



Exploiting second-order dissimilarity representations for hierarchical clustering and visualization

Helena Aidos¹

Received: 1 December 2019 / Accepted: 18 April 2022 / Published online: 11 May 2022

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2022

Abstract

The representation of objects is crucial for the learning process, often having a large impact on the application performance. The dissimilarity space (DS) is one of such representations, which is built by applying a dissimilarity measure between objects (*e.g.*, Euclidean distance). However, other measures can be applied to generate more informative data representations. This paper focuses on the application of second-order dissimilarity measures, namely the Shared Nearest Neighbor (SNN) and the Dissimilarity Increments (Dinc), to produce new DSs that lead to a better description of the data, by reducing the overlap of the classes and by increasing the discriminative power of features. Experimental results show that the application of the proposed DSs provide significant benefits for unsupervised learning tasks. When compared with Feature and Euclidean space, the proposed SNN and Dinc spaces allow improving the performance of traditional hierarchical clustering algorithms, and also help in the visualization task, by leading to higher area under the precision/recall curve values.

Keywords Dissimilarity representation · Dissimilarity increments · Shared nearest neighbor · Geometrical complexity · Clustering · Visualization

1 Introduction

Many different algorithms and methods have been proposed to solve multiple learning problems (Duda et al. 2001; Jain et al. 2000), usually supported on traditional feature-based data representations. Hence, many approaches disregard the way data is structured and presented to the learning algorithms. However, this may constrain the

Responsible editor: Johannes Fürnkranz, Ian Davidson.

✉ Helena Aidos
haidos@ciencias.ulisboa.pt

¹ LASIGE, Departamento de Informática, Faculdade de Ciências, Universidade de Lisboa, Lisbon, Portugal

effectiveness of such algorithms, since an efficient data representation may increase its discrimination power, decrease class overlap, and transform class boundaries, making it easier for the learning algorithms.

Typically, objects are represented by a set of features, which should be able to characterize them and be relevant to discriminate among the classes. The most popular representation is the Euclidean space (Duda et al. 2001). However, defining features to obtain a complete description of objects and with a high discriminant power may be a challenging task, since different objects may have the same feature-based representation, but still be different due to characteristics not considered in the chosen feature set. That difficulty emerges due to high-dimensional data or the need to describe objects using continuous and categorical variables, which may cause class overlap, leading to poor learning performances.

A dissimilarity representation, constructed from the comparison of objects, may be used to overcome the limitations raised by feature vectors (Pekalska and Duin 2002), solving the problem of class overlap, since only identical objects have a dissimilarity of zero.

Moreover, the dissimilarity representation has some attractive and promising advantages: the potential informativeness of non-Euclidean spaces (Duin and Pekalska 2010) or even non-metric dissimilarity measures (Duin and Pekalska 2010; Plasencia-Calaña et al. 2013). A large set of advantages of the dissimilarity representation, as an alternative to the classical feature approach, was investigated within the EU-FP7 SIMBAD¹ project, as reported in Pelillo (2013).

The construction of a dissimilarity representation can be based on metric learning (Duin et al. 2014) or a previously existing distance measure or function, which can discriminate between relevant and irrelevant objects, where relevant objects are defined as the ones closer to the reference object than irrelevant ones. Hence, relevant objects are the ones belonging to the same class as the reference, whereas irrelevant objects belong to a different class. In pattern recognition, dissimilarities have been used for many applications, *e.g.* cluster analysis (Theodoridis and Koutroumbas 2009) and visualization (Lee and Verleysen 2010). Moreover, based on the work of Pekalska and Duin (2005), some classification methods for dissimilarity data have been proposed (Duin and Pekalska 2010; Schleif et al. 2012; Chen et al. 2009), which are useful to tackle problems in computer vision, bioinformatics, among others (Liao and Noble 2003). More specifically, dissimilarity representations have been used as a labeled graph classifier in an optimized dissimilarity space embedding system (Livi 2017), or as a dissimilarity-based ensemble classification method for multiple instance learning (Cheplygina et al. 2016). Dissimilarity approaches are considered in the design of signature verification systems (Eskander et al. 2013; Batista et al. 2010); in frameworks that allow any person re-identification, (Satta et al. 2012); in systems for pose-based human action recognition (Theodorakopoulos et al. 2014); or even, in odour classification (Bicego 2005). Moreover, dissimilarity spaces have been used in biology-related applications, namely protein function classification (De Santis et al. 2018) or radiomics data classification (Cao et al. 2018). Also, dissimilarity-based classification techniques are used to schizophrenia classification using MRI (Ulas et al. 2011), and for ECG

¹ Similarity-based Pattern Analysis and Recognition project: <http://simbad-fp7.eu>.

biometrics (Marques et al. 2015; Batista et al. 2018). The dynamic time warping distance was used to construct a dissimilarity space to classify seismic volcanic patterns (Orozco-Alzate et al. 2015). Finally, frameworks based on multiple classifiers on dissimilarity space are proposed either to identify images of forest species (Martins et al. 2015) or for text categorization (Pinheiro et al. 2017).

Such works are based on first-order or primary dissimilarity measures, e.g. the Euclidean distance, to construct the dissimilarity representation, either by direct comparison of objects or by computing it over feature vectors. However, most first-order measures are sensible to variations within data distribution or the dimensionality of the data space. An interesting alternative are second-order or secondary measures, often based on rankings induced by a specified first-order measure. For instance, one can consider the shared nearest neighbor similarity measure proposed by Jarvis and Patrick (1973) and its variants. This measure was used to improve clustering, to find the most representative items in a set of objects (Ertöz et al. 2003), to enhance audio similarity (Pohle et al. 2006), or for outlier detection (Jin et al. 2006). Finally, in the context of topic segmentation, a third-order similarity measure or weighted second-order measure was proposed (Moreno et al. 2013).

We herein propose two novel dissimilarity representations of data, based on second-order measures: one consists of the information given by shared nearest neighbors (Jarvis and Patrick 1973), and the other consists of triplets of nearest neighbors (Fred and Leitão 2003). The first space, called *Shared Nearest Neighbor space* (SNN space), is built upon the concept of “overlap” between the neighborhoods of two objects. The other space, called *Dissimilarity Increments space* (Dinc space), is built by the increment in dissimilarity between an object and a set of representative objects (composed by edges between prototypes and its nearest neighbor).

Although the herein proposed Dinc space has some similarities with the feature lines approach proposed in Orozco-Alzate et al. (2009), the later representation has limited applicability to correlated datasets with a moderately nonlinear structure, as specifically pointed out by the authors. In contrast, the dissimilarity increments measure yields different values for each configuration of objects mentioned, and the proposed Dinc space shows to provide good representations for a wide set of unsupervised learning applications, including clustering and visualization (see Sects .5 and 6 , respectively). Thus, the main contributions of this paper are:

- The development of two new dissimilarity spaces based on already existing second-order dissimilarity measures. Although second-order dissimilarity measures have been used to compare objects in clustering or classification algorithms to improve accuracy (Aidos and Fred 2012; Aidos et al. 2012; Jarvis and Patrick 1973), to the best of our knowledge, they have not been used to construct feature-based dissimilarity representations.
- A characterization of each second-order dissimilarity space by applying measures of geometrical complexity of classification problems (Ho et al. 2006), thus providing insightful information about the geometry, topology, and density of the proposed spaces. By following the work from Pekalska and Duin (2005) a set of measures are applied to evaluate the metricity and the Euclidean behavior of each space.

- An evaluation of the computational cost to build each space when varying the number of objects of the original feature space. This study shows that, although the proposed second-order dissimilarity representations are more time consuming than their counterparts (the first-order dissimilarity measures), the proposed data representations are still computationally feasible while bringing advantages over the first-order dissimilarity space.
- Validation of the proposed spaces in the context of clustering problems by relying on traditional hierarchical techniques. Experimental results on a set of 20 datasets with different sizes and characteristics show average 9.9% relative improvement in the consistency index regarding the original feature space and 8.6% relative improvement regarding the Euclidean space.
- Validation of the proposed spaces in different types of datasets, namely, time-series, categorical and graph data. Experimental results show average 3.1%, 11.9% and 2.4% relative improvement for time-series, categorical and graph data, respectively, in the consistency index regarding the original feature space. Moreover, there are a relative improvement with SNN space of 14.1%, 18.2% and 2.3% in the consistency index for time-series, categorical and graph data, respectively, regarding the original feature space.
- Evaluation of the proposed spaces for visualization problems. Hence, four embedding algorithms are used to reveal that, regarding the original feature space, the proposed second-order spaces allow for an area under the curve (AUC) relative improvement of around 26.8%.

