensional vector) – for more details, analyses and alternatives for representation, please refer to [40].

In all experiments, to train the Random Forest we used the approach of [12], [13]. The binary decision trees are built with two classes: the positive class, which contains the points to be clustered, and a synthetically generated negative class, obtained with random sampling from marginals (the negative class has the same number of objects as the positive).

In all experiments related to the first group of datasets, each tree is built by randomly selecting 80% of training set – we choose this ratio in order to keep a good compromise between the number of objects used to train the tree and the diversity of them given by the randomization. For the selection of the best split, we employed two options: select it from all the features ("All"), or select it among a random 50% of the features ("Half"). This second option reduces the possible choices, but at the same time enhances the randomization inside each tree. As split criterion we used the classical Gini criterion, and the splitting process is stopped only when a node contains either one element or only objects with the same label. Finally, we report experiments with different forests sizes, namely 50, 100 and 200 trees.

For every configuration (number of trees - features), Random Forests have been trained 30 times, each one representing the starting point from which to compute the similarities. Given the similarity/dissimilarity, clustering is performed with four different methods: spectral clustering [42], a typical choice in more recent RF-clustering works (e.g., [12]), using the Ng-Jordan-Weiss normalized version [42], and repeating the inner k-means 20 times[8], Affinity Propagation [43], a renown distance-based clustering approach[9], and two Hierarchical Clustering schemes, the Complete-Link and Ward-Link (implementation of the Statistics and Machine Learning Toolbox of Matlab).

In the second group of datasets we perform all experiments using a single fixed parametrization. In particular, to cope with the high dimension of the dataset, we train each

[3] The first variant in [12] (*ClustRF-Bi*) coincides with s^{Shi}.

[4] http://profs.scienze.univr.it/∼bicego/code.html

[5] https://archive.ics.uci.edu/ml/datasets.php

[6] In particular we used "Dataset 1" from mason.gmu.edu/∼lzhao9/materials/data/UAV/

[7] We thank J.M. Londoño-Bonilla and the Observatorio Vulcanológico y Sismológico de Manizales, Colombia for the data.

[8] https://github.com/areslp/matlab/blob/master/spectral_clustering/SpectralClustering.m

[9] The version we used allows setting the number of clusters, see http://www.psi.toronto.edu

TABLE 1
Details of the datasets employed for testing.

— Group 1 —

Name	Description	#objects	#features	#cluster	#obj per cluster
Iris	Types of iris plant from sepal and petal features	150	4	3	50,50,50
Wine	Origin of wines from chemical features	178	13	3	59,71,48
Glass	Glass type from oxide content	214	9	4	70,76,17,51
WBC	Detection of Breast Cancer	683	9	2	444,239
BTissue	Types of Breast tissues from impedance features	106	9	6	21,15,18,16,14,22
Heart	Types of heart diseases	297	13	2	160,137
Lung	Types of Lung cancers	32	54	3	9,13,10
Parkinsons	Presence of Parkinsons from voices	195	22	2	48,147
Auto-mpg	Levels of city-cycle fuel consumption	398	6	2	229,169
Pima	Presence of Diabetes	768	8	2	268,500
Volcano	Types of seismic volcanic events	1078	65	5	153,333,237,251,104
Energy	Personal usage of electricity in houses	1500	24	50	30 each

— Group 2 —

Name	Description	#objects	#features	#cluster	#obj per cluster
Isolet	Types of Spoken letters	7797	617	26	min:298 - max:300
Gas	Types of gas from electronic noses	13910	128	6	min:1641 - max:3009
UAV	Unmanned Aerial Vehicle Intrusion Detection	19380	54	2	8663,10717

tree with a small number of randomly picked samples (256, in our experiments). Training with few random samples is a solution widely and successfully applied in other RF scenarios like outlier detection [23]. Very recently, this scheme has been successfully employed also for RF-clustering [44]. Training with few samples permits to have more compact trees, thus speeding up both the RF training phase and the distance computation. The other parameters are chosen as follows: Gini criterion is used for computing the split threshold; the best split is selected among 50% of the features; the size of the forests is fixed to 100 trees; spectral clustering is employed to get the final result (for the motivations behind these choices see the analysis of parameters proposed in Section 4.4). Also in this case experiments were repeated 30 times.

