An Interesting Property of Random Forest Distances with Respect to the Curse of Dimensionality

Manuele Bicego(✉) and Ferdinando Cicalese

University of Verona, Verona, Italy
{manuele.bicego,ferdinando.cicalese}@univr.it

Abstract. Random forest distances represent a powerful class of data-dependent similarity measures whose usefulness has been shown in many different scenarios. In this paper, we discuss an interesting property of these measures with respect to the curse of dimensionality, i.e., the set of problems that may arise when the feature space is too large with respect to the number of available objects. Starting from a recent theoretical characterization of two RF-distances defined on an ensemble of Extremely Randomized Trees (ERT), we provide some empirical evidence that such distances are indeed robust to the curse of dimensionality, improving their performances when increasing the dimensionality of the space. Further, we empirically show that this behavior is not restricted to the ERT-based RF-distances, but in general, it also holds with alternative training schemes.

Keywords: Random Forest distances · Curse of dimensionality · RatioRF

1 Introduction

Random Forests (RF) [13] represent successful classification and regression tools, based on an ensemble of decision trees. In recent years, RFs have also been successfully exploited to derive meaningful measures of (dis)similarity [4], to be employed in different types of contexts, like nearest-neighbor classification, distance-based anomaly detection, and mostly, Random Forest Clustering [4,8,21,24,26]. In a RF-distance, the main idea is that it is possible to measure the similarity between two objects by looking at the way they traverse each tree of the forest. For example, in the simplest and most used RF-distance defined by Breiman [13,21], two objects are highly similar if, for many of the trees of the forest, they fall in the same leave, since they are answering in the same way to all the tests found in the root-to-leave path. This concept has be refined along different directions, leading to several definitions of RF-based similarity measures [4,8,24,26].

A. Torsello et al. (Eds.): S+SSPR 2024, LNCS 15444, pp. 188–198, 2025.
https://doi.org/10.1007/978-3-031-80507-3_19

Typically, to derive a RF-distance, we follow two steps: in the first, we learn a forest using the available data; in the second, we exploit the trained forest to define the similarity for a pair of objects x, y, by using the information contained in the paths x, y traverse in each tree. For what concerns the training, a widely adopted solution is to use *Extremely Randomized Trees* (ERT) [16], a class of decision trees in which randomization is taken to the extreme: in every node, the test is defined by choosing a random feature and a random threshold in the domain of that feature. This training scheme has been largely and successfully employed for RF-distance computation [4,9,24], being an unsupervised, simple, and efficient way to derive a forest, is shown to outperform alternatives even in supervised scenarios [23]. Recently, a theoretical justification of the good behavior of some ERT-based RF-distances has been proposed in [10]: in particular, assuming there exists a "true" distance, and assuming that there is a proper representation (i.e., a vectorial representation of the objects which satisfies the *Compactness Hypothesis* formulated by Arkadev and Braverman in 1967 [3]), [10] shows that near objects in the true distance are, with high probability, also near in the RF-distances computed with ERT forests. In the derivation of the theorems in [10], the assumption is that the tests on the same root-to-leaf path are on distinct features: if not making this restricting assumption and assuming that features can be reused several times on the same root-to-leaf path, [10] suggested a worst approximation guarantee of the RF-distance with respect to the true distance.

In this paper, we show another interesting property of these RF-distances. We started from the following reinterpretation of the findings in [10]: in problems with few features, the probability of re-using the same feature in different tests of the tree is higher than in problems with many features, since in each node of an ERT the feature on which to compute the split is randomly chosen among the available features. Therefore, we may argue that having more features would permit us to get a better approximation of the RF-distances with respect to the true distance. This intriguing property would be clearly in contrast with the Curse of Dimensionality, which says that in too high dimensional spaces, measures of similarity become meaningless: due to the increasing sparsity of the data, under reasonable assumptions, the ratio between the distances of the nearest and farthest neighbors tends to 1 for a variety of distance functions (the so-called concentration effect) [1,22]. In this paper, we empirically investigate this issue, providing some evidence that this appealing behavior for RF-distances is actually true, i.e., a better behavior is obtained when using more features.