The remainder of this paper is organized as follows: Section 2 explains how to build dissimilarity spaces, and proposes two new spaces based on second-order measures—the shared nearest neighbor and the dissimilarity increments. Section 3 describes the datasets used throughout this paper. Section 4 presents a characterization for each dissimilarity space. Sections 5 and 6 evaluate the application of the proposed dissimilarity spaces for clustering and visualization tasks, respectively, while the final remarks and conclusions are drawn in Sect. 7.

2 Dissimilarity representation

To introduce the proposed dissimilarity representations, consider an input dataset $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ of n generic objects. Although such an object can represent an image, a signal, or any other object type, without loss of generality, it is herein assumed that \mathbf{x}_i is a feature vector in \mathbb{R}^p , represented as $\mathbf{x}_i = [x_{i1} \dots x_{ip}]$. Furthermore assume that the cardinality of a set is defined as $\text{card}(\cdot)$, such that $\text{card}(\mathcal{X}) = n$. Also let $\mathcal{R} = \{\mathbf{o}_1, \dots, \mathbf{o}_m\}$ be the set of representative objects, with $\text{card}(\mathcal{R}) = m$, such that $\text{card}(\mathcal{R}) \leq \text{card}(\mathcal{X})$. The set of representative objects \mathcal{R} can be either a subset of \mathcal{X} (i.e., $\mathcal{R} \subseteq \mathcal{X}$) or a set of prototypes obtained through some prototype selection method, e.g. (García et al. 2012; Calvo-Zaragoza et al. 2016).

A dissimilarity space (Pekalska and Duin 2005) can be defined as the data-dependent mapping $D(\cdot, \mathcal{R}) : \mathcal{X} \rightarrow \mathbb{R}^m$. Each object \mathbf{x}_i is described by an m -dimensional vector

$$D(\mathbf{x}_i, \mathcal{R}) = [d(\mathbf{x}_i, \mathbf{o}_1) \dots d(\mathbf{x}_i, \mathbf{o}_m)], \quad (1)$$

where $d(\cdot, \cdot)$ represents a dissimilarity measure. The *dissimilarity space* is therefore characterized by the $n \times m$ dissimilarity matrix D , where $D(\mathbf{x}_i, \mathcal{R})$ is the i -th row of D . Moreover, each vector \mathbf{o}_i represents a direction from the dissimilarity space, whose dimension is $\text{card}(\mathcal{R})$. For simplicity, assume that \mathcal{R} corresponds to the entire set \mathcal{X} , meaning that the dissimilarity space is represented as an $n \times n$ dissimilarity matrix. Hence, the dissimilarity space is defined as a vector space Y , where the i -th element corresponds to the vector $D(\mathbf{x}_i, \mathcal{R})$.

When taking $d(\cdot, \cdot)$ in (1) as the Euclidean distance

$$d(\mathbf{x}_i, \mathbf{o}_j) = \left(\sum_{l=1}^p (x_{il} - o_{jl})^2 \right)^{1/2}, \quad (2)$$

we obtain the so-called Euclidean dissimilarity space, which will be herein referred as *Euclidean Space*, for simplicity.

In this paper, two different spaces are proposed based on second-order dissimilarity measures, namely the *shared nearest neighbor space* and the *dissimilarity increments space*. The following subsections introduce such spaces.

2.1 Shared nearest neighbor space

Traditional (dis)similarity measures are pairwise or first-order measures, which means they are computed over pairs of objects. By relying on one of such first-order (dis)similarity measures, second-order measures can be defined (such as those based on rankings). In this manuscript, the first of such measures being presented builds on the shared nearest neighbor (SNN) information (Jarvis and Patrick 1973), which is built upon the concept of “overlap” between the neighborhoods of object pairs. The neighborhoods of each object are determined by any first-order measure (e.g., the L_p norm or the cosine similarity) and do not even impose the requirement for data objects to be represented as vectors. By relying on such a concept, a dissimilarity representation of data is herein proposed, which discovers natural groupings of different point densities, and handles noise and outliers.

To present such dissimilarity representation of data, let $NN_k(\mathbf{x}_i) \subseteq \mathcal{X}$ be the set of k -nearest neighbors ($k \in \mathbb{N}$) of $\mathbf{x}_i \in \mathcal{X}$, determined by a first-order (dis)similarity measure. The overlap between objects \mathbf{x}_i and \mathbf{o}_j is defined to be the intersection size

$$SNN_k(\mathbf{x}_i, \mathbf{o}_j) = \text{card}(NN_k(\mathbf{x}_i) \cap NN_k(\mathbf{o}_j)), \quad (3)$$

which produces a similarity measure between pairs of objects. Although there are several ways to transform it into a dissimilarity measure, a linear approach is herein adopted, such that the dissimilarity between an object \mathbf{x}_i and a object representative \mathbf{o}_j is defined as:

$$D(\mathbf{x}_i, \mathbf{o}_j) = k - SNN_k(\mathbf{x}_i, \mathbf{o}_j). \quad (4)$$

Hence, a dissimilarity matrix D is obtained that represents a dissimilarity space, which will be henceforth referred to as *SNN Space*.

The characteristics of the proposed dissimilarity space are evaluated by considering the Euclidean distance as the first-order dissimilarity measure to obtain $NN_k(\mathbf{x}_i)$, the set of nearest neighbors of an object \mathbf{x}_i . However, it should be noticed that depending on application characteristics and requirements, other application-specific measures can potentially be employed.

2.2 Dissimilarity increments space

The dissimilarity increments (DIs) is another second-order measure, built upon the concept of triplets of points, that can be used to construct a dissimilarity space (Aidos and Fred 2015b). However, to properly present such a space let us begin with the definition and properties of this measure. The actual definition of the DIs space, herein simply referred to as *Dinc Space*, will be performed in Sect. 2.2.2.

2.2.1 Dissimilarity increments: definition and properties

Assume that $(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k)$ is a triplet of nearest neighbors in \mathcal{X} , obtained as

$$\begin{aligned} (\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) & \text{— nearest neighbor triplet} \\ \mathbf{x}_j & : j = \arg \min_l \{d(\mathbf{x}_l, \mathbf{x}_i), l \neq i\} \\ \mathbf{x}_k & : k = \arg \min_l \{d(\mathbf{x}_l, \mathbf{x}_j), l \neq i, l \neq j\}. \end{aligned}$$

The *dissimilarity increments* (Fred and Leitão 2003) between neighboring patterns is defined as

$$d_{\text{inc}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = |d(\mathbf{x}_i, \mathbf{x}_j) - d(\mathbf{x}_j, \mathbf{x}_k)|, \quad (5)$$

where $d(\cdot, \cdot)$ represents the pairwise dissimilarity between two objects, which can be obtained by applying any first-order dissimilarity measure. As in the case of the SNN space, multiple measures could be adopted depending on application characteristics. However, to make the comparison between alternative data representation spaces consistent, the Euclidean distance is adopted.

The advantage of the DIs regarding the first-order pairwise measure is that it provides relevant information about the structure of a dataset. However, before presenting such space, it is important to remark the DIs properties (Aidos and Fred 2015a):

- The DIs is non-negative, hence $d_{\text{inc}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) \geq 0$ and $d_{\text{inc}}(\mathbf{x}_i, \mathbf{x}_j, \mathbf{x}_k) = 0 \implies d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_k)$
- It works with different data representations. The increment can be computed using a feature space and some (dis)similarity measure or using a dissimilarity representation of the objects when no features are available.

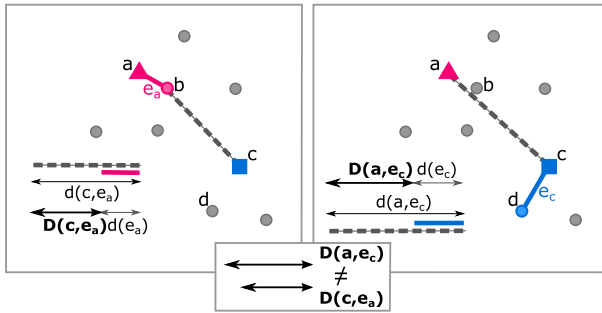


Fig. 1 Example of a set of objects to illustrate how to compute elements from the Dinc space D and to demonstrate its asymmetry. If a is a prototype, e_a is the representative object constructed as an edge between a and its nearest neighbor b . Then, $D(c, e_a)$ is the dissimilarity increment between c and the representative object, e_a , computed from (7). $D(c, e_a) \neq D(a, e_c)$ since different triplets of patterns are used to compute D . Adapted from Aidos and Fred (2015b)

- It has a smooth evolution. Inside a cluster or class, abrupt changes in the DIs should not occur. If abrupt changes take place, it indicates that we are in the presence of a different cluster or class.
- It identifies sparse clusters. While most of the distances used in the literature discard samples that are far apart in a sparse cluster, this measure can easily identify those patterns as belonging to the same cluster.
- It is invariant to shape features or orientation. The dissimilarity increment can be applied to identify the clusters with odd shapes since it only takes into account the nearest neighbors.

