On the Good Behaviour of Extremely Randomized Trees in Random Forest-Distance Computation

Manuele Bicego(✉) and Ferdinando Cicalese

Computer Science Department, University of Verona, Verona 37134, Italy
{manuele.bicego,ferdinando.cicalese}@univr.it

Abstract. Originally introduced in the context of supervised classification, ensembles of *Extremely Randomized Trees* (ERT) have shown to provide surprisingly effective models also in unsupervised settings, e.g., for anomaly detection (via Isolation Forests) and for distance computation. In this paper, we focus on this latter application of ERT, namely in the context of Random Forest (RF) distance computation. We aim at narrowing the gap between the established empirical evidence of the good behaviour of ERT and the still limited theoretical understanding of their (somehow) surprisingly good performance when compared to more involved methodologies. Our main contribution is the following: we assume the existence of a proper representation on a given domain, i.e., a vectorial representation of the objects which satisfies the *Compactness Hypothesis* formulated by Arkadev and Braverman in 1967. Under such a hypothesis, given the "true" distance between two objects, we show how to derive a bound on the approximation guaranteed by two main RF-distances obtained by employing ensembles of ERTs, with respect to such "true" distance. In other words, we show that there exists a constant c such that if two objects are ϵ-close in the true distance, then with high probability they are $(c \cdot \epsilon)$-close in the RF-distances computed with ERT forests.

Keywords: Random Forest-distances · Extremely Randomized Trees · Compactness hyphotesis

1 Introduction

Random Forests (RF) [7,10] are ensembles of decision trees [21], successfully applied in Pattern Recognition and Machine Learning as models for regression and classification, and more recently, for other tasks such as clustering or outlier detection [17,18,23]. Another exploitation of Random Forest, less investigated than previous ones, is for distance computation: starting from the seminal work of Breiman [7], it has been shown that powerful data-dependent distances can be extracted from RFs: the main idea is that it is possible to assess the similarity between two objects by looking at the way they answer to the tests of the different trees. In the basic version of [7,23], the similarity is proportional to the number

D. Koutra et al. (Eds.): ECML PKDD 2023, LNAI 14172, pp. 645–660, 2023.
https://doi.org/10.1007/978-3-031-43421-1_38

of times, over the total number of trees of the forest, the two objects reach the same leaf – thus answering in the same way to all questions in the path. Many different extensions have been proposed, exploiting different aspects like tests in the paths [32], probability masses [2,29] or other concepts [6].

Typically, the RF-distance is extracted in two steps: i) a forest is learned from the available data and ii) the distance is defined given the trained forest. The solution of the first step typically depends on the task: the forest can be derived in a supervised way (i.e. using the labels, as in supervised RF-distances of [7,23]), or in an unsupervised way. In this last case, different context specific solutions can be adopted (e.g. [23] or [5] for clustering), but a widely applied option is to use *Extremely Randomized Trees* (ERT) [16]. An ERT is a classic decision tree as CART [8], i.e. a binary tree in which questions in each node are defined with thresholds on a single feature. The difference with respect to CART is the way in which the tree is trained: instead of looking for "optimized" questions, the ERT is built in a completely random way: at each node, the question is defined by choosing a random feature and a random threshold inside the domain of that feature. These trees have shown to be surprisingly good for classification [16], but also for unsupervised domains, like anomaly detection [17,18] or clustering [5,20,24].

In the context of RF-distance computation, ERT have been largely and successfully employed, for example in [2–5,27,29]. Training forests for RF-distance computation with ERT is attractive for many different reasons: i) it is an unsupervised, simple and efficient way to derive a forest: distances based on these forests have shown to outperform many other distances, also in semi-supervised settings [31]; ii) in the clustering case, authors of [5] have shown that this option is competitive with alternative more sophisticated strategies on datasets of moderate size, and better than all alternatives on datasets of larger size; iii) in the classification case, it has been shown in [28] that, when used with Support Vector Machines, distances computed from ERT are significantly better than RF-distances computed from supervisedly trained RF [9,11], probably because of a reduced risk of overtraining.

From a theoretical point of view, only few studies characterize the good performances of RF-distances. For example, [9] shows how to derive a properly defined kernel from the RF-distance: such starting work has been further extended and integrated in [11,22]. In these studies, the RF-distances were all based on supervisedly trained Random Forests. More in relation to ERT-based RF-distances, Ting and colleagues showed in recent papers such as [25,26,28] some theoretical properties of Isolation Kernels (kernels extracted from Isolation Forests, i.e. ERT-ensembles): for example it has been theoretically shown that the Isolation Kernel assigns a higher similarity to two points being in a sparse region than to two points of the same inter-point distance in a dense region, which is the main motivation behind the derivation of these data-dependent measures.

In this paper we make one step forward along this direction, proposing a novel theoretical characterization of RF-distances built from forests of ERT, aimed at providing evidences of the motivation behind the success of such RF distances in characterizing the distance between objects. In particular, in the

paper we theoretically show that under some assumptions, if two objects x and y have a small "true distance", then also their RF-distance, built starting from a RF defined with ERT, is small. We provide such theoretical characterization for the original RF-distance introduced by Breiman [7,23] and for the recent RatioRF distance [6]. To show that, we assume that there is a representation which satisfies the "Compactness Hypothesis" of Arkadev and colleagues [1]; then, we derive a bound on the RF-distance, computed with ERT-forests based on such representation, with respect to the true distance. We also provide some simulations to understand the different aspects of the bounds, also suggesting a procedure to derive the minimum number of trees of the forest needed to get a given probabilistic guarantee.