In particular, we show our point with classification tests and *feature curves*, namely curves measuring the generalization error of a classifier exploiting a RF-distance defined on a varying number of features. We used the Nearest Neighbor rule, measuring the generalization error with the Leave One Out error protocol. We started from some datasets for which the feature curves, computed using the classic Euclidean Distance, show the expected curse of dimensionality behavior (i.e., the error starts to increase after a certain critical number of features). We then compute on the same datasets the feature curves using some well-kwown RF-distances, showing that instead the generalization error keeps on having

a decreasing trend also when adding more and more features. Furthermore, we provide evidence of two other interesting facts: i) this behavior is not restricted to the distances for which the theorems in [10] hold, but also for other RF-distances; ii) the behavior mainly persists even if we use training schemes alternative to ERTs, such as those typically employed for clustering [9,21].

The rest of the paper is organized as follows: in Sect. 2 we recap the main ideas behind Random Forest distances, whereas in Sect. 3 we summarize the issues related to the curse of dimensionality, especially in the distance case. Then, Sect. 4 contains the main empirical findings, and, finally, Sect. 5 concludes the paper with some discussion.

2 Random Forest Distances

Random Forest distances represent powerful and flexible data dependent measure of similarity [4], shown to be very useful in different tasks [8,9,24,26], also in the presence of missing data [11]. Typically, a RF-distance is defined in two steps: i) an RF is trained on the available data; ii) a distance between two objects is defined through the trained RF, typically by making the two objects traverse all the trees of the trained Forest, and comparing the answers they provide.

Let us summarize here the two distances employed in [10]. Given an object x, and a tree t, let us call $\ell_t(x)$ the leaf reached by the object x 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 measure, which we call *Shi*, is the RF distance introduced by Breiman in his seminal paper [13] and then employed by Shi and colleagues for Random Forest Clustering [21]. The distance simply counts in how many trees, over the total number of trees, two objects do not fall in the same leave:

$$d_{Shi}(x,y) = \sqrt{\frac{1}{T}\sum_t (1 - \mathrm{I}(\ell_t(x), \ell_t(y)))} \tag{1}$$

where T represents the number of trees in the forest, and $\mathrm{I}(a,b)$ is the indicator function:

$$\mathrm{I}(a,b) = \begin{cases} 1 & \text{if} a = b \\ 0 & \text{if} a \neq b \end{cases} \tag{2}$$

The second measure is the recent RatioRF [8], which compares two objects on the basis of the answers they provide to all the tests contained in the two paths $P_t(x)$ and $P_t(y)$. In more detail, calling S_t^x the set of tests in the path $P_t(x)$ of x, we define as S_t^{xy} the set of all tests in the two paths, $S_t^{xy} = S_t^x \cup S_t^y$, and A_t^{xy} as the set of tests in S_t^{xy} on which x and y agree. The RatioRF distance is then defined as

$$d_{RRF}(x,y) = \sqrt{1 - \frac{1}{T}\sum_t \frac{|A_t^{xy}|}{|S_t^{xy}|}} \tag{3}$$

For these two distances, results in [10] show that when i) the Compactness Hypothesis [3] holds, i.e. the problem representation is such that similar objects

have close representations (i.e. they are close for most of the features), and ii) the forest is composed by Extremely Randomized Trees [16], then 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.

3 Curse of Dimensionality

Curse of dimensionality [2] can be defined as "the severe difficulty that can arise in spaces of many dimensions" [12]. This phenomenon, first spotted by Bellman in [6], has gained a renewed interest in recent years, mainly due to the surprising findings in the neural network community related to the so-called double descent [5,19]. In a few words, the curse of dimensionality represents a set of problems that may arise when employing PR/ML tools in problems with too many features with respect to the training objects. A dual (and more general) version of the phenomenon postulates that the Curse of Dimensionality may arise when the model is *too complex* with respect to the number of training objects, like for example, when employing a polynomial model with a too large degree – within this more general view, the number of features is no more than a measure of "complexity" of a model.

Among the different problems that may occur, let us cite two examples: i) when increasing the dimension, keeping fixed the number of objects, the space becomes almost empty, making density estimation impossible; ii) a model trained in a too high dimensional space –or, more intuitively, a too complex model– can be easily overtrained, not being able to generalize. Also distances have been largely studied with respect to the curse of dimensionality: in this case, the main result is the so-called "concentration effect" [1,22], which says that a distance that depends on too many independent features – i.e. a distance in a too high dimensional space – is almost constant. Some authors [7,17,20] also discussed the effects of this problem on the computation of the Nearest Neighbor (or the K-neaerest Neighbor) in high dimensional spaces, postulating the conditions under which such computation is meaningful. These aspects may affect the generalization capabilities of distance-based classifiers (like K-Nearest Neighbor): luckily, the concentration effect in practical cases is not as severe as the theory says, the generalization depending on the intrinsic dimensionality of the dataset [18], and, crucially, also on the distance. In this respect, [15] showed that the concentration does not derive from the finiteness of the dataset, but is an intrinsic property of the distance; in this paper, we provide some more empirical findings on this aspect, showing that RF-distances are robust with respect to the dimension of the feature space.