2.2.2 Dinc space

Based on the previous definition and properties of dissimilarity increment, it is possible to build a novel second-order dissimilarity space. Hence, like in the previous cases, each object in the second-order space is described by an n -dimensional dissimilarity vector $D(\mathbf{x}_i, \mathcal{R})$, which is computed by evaluating the dissimilarity increment between each object \mathbf{x}_i and the set of representative objects, $\{\mathbf{o}_1, \dots, \mathbf{o}_m\} \in \mathcal{R}$. For the DIs space, each representative object \mathbf{o}_j is constructed by considering the edge between a prototype \mathbf{v}_j (a sample of the dataset) and its nearest neighbor $\mathbf{x}_{\mathbf{v}_j}$. Thus, the distance between an object \mathbf{x}_i and the representative object \mathbf{o}_j is given by

$$d(\mathbf{x}_i, \mathbf{o}_j) = \min\{d(\mathbf{x}_i, \mathbf{v}_j), d(\mathbf{x}_i, \mathbf{x}_{\mathbf{v}_j})\}, \quad (6)$$

and the (i, j) -th element of our Dinc space is defined as

$$D(\mathbf{x}_i, \mathbf{o}_j) = |d(\mathbf{x}_i, \mathbf{o}_j) - d(\mathbf{o}_j)|. \quad (7)$$

By adopting such Dinc space representation, it is ensured that the matrix D is non-negative (from (7)) and asymmetric. To illustrate the asymmetric property, consider a set of objects distributed as shown in Fig. 1(left). If a is a prototype, then e_a is an edge

between a and its nearest neighbor b , and, accordingly, e_a will be the representative object. Now, the dissimilarity increment between c and the representative object, e_a , is $D(c, e_a)$. On the other hand, when c is a prototype (see Fig. 1(right)), the representative object, e_c , is the edge between c and its nearest neighbor d , and, thus, $D(a, e_c)$ is the dissimilarity increment between a and the representative object. Therefore, $D(c, e_a) \neq D(a, e_c)$ (see Fig. 1).

3 Datasets

The remainder of this paper is focused on characterizing the proposed dissimilarity spaces and applying them in unsupervised problems. In that sense, a total of 20 benchmark datasets is used from the UCI Machine Learning Repository². The considered datasets were chosen to represent a large set of problems, including those with different space dimensionalities, number of samples and/or classes, and covering balanced and unbalanced classes.

Such diversity can also be perceived from the dimensionality of the feature space (ranging from 4 to 4096), and by the number of classes in the datasets, which ranges from 2 to 18. Moreover, while the classes in some datasets have a uniform size, in others the classes size are highly heterogeneous. To quantify such an effect, we use the coefficient of variation (CV) defined as the ratio between the standard deviation and the mean of the number of objects per class, i.e.,

$$CV = \frac{\left(\frac{1}{nc-1} \sum_{i=1}^{nc} (\text{card}(c_i) - \mu)^2\right)^{1/2}}{\frac{1}{nc} \sum_{i=1}^{nc} \text{card}(c_i)}, \quad (8)$$

where nc is the total number of classes and $\text{card}(c_i)$ is the number of elements in class c_i . Hence, CV is a measure of the dispersion degree of a random distribution and is a dimensionless number that allows the comparison of populations that have significantly different mean values. In general, a CV of zero means that the dataset is composed of balanced classes, whereas higher values of CV mean that the dataset is composed of a set of classes with a great variability in the number of objects. As can be concluded by analyzing the fifth column of Table 1, the datasets used in this paper have a CV between 0 and 0.811, indicating a high variability in some datasets.

4 Characterization of dissimilarity spaces

To evaluate the quality of the spaces proposed and defined in Sect. 2, a set of geometrical complexity measures proposed by Ho et al. (2006), and Duin and Pekalska (2010) will be used to characterize such spaces regarding metricity and non-Euclideaness, geometry, topology, and class separability. Furthermore, an analysis of time complexity to build each dissimilarity space is performed to evaluate their applicability in time-constrained applications.

² <http://archive.ics.uci.edu/ml>.

Table 1 Summary of the datasets used in the analysis of dissimilarity spaces

	N	p	nc	CV	$Nmin$	$Nmax$	$Navg$
1. 80x	45	8	3	0	15	15	15
2. austra	690	15	2	0.156	307	383	345
3. auto-mpg	398	6	2	0.213	169	229	199
4. biomed	194	5	2	0.437	67	127	97
5. breast	683	9	2	0.424	268	500	341.5
6. derm	366	11	6	0.499	20	112	61
7. diabetes	768	8	2	0.427	268	500	384
8. ecoli	272	7	3	0.519	52	143	90.7
9. german	1000	24	2	0.566	300	700	500
10. heart	297	13	2	0.110	137	160	148.5
11. house-votes	232	16	2	0.098	108	124	116
12. imox	192	8	4	0	48	48	48
13. iris	150	4	3	0	50	50	50
14. kimia	216	4096	18	0	12	12	12
15. liver	345	6	2	0.225	145	200	172.5
16. lvqdata	3655	20	13	0.811	106	956	281.2
17. mfeat-pix	2000	240	10	0	200	200	200
18. sonar	208	60	2	0.095	97	111	104
19. std-yeast	384	17	5	0.440	52	135	76.8
20. wine	178	13	3	0.194	48	71	59.3

N is the number of samples, p the dimension of the feature space, nc the number of classes, and CV the coefficient of variation given by (8). $Nmin = \min_{i=1,\dots,nc} \{\text{card}(c_i)\}$, $Nmax = \max_{i=1,\dots,nc} \{\text{card}(c_i)\}$ and

$Navg = \sum_{i=1}^{nc} \text{card}(c_i)/nc$, with $\text{card}(c_i)$ the number of samples in c_i

4.1 Non-Euclidean and non-metric spaces

A pseudo-Euclidean space is a vector space defined by the Cartesian product of two real spaces: $\mathcal{E} = \mathbb{R}^r \times \mathbb{R}^s$ (Pekalska and Duin 2005). This space is equipped with an inner product defined as $\langle \mathbf{x}, \mathbf{y} \rangle_{\mathcal{E}} = \mathbf{x}^T \mathcal{J}_{rs} \mathbf{y}$, where $\mathcal{J}_{rs} = [I_{r \times r} \ 0; 0 \ -I_{s \times s}]$, and I is the identity matrix. This inner product is positive definite on \mathbb{R}^r and negative definite on \mathbb{R}^s . It is possible to embed a symmetric $n \times n$ dissimilarity matrix D into a $(n-1)$ -dimensional pseudo-Euclidean space, by eigenvalue decomposition. However, that eigenvalue decomposition produces r positive and s negative eigenvalues, such that $r + s = n - 1$, and the corresponding eigenvectors. Accordingly, a measure of the non-Euclidean behavior of the data, the *negative eigenfraction* (NEF) (Duin and Pekalska 2010), is defined as

$$NEF = \frac{\sum_{i=r+1}^{r+s} |\lambda_i|}{\sum_{i=1}^{r+s} |\lambda_i|}, \quad (9)$$

Table 2 Non-metric and non-Euclidean measures for the three dissimilarity spaces considered in this paper: Euclidean (Eucl), shared nearest neighbor (SNN), and dissimilarity increments (Dinc)

	Trineq			NEF		
	Eucl	SNN	Dinc	Eucl	SNN	Dinc
Med	0	0	0.047	4e-16	0.135	0.191
Max	0	0	0.141	2e-15	0.157	0.281
Min	0	0	0.004	0	0.081	0.054
Q1	0	0	0.029	2e-16	0.124	0.148
Q3	0	0	0.073	9e-16	0.141	0.243

Trineq stands for the fraction of triangle inequalities violations, and *NEF* for negative eigenfraction. The values presented correspond to the median (Med), maximum (Max), minimum (Min), first and third quartiles (Q1 and Q3, respectively)

with λ_i the eigenvalues obtained from the eigenvalue decomposition. This measure indicates the amount of non-Euclidean influence in the dissimilarity matrix since it takes values in $[0, 1]$, and $NEF = 0$ indicates Euclidean behavior.