The remainder of the paper is organized as follows: in Sect. 2 we provide the basic notation, whereas we present our main results in Sect. 3. We show some numerical simulations in Sect. 4, and we discuss our findings and conclude the paper in Sect. 5.

2 Background

In this section, we introduce the basic concepts needed to understand our main results. In the more general formulation [8], given a vectorial representation of d features, a decision tree t is a *complete* binary tree in which each internal node j is associated to a test $\theta_j = (\nu_j, f_j)$, where ν_j is a threshold on a feature f_j; the two edges which link the node to the children represent the two possible results of the binary test $\theta_j = (\nu_j, f_j)$: an object $\mathbf{x} = [x_1, .., x_d]$ takes the left branch if $x_{f_j} < \nu_j$, the right one otherwise. Typically, decision trees are learned starting from a training set X, used to determine, at each node j, the optimal test $\theta_j = (\nu_j, f_j)$. Extremely Randomized Trees [16] are decision trees characterized by a high degree of randomness: in their extreme version, there is no optimization, and the tests $\theta_j = (\nu_j, f_j)$ are defined completely at random. More in detail, given a training set X, the training follows a recursive procedure: in a given node, i) a feature f_j is randomly chosen among the d features, ii) the threshold ν_j is uniformly sampled from the domain of the objects of X arrived at that node, and iii) the objects are propagated to the left or the right node according to the test. This recursive procedure is repeated until a node contains a single object or a maximum depth is reached. The ERT-Random Forest is then obtained following the standard procedure [7]: M different ERT are built starting from random subsamples of the problem training set. ERTs have been shown to be successful in different contexts, such as classification [16], distance computation [2–4,27,29], clustering [5,20,24] and anomaly detection - where ERTs are referred to as Isolation Trees, leading to Isolation Forests [17,18], one of the most powerful anomaly detection techniques ever introduced according to [12,15].

2.1 RF-Distances

Breiman was the first to point out that it is possible to derive highly descriptive data-dependent measures of similarity from Random Forests [7]. After his

seminal work, many other powerful RF-distances have been presented (see, e.g., [2–6,23,27,29,32]) and proven to be very effective in a range of different applications such as classification, clustering, outlier detections and others. In all these measures, the main idea is that the relation between two objects x and y can be quantified by i) making the two objects traverse all trees of the trained Forest, and ii) comparing the answers they provide.

In this paper, we focus on two distances, briefly summarized in the following. Given a tree t, and an object x, let us denote as $\ell_t(x)$ the leaf where the object x falls after traversing the tree t. Let us also denote as $P_t(x)$ the *path* of x from the root to its leaf. The first distance, which we call *Shi* [7,23], represents the RF distance originally introduced by Breiman [7] and then exploited by Shi and colleagues for Random Forest Clustering [23]. The distance is firstly defined at the tree level by postulating the similarity between two objects x and y as 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), 0 otherwise. We have:

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

Given the similarity, the distance based on a forest of M trees is then defined as[1]:

$$d_{Shi}(x,y) = 1 - \frac{1}{M} \sum_t \text{Shi}_t(x,y) \tag{2}$$

The second distance is the recently introduced RatioRF measure [6], a RF-distance defined on a set-based interpretation of the Tversky definition of similarity [30]. For simplicity, let us introduce here only the basic mechanism, referring to [6] the readers interested in the contextualization into the Tversky theory. Basically, within the RatioRF measure, two objects are compared on the basis of their answers to all the tests contained in the two paths $P_t(x)$ and $P_t(y)$ – these being the sole tests needed to characterize x and y. More in detail, let us call S_t^{xy} the set containing all tests in the two paths, i.e. $S_t^{xy} = S_t^x \cup S_t^y$, where S_t^x is the set of tests $\{\theta_{\text{root}}, \cdots, \theta_{\ell_t(x)}\}$ in the path $P_t(x)$. Let us denote as A_t^{xy} the set of tests in S_t^{xy} for which x and y provide the same answer. At tree level, the RatioRF similarity between x and y is defined by:

$$\text{RRF}_t(x,y) = \frac{|A_t^{xy}|}{|S_t^{xy}|} \tag{3}$$

where $|\cdot|$ denote the cardinality of a set. Given this similarity, the distance based on a forest of M trees is then defined as[2]:

$$d_{RRF}(x,y) = 1 - \frac{1}{M} \sum_t \text{RRF}_t(x,y) \tag{4}$$

[1] Please note that to ease the computation this formulation is the squared version of the original formulation of the distance, as given in [23].

[2] Also in this case, to simplify the computation, we remove the squared root from the original definition of the distance given in [6].

3 Main Result

We want to show that if two objects x and y have a small true distance, then their RF-distance, built starting from a RF defined with ERT, is also small. More precisely, if we denote by $d^*(x, y)$ the true distance between x and y and by $d^R(x, y)$ the distance obtained with a Random Forest R built with *Extremely Randomized Trees*, e.g., as in (2) and in (4), then our goal is to show that

$$d^*(x, y) \leq \epsilon \quad \Rightarrow \quad Pr\left(d^R(x, y) \geq (1 + \delta_\epsilon)\epsilon\right) \leq P \tag{5}$$

where ϵ, δ, and P are small numbers, i.e., with high probability the distance computed via the ERT random forest is a good approximation of the original distance.