4 Experiments

In this section our empirical analysis is provided. We first introduce the experimental details, followed by results, comments and extensions.

Table 1. Summary of the employed Datasets

Name	Source	# objects	# features	# classes
ACSF1	UCR-TS (ACSF1)	200	1460	10
BeetleFly	UCR-TS (BeetleFly)	40	512	2
BirdChicken	UCR-TS (BirdChicken)	40	512	2
Gait	UCI-ML (Gait Classification)	48	321	16
Gastro-WL	UCI-ML (Gastrointestinal Lesions - White Light)	76	698	3
Gastro-NBI	UCI-ML (Gastrointestinal Lesions - Narrow Band Imaging)	76	698	3
HouseTwenty	UCR-TS (HouseTwenty)	159	2000	2
Herring	UCR-TS (Herring)	128	512	2
ORL-2Sub	Kaggle (ORL face dataset)	20	4096	2

4.1 Experimental Details

As we have seen in Sect. 3, the curse of dimensionality is defined as the set of problems that may occur when *the number of features is too large with respect to the number of objects*. Thus, to observe this effect, we searched for some problems with a relatively small number of objects and a remarkably high number of features: this is a very common scenario in biomedical domains, where few patients are typically characterized by a very large number of features (think, for example, to spectra or expression data). We also employed some datasets from sequence benchmarks, since in typical scenarios like the UCR Time series classification Archive [14] all sequences have the same length, and can be then considered as vectors – actually in such scenarios all the baseline errors are computed with the Nearest Neighbor rule and the Euclidean Distance. All the datasets are summarized in Table 1, where in the "source" column UCI-ML indicates the UCI Machine Learning Repository[1] CVR-TS the UCR Time Series Classification Archive[2] and Kaggle the Kaggle repository[3] All datasets were used as provided, except for the ORL-2Sub; in this case, the images have been resized to 64×64 and vectorized, and we used only subject 1 and subject 36, representing the most difficult pair of subjects to be classified, according to a LOO nearest neighbor evaluation.

To estimate the feature curves, we compute the Leave One Out error of the nearest neighbor classifier which uses in input the distance. LOO errors have been computed with the number of features n equal to 4, 8, 16, 32, ... and so on up to the whole dimension of the dataset. For every such choice of n, n random features have been extracted and used. In the subsequent step (i.e. when

[1] https://archive.ics.uci.edu/.
[2] https://www.cs.ucr.edu/~eamonn/time_series_data_2018/.
[3] https://www.kaggle.com/datasets.

extracting $2n$ features), the previous set has been maintained (i.e., the set with $2n$ features is obtained by adding n novel random features to the existing set) – this being in line with classical settings in learning curves [25]. RF distances have been computed using forests with 100 trees, each built on a random 50% of the dataset, and grown until its maximum depth. The whole procedure has been repeated 500 times, and averaged results (with standard errors of the mean) are computed. To ensure a fair comparison among distances, the same random selection of features has been used for all distances.

4.2 Results

As a preliminary result, we show in Fig. 1 the feature curves computed by using the Euclidean Distance. In this case, it is evident that the NN based on Euclidean distance suffers from the curse of dimensionality, showing the classical U-shaped curve: at the beginning, adding more features is useful, but after a certain level, it makes the classifier more confused.