Table 2 presents the results of two measures over the 20 datasets from Sect. 3: one corresponds to the fraction of triangle inequalities violations (*Trineq*), and the other to the negative eigenfraction (*NEF*). The first measure provides information about the metricity of the dissimilarity matrix, while *NEF* tells us if the dissimilarity matrix has an Euclidean behavior or not. Recall that since the Dinc space is asymmetric (see Sect. 2.2.2), in this section, it is made symmetric by averaging with its transpose and normalized by the average off-diagonal dissimilarity to compute the corresponding eigenvalues.

As expected the Euclidean space is metric and has a Euclidean behavior, since *NEF* has a near zero median value (in practice the non-zero values arise due to noise or errors; consequently, the negative eigenvalues may be neglected (Duin and Pekalska 2010)). On the other hand, the two second-order dissimilarity spaces are non-Euclidean, with a median *NEF* value of 0.135 and 0.191 for SNN space and Dinc space, respectively. Thus, the negative eigenvalues in those spaces may be relevant and used to improve learning tasks (Duin and Pekalska 2010). Moreover, analyzing the fraction of triangle inequalities violations, we can conclude that the Dinc space is non-metric, while the SNN space is a metric space.

4.2 Measures of geometrical complexity

To characterize the geometrical complexity of classification problems, Ho et al. (2006) developed some measures based on the analysis of different classifiers, leading to a better understanding of the geometry, topology, and density of manifolds. Accordingly, those measures will be applied to the dissimilarity spaces to understand if the learning process becomes easier in those spaces than in the original feature space. The considered set of metrics is categorized and presented in the following paragraphs.

Measures of overlaps in feature values from different classes focus on how good the features are in separating the classes, by examining the range and spread of values in the dataset within each class and checking for overlaps among different classes.

Two measures are considered: (i) maximum Fisher's discriminant ratio (**F1**), and (ii) collective feature efficiency (**F2**). F1 computes the maximum discriminant power of each feature, with high values indicating that at least one of the features provides valuable information to ease the process of separating the objects among different classes. F2 computes the discriminative power of all the features.

Measures of separability of classes evaluate, based on the existence and shape of the class boundary, to what extent two classes are separable: (i) training error of a linear classifier (**L2**), (ii) ratio of average intra/inter-class nearest neighbor distance (**N2**), (iii) fraction of points on the class boundary (**N1**), and (iv) leave-one-out error rate of the one-nearest neighbor classifier (**N3**). L2 allows evaluating if the classes of the training data are linearly separable, by returning the error rate of such classifier. N2 compares the within class distances to the nearest neighbors of other classes, with higher values indicating that samples of the same class are dispersed. N1 gives an estimate of the length of the class boundary, and high values indicate that the majority of the points lay closely to the class boundary. N3 verifies how close the objects of different classes are, with lower values showing that there is a high gap in the class boundary.

Measures of geometry, topology, and density of manifolds characterize classes, assuming that each class is composed by a single or multiple manifolds, and their shape and position determines how well two classes are separated. Two measures are used to evaluate these characteristics: (i) nonlinearity of a linear classifier (**L1**), and (ii) nonlinearity of the one-nearest neighbor classifier (**N4**). L1 measures, for linearly separable problems, the alignment of the decision surface of linear classifiers with the class boundary. N4 measures the alignment of the nearest neighbor boundary with the shape of the gap or overlap between the convex hulls of the classes.

It should be noticed that some of the presented measures are designed for two-class problems. To overcome this issue, the average across all cases is considered whenever the dataset is composed of more than two classes.

4.2.1 SNN space: influence of the number of neighbors k

To analyze the proposed spaces, we start by studying the influence of the number of nearest neighbors (k) in the SNN space. Hence, several values of k were considered, namely $\{5, 8, 11, 14, 17, 20, 23, 26\}$. For each case, the value of the presented geometrical measures was computed and a Nemenyi test (Demšar 2006) was performed with a significance level of $\alpha = 0.10$, to evaluate whether each k -SNN space is significantly different from another. Figure 2 visually presents the results of the Nemenyi test for each geometrical measure, by considering the mean rank over the datasets for each k -SNN space (lower rank is better). The red line (critical difference) allows determining which points are significantly different, i.e., any point outside the corresponding critical difference area is significantly different from the ones inside the red line (lower rank better, higher rank worse).

When analyzing the results for the measures of overlap (Fig. 2(top)), it can be observed that there was no statistically significant difference inside three different groups of F1 measure of k -SNN space, with $k \in \{5, 8, 11\}$, $\{11, 14, 17\}$, and $\{14, 17, 20, 23, 26\}$. Moreover, in F2 measure, one can notice that $k = 5$ is statistical

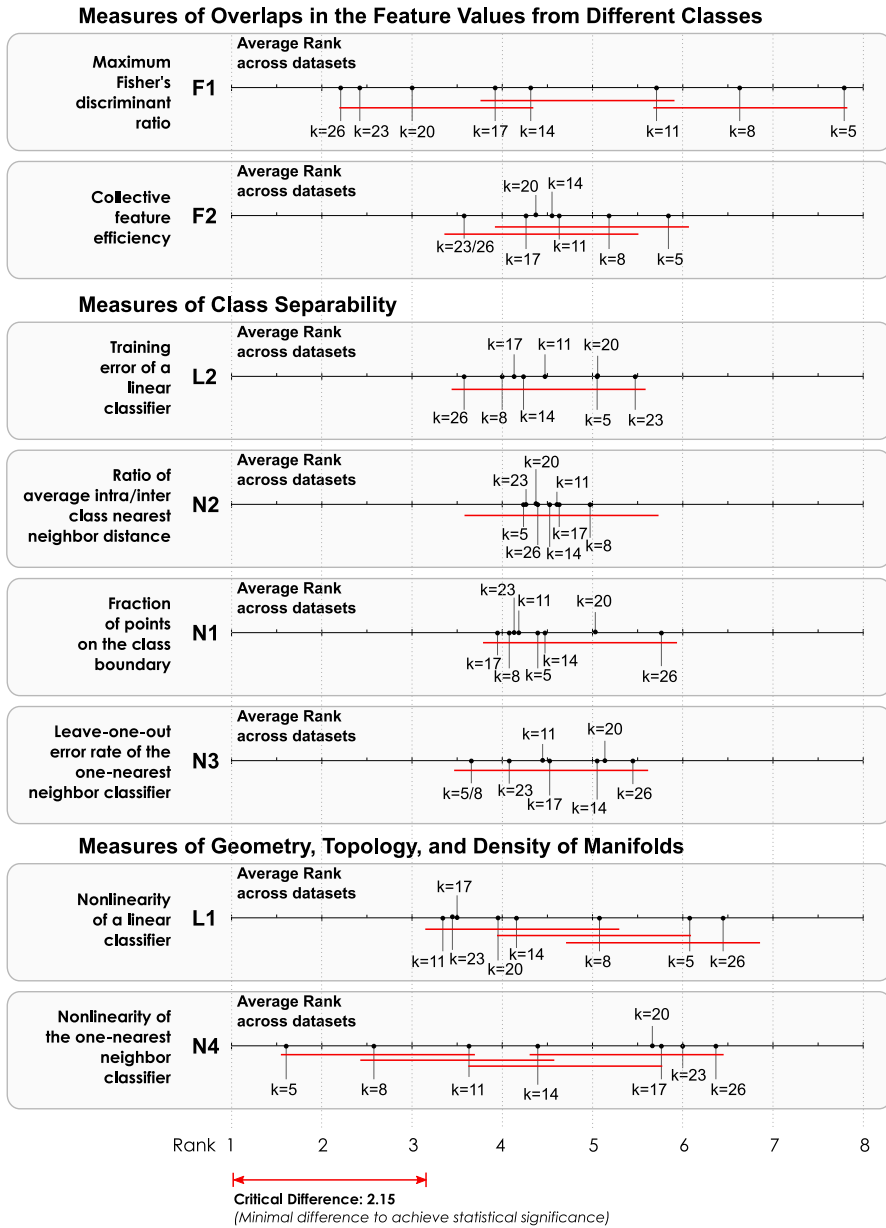


Fig. 2 Comparison of SNN space with different k values ($k \in \{5, 8, 11, 14, 17, 23, 26\}$) with the Nemenyi test. The values are mean ranks over the datasets, and the lower value is better. All spaces with ranks outside the marked interval (red line) are significantly different ($\alpha = 0.10$) from the ones inside the marked interval

different from $k = 23, 26$. In fact, the significant difference occurs between small and large k values, with small k values having the worst ranks, indicating that these spaces have a lower discriminant power of features.

For the measures of class separability (Fig. 2(middle)), one can observe no statistically significant difference between the several spaces.

Finally, when considering the measures of geometry (Fig. 2(bottom)), it can be observed that a significantly better result could only be obtained for the N4 measure with $k = 5, 8, 11$. While for the L1 measure, the best results are achieved with $k \in \{11, 23, 17, 20, 14, 8\}$ with no statistical difference among the spaces.