Results are described in the following subsections. We present 5 different analyses: the first 4 are based on experiments on the first group of datasets, whereas the last (Subsection 4.6) contains the analysis on larger scale problems. More in detail, in Subsection 4.2 we analyze the averaged accuracies according to different aspects (clustering methods, datasets, and parametrizations of the forests); in Subsection 4.3 we then present a deep comparison between the proposed measure and the most direct competitor s^{Zhu2}; subsequently, in Subsection 4.4 we perform an analysis of the impact of the different parameters on the proposed framework; finally, in Subsection 4.5 we propose some results using an automatic selection of the best parametrization.

4.2 Analysis 1: averaged accuracies

In this section we report the averaged accuracies related to the 12 datasets of the first group: first, we performed the average over all parameter configurations, repetitions, clustering methods and datasets. In total, 8640 experiments: 6 forest parametrizations × 30 repetitions × 4 clustering methods × 12 datasets. To investigate the different aspects, we also provided aggregated results with respect to: varying the datasets (720 experiments: 6 × 30 × 4), varying the clustering methods (2160 experiments: 6 × 30 × 12); and varying the parametrizations (1440 experiments: 30 × 4 × 12).

In order to have a statistically significant comparison, for every aspect we performed a paired t-test (significance 0.05) to compare the accuracies of the best performing measure with the others. Such results are reported in Table 2 for both the Purity and the Adjusted Rand Index. A bold value indicates that the best value is larger than the alternatives with a statistical significance.

What we can immediately observe is that our proposed novel measure largely outperforms all the competitors, with a statistically significant difference with respect to: the total average; all the different clustering methods; and all the different parametrizations.

With respect to the different datasets, the proposed measure outperforms all the competitors except in Glass (for the ARI index), Parkinsons (purity), Pima (purity) and Auto-mpg (both purity and ARI). In these cases, however, no measure is significantly better than the others.

For what concerns the alternative distances, we can observe that on average d^{Aryal} performs better than the others; s^{Zhu2}, s^{Zhu3} and d^{Ting} perform equally, whereas d^{Shi} is on average inferior. However, this last distance is particularly accurate in high-dimensional datasets, like Lung, Parkinsons, Volcano and Energy. This is also confirmed by the analysis reported with automatic versions – Section 4.5.

4.3 Analysis 2: deep comparison of $\mathrm{RatioRF}$ and s^{Zhu2}

We now report a comparison between RatioRF and its direct competitor s^{Zhu2}, using the datasets in the first group. As before, we focus on different clustering techniques, different datasets and different forest parametrizations.

Given one particular aspect, we compared, among all experiments involving such aspect, how many times the accuracy obtained with $\mathrm{RatioRF}$ is larger/lower/equal than that of s^{Zhu2}. Results are reported in Table 3, for both ARI and Purity.

Every row of the table indicates the comparisons for a particular aspect: for example, when comparing $\mathrm{RatioRF}$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TKDE.2021.3086147, IEEE Transactions on Knowledge and Data Engineering

TABLE 2
Averaged clustering accuracies (see text). A bold value indicates a statistically significant better result with respect to the other measures.