3.1 Step 1: The Representation

We start our derivation by assuming that there exists a *proper representation* for our problem. To instantiate the concept of *proper representation*, we resort to the "Compactness Hypothesis", formulated by Arkadev and colleagues in 1967 [1], and then developed by Duin and Pekalska in [13]. Within such hypothesis, a representation is *proper* if two objects which are near in the real world are also near in the representation space. Arkadev and colleagues, together with Duin and Pekalska, showed that generalization is not possible if this hypothesis is not fulfilled. Please note that the definition only implies that similar objects have similar representations, and not that *dissimilar* objects have *dissimilar* representation. If this last is also true, then they refer to *true* representations, for which even a simple boundary-based classifier can permit zero-error classification. The principle is defined in a vague way (simply stating that near objects should have near representations), and can be formalized in different ways, depending on the goal: for example, Duin in [14] defined a measure to quantify the compactness of a given representation in case of nearest neighbour classification.

Here we provide the following formalization: let us assume that we have a representation based on a set of features F, i.e. every object x of our problem is encoded with an $|F|$-dimensional vector $rep(x) = [x_1, \cdots, x_{|F|}] \in [0, 1]^{|F|}$. We assume, for the sake of the presentation, that the components of this vector are normalized to values in $[0, 1]$. Now, we formalize the property, for the representation, to be *proper*, i.e. to satisfy the "Compactness Hypothesis": if two objects x and y of our problem have a low distance, their representation $rep(x) = [x_1, \cdots, x_{|F|}]$ and $rep(y) = [y_1, \cdots, y_{|F|}]$ should be close. More precisely, we work with the following parameterized and more quantitative notion of a proper representation.

Definition 1. *Fix numbers $\theta, \epsilon \in [0, 1]$. The representation $z \mapsto rep(z) = [z_1, \cdots, z_{|F|}]$ is (θ, ϵ)-**proper** with respect to a (true) distance d^* if the following condition holds:*

$$\forall x, y \ s.t. \ d^*(x, y) \leq \epsilon \quad \exists \tilde{F} \subset F \ s.t. \ \begin{cases} |\tilde{F}| \geq (1 - \theta)|F| \\ \forall f \in \tilde{F}, |x_f - y_f| \leq \epsilon. \end{cases} \tag{6}$$

We say that the representation is θ-proper, if it satisfies (6) for all $\epsilon > 0$. *Finally, we say that the representation is simply proper if it is θ-proper for some $\theta < 0.1$.*

3.2 Step 2: The Bounds

We can now show our probabilistic bounds on the approximation guarantee achievable by a ERT-based RF-distance computed over a proper representation. We assume that the forest is built based on a (θ, ϵ)-proper representation, over a set F of features with values in $[0, 1]$. We let M denote the number of trees in the forest, and we assume that each tree has height h. For the sake of the analysis it is easier to think that in each tree, each leaf is at depth h, although all the arguments remain valid under the hypothesis that h is an upper bound on the maximum depth of a leaf.

Theorem 1 (Shi distance). *Given two objects x and y, whose true distance is $d^*(x, y) \leq \epsilon$ and assuming an (θ, ϵ)-proper representation $rep(z)$, according to def. 1, let $d^R(x, y)$ be the RF-distance computed with Eq. (2) on the representation $rep(x), rep(y)$, starting from a forest of M ERT trees. Then, for all $\delta \in (0, 1]$ and $\delta_\epsilon \geq 0$ such that*

$$(1 + \delta_\epsilon)\epsilon \geq (1 + \delta)\left[1 - ((1 - \theta)(1 - \epsilon))^h\right] \tag{7}$$

it holds that

$$Pr\left(d^R(x, y) \geq (1 + \delta_\epsilon)\epsilon\right) < \exp\left(\frac{-M\delta^2\left[1 - ((1 - \theta)(1 - \epsilon))^h\right]}{3}\right) \tag{8}$$

where $\exp(x)$ indicates e^x. The theorem says that under the condition stated in Eq. (7), the probability that the RF-distance is far away from the true distance – according to δ_ϵ – can be made as small as possible by increasing the number of the trees M of the forest.

Theorem 2 (RatioRF distance). *Given two objects x and y, which true distance is $d^*(x, y) \leq \epsilon$ and assuming an (θ, ϵ)-proper representation $rep(z)$, according to def. 1, let $d^R(x, y)$ be the RF-distance computed with Eq. (4) on the representation $rep(x), rep(y)$, starting from a forest of M ERT trees. Then, for all $\delta \in (0, 1]$ and $\delta_\epsilon \geq 0$ such that*

$$(1 + \delta_\epsilon)\epsilon \geq (1 + \delta)\left[1 - ((1 - \theta)(1 - \epsilon))\right] \tag{9}$$

it holds that

$$Pr(d^{RR}(x, y) \geq 2(1 + \delta_\epsilon)\epsilon) \leq 2\exp\left(\frac{-Mh\delta^2\left[1 - (1 - \theta)(1 - \epsilon)\right]}{3}\right) \tag{10}$$

3.3 The Proofs

Both proofs are based on the Chernoff bound (see, e.g., [19]). Among the several variants of the bound available, we use the following version: let $X_1, X_2, \ldots X_n$ be independent Poisson trials with $Pr[X_i = 1] = p_i$. Let X be the sum of the X_i, and let μ be an upper bound on $E[X]$, i.e., $E[X] \leq \mu$. Then, for any $\delta \in (0, 1]$ we have that:

$$Pr(X > (1 + \delta)\mu) < \exp\left(\frac{-\mu\delta^2}{3}\right) \tag{11}$$

Let us derive the proof for Theorem 1.