Since this paper focuses on evaluating the proposed spaces in a more general way, in the remainder of the manuscript, we will assume that the SNN space is built with $k = 20$ to obtain a good balance between all of the referred measures. However, any choice between 14 and 23 will be acceptable to construct the SNN space. Finally, choosing lower values of k (e.g., 5, 8, or 11) will have an impact in the nonlinearity of the one-nearest-neighbor classifier (N4 measure), however, selecting one of such values would result in a poor performance according to several other measures.

4.2.2 Analysis of geometrical complexity measures

Figure 3 presents the results of the geometrical complexity measures described in this section, for the Feature space and the three dissimilarity spaces considered in this paper. Note that the SNN space has higher median value of F1 comparing to the other spaces, meaning that at least one feature has a high discriminant power, which turns the problem of separating the classes easier. Moreover, all three dissimilarity spaces have a high discriminant power of features in separating the classes (higher values of F2). Both Dinc and Eucl spaces have lowest values of F1 and highest values of F2, compared to the SNN space, indicating that a single feature does not have a high discriminant power in separating the classes, but the union of some features together have. Note also that there is a significantly statistical difference between Dinc space and the Feature space for F2.

When considering the measures of class separability, in particular measures L2 and N2, it can be observed that the Feature space have higher values than the dissimilarity spaces, especially the second-order dissimilarity spaces (Dinc and SNN). This means that the training data of the Feature space is not linearly separable and the objects from a class are dispersed, while in the second-order spaces, the classes boundaries become more linear and the classes denser. Moreover, the Euclidean and SNN spaces have higher values of N1 and N3, which means that, when compared to the Feature space, most of the objects lay closer to the class boundary and the gap between classes is lower.

Finally, the Feature space has higher values of L1, *i.e.* the alignment of the decision surface of linear classifiers with the class boundary is difficult. On the other hand, second-order dissimilarity measures have lower values of L1, since the classes are more linearly separable and denser. Furthermore, for L1, it is observed that the Dinc and SNN spaces are statistically different from the Feature space, and SNN is also statistically different from the Euclidean space. Thus, overall the datasets are better described in the second-order dissimilarity spaces, even with the increase of dimensionality on



Fig. 3 Measures to characterize the geometrical complexity of classification problems in the original feature space (Feat), and in dissimilarity spaces, namely Euclidean space (Eucl), SNN space (SNN) and Dinc space (Dinc). Comparison of SNN and Dinc spaces against the others two spaces with Bonferroni-Dunn test. The values are mean ranks over the datasets, and the lower value is better. All spaces with ranks outside the marked interval are significantly different ($\alpha = 0.10$) from the control space, either SNN space (red line) or Dinc space (blue line). The values presented on the right correspond to median values over all datasets, and high values for F1 and F2 is better, while lower values for the remaining measures is better

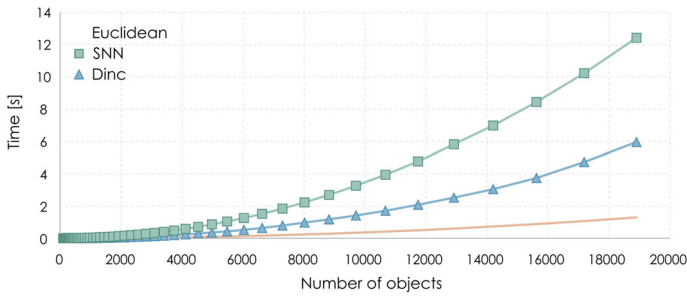


Fig. 4 Time complexity to build each dissimilarity space for varying number of points. The datasets were randomly generated from a 20-dimensional standard normal distribution

those spaces, compared to the Feature space. Finally, there exists less overlap between the classes from the training data of second-order dissimilarity spaces, which may facilitate the learner to separate the samples of different classes.

4.3 Time complexity

A final characterization of the spaces takes into account its application in real problems, and in particular, its impact in the algorithm execution time. Since the proposed dissimilarity representations are based on second-order dissimilarity measures, which are constructed on top of a first-order measure, it is expected for these spaces to be characterized with longer execution times. However, to properly characterize this effect, a simple program was written in C language to build the Euclidean, Dinc and SNN spaces for different datasets, according to the definitions presented in Sect. 2, and compiled using the Intel Compiler 13.0, with auto-parallelizer and -O3 flag enabled. The datasets are composed of normally distributed objects (zero mean vector, covariance equal to the identity matrix) on a 20-dimensional space, with the number of objects (n_z) for dataset z being determined according to $n_z = 100 \times 1.1^z$ ($z = \{0, \dots, 55\}$).

The devised benchmark was run on an Intel Core i7 5960X operating at 3GHz (32GB of RAM memory) running Fedora 21, with the execution times being accurately measured by relying on system functions to access hardware performance counters. The execution time for each dataset is presented in Fig. 4, by considering an average across five runs and 20 neighbors for the SNN space.

As it can be observed, by analyzing Fig. 4, the generation of the Dinc and SNN spaces take more time than the generation of the Euclidean space. Nonetheless, even for the largest dataset, composed of 18905 objects, only 6.0s and 12.4s are required to generate the Dinc and SNN spaces, respectively, which correspond to a time of around 0.32ms and 0.66ms per object on the second-order spaces. Hence, it is not expected that the increase in execution time constrains the application of the proposed dissimilarity spaces on large datasets, especially when considering that they provide substantial quality improvements in several learning tasks, as shown with the presented space characterization and complemented by the set of experimental results in Sects. 5 and 6.

5 Clustering in dissimilarity spaces

Although the previously presented dissimilarity space characterization already shows significant differences between the proposed second-order spaces and the feature and Euclidean spaces, a more in-depth evaluation is provided in this section by evaluating each space when applied to typical hierarchical clustering problems. Hence, we applied seven linkage algorithms to the mentioned datasets (see Sect. 3), namely: unweighted pair group method average or average-link (AL), weighted pair group method average or weighted-link (WeL), unweighted pair group method centroid or centroid-link (CeL), weighted pair group method centroid or median-link (MeL), single-link (SL), complete-link (CL) and Ward's linkage (WL) (Theodoridis and Koutroumbas 2009). Additionally, we set the number of clusters in each clustering algorithm as being equal to the true number of classes (see Table 1).

5.1 Evaluation metrics

The performance of each partition given by a clustering algorithm is assessed through the *consistency index* (CI) (Fred 2001), which is the percentage of correctly classified samples, i.e.,

$$CI(P, P^{gt}) = \frac{1}{N} \sum_{k'=match(k)} |C_k \cap C_{k'}^{gt}|, \quad (10)$$

where $P = \{C_1, \dots, C_K\}$ is the partition given by a clustering algorithm and $P^{gt} = \{C_1^{gt}, \dots, C_{K'}^{gt}\}$ the true labelling.

Moreover, the normalized mutual information (NMI) is used to measure the information shared between P , the partition given by a clustering algorithm and the true labelling, while the adjusted rand index (ARI) measures the similarity, corrected-by-chance, between the two partitions.

5.2 Clustering results for fixed k in SNN space

Figure 5 presents a comparison of the average CI between feature and dissimilarity representations, the number of datasets won by each representation and by each clustering algorithm as well as the average clustering improvement of the spaces, for the cases where the winner leads to a better classification than the remaining representations.

When analyzing the presented results, it can be observed that, for the average-link, the SNN space has the highest average CI (74.2%) and Dinc space the second highest score (68.1%). Also, SNN is better in 50% of the datasets (10 out of 20) with an average improvement over the remaining spaces on those datasets of 17.3%. On the other hand, Dinc space is better in 15% of the datasets and an average improvement on CI over those datasets of 7.6%. In contrast, the original feature space has the lowest average CI (65.1%), is better on 25% of the datasets and has an average improvement on CI on those datasets of 6.9%.