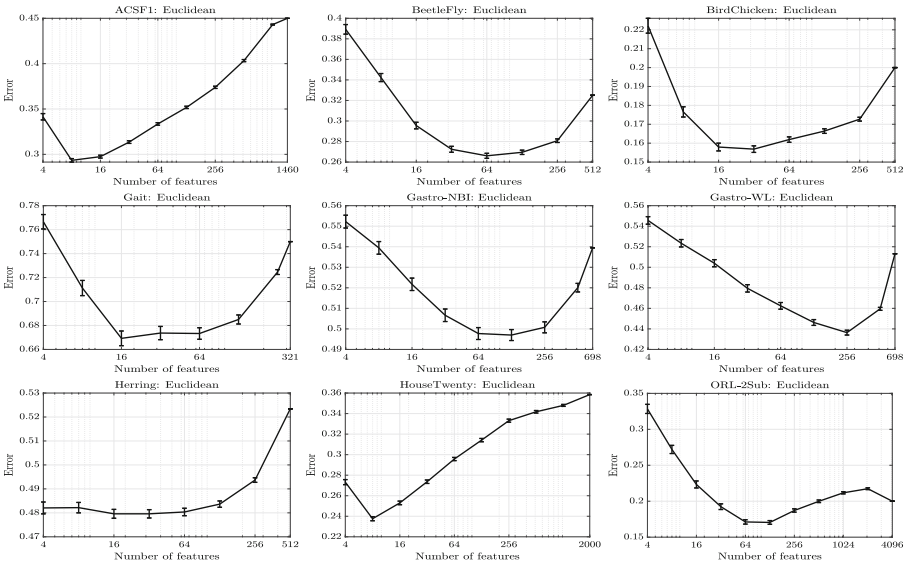

Fig. 1. Feature curves with Euclidean Distances

Then we computed the curves for the two RF-distances considered in the theoretical study of [10], namely the Shi distance and the RatioRF distance. Results are shown in Fig. 2: it seems evident that the curse of dimensionality is not an issue anymore, as the generalization error appears always to decrease when increasing the dimensionality of the dataset. These results confirm and extend the theoretical findings of [10], which hold under the restricting assumptions of the presence of a "proper representation" and a "true" distance: with

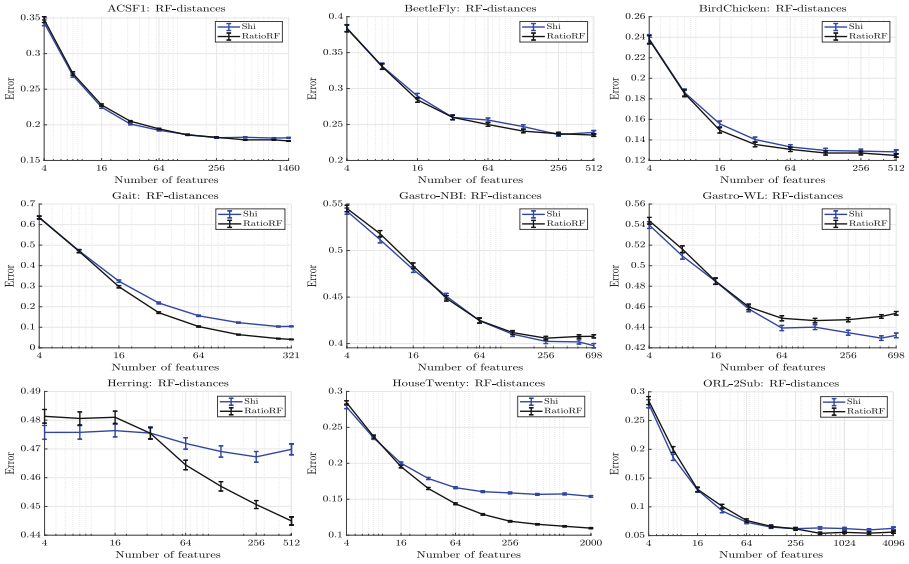

Fig. 2. Feature curves with RatioRF and Shi distances

more features, we have a better approximation of the true distance with the RF-distance, which, empirically, corresponds to a better generalization of a classifier exploiting it.

Let us extend the scope of the empirical analysis: theorems in [10] have been shown for two RF-distances (Shi and RatioRF) built on forests trained with the ERT paradigm, and one may wonder if the good behavior with respect to the curse of dimensionality depends on the particular distance and training employed, or if it generalizes to other RF-distances and other training schemes. To investigate this aspect, we first repeated the same set of experiments using RF-distances alternative to Shi and RatioRF, computed again on ERT ensembles: the "Zhu2" distance, the second variant introduced in [26] (called *ClustRF-Strct-Unfm* in such paper), the "Ting" measure, a mass-based Random Forest distance defined in [24], and the "Aryal" measure, a more recent RF-distance defined in [4] which implements an extension of the class of m_p distances.