Proof (Theorem 1). Given x and y, let us define the random variable $X_t \in \{0, 1\}$ as:

$$X_t = \begin{cases} 1 & \text{if } x \text{ and } y \text{ fall in different leaves in the tree } t \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Given this definition, the RF-distance defined in Eq. (2) can be written as:

$$d^R(x, y) = \frac{1}{M} \sum_{t=1}^{M} X_t \tag{13}$$

Assume that a proper representation is given that uses the set of features F. The probability that $X_t = 0$ is the probability that the two objects fall in the same leaf, i.e. that they follow the same root-to-leaf path in the tree t, answering in the same way to all the questions along such a path. Let us consider the root. The probability that, in the root, two objects take the same branch is 1 minus the probability that they are separated, i.e. that the chosen threshold is exactly between their value on the feature tested in the root. Considering that the threshold is randomly chosen in the domain, if the feature f used in the test belongs to $\tilde{F}$, then $|x_f - y_f| \leq \epsilon$, thus the probability that this happens is $\geq 1 - \epsilon$. Considering that there are at least $(1 - \theta)|F|$ such features, then the probability that x and y answer in the same way is $\geq (1 - \theta)(1 - \epsilon)$.

Now, let us consider the second node of the path. The reasoning is exactly the same, but for the fact that the domain of the feature tested at this node might be different from $[0, 1]$. Actually, if the feature used for the split is the same as the feature used in the root, then the domain is reduced: if τ is the threshold used in the root, the domain is $[0, \tau]$ or $[\tau, 1]$, depending on whether x and y took the left or the right path. However, if the feature in the second node of the path is different from the one used in the root, then the probability that x and y answer in the same way is again $\geq (1 - \theta)(1 - \epsilon)$. For the simplicity in the treatment let us for now assume that, on every root-to-leaf path, every split is done on a distinct feature, so that we can consider the threshold is always randomly chosen over the whole domain $[0, 1]$. We remark that this is not such a strong assumption since i) ERT trees used for RF-distances are very short — typically each tree is built with 128 or 256 objects, and the max depth is set to $\log(n)$, i.e., 7 or 8; ii) the contexts in which ERT-based RF-distances are more

suitable are those characterized by high dimensional spaces, which implies that over a short path the probability of randomly choosing twice the same feature is very small. We will provide some arguments on the relaxation of this assumption in Sect. 5.

Since the path followed by x, has length at most h, we have that

$$Pr(X_t = 0) \geq [(1-\theta)(1-\epsilon)]^h \tag{14}$$

hence for the expected value $E[X_t]$ we have that:

$$E[X_t] = 0 \cdot Pr(X_t = 0) + 1 \cdot Pr(X_t = 1) \leq 1 - [(1-\theta)(1-\epsilon)]^h \tag{15}$$

First, let us rewrite the probability in the left part of Eq. (8) by using the definition of the Shi distance provided in Eq. (13):

$$Pr\left(d^R(x,y) \geq (1+\delta_\epsilon)\epsilon\right) = Pr\left(\frac{1}{M}\sum_{t=1}^{M} X_t \geq (1+\delta_\epsilon)\epsilon\right) = Pr\left(\sum_{t=1}^{M} X_t \geq M(1+\delta_\epsilon)\epsilon\right)$$

Let $\mu = M\left[1 - ((1-\theta)(1-\epsilon))^h\right]$. Then, using (15), the Expected value of the variable $X = X_1 + X_2 + \ldots + X_M$ satisfies the inequality

$$E[X] = E\left[\sum_{t=1}^{M} X_t\right] = \sum_{t=1}^{M} E[X_t] \leq M\left[1 - ((1-\theta)(1-\epsilon))^h\right] = \mu \tag{16}$$

Now, from (9) it follows that $M(1+\delta_\epsilon)\epsilon \geq (1+\delta)\mu$ and by the Chernoff bound above (see Eq. (11)), we have

$$\begin{aligned} Pr\left(d^R(x,y) \geq (1+\delta_\epsilon)\epsilon\right) &= Pr\left(\sum_{t=1}^{M} X_t \geq M(1+\delta_\epsilon)\epsilon\right) \leq Pr\left(\sum_{t=1}^{M} X_t \geq (1+\delta)\mu\right) \\ &< \exp\left(\frac{-\mu\delta^2}{3}\right) = \exp\left(\frac{-M\delta^2\left[1 - ((1-\theta)(1-\epsilon))^h\right]}{3}\right), \end{aligned}$$

□

Similarly we can provide the proof of the Theorem 2

Proof (Theorem 2). Recall that we assume a Forest R with M ERT trees, built on the feature set F of a proper representation, where each feature has domain $[0,1]$. Recall also that, given x and y, the RatioRF distance is computed by considering the set S_t^{xy} of the tests on the two root-to-leaf paths followed by x and y in the tree t. We start by defining the random variable $X_t^i \in \{0,1\}$, for each tree $t = 1, \ldots, M$ and each test i in the set S_t^{xy}, i.e., $1 \leq i \leq |S_t^{xy}|$:

$$X_t^i = \begin{cases} 1 & \text{if } x \text{ and } y \text{ give a different answer to the test } i \text{ in the set } S_t^{xy} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

Given this definition, the RF-distance can be reformulated as:

$$d^{RR}(x,y) = \frac{1}{M}\sum_{t=1}^{M}\frac{\sum_{i=1}^{|S_t^{xy}|} X_t^i}{|S_t^{xy}|} \tag{18}$$