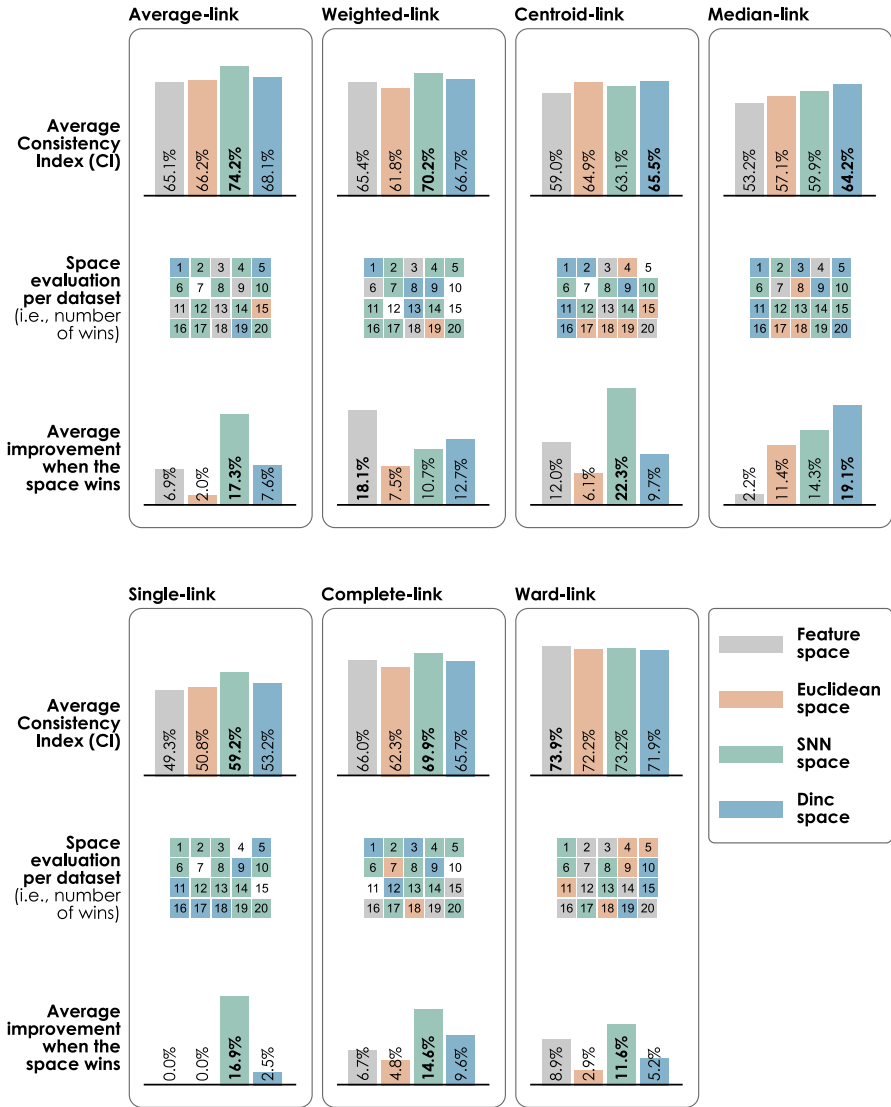


Fig. 5 Comparisons between each dissimilarity representation and the feature space, in seven clustering algorithms. On the top, average consistency index over the 20 datasets. In the middle, the number of wins for each space, each number corresponds to a dataset in Table 1, and the color indicates the best space for a given dataset (white color means ties between two or more spaces), according to the clustering algorithm. On the bottom, the average improvement in consistency index on the datasets won by one space over the remaining spaces (Color figure online)

Similar conclusions can be taken for weighted-link, single-link and complete-link, *i.e.*, SNN has the highest average CI, with 70.2%, 59.2% and 69.9%, respectively. Also, it is better in the majority of the datasets, winning 45% in weighted-link, 55% in single-link, and 45% in complete-link. The only difference in results is in the average improvement on CI on those datasets. For instance, for the weighted-link, the feature space has highest values, leaving the SNN space with the third highest improvement of 10.7%. On the other two algorithms the SNN space has again the highest average improvement on CI over the remaining spaces.

Regarding centroid-link and median-link, the Dinc space attains the highest average CI value, 65.5% and 64.2%, respectively. Also, the Dinc space is better in 25% and 35% of the datasets using centroid-link and weighted-link, respectively, and an average improvement of CI over those datasets of 9.7% and 19.1%. Accordingly, the Euclidean and SNN space are the second best representation when using centroid-link and median-link, respectively. SNN wins 25% and 40% of the datasets with centroid-link and median-link, respectively.

Finally, in ward's linkage, feature space is the one with the highest average CI value (73.9%), winning 35% of the datasets, followed by SNN space with an average CI of 73.2% and winning 25% of the datasets. However, the average improvement over those datasets is higher for SNN space (11.6%) compared to the feature space (8.9%).

Table 3 presents the results for the average NMI and average ARI for each space and each clustering algorithm. As it can be observed, the NMI and ARI scores are highly consistent with those obtained with the CI. In particular, the SNN space has the highest CI, NMI and ARI scores for average-link, weighted-link, single-link and complete-link, while the Dinc space attains the highest CI, NMI and ARI for centroid-link and median-link. The ward's linkage algorithm has the highest CI for the feature space, however the highest NMI and ARI belongs to the SNN space, with the difference in the three metrics for the two spaces being very small. Moreover, the best pair space and clustering algorithm is the SNN space with the average-link, with scores of 74.2%, 0.475 and 0.461 for the CI, NMI and ARI, respectively.

Overall, the proposed second-order spaces have a better average CI, NMI and ARI, and average improvement when the space wins compared to the other spaces for almost all clustering algorithms. Also, the number of datasets won by each space is higher for SNN and Dinc or Euclidean spaces in almost all clustering algorithms. The only exception is for ward's linkage where the feature space presents higher average CI values and has won 35% of the datasets, even though SNN space presents identical but better NMI and ARI values. Finally, the SNN space presents the better results in average-link, weighted-link, single-link and complete-link, while in centroid-link and median-link, the Dinc space shows a better performance.

To complete the above analysis, Table 4 presents, for each dataset, the best CI values in the dissimilarity spaces, the corresponding clustering algorithm, and space. Also, it presents the CI values in the feature space for the same clustering algorithm and the gain of the dissimilarity representation over the feature space. We computed the gain in the CI, for each dataset i , as

$$Gain(i) = \frac{CI^{(y)}(P, P^{gt}) - CI^{Feat}(P, P^{gt})}{CI^{Feat}(P, P^{gt})}, \quad (11)$$

Table 3 Average consistency index (CI), normalized mutual information (NMI) and adjusted rand index (ARI) for each dissimilarity representation and the feature space, in seven hierarchical clustering algorithms

		AL	WeL	CeL	MeL	SL	CL	WL
Feature space	CI	65.1%	65.4%	59.0%	53.2%	49.3%	66.0%	73.9%
	NMI	0.371	0.374	0.301	0.208	0.127	0.370	0.454
	ARI	0.311	0.308	0.226	0.117	0.045	0.302	0.426
Euclidean space	CI	66.2%	61.8%	64.9%	57.1%	50.8%	62.3%	72.2%
	NMI	0.400	0.332	0.387	0.265	0.144	0.327	0.447
	ARI	0.332	0.250	0.307	0.168	0.068	0.241	0.413
SNN space	CI	<u>74.2%</u>	70.2%	63.1%	59.9%	59.2%	69.9%	73.2%
	NMI	<u>0.475</u>	0.462	0.307	0.245	0.298	0.426	0.455
	ARI	<u>0.461</u>	0.409	0.266	0.198	0.231	0.371	0.428
Dinc space	CI	68.1%	66.7%	65.5%	64.2%	53.2%	65.7%	71.9%
	NMI	0.414	0.395	0.394	0.343	0.186	0.367	0.438
	ARI	0.356	0.318	0.319	0.276	0.105	0.290	0.394

Best results for each clustering algorithm are shown in bold and the best overall are highlighted in underline

with (\mathcal{V}) representing one of the dissimilarity spaces (Euclidean, SNN or Dinc), P corresponds to the partition given by one of the clustering algorithms, {AL, WeL, CeL, MeL, SL, CL, WL}, and P^{gt} corresponds to the ground truth partition.

As can be seen, second-order dissimilarity spaces are the best approach overall, with the SNN space winning 11 times, the Dinc space 5 times, the Euclidean space 3 times and the original feature space wins 3 times (as can be seen by the negative gain). In particular, the gain over the feature space can go as high as 100.7% (see dataset 17), since the SNN space with AL has a CI of 96.2%, while in the feature space, it is 47.9% with the same algorithm. Moreover, the best clustering algorithms are AL and WL chosen 6 times each, followed by WeL and MeL chosen 4 times each.

5.3 Automated k -selection for the SNN space

In the previous section the number of neighbors k to construct the SNN space was fixed to 20. However, the value of k can be automatically selected by relying on an internal validation metric (e.g., silhouette or Calinski-Harabasz indexes). Hence, in this section, different SNN spaces, corresponding to $k \in \{5, 8, 11, 14, 17, 20, 23, 26\}$, are constructed and all seven hierarchical clustering algorithms are applied to each space. Afterwards, the value of k is selected according to the silhouette index, as this is one of the most widely used internal metrics.