	(Purity)						(ARI)					
	d^{Shi}	s^{Zhu2}	s^{Zhu3}	d^{Ting}	d^{Aryal}	RatioRF	d^{Shi}	s^{Zhu2}	s^{Zhu3}	d^{Ting}	d^{Aryal}	RatioRF
All Experiments												
	0.628	0.663	0.655	0.659	0.671	**0.704**	0.218	0.310	0.302	0.296	0.311	**0.373**
Different Clustering Methods												
SC	0.668	0.690	0.684	0.681	0.696	**0.725**	0.274	0.349	0.340	0.332	0.349	**0.408**
AP	0.661	0.688	0.684	0.681	0.690	**0.716**	0.263	0.340	0.336	0.324	0.336	**0.381**
HC-CL	0.541	0.607	0.606	0.611	0.622	**0.666**	0.104	0.232	0.241	0.227	0.243	**0.325**
HC-W	0.641	0.666	0.646	0.663	0.675	**0.709**	0.231	0.317	0.290	0.302	0.317	**0.380**
— Different Datasets —												
Iris	0.671	0.819	0.838	0.812	0.826	**0.878**	0.350	0.622	0.647	0.604	0.627	**0.707**
Wine	0.801	0.797	0.829	0.754	0.793	**0.899**	0.572	0.566	0.622	0.488	0.557	**0.743**
Glass	0.530	0.547	0.550	0.535	0.542	**0.558**	0.126	0.181	0.168	0.159	0.166	0.175
WBC	0.720	0.939	0.931	0.928	0.931	**0.970**	0.174	0.780	0.751	0.744	0.752	**0.883**
BTissue	0.577	0.568	0.576	0.566	0.574	**0.597**	0.353	0.341	0.354	0.333	0.347	**0.379**
Heart	0.674	0.674	0.695	0.659	0.677	**0.748**	0.149	0.149	0.180	0.126	0.150	**0.252**
Lung	0.518	0.498	0.510	0.492	0.501	**0.530**	0.091	0.063	0.083	0.053	0.066	**0.113**
Parkinsons	0.754	0.755	0.754	0.756	0.755	0.754	0.141	0.142	0.143	0.129	0.140	**0.154**
Auto-mpg	0.645	0.778	0.756	0.778	0.794	0.802	0.094	0.331	0.287	0.331	0.364	0.382
pima	0.654	0.658	0.658	0.658	0.680	0.659	0.018	0.046	0.043	0.053	0.029	**0.066**
Volcano	0.433	0.437	0.432	0.442	0.442	**0.470**	0.128	0.135	0.134	0.134	0.136	**0.164**
Energy	0.555	0.481	0.326	0.525	0.535	**0.584**	0.422	0.360	0.209	0.400	0.401	**0.463**
— Different Random Forest Parametrizations —												
50 Trees - Half	0.626	0.672	0.669	0.669	0.683	**0.713**	0.208	0.323	0.316	0.311	0.332	**0.384**
50 Trees - All	0.614	0.643	0.643	0.636	0.642	**0.689**	0.203	0.280	0.286	0.263	0.274	**0.354**
100 Trees - Half	0.637	0.683	0.667	0.682	0.693	**0.716**	0.224	0.338	0.314	0.330	0.343	**0.388**
100 Trees - All	0.622	0.646	0.640	0.640	0.648	**0.696**	0.217	0.287	0.286	0.272	0.283	**0.365**
200 Trees - Half	0.640	0.682	0.665	0.678	0.705	**0.715**	0.233	0.339	0.316	0.323	0.344	**0.387**
200 Trees - All	0.627	0.650	0.643	0.647	0.653	**0.695**	0.224	0.291	0.292	0.279	0.291	**0.363**

and s^{Zhu2} on the Iris dataset (first line of the "Different Datasets" part of Table 3), we can see that in 71.2% of the 720 experiments (6 forest parametrizations × 30 repetitions × 4 clustering methods) the accuracy obtained with $\mathrm{RatioRF}$ is larger than the accuracy obtained with s^{Zhu2}, whereas in 25.4% is lower, and in 3.3% they are equal. This represents an alternative direct comparison since we analyse the behaviour of the two measures given *exactly the same starting point* (the trained Random Forest, the clustering method or the dataset).

From the table we can observe that $\mathrm{RatioRF}$ compares very favourably with s^{Zhu2}: on average, in 65.8% of the experiments $\mathrm{RatioRF}$ has larger purity than s^{Zhu2}, whereas this increases to 73.5% when considering the ARI. This superiority can be found in all aspects, with the exception of some datasets; remarkably, in real world challenging high dimensional problems (like Volcano and Energy) our measure largely outperforms s^{Zhu2}: in more than 80% of the cases for Volcano, and almost everywhere for Energy.

4.4 Analysis 3: impact of parameters

In this section we focus on experiments meant to assess the impact of the parameters on the proposed framework. We report a comparative analysis of the different options relative to each aspect of the proposed RF clustering pipeline (number of trees, feature subsampling and clustering method), using the $\mathrm{RatioRF}$ distance and the datasets of the first group.

For each aspect we compute the average of the purity and ARI values of the different alternatives by varying all other aspects: for example, when analysing the number of trees (first three columns of Table 4), we compute the average of all results obtained with 50, 100 and 200 trees, for all different feature subsampling, clustering methods, and repetitions, thus resulting, for each dataset, in 240 values (2 feature subsamplings × 4 clusterings × 30 repetitions). The row "Average" contains the average over all datasets (i.e., in the case of trees of 240 × 12 = 2880 values).

The analysis is reported in Table 4, for purity (top) and ARI (bottom). Also in this table, for each dataset, a bold value indicates the best option which has a statistically significant difference with respect to the others, according to an unpaired t-test with significance 0.05.

Different observations can be derived from the table. The first one is that the number of trees does not represent a crucial parameter in the proposed approach: in almost all the experiments the averaged accuracy does not vary significantly among the different possibilities. The only exception is the Energy dataset, where performances with 200 trees are better than the alternatives. Energy is a dataset in which we have many clusters (50) and many objects (1500), and probably a larger number of trees is needed to unravel the underlying complexity (please note the poor result with 50 trees).