Then, we repeated the experiments with an alternative training scheme, the so-called "negative-sampling" method, which represents the first and most classic solution for computing RF distances for RF-clustering [9,21,26]. Within this paradigm, a classification forest is trained on a two classes problem: one class contains the training points, the second contains a set of synthetically generated points, obtained by random sampling from the product of empirical marginal distributions (to remove the dependency between features). The results for alterna-

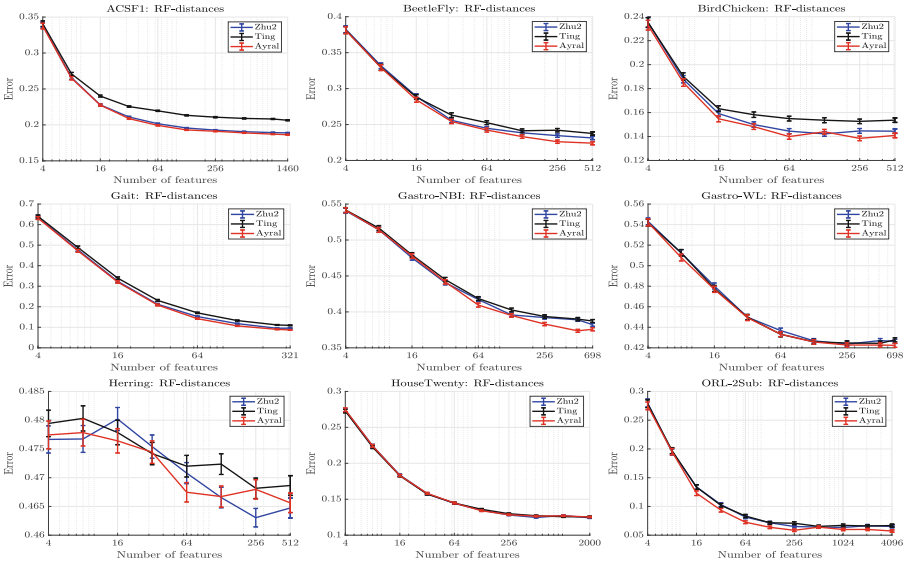

Fig. 3. Feature curves with other RF distances, defined on ERT forests.

tive distances, based on ERT forests, are reported in Fig. 3, whereas the results with the negative-sampling training schemes are shown in Figs. 4, for all RF-distances (in this case, due to the high computational requirements of this training scheme, we repeated the experiments only 100 times). For what concerns the ERT trained distances, we can observe that also in this case we have a good behavior of generalization errors, which do not show the classical U-shape of curse of dimensionality. When changing the training scheme, we can observe a generally good trend, even if, in some cases, not as good as when trained with ERT (see, for example, the ACSF1 and the Herring dataset).

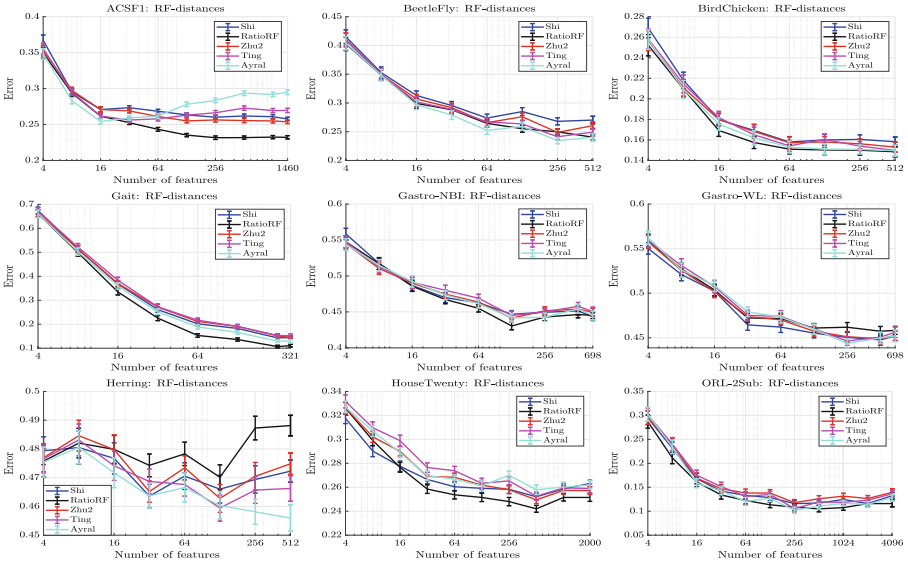

Fig. 4. Feature curves with all RF distances with the "Sample Negative" training scheme.