Also in this case we can estimate the probability that $X_t^i = 0$. Under the assumption—see the discussion above in the part about the Shi distance—that all thresholds on the same root-to-leaf path are uniformly chosen in $[0,1]$, i.e., the features of the tests on the same root-to-leaf path are distinct, we have that for each t and i,:

$$Pr(X_t^i = 0) \geq (1-\theta)(1-\epsilon) \tag{19}$$

hence,

$$E[X_t^i] \leq [1-(1-\theta)(1-\epsilon)] \tag{20}$$

We will now start with a reformulation of (18). Recall, from Sect. 2, that S_t^x and S_t^y represent the set of tests on the root-to-leaf paths associated to x and y, respectively. Notice that on each node ν of the common part, $S_t^x \cap S_t^y$, of the these two paths–apart from the node where they separate– we have that $X_t^\nu = 0$. Then, from (18) it follows that

$$d^{RR}(x,y) \leq \frac{1}{M}\sum_{t=1}^{M}\frac{\sum_{i=1}^{|S_t^x|} X_t^i + \sum_{i=1}^{|S_t^y|} X_t^i}{|S_t^{xy}|} \tag{21}$$

Let us define

$$d_x^{RR}(y) = \frac{1}{M}\sum_{t=1}^{M}\frac{\sum_{i=1}^{|S_t^x|} X_t^i}{|S_t^x|}, \qquad d_y^{RR}(x) = \frac{1}{M}\sum_{t=1}^{M}\frac{\sum_{i=1}^{|S_t^y|} X_t^i}{|S_t^y|}. \tag{22}$$

Then, from (21) and (22) we have $d^{RR}(x,y) \leq d_x^{RR}(y) + d_y^{RR}(x)$. Hence,

$$Pr(d^{RR}(x,y) \geq 2(1+\delta_\epsilon)\epsilon) \leq Pr(d_x^{RR}(y) \geq (1+\delta_\epsilon)\epsilon) + Pr(d_y^{RR}(x) \geq (1+\delta_\epsilon)\epsilon) \tag{23}$$

where the inequality follows by noticing that, for every $a > 0$ the event $A = \{d^{RR}(x,y) > 2a\}$ implies at least one of the events: $B = \{d_x^{RR}(y) > a\}$, or $C = \{d_y^{RR}(x) > a\}$. I.e., we are using

$$A \subseteq B \cup C \Rightarrow Pr(A) \leq PR(B \cup C) \leq Pr(B) + Pr(C)$$

Under the assumption made above that both paths of x and y are of fixed length h, and that this length is the same for all trees, we can simplify the Eq. (22) as

$$d_x^{RR}(y) = \frac{\sum_{t=1}^{M}\sum_{i=1}^{h} X_t^i}{Mh}, \qquad d_y^{RR}(x) = \frac{\sum_{t=1}^{M}\sum_{i=1}^{h} X_t^i}{Mh} \tag{24}$$

Using (20), we can define an upper bound μ on the expected value of the sum in the enumerator of (24) as follows:

$$\mu = Mh(1-(1-\theta)(1-\epsilon)) \geq \sum_{i=1}^{M}\sum_{i=1}^{h} E[X_t^i].$$

From (9) it follows that $Mh(1+\delta_\epsilon)\epsilon \geq (1+\delta)\mu$. Hence, by the Chernoff bound, we have

$$Pr\left(d_x^{RR}(y) \geq (1+\delta_\epsilon)\epsilon\right) = Pr\left(\sum_{t=1}^{M}\sum_{i=1}^{h} X_t^i \geq Mh(1+\delta_\epsilon)\epsilon\right)$$
$$\leq Pr\left(\sum_{t=1}^{M}\sum_{i=1}^{h} X_t^i \geq (1+\delta)\mu)\right) < \exp\left(\frac{-\mu\delta^2}{2+\delta}\right) = \exp\left(\frac{-Mh\delta^2\left[1-(1-\theta)(1-\epsilon)\right]}{3}\right)$$

Analogously, we also obtain

$$Pr\left(d_y^{RR}(x) \geq (1+\delta_\epsilon)\epsilon\right) \leq \exp\left(\frac{-Mh\delta^2\left[1-(1-\theta)(1-\epsilon)\right]}{3}\right) \tag{25}$$

Therefore, recalling (23) we have the desired result

$$Pr(d^{RR}(x,y) \geq 2(1+\delta_\epsilon)\epsilon) \leq 2\exp\left(\frac{-Mh\delta^2\left[1-(1-\theta)(1-\epsilon)\right]}{3}\right) \tag{26}$$

□

4 Understanding the Bounds

The two bounds say that the probability that the RF-based distance is significantly larger than the true distance can be made as small as possible by increasing the number of trees of the forest, as long as the conditions in Eq. (7) and in Eq. (9) are satisfied. In this section we discuss the relationship between δ_ϵ, δ and the size M of the forest established by the bounds and the conditions. The parameters ϵ and θ are given by the proper representation available. We let P^* be a desired upper bound on the probability in the right-hand side of Eqs. (8) and (10), i.e., P^* is an upper bound on the probability that the RF-based distance d^R is not a good approximation of the true distance d^*.

4.1 The Number of Trees M

From Eqs. (8) and (10) we can compute the minimum number of trees—henceforth denoted by $M_{\min}$—needed to guarantee the upper bound P^*.

Shi Distance. Let $\mu_t = \left[1-((1-\theta)(1-\epsilon))^h\right]$. Then, the upper bound P^* is guaranteed by requiring $\exp\left(\frac{-M\delta^2\mu_t}{3}\right) \leq P^*$ which implies $M \geq \frac{-3\ln P^*}{\mu_t\delta^2}$.