Table 5 presents the average CI, NMI and ARI for the best SNN space according to the silhouette index. As can be seen, the three metrics have increased their corresponding values when the number of neighbors to construct the SNN space is selected with the silhouette index. For instance, the CI values increased on average 2.5% regarding the SNN space with fixed $k = 20$. Naturally, if the value of k is optimally selected

Table 4 Best algorithm and dissimilarity space according to the consistency index (CI, in %), and CI in the feature space for the same clustering algorithms

Dataset	Algorithm	Best approach in dissimilarity space		Feature Space CI (%)	Gain
		Space	CI (%)		
1. 80x	WL	SNN	86.7%	57.8%	50.0%
2. austra	CL	SNN	62.8%	55.7%	12.8%
3. auto-mpg	MeL	Dinc	81.7%	72.6%	12.5%
4. biomed	AL/WeL	SNN	86.1%	66.0%/69.1%	30.5%/24.6%
5. breast	WL	Eucl	97.5%	96.6%	0.9%
	WeL	SNN	97.5%	68.5%	42.3%
6. derm	AL	SNN	96.4%	80.3%	20.1%
7. diabetes	CL	Eucl	66.0%	65.0%	1.6%
8. ecoli	WeL	Dinc	96.7%	82.0%	17.9%
9. german	CL/CeL/MeL	Dinc	71.3%	70.3%/71.0%/70.5%	1.4%/0.4%/1.1%
10. heart	AL/MeL	SNN	59.6%	53.5%/53.5%	11.3%/11.3%
	WL	Dinc	59.6%	53.9%	10.6%
11. house-votes	AL	SNN	90.5%	91.4%	-0.9%
12. imox	WL	Dinc	85.9%	91.7%	-6.3%
13. iris	SL	SNN	97.3%	68.0%	43.1%
14. kimia	WL	Eucl	90.6%	85.6%	-5.9%
15. liver	MeL	SNN	58.0%	56.5%	2.6%
16. lvsqdata	AL	SNN	83.9%	69.7%	20.3%
17. mfeat-pix	AL	SNN	96.2%	47.9%	100.7%
18. sonar	CL	Eucl	55.8%	51.0%	9.4%
19. std-yeast	WL	Dinc	70.8%	66.9%	5.8%
20. wine	WeL	SNN	71.9%	56.2%	28.0%

The gain (in %) of the best approach in the dissimilarity space over the feature space is computed according to (11)

even better results can be achieved. In particular, the CI values can go as high as 79.1% on average with the weighted-link (an average increase of 3.5%). Similar conclusions can be observed from the other two metrics, NMI and ARI. Overall the best algorithm is still the average-link, however, the ward's link, when k is chosen with the silhouette score, has a CI of 74.3%, which is better than the one obtained by the feature space in Table 3.

5.4 Other dataset types

To better understand the advantages of such second-order spaces in hierarchical clustering algorithms, a new set of datasets are used. These consist in categorical data retrieved from the UCI Machine Learning Repository (car, chess, lymphography and nursery), multivariate time-series data from Baydogan and Runger (2016) (CMU_MOCAP_S16, Japanese Vowels, KickvsPunch and Libras) and graph data from Rossi and Ahmed (2015) (citeseer, DD497 and TerroristRel). From the previous section and as reported in the literature, the average-link and ward's linkage are generally the best linkage algorithms for real-world datasets, usually outperforming the remaining hierarchical clustering algorithms. Hence, only this two algorithms are applied, but over four different spaces:

- the *original feature space*, but using application-specific measures to determine the merge of clusters;
- a *first-order space*, where an application-specific measure is used to obtain the dissimilarity space, and then the Euclidean distance is used to determine the merge of clusters;
- the *SNN and Dinc spaces*, where an application-specific measure is used as first-order dissimilarity, and then the Euclidean distance is used to determine the merge of clusters.

The selected application-specific measures depends on the type of data, namely the Gower distance³ is used for categorical data; the Dynamic Time Warping (DTW) distance (Tavenard et al. 2020) for time-series data; and the SimRank similarity (Jeh and Widom 2002) for graph data. Finally, for SNN space, the best number of neighbors to obtain the space is chosen accordingly to the silhouette index (also from the set {5, 8, 11, 14, 17, 20, 23, 26}), as mentioned in the previous section.

Table 6 presents the results of average CI, NMI and ARI in the three different types of data mentioned above, namely time-series data, categorical data and graph data. As can be seen, SNN space is robust to the type of data used, outperforming the remaining space representations of data. The only exception occurs in the graph data when average-link is applied, where Dinc space as a higher CI value, however, the difference between Dinc and SNN spaces is small. Furthermore, the CI gain of SNN space over the feature representation can go up to 23.1% for the Ward's linkage in categorical data, while in graph data, the gain can only go up to 2.3% for the average-link.

³ From the package Python gower: <https://pypi.org/project/gower/>.

Table 5 Average consistency index (CI), normalized mutual information (NMI) and adjusted rand index (ARI) for the SNN space when $k = 20$, when k is chosen according to the silhouette score (SS), and for the best k for each dataset, in seven hierarchical clustering algorithms

	AL	WeL	CeL	MeL	SL	CL	WL
SNN space (best k)	CI	79.0%	79.1%	67.6%	68.8%	64.2%	76.1%
	NMI	0.497	0.502	0.347	0.376	0.344	0.469
	ARI	0.510	0.497	0.309	0.336	0.271	0.437
SNN space (k from SS)	CI	76.4%	75.1%	64.1%	64.9%	61.1%	71.4%
	NMI	0.480	0.466	0.321	0.329	0.319	0.446
	ARI	0.475	0.442	0.282	0.292	0.250	0.404
SNN space (k fixed at 20)	CI	74.2%	70.2%	63.1%	59.9%	59.2%	69.9%
	NMI	0.475	0.462	0.307	0.245	0.298	0.426
	ARI	0.461	0.409	0.266	0.198	0.231	0.371

Table 6 Average consistency index (CI), normalized mutual information (NMI) and adjusted rand index (ARI) for each dissimilarity representation and the feature space, in average-link (AL) and Ward's linkage (WL)

		Time-series data		Categorical data		Graph data	
		AL	WL	AL	WL	AL	WL
Feature space	CI	56.9%	66.8%	50.1%	46.4%	37.1%	38.2%
	NMI	0.369	0.493	0.095	0.044	0.023	0.116
	ARI	0.231	0.376	0.053	0.020	0.000	-0.005
First-order space	CI	52.1%	62.1%	47.9%	48.8%	37.8%	38.5%
	NMI	0.312	0.411	0.057	0.064	0.043	0.113
	ARI	0.173	0.292	0.035	0.047	-0.005	-0.016
SNN space	CI	69.7%	70.6%	56.8%	57.1%	38.0%	39.0%
	NMI	0.532	0.528	0.145	0.140	0.047	0.072
	ARI	0.396	0.392	0.098	0.094	0.003	0.063
Dinc space	CI	51.5%	62.7%	52.5%	49.3%	38.8%	38.4%
	NMI	0.297	0.416	0.109	0.057	0.040	0.044
	ARI	0.161	0.294	0.073	0.050	0.004	0.004

First-order dissimilarity spaces are obtained by using an application-specific measure, namely Gower distance for categorical data, the dynamic time warping for time-series and the SimRank for graph data. Best results for each clustering algorithm are shown in bold

6 Visualization of dissimilarity spaces

To better understand the advantages of the proposed dissimilarity representations, we apply embedding methods to each dissimilarity representation as well as to the original Feature space. Specifically, methods following different paradigms were applied (Lee and Verleysen 2010): variance preservation (such as principal component analysis—PCA), distance preservation (such as Isomap), topological mapping (such as locally linear embedding—LLE), and similarity preservation (such as t-distributed stochastic neighbor embedding—t-SNE).

The performance of each embedding method for visualizing the data was assessed with the precision and recall. The curves were plotted by fixing the 20 nearest neighbors of a point in the original data as the set of relevant items, and then varying the number of neighbors retrieved from the visualization between 1 and 100, plotting mean precision and recall for each number. The average area under the precision/recall curves (AUC) over the 20 datasets presented in Sect. 3 was used to make pairwise comparisons of spaces in each method. Moreover, trustworthiness and continuity are used to measure visualization quality, since they are a trade-off between precision and recall (Kaski et al. 2003). A visualization is said to be trustworthy when the nearest neighbors of a point in the low-dimensional space are also close in the original space. On the other hand, a visualization on the lower space is said to be continuous when points near to each other in the original space are also nearby in the output space. These two measures are computed using 20 nearest neighbors of a point.