5 Discussion and Conclusion

In this paper, we provided some empirical evidence that RF-distances are not affected by the curse of dimensionality, especially if trained with the ERT scheme. This represents another empirical confirmation of the goodness of these distances in facing ML/PR problems. From a methodological perspective, why should RF-distances be less prone to the curse of dimensionality? The first explanation derives from the investigation of [10], which focused on showing that the distance computed by the forest does not diverge too much from the actual distance (assumed to be well represented in the data), and in fact indicating that high dimensionality data are more likely to guarantee this. More than this, our intuition is that the natural function of a decision tree is to separate objects, hence, to possibly magnify the distance even between very similar objects. Quantitatively, a few tests with different result over a reasonably short tree (compared to a possibly very high total number of features) translate into a significant difference. In this sense, a decision tree-based distance is less likely to become "uniform", i.e., it is less prone to suffer from the concentration effect.

Acknowledgments. M.B. would like to thank Tom Viering (TUDelft) for suggesting the connection between the work in [10] and the curse of dimensionality.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Aggarwal, C.C., Hinneburg, A., Keim, D.A.: On the surprising behavior of distance metrics in high dimensional space. In: Proceedings of the International Conference on Database Theory (ICDT), pp. 420–434 (2001)
2. Altman, N., Krzywinski, M.: The curse (s) of dimensionality. Nat. Methods **15**(6), 399–400 (2018)
3. Arkadev, A.G., Braverman, E.M.: Teaching computers to recognize patterns. Transl. from the Russian by W. Turski and JD Cowan. Academic (1967)
4. 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)
5. Belkin, M., Hsu, D., Ma, S., Mandal, S.: Reconciling modern machine-learning practice and the classical bias-variance trade-off. Proc. Natl. Acad. Sci. **116**(32), 15849–15854 (2019)
6. Bellman, R.: Adaptive Control Processes: A Guided Tour. Princeton University Press, Princeton (1961)
7. Beyer, K., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is "nearest neighbor" meaningful? In: Proceedings of the International Conference on Database Theory (ICDT), pp. 217–235. (1999)
8. 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)
9. 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)
10. Bicego, M., Cicalese, F.: On the good behavior of extremely randomized trees in random forest-distance computation. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD), pp. 645–660 (2023)
11. Bicego, M., Cicalese, F.: Computing random forest-distances in the presence of missing data. ACM Trans. Knowl. Discov. Data **18**(7) (2024)
12. Bishop, C.M.: Pattern Recognition and Machine Learning. Springer, Heidelberg (2006)
13. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
14. Dau, H.A., et al.: The UCR time series archive. IEEE/CAA J. Autom. Sinica **6**(6), 1293–1305 (2019)
15. François, D., Wertz, V., Verleysen, M.: The concentration of fractional distances. IEEE Trans. Knowl. Data Eng. **19**(7), 873–886 (2007)
16. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. Mach. Learn. **63**(1), 3–42 (2006)
17. Hinneburg, A., Aggarwal, C.C., Keim, D.A.: What is the nearest neighbor in high dimensional spaces? In: Proceedings of the International Conference on Very Large Data Bases, pp. 506–515 (2000)
18. Kpotufe, S.: k-NN regression adapts to local intrinsic dimension. In: Advances in Neural Information Processing Systems, vol. 24 (2011)
19. Loog, M., Viering, T., Mey, A., Krijthe, J.H., Tax, D.M.: A brief prehistory of double descent. Proc. Natl. Acad. Sci. **117**(20), 10625–10626 (2020)
20. Radovanovic, M., Nanopoulos, A., Ivanovic, M.: Hubs in space: popular nearest neighbors in high-dimensional data. J. Mach. Learn. Res. **11**(sept), 2487–2531 (2010)

21. Shi, T., Horvath, S.: Unsupervised learning with random forest predictors. J. Comput. Graph. Stat. **15**(1), 118–138 (2006)
22. Talagrand, M.: A new look at independence. Ann. Probab. 1–34 (1996)
23. Ting, K.M., Zhu, Y., Zhou, Z.H.: Isolation kernel and its effect on SVM. In: Proceedings of the International Conference on Knowledge Discovery & Data Mining (KDD), pp. 2329–2337 (2018)
24. Ting, K., Zhu, Y., Carman, M., Zhu, Y., Zhou, Z.H.: Overcoming key weaknesses of distance-based neighborhood methods using a data dependent dissimilarity measure. In: Proceedings of the International Conference on Knowledge Discovery and Data Mining (KDD), pp. 1205–1214 (2016)
25. Viering, T., Loog, M.: The shape of learning curves: a review. IEEE Trans. Pattern Anal. Mach. Intell. **45**(6), 7799–7819 (2022)
26. 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)