The last expression is minimized by $\delta = 1$, hence we have:

$$M_{\min}^{Sh} = \frac{-3 \ln P^*}{\mu_t} = \frac{-3 \ln P^*}{1 - ((1-\theta)(1-\epsilon))^h}. \tag{27}$$

RatioRF Distance. Analogously, we can compute the minimal number of trees $M_{\min}^{RR}$, necessary for guaranteeing the upper bound P^* on the probability in the right side of Eq. (10), when the RatioRF distance is used. Let $\mu_t^i = [1 - (1-\theta)(1-\epsilon)]$ be the upper bound we obtained on the expected value of X_t^i as defined in Eq. (20). Proceeding as for the Shi distance, we get the condition $M \geq \frac{-3 \ln(0.5 P^*)}{h \mu_t^i \delta^2}$, from which (with $\delta = 1$) we have

$$M_{\min}^{RR} = \frac{-3 \ln(0.5 P^*)}{h \left[1 - (1-\theta)(1-\epsilon)\right]} \tag{28}$$

For a better visualization of the relationship between $M_{\min}^{Sh}$ and $M_{\min}^{RR}$ and the parameter ϵ, in Fig. 1(a), we provide such plots, for increasing $\epsilon \in [0.05, 0.5]$, assuming the other parameters in (27) and (28) fixed to $\theta = 0.01, P^* = 0.05$, and $h = 8$ which represents the expected height of the trees built on 256 samples (see e.g. [2,6]). As empirically accepted, these plots show that small forests are indeed sufficient. We can also observe that the curve for RatioRF is drastically better, especially for larger ϵ.

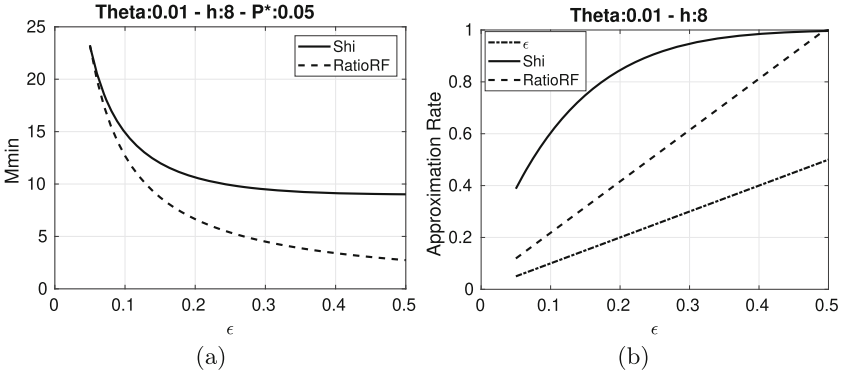

Fig. 1. (a) $M_{\min}^{Sh}$ and $M_{\min}^{RR}$ versus ϵ; (b); Approximation versus ϵ.

4.2 The Approximation δ_ϵ

From the conditions in Eqs. (7) and (9) we estimate the minimum δ_ϵ, which constraints the best approximation guarantee one can probabilistically achieve with the RF-distance.

Shi Distance. From the condition (7) we have that δ_ϵ is lower bounded as $\delta_\epsilon \geq \frac{(1+\delta)\left[1-((1-\theta)(1-\epsilon))^h\right]}{\epsilon} - 1$, and its minimum value—denoted by $\delta^{Sh}_{\epsilon(\min)}$— is achieved with $\delta \mapsto 0$, that is

$$\delta_\epsilon > \delta^{Sh}_{\epsilon(\min)} = \frac{\left[1-((1-\theta)(1-\epsilon))^h\right]}{\epsilon} - 1 \tag{29}$$

RatioRF Distance. Analogously, for the best approximation guarantee achievable in the case of the RatioRF distance, as given by the minimum value of δ_ϵ in condition (9), here denoted $\delta^{RR}_{\epsilon(\min)}$, we have:

$$\delta_\epsilon > \delta^{RR}_{\epsilon(\min)} = \frac{[1-(1-\theta)(1-\epsilon)]}{\epsilon} - 1 = \frac{\theta(1-\epsilon)}{\epsilon} \tag{30}$$

On the basis of (29) and (30), in Fig. 1(b), we plot the (best possible) approximation $(1+\delta^{Sh}_{\epsilon(\min)})\epsilon$ computed with the Shi distance and $2(1+\delta^{RR}_{\epsilon(\min)})\epsilon$ computed with the RatioRF distance as a function of the parameter ϵ taken as an estimate of the true distance. Also in this case the remaining parameters are fixed to $\theta = 0.01, P^* = 0.05$ and $h = 8$.

4.3 Using the Bounds for Estimating the Size of the Forest

Let us conclude our treatment with some practical considerations on how, in a given problem, the bounds can be used to compute the minimal number of trees required to get the guarantee with a given probability P^* and a given approximation parameter δ_ϵ. The procedure is described in the following. Please note that we repeat and summarize some of the formulas shown before, in order to have a clear comparison between the two distances.

Step 1. Fix the required approximation on the distance δ_ϵ. Important, the conditions in Eqs. (29) and (30) should hold:

$$\text{Shi:} \quad \delta_\epsilon > \delta^{Sh}_{\epsilon(\min)} = \frac{\left[1-((1-\theta)(1-\epsilon))^h\right]}{\epsilon} - 1 \tag{31}$$

$$\text{RatioRF:} \quad \delta_\epsilon > \delta^{RR}_{\epsilon(\min)} = \frac{\theta(1-\epsilon)}{\epsilon} \tag{32}$$

It is possible that for some choices of θ, ϵ, h the corresponding δ_ϵ is too high.