A second observation regards the different options for feature sampling. We can note that on average using half of the features when choosing the split represents a better choice, probably because of an increased randomization of the forests. Even if there is not a clear trend linked to dataset characteristics (like number of objects/features/clusters),

TABLE 3
Pairwise comparison between $\mathrm{RatioRF}$ and s^{Zhu2}.

	N. Tests	(Purity) Win/Lose/Draw	(ARI) Win/Lose/Draw
All Experiments			
	8640	65.8%/18.1%/16.1%	73.5%/25.1%/1.4%
Different Clustering Methods			
SC	2160	69.5%/15.0%/15.4%	77.1%/21.4%/1.5%
AP	2160	64.1%/20.9%/15.0%	70.4%/28.6%/1.0%
HC-CL	2160	65.9%/18.2%/15.9%	75.5%/23.1%/1.3%
HC-W	2160	63.5%/18.2%/18.3%	70.9%/27.4%/1.7%
Different Datasets			
Iris	720	71.2%/25.4%/3.3%	72.2%/27.2%/0.6%
Wine	720	87.9%/8.8%/3.3%	89.9%/10.0%/0.1%
Glass	720	57.4%/38.8%/3.9%	44.3%/55.7%/0.0%
WBC	720	84.2%/12.4%/3.5%	84.9%/14.0%/1.1%
BTissue	720	68.1%/21.5%/10.4%	72.9%/26.8%/0.3%
Heart	720	82.1%/15.1%/2.8%	83.3%/16.5%/0.1%
Lung	720	57.4%/16.0%/26.7%	72.2%/21.5%/6.2%
Parkinsons	720	6.2%/9.4%/84.3%	53.1%/43.9%/3.1%
Auto-mpg	720	61.8%/29.4%/8.8%	63.7%/31.2%/5.0%
pima	720	30.4%/23.2%/46.4%	63.6%/36.4%/0.0%
Volcano	720	82.5%/17.2%/0.3%	82.2%/17.8%/0.0%
Energy	720	100.0%/0.0%/0.0%	99.4%/0.6%/0.0%
Different Random Forest Parametrizations			
50 Trees - Half	1440	65.8%/18.7%/15.5%	72.2%/27.1%/0.8%
50 Trees - All	1440	67.4%/17.2%/15.3%	74.7%/23.6%/1.7%
100 Trees - Half	1440	61.7%/20.6%/17.6%	70.6%/28.3%/1.2%
100 Trees - All	1440	70.5%/11.8%/17.7%	79.4%/18.6%/1.9%
200 Trees - Half	1440	62.6%/22.8%/14.5%	70.2%/28.9%/0.9%
200 Trees - All	1440	66.5%/17.4%/16.1%	73.8%/24.4%/1.8%

TABLE 4
Analysis of the impact of the parameters.

Purity

	N. Trees			Feat. Sampling		Clustering			
Dataset	50	100	200	Half	Full	SC	AP	HC-CL	HC-W
Iris	0.877	0.879	0.877	0.874	**0.882**	**0.888**	0.876	0.867	0.881
Wine	0.894	0.905	0.898	**0.920**	0.878	**0.943**	0.901	0.819	0.933
Glass	0.559	0.559	0.557	**0.568**	0.549	0.574	**0.584**	0.518	0.557
WBC	0.970	0.970	0.970	0.970	0.970	**0.974**	0.972	0.964	0.970
BTissue	0.600	0.601	0.590	**0.613**	0.580	**0.610**	0.603	0.575	0.600
Heart	0.747	0.747	0.749	**0.752**	0.743	0.764	**0.772**	0.715	0.740
Lung	0.532	0.529	0.530	**0.550**	0.510	**0.542**	0.530	0.524	0.525
Parkinsons	0.754	0.754	0.754	0.753	**0.755**	0.753	0.753	0.755	0.755
Auto-mpg	0.803	0.801	0.801	0.788	**0.815**	**0.847**	0.808	0.747	0.804
pima	0.659	0.660	0.659	**0.662**	0.657	**0.668**	0.663	0.654	0.651
Volcano	0.469	0.472	0.470	0.468	0.472	0.481	0.487	0.430	0.483
Energy	0.551	0.595	**0.605**	**0.660**	0.507	**0.658**	0.641	0.422	0.613
Average	0.701	0.706	0.705	**0.715**	0.693	**0.725**	0.716	0.666	0.709

ARI

	N. Trees			Feat. Sampling		Clustering			
Dataset	50	100	200	Half	Full	SC	AP	HC-CL	HC-W
Iris	0.705	0.710	0.705	0.698	**0.716**	0.721	0.698	0.693	0.716
Wine	0.729	0.756	0.745	**0.788**	0.699	**0.836**	0.730	0.590	0.818
Glass	0.172	0.179	0.172	0.176	0.174	**0.195**	0.172	0.157	0.174
WBC	0.883	0.882	0.883	0.883	0.882	**0.897**	0.890	0.861	0.882
BTissue	0.380	0.387	0.371	**0.398**	0.360	0.385	0.382	0.362	0.386
Heart	0.251	0.251	0.253	**0.261**	0.242	0.279	**0.296**	0.198	0.234
Lung	0.116	0.109	0.113	**0.138**	0.089	**0.129**	0.107	0.111	0.105
Parkinsons	0.155	0.152	0.154	0.155	0.152	0.154	**0.165**	0.142	0.154
Auto-mpg	0.386	0.381	0.380	0.353	**0.412**	**0.483**	0.388	0.269	0.390
pima	0.065	0.069	0.064	**0.074**	0.058	**0.098**	0.083	0.047	0.036
Volcano	0.161	0.165	0.167	0.163	0.165	0.178	0.171	0.135	0.174
Energy	0.421	0.475	**0.494**	**0.548**	0.379	**0.537**	0.496	0.333	0.487
Average	0.369	0.376	0.375	**0.386**	0.361	**0.408**	0.381	0.325	0.380

we can observe that, reasonably, the superiority of the option "Half" is more evident for datasets with a larger number of features, such as Lung, Wine and Energy.

Finally, concerning the clustering method, it seems evident that spectral clustering is the best choice for Random Forest-based clustering, being superior to the alternatives in 8 cases over 12 (note that in 2 cases there is not a winning option). This confirms recent trends in RF-clustering literature [12]. However, we note that also Affinity Propagation works very well. This alternative (hardly investigated in RF-clustering literature) also permits to automatically detect the number of clusters (even if in our experiments this number has been fed directly in input). For what concerns hierarchical clustering, we can observe that the complete link variant represents the worst choice for many datasets, whereas the Ward Link reaches reasonable results almost everywhere.

Summarizing, we think that the following parametrization for the proposed approach would permit to get reasonable results in many situations: build forests with 100 trees, selecting half features for splitting, and use Spectral Clustering to get the final clustering. If the problem appears to be too complex (e.g. high number of clusters), try to increase the number of trees.

4.5 Analysis 4: automatic versions

The large scope analyses presented in the previous sections showed that the proposed measure is consistently better than the literature alternatives, for a large range of parametrizations. In this section we report a practical analysis, involving "automatic versions", i.e. versions where the parameters of the clustering procedure (number of trees, splitting features, or clustering algorithm) are chosen in an automatic manner.

Given a dataset, we computed the clustering with different parametrizations, selecting the best one using the *silhouette index* [45], a well known internal index used to evaluate clustering results. We chose this index for two reasons: i) it is based on distances (a high index indicates that points inside the same cluster are very similar, points from different clusters are very dissimilar), so that it can be directly computed using the RF similarities; ii) a recent large scope experimental evaluation [46] showed that it is one of the best internal criterion for assessing clustering quality. For each dataset and each distance, we selected the configuration leading to the highest silhouette (among all parametrizations, clustering methods and repetitions). Such values are reported in Table 5.

We can observe that, on average, also in this case our measure largely outperforms the alternatives, with an improvement over the second best measure (s^{Zhu3}) which is 0.05 for purity and 0.08 for ARI. For what concerns the different datasets, for Purity, in 7 datasets our measure represents the best choice, in other 2 the best is d^{Aryal}, in 1 d^{Shi} whereas in 2 datasets all measures lead to the same accuracy; similar reasonings can be done w.r.t. the ARI index.

Interestingly, whenever our measure is not the best choice the measure which works best is not always the same for different datasets. In other words, specific datasets *prefer* specific measures (e.g., d^{Shi} seems to be very suitable for high dimensional datasets, while performing poorly on others), while our measure represents a good choice *in general*.

4.6 Analysis 5: larger scale problems

In this section we show the results obtained with the datasets of group 2, i.e. with datasets of larger size. As described before, we used a single parametrization, in particular the one suggested at the end of the analysis reported in Subsection 4.4: 100 trees, selecting half of the features for the split, and using Spectral Clustering to get the final result. As already done for the other analyses, for each dataset we performed a paired t-test (significance 0.05) to compare the accuracies of the best performing measure with the others.

The results are shown in Table 6 for Purity and Adjusted Rand Index: also in this case, if present, a bold value indicates that the best value is larger than the alternatives with a statistical significance.

It is interesting to observe that also here $\mathrm{RatioRF}$ outperforms the alternative measures with statistical significance. For what concerns the different alternative measures, we have a confirmation that d^{Shi} is very adequate for high dimensional datasets: in all these experiments this distance reaches accuracies which are comparable to those obtained with more sophisticated distances.

5 Conclusions

In this paper we proposed $\mathrm{RatioRF}$, a Random Forest-based similarity measure for clustering. $\mathrm{RatioRF}$ represents a novel way of extracting the information contained in the trees of a RF to implement Tversky's ratio model definition of similarity. We present both a structural and empirical comparison of RatioRF with analogous measures in the literature. The extensive experimental results suggest that RatioRF might be a valid alternative to the state-of-the-art ones, with respect to different aspects, like forest parametrizations, clustering algorithms or datasets.

References

[1] A. Tversky, "Features of similarity," *Psychological review*, vol. 84, no. 4, p. 327, 1977.

[2] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.

[3] A. Criminisi, J. Shotton, and E. Konukoglu, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Foundations and Trends in Computer Graphics and Vision*, vol. 7, no. 2-3, pp. 81–227, 2012.

[4] G. Biau and E. Scornet, "A random forest guided tour," *TEST*, vol. 25, no. 2, pp. 197–227, 2016.

[5] L. Kuncheva, *Combining pattern classifiers: methods and algorithms*. John Wiley & Sons, 2004.

[6] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*, 1st ed. Chapman & Hall/CRC, 2012.

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

[8] D. Lin, "An information-theoretic definition of similarity," in *Proc. Int. Conf. on Machine Learning (ICML98)*, vol. 98, no. 1998, 1998, pp. 296–304.

[9] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language," *J. Artif. Int. Res.*, vol. 11, no. 1, pp. 95–130, Jul. 1999.

TABLE 5
Results with automatic versions.

	(Purity)						(ARI)					
Dataset	d^{Shi}	s^{Zhu2}	s^{Zhu3}	s^{Ting}	d^{Aryal}	RatioRF	d^{Shi}	s^{Zhu2}	s^{Zhu3}	s^{Ting}	d^{Aryal}	RatioRF
Iris	0.819	0.705	0.846	0.658	0.658	**0.973**	0.551	0.473	0.617	0.390	0.390	**0.922**
Wine	0.785	0.610	**0.949**	0.610	0.605	**0.949**	0.450	0.400	**0.854**	0.400	0.397	**0.854**
Glass	0.559	0.521	0.582	0.526	0.516	**0.587**	0.208	0.215	0.217	**0.220**	0.211	0.214
WBC	0.650	0.968	0.968	0.930	0.937	**0.977**	-0.032	0.874	0.874	0.737	0.763	**0.908**
BTissue	0.438	0.438	0.362	0.448	**0.552**	0.419	0.230	0.267	0.218	0.228	**0.344**	0.252
Heart	0.618	0.537	0.625	0.696	0.659	**0.794**	0.053	0.000	0.060	0.152	0.099	**0.345**
Lung	**0.581**	0.549	0.549	0.549	0.549	**0.581**	**0.160**	0.111	0.111	0.111	0.111	**0.160**
Parkinsons	0.753	0.753	0.753	0.753	0.753	0.753	0.195	-0.098	-0.096	-0.098	0.010	-0.098
Auto-mpg	0.670	0.678	0.665	0.668	**0.811**	0.678	0.110	0.121	0.103	0.107	**0.386**	0.121
pima	0.651	0.651	0.651	0.651	0.651	0.651	0.005	0.005	0.005	0.004	0.004	**0.019**
Volcano	0.351	0.326	0.341	0.338	0.320	**0.362**	**0.088**	0.007	0.048	0.056	0.010	0.087
Energy	**0.401**	0.375	0.168	0.318	0.324	0.388	0.209	0.207	0.030	0.202	0.197	**0.209**
Average	0.606	0.593	0.622	0.595	0.557	0.676	0.185	0.215	0.253	0.209	0.243	0.333

TABLE 6
Results with datasets of group 2

Distance	(Purity)						(ARI)					
	d^{Shi}	s^{Zhu2}	s^{Zhu3}	d^{Ting}	d^{Aryal}	RatioRF	d^{Shi}	s^{Zhu2}	s^{Zhu3}	d^{Ting}	d^{Aryal}	RatioRF
Isolet	0.495	0.498	0.494	0.489	0.500	**0.523**	0.373	0.374	0.370	0.361	0.379	**0.408**
Gas	0.453	0.463	0.463	0.455	0.449	**0.487**	0.197	0.206	0.205	0.198	0.195	**0.233**
UAV	0.963	0.953	0.960	0.954	0.955	**0.965**	0.857	0.823	0.846	0.824	0.829	**0.865**

[10] S. Santini and R. Jain, "Similarity measures," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 871–883, 1999.

[11] L. Cazzanti and M. R. Gupta, "Information-theoretic and set-theoretic similarity," in *Proc. Int. Symposium on Information Theory*. IEEE, 2006, pp. 1836–1840.

[12] X. Zhu, C. Loy, and S. Gong, "Constructing robust affinity graphs for spectral clustering," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition, CVPR 2014*, 2014, pp. 1450–1457.

[13] T. Shi and S. Horvath, "Unsupervised learning with random forest predictors," *Journal of Computational and Graphical Statistics*, vol. 15, no. 1, pp. 118–138, 2006.

[14] K. Ting, Y. Zhu, M. Carman, Y. Zhu, and Z.-H. Zhou, "Overcoming key weaknesses of distance-based neighbourhood methods using a data dependent dissimilarity measure," in *Proc. Int. Conf. on Knowledge Discovery and Data Mining*, 2016, pp. 1205–1214.

[15] S. Aryal, K. Ting, T. Washio, and G. Haffari, "A comparative study of data-dependent approaches without learning in measuring similarities of data objects," *Data Min. Knowl. Discov.*, vol. 34, no. 1, pp. 124–162, 2020.

[16] K. R. Gray, P. Aljabar, R. A. Heckemann, A. Hammers, and D. Rueckert, "Random forest-based similarity measures for multi-modal classification of alzheimer's disease," *NeuroImage*, vol. 65, pp. 167 – 175, 2013.

[17] T. Shi, D. Seligson, A. Belldegrun, A. Palotie, and S. Horvath, "Tumor classification by tissue microarray profiling: Random forest clustering applied to renal cell carcinoma," *Modern Pathology*, vol. 18, pp. 547–557, 2005.

[18] M. C. Abba, H. Sun, K. A. Hawkins, J. A. Drake, Y. Hu, M. I. Nunez, S. Gaddis, T. Shi, S. Horvath, A. Sahin, and C. M. Aldaz, "Breast cancer molecular signatures as determined by sage: Correlation with lymph node status," *Molecular Cancer Research*, vol. 5, no. 9, pp. 881–890, 2007.

[19] S. I. Rennard, N. Locantore, B. Delafont, R. Tal-Singer, E. K. Silverman, J. Vestbo, B. E. Miller, P. Bakke, B. Celli, P. M. Calverley *et al.*, "Identification of five chronic obstructive pulmonary disease subgroups with different prognoses in the eclipse cohort using cluster analysis," *Annals of the American Thoracic Society*, vol. 12, no. 3, pp. 303–312, 2015.

[20] C. Xiong, D. Johnson, R. Xu, and J. J. Corso, "Random forests for metric learning with implicit pairwise position dependence," in *Proc. Int. Conf. on Knowledge discovery and data mining*, 2012, pp. 958–966.

[21] D. Johnson, C. Xiong, and J. Corso, "Semi-supervised nonlinear distance metric learning via forests of max-margin cluster hierarchies," *IEEE Trans. on Knowledge and Data Engineering*, vol. 28, no. 4, pp. 1035–1046, 2016.

[22] Y. Dong, B. Du, and L. Zhang, "Target detection based on random forest metric learning," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 8, no. 4, pp. 1830–1838, 2015.

[23] F. Liu, K. Ting, and Z. Zhou, "Isolation forest," in *Proc. of Int. Conf. on Data Mining*, 2008, pp. 413–422.

[24] S. Aryal, K. Ting, G. Haffari, and T. Washio, "Mp-dissimilarity: a data dependent dissimilarity measure," in *Proc. of Int. Conf. on Data Mining (ICDM)*, 2014, pp. 707–712.

[25] H. Blockeel, L. D. Raedt, and J. Ramon, "Top-down induction of clustering trees," in *Proc. Int. Conf. on Machine Learning (ICML 1998)*, 1998, pp. 55–63.

[26] B. Liu, Y. Xia, and P. Yu, "Clustering through decision tree construction," in *Proc. Int. Conf. on Information and Knowledge Management - CIKM*, 2000, pp. 20–29.

[27] J. Basak and R. Krishnapuram, "Interpretable hierarchical clustering by constructing an unsupervised decision tree," *IEEE Trans. on Knowledge and Data Engineering*, vol. 17, no. 1, pp. 121–132, 2005.

[28] F. Moosmann, B. Triggs, and F. Jurie, "Fast discriminative visual codebooks using randomized clustering forests," in *Advances in Neural Information Processing Systems 19*, 2006, pp. 985–992.

[29] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[30] J. Shotton, M. Johnson, and R. Cipolla, "Semantic texton forests for image categorization and segmentation," in *Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR 2008)*, 2008.

[31] F. Perbet, B. Stenger, and A. Maki, "Random forest clustering and application to video segmentation," in *Proc. British Machine Vision Conference, BMVC 2009*, 2009, pp. 1–10.

[32] M. Bicego, "K-random forests: a K-means style algorithm for random forest clustering," in *Proc. Int. Joint Conf. on Neural Networks (IJCNN2019)*, 2019.

[33] D. Yan, A. Chen, and M. Jordan, "Cluster forests," *Computational Statistics & Data Analysis*, vol. 66, pp. 178–192, 2013.

[34] J. Gower, "Metric and euclidean properties of dissimilarity coefficients," *Journal of Classification*, vol. 3, pp. 5–48, 1986.

[35] Y. Lin and Y. Jeon, "Random forests and adaptive nearest neighbors," *Journal of the American Statistical Association*, vol. 101, no. 474, pp. 578–590, 2006.

[36] C. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[37] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, pp. 193–218, 1985.

[38] L. Zhao, A. Alipour-Fanid, M. Slawski, and K. Zeng, "Prediction-time efficient classification using feature computational dependencies," in *Proc. of ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining*, 2018, pp. 2787–2796.

[39] M. Orozco-Alzate, P. Castro-Cabrera, M. Bicego, and J. Londoño-Bonilla, "The dtw-based representation space for seismic pattern classification," *Computers & Geosciences*, vol. 85, pp. 86–95, 2015.

[40] M. Bicego, A. Farinelli, E. Grosso, D. Paolini, and S. Ramchurn, "On the distinctiveness of the electricity load profile," *Pattern Recognition*, vol. 74, pp. 317–325, 2018.

[41] "https://archive.ics.uci.edu/ml/datasets/electricityloaddiagrams20112014."

[42] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[43] B. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, p. 972976, 2007.

[44] M. Bicego, "Dissimilarity random forest clustering," in *Proc. Int. Conf. on Data Mining (ICDM)*, 2020, pp. 936–941.

[45] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, 1987.

[46] O. Arbelaitz, I. Gurrutxaga, J. Muguerza, J. Perez, and I. Perona, "An extensive comparative study of cluster validity indices," *Pattern Recognition*, vol. 46, no. 1, pp. 243 – 256, 2013.

Manuele Bicego Manuele Bicego is an associate professor at the Computer Science Dept. of the Univ. of Verona (Italy) since 2017. His research interests are in statistical pattern recognition and bioinformatics, e.g. on the probabilistic modelling, representation, clustering and biclustering of biological data. He is author of more than 130 papers, published in international journals, edited books and conferences. He is AE of Pattern Recognition, and PC member of many different international conferences and workshops related to his research interests.

Ferdinando Cicalese Ferdinando Cicalese has been associate professor of the Computer Science dept, at Univ. of Verona (Italy) since 2014. He received the Masters and PhD degrees in computer science from University of Salerno (Italy) in 1995 and 2001, respectively. From 2001 to 2014 he was first assistant professor and then associate professor at University of Salerno and from 2004 to 2009 he was research group leader at Bielefeld University (Germany). His research interests are in the area of algorithms and complexity, (with a special emphasis on combinatorial search algorithms and decision tree construction optimization), information theory and fault tolerant error-correction codes. Dr. Cicalese is the recipient of the 2004 Sofja Kovalevskaja award from the Humboldt Foundation and the Germany BMBF. He is author of more than 100 scientific publications, including a Springer monograph on fault tolerant search algorithms, in 2011. Dr. Cicalese has been guest editor of international journals and PC member and program chair of several international conferences.

Antonella Mensi Antonella Mensi received her Master degree in Medical Bioinformatics from the Univ. of Verona in 2018, with a thesis on protein remote homology detection via multiple instance learning and dissimilarity-based representation. Since 2018, she is a Ph.D. student in Computer Science at the Computer Science Dept. of the Univ. of Verona. Her research interests include statistical pattern recognition, Random Forests and bioinformatics.