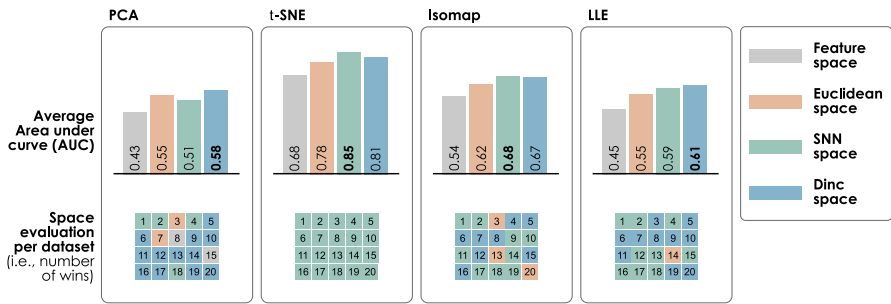


Fig. 6 Comparison between each dissimilarity representation and the feature space, in four embedding algorithms. On the top, average area under the curve (AUC) over the 20 datasets. On the bottom the number of wins for each space, each number corresponds to a dataset in Table 1

The parameters of all embedding methods were chosen such as to maximize the *F*-measure, which is the harmonic mean of the precision and recall. Thus, the parameter space was swept to find the best resulting embedding. Many of the methods have a parameter *k* denoting the number of nearest neighbors for constructing a neighborhood graph; for each method and each dataset we tested values of *k* from the set {5, 9, 13, 17, 21, 25, 29, 33, 37} and chose the value that produced the best *F*-measure. Methods that may have local optima were run five times with different random initializations and the best run, regarding the *F*-measure, was selected.

Figure 6 presents, for each embedding method, a comparison between feature and each dissimilarity representation according to the average AUC over the datasets, as well as the datasets won by each space.

When analyzing the presented results, it can be observed that, for PCA and LLE, Dinc space has the highest average AUC (with 0.58 and 0.61, respectively) and wins more than 50% of the datasets in each embedding algorithm. For the remaining algorithms, SNN space has highest average AUC (with 0.85 for t-SNE and 0.68 for Isomap) and wins more than 45% of datasets, in fact with t-SNE, SNN has the highest AUC for all datasets, i.e., it wins all the datasets. Moreover, the feature space has the lowest average AUC values for all the algorithms compared to any dissimilarity representation considered. In fact, only a maximum of 20% of the datasets are winning by feature and Euclidean space for PCA, and a lower percentage is achieved in the remaining algorithms (where the feature space never gets a higher AUC in any dataset).

To further analyze the visualization, the trustworthiness and continuity measures were computed and the results are presented in Fig. 7. Note that the best performance is located at the top right corner of the plot. As can be seen, t-SNE combined with the SNN or Dinc spaces is the best method for visualizing the data, followed by t-SNE with Euclidean space and Isomap in both second-order dissimilarity spaces (SNN and Dinc). The worst combination of embedding method and data representation is PCA or LLE with the original feature space. This results corroborate the ones in Fig. 6 since the application of PCA and LLE over the original feature space achieves the lowest average AUC of all (with 0.43 and 0.45, respectively). Also, t-SNE combined with second-order spaces has the highest average AUC of all (with 0.85 for SNN space

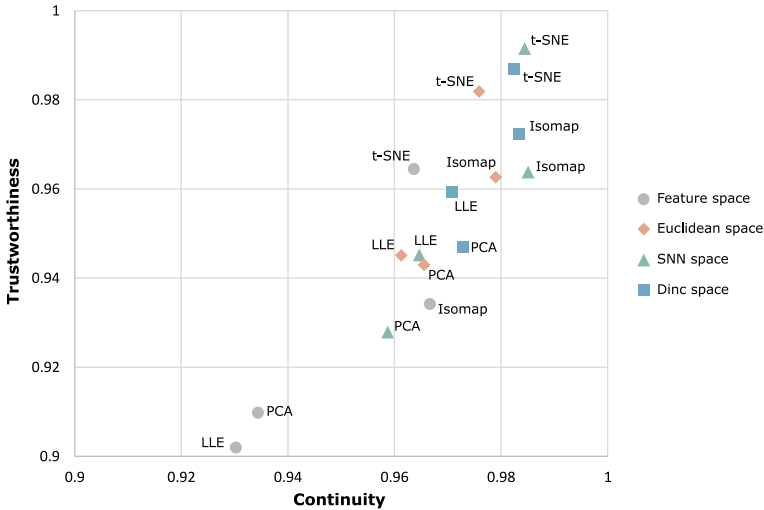


Fig. 7 Average trustworthiness and continuity over the 20 datasets plotted for the feature space and each dissimilarity representation, in the four embedding methods considered. The best performance is in the top right corner

and 0.81 for Dinc space) consistent with the position in the top right corner of the trustworthiness and continuity plot.

To complete the above analysis, Table 7 presents, for each dataset, the best and second best AUC values in the dissimilarity spaces, the corresponding embedding method, and space. Also, it presents the best AUC values in the feature space with the corresponding embedding method and the AUC gain of the dissimilarity representation over the feature space. Similar to the gain in the CI (see (11)), the gain in the AUC, for each dataset i , is given by

$$Gain_{AUC}(i) = \frac{AUC^{(y)}(\xi) - AUC^{Feat}(\zeta)}{AUC^{Feat}(\zeta)}, \tag{12}$$

with (y) is one of the dissimilarity spaces (Euclidean, SNN or Dinc), and ξ and ζ corresponds to the best of the embedding methods considered.

As can be seen, the SNN space is the best dissimilarity representation achieving highest AUC when combined with t-SNE. The only exception occur for 80x, auto-mpg and wine datasets, where Isomap is the best embedding method and the Euclidean space is the best dissimilarity representation for two of those datasets. The gain in AUC over the best algorithm in the feature space can go up to 75.7%. Moreover, the Dinc space is the second best dissimilarity representation also combined with t-SNE. In this case, only kimia dataset is better with the Euclidean space, and 80x, auto-mpg, biomed and wine have better results with the Dinc space and Isomap as embedding method. Also, the gain in AUC over the best algorithm in feature space can go up to 53.2%.

Table 7 Best and second best algorithm and dissimilarity space according to the area under the curve (AUC), and best AUC in the feature space and corresponding embedding method

Dataset	Best alg. in Feat Space			Best approach in dissimilarity space			2nd best approach in dissimilarity space				
	Alg.	AUC	Space	Alg.	AUC	AUC Gain	Alg.	Space	AUC	AUC Gain	
1. 80x	Isomap	0.813	Isomap	Isomap	SNN	0.928	14.2%	Isomap	Dinc	0.905	11.2%
2. austra	t-SNE	0.537	t-SNE	t-SNE	SNN	0.774	44.2%	t-SNE	Dinc	0.662	23.3%
3. auto-mpg	t-SNE	0.926	Isomap	Isomap	Eucl	0.949	2.5%	Isomap	Dinc	0.949	2.5%
4. biomed	Isomap	0.794	t-SNE	t-SNE	SNN	0.909	14.5%	Isomap	Dinc	0.886	11.7%
5. breast	t-SNE	0.551	t-SNE	t-SNE	SNN	0.792	43.6%	t-SNE	Dinc	0.764	38.5%
6. derm	t-SNE	0.699	t-SNE	t-SNE	SNN	0.862	23.2%	t-SNE	Dinc	0.815	16.6%
7. diabetes	t-SNE	0.689	t-SNE	t-SNE	SNN	0.843	22.4%	t-SNE	Dinc	0.833	21.0%
8. ecoli	t-SNE	0.691	t-SNE	t-SNE	SNN	0.848	22.8%	t-SNE	Dinc	0.847	22.6%
9. german	t-SNE	0.662	t-SNE	t-SNE	SNN	0.818	23.5%	t-SNE	Dinc	0.758	14.5%
10. heart	t-SNE	0.674	t-SNE	t-SNE	SNN	0.851	26.3%	t-SNE	Dinc	0.814	20.9%
11. house-votes	t-SNE	0.553	t-SNE	t-SNE	SNN	0.851	53.8%	t-SNE	Dinc	0.835	51.1%
12. imox	t-SNE	0.744	t-SNE	t-SNE	SNN	0.876	17.7%	t-SNE	Dinc	0.840	13.0%
13. iris	t-SNE	0.857	t-SNE	t-SNE	SNN	0.937	9.3%	t-SNE	Dinc	0.928	8.2%
14. kimia	t-SNE	0.777	t-SNE	t-SNE	SNN	0.894	15.1%	t-SNE	Eucl	0.853	9.9%
15. liver	t-SNE	0.629	t-SNE	t-SNE	SNN	0.819	30.3%	t-SNE	Dinc	0.815	29.6%
16. lvsdata