Step 2. Compute the corresponding $\delta_{(\max)}$, i.e. the largest δ for which the validity conditions in Eqs. (7) and (9) for the bounds hold. Please note that we are looking for the maximum δ since this would permit to get the minimal amount of trees

$$\text{Shi:} \quad \delta^{Sh}_{(\max)} = \frac{(1+\delta_\epsilon)\epsilon}{1-((1-\theta)(1-\epsilon))^h} - 1 \tag{33}$$

$$\text{RatioRF:} \quad \delta^{RR}_{(\max)} = \frac{(1+\delta_\epsilon)\epsilon}{1-(1-\theta)(1-\epsilon)} - 1 \tag{34}$$

Step 3. Compute the minimum number of trees for which the bound holds for a given probability P^* and for the given approximation level δ_ϵ, which is:

$$\text{Shi:} \quad M_{\min}^{Sh} = \frac{-3 \ln P^*}{\left[1 - ((1-\theta)(1-\epsilon))^h\right] (\delta_{(\max))}^{Sh})^2} \tag{35}$$

$$\text{RatioRF:} \quad M_{\min}^{RR} = \frac{-3 \ln(0.5P^*)}{h\,[1 - (1-\theta)(1-\epsilon)]\,(\delta_{(\max)}^{RR})^2} \tag{36}$$

In Fig. 2 we provide the number of required trees for different values of δ_ϵ. Also in this case let us keep fixed $\theta = 0.01$ and $P^* = 0.05$, and let us vary ϵ in the interval [0.05-0.5], with step 0.01, with $h = 8$. The behaviour is reported in Fig. 2. Please note that we report only the values for RatioRF, since $\delta_{\epsilon(\min)}$ for Shi is always larger than the required approximation δ_ϵ. This provides a further theoretical confirmation of the superiority of the RatioRF measure with respect to the Shi distance: for the latter measure, with the analysed configuration of θ, P^*, it is not possible to have a reasonably low approximation rate.

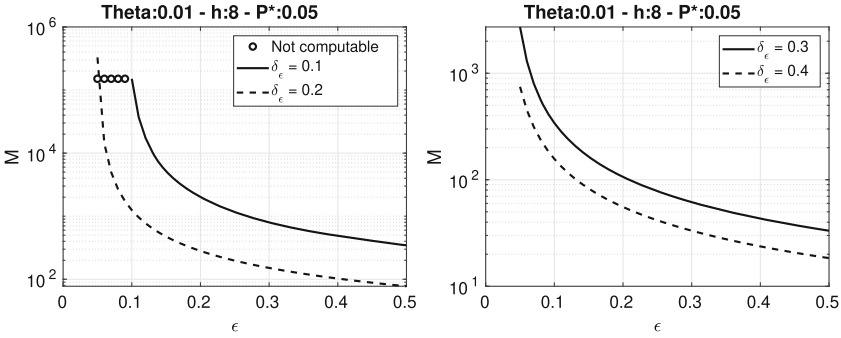

Fig. 2. Number of required trees, for a different approximation levels δ_ϵ. A circle denotes a value of ϵ for which the condition on δ_ϵ was not satisfied (i.e. $\delta_\epsilon < \delta_{\epsilon(\min)}$)

5 Discussion and Final Remarks

In the paper we have shown that, given a proper representation, it is possible to approximate the true distance between two objects with a ERT-based RF-distance, an approximation which is guaranteed by choosing a sufficiently large number of trees. This provides a theoretical confirmation of the widely assessed empirical efficacy of such ERT-based RF distances. Moreover, we have also shown that the approximation rate is drastically better for the RatioRF distance, thus confirming the empirical results shown in [6].

In the presentation of our result, we assumed that the tests on the same root-to-leaf path are on distinct features, so that the thresholds are all uniformly randomly sampled in $[0, 1]$. Let us now concentrate on the general case, i.e. we can use many times the same feature. We will concentrate on the Shi distance, similar reasoning can also be applied in the RatioRF case. Suppose that two objects are at distance less than ϵ; suppose first that all tests are on the same feature f, and that $f \in \tilde{F}$. In this case, the probability that the two objects are not split in the first i tests is $\geq \max\{0, 1 - (2^i - 1)\epsilon\}$.

If we now assume that, along a root to leaf path (of length h), feature f is tested h_f times, we have:

$$P(X_t = 0) \geq \prod_{i=1}^{h_f} \max\left\{0, \left[1 - (2^i - 1)\epsilon\right]\right\}$$

Note that this probability is smaller than the probability defined in Eq. (14), possibly becoming 0: this is reasonable, since if we continue to split on the same feature f we continue to reduce the domain, which will be at a certain level so small that x and y are split with probability 1.

If we have $|F|$ features, each one used h_f times in the path, then the probability in (14) can be written as

$$Pr(X_t = 0) \geq (1 - \theta)^h \prod_{f=1}^{|F|} \left[\prod_{i=1}^{h_f} \max\left\{0, \left[1 - (2^i - 1)\epsilon\right]\right]\right] \tag{37}$$

From this, we have that the expected value $E[X_t]$ satisfies:

$$E[X_t] \leq 1 - \left((1 - \theta)^h \prod_{f=1}^{|F|} \left[\prod_{i=1}^{h_f} \max\left\{0, \left[1 - (2^i - 1)\epsilon\right]\right]\right]\right) \tag{38}$$

which can now be used in the place of the one defined in Eq. (15) to complete the proof also in this case.

Rather than re-deriving the resulting (much more involved) bound, we limit ourselves to observe that the bound worsens with the decrease of the expected value. On the other hand, this analysis provides additional interesting information. The fact that, when features are expected to be reused several times on the same root-to-leaf path induces a worst approximation guarantee of the RF-distance with respect to the true distance provides a theoretical justification for the empirical evidence that for most of the ERT-based RF-distances the good results are obtained i) by using small trees – e.g. [6,25,26,28] and ii) on datasets with a large number of features – e.g. [5].

Ethical Statement. We do not see any evident ethical implication of our submission. Our paper is mainly theoretical, not involving the collection and processing of personal data, or the inference of personal information. We do not see any potential use of our work for military applications.

References

1. Arkadev, A.G., Braverman, E.M.: Teaching Computers to Recognize Patterns. Academic, Transl. from the Russian by W. Turski and J.D. Cowan (1967)
2. Aryal, S., Ting, K., Washio, T., Haffari, G.: A comparative study of data-dependent approaches without learning in measuring similarities of data objects. Data Min. Knowl. Discov. **34**(1), 124–162 (2020)
3. Aryal, S., Ting, K.M., Haffari, G., Washio, T.: Mp-Dissimilarity: a data dependent dissimilarity measure. In: 2014 IEEE International Conference on Data Mining, pp. 707–712. IEEE (2014)
4. Aryal, S., Ting, K.M., Washio, T., Haffari, G.: Data-dependent dissimilarity measure: an effective alternative to geometric distance measures. Knowl. Inf. Syst. **53**(2), 479–506 (2017). https://doi.org/10.1007/s10115-017-1046-0
5. Bicego, M., Escolano, F.: On learning random forests for random forest-clustering. In: Proceedings of the International Conference on Pattern Recognition (ICPR), pp. 3451–3458. IEEE (2021)
6. Bicego, M., Cicalese, F., Mensi, A.: RatioRF: a novel measure for random forest clustering based on the Tversky's ratio model. IEEE Trans. Knowl. Data Eng. **35**(1), 830–841 (2023)
7. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
8. Breiman, L., Friedman, J., Olshen, R., Stone, C.: Classification and regression trees. Wadsworth (1984)
9. Breiman, L.: Some infinity theory for predictor ensembles. Tech. Rep. CiteSeer (2000)
10. Criminisi, A., Shotton, J., Konukoglu, E.: Decision forests: a unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. Found. Trends Comput. Graph. Vis. **7**(2–3), 81–227 (2012)
11. Davies, A., Ghahramani, Z.: The random forest Kernel and other Kernels for big data from random partitions. arXiv preprint arXiv:1402.4293 (2014)
12. Domingues, R., Filippone, M., Michiardi, P., Zouaoui, J.: A comparative evaluation of outlier detection algorithms: experiments and analyses. Pattern Recognit. **74**, 406–421 (2018)
13. Duin, R.P., Pekalska, E.: Dissimilarity representation for pattern recognition. Foundations and applications, vol. 64. World scientific (2005)
14. Duin, R.: Compactness and complexity of pattern recognition problems. In: Proceedings of the International Symposium on Pattern Recognition "In Memoriam Pierre Devijver", pp. 124–128. Royal Military Academy (1999)
15. Emmott, A.F., Das, S., Dietterich, T., Fern, A., Wong, W.K.: Systematic construction of anomaly detection benchmarks from real data. In: Proceedings of the ACM SIGKDD Workshop on Outlier Detection and Description, pp. 16–21 (2013)
16. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**(1), 3–42 (2006)

17. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422. IEEE (2008)
18. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation-based anomaly detection. ACM Trans. Knowl. Discov. Data (TKDD) **6**(1), 1–39 (2012)
19. Mitzenmacher, M., Upfal, E.: Probability and computing: randomized algorithms and probabilistic analysis. Cambridge University Press (2005)
20. Moosmann, F., Triggs, B., Jurie, F.: Fast discriminative visual codebooks using randomized clustering forests. In: Advances in Neural Information Processing Systems 19, pp. 985–992 (2006)
21. Quinlan, J.: C4.5: programs for machine learning. Morgan Kaufmann Publishers Inc. (1993)
22. Scornet, E.: Random forests and Kernel methods. IEEE Trans. Inf. Theory **62**(3), 1485–1500 (2016)
23. Shi, T., Horvath, S.: Unsupervised learning with random forest predictors. J. Comput. Graph. Stat. **15**(1), 118–138 (2006)
24. Shotton, J., Johnson, M., Cipolla, R.: Semantic texton forests for image categorization and segmentation. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2008) (2008)
25. Ting, K.M., Wells, J.R., Washio, T.: Isolation Kernel: the X factor in efficient and effective large scale online kernel learning. Data Min. Knowl. Disc. **35**(6), 2282–2312 (2021)
26. Ting, K.M., Xu, B.C., Washio, T., Zhou, Z.H.: Isolation distributional Kernel: a new tool for kernel based anomaly detection. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 198–206 (2020)
27. Ting, K.M., Zhu, Y., Carman, M., Zhu, Y., Washio, T., Zhou, Z.H.: Lowest probability mass neighbour algorithms: relaxing the metric constraint in distance-based neighbourhood algorithms. Mach. Learn. **108**, 331–376 (2019). https://doi.org/10.1007/s10994-018-5737-x
28. Ting, K.M., Zhu, Y., Zhou, Z.H.: Isolation Kernel and its effect on SVM. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 2329–2337 (2018)
29. Ting, K., Zhu, Y., Carman, M., Zhu, Y., Zhou, Z.H.: 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, pp. 1205–1214 (2016)
30. Tversky, A.: Features of similarity. Psychol. Rev. **84**(4), 327 (1977)
31. Wells, J.R., Aryal, S., Ting, K.M.: Simple supervised dissimilarity measure: bolstering iForest-induced similarity with class information without learning. Knowl. Inf. Syst. **62**, 3203–3216 (2020)
32. Zhu, X., Loy, C., Gong, S.: Constructing robust affinity graphs for spectral clustering. In: Proceedings of the International Conference on Computer Vision and Pattern Recognition, CVPR 2014, pp. 1450–1457 (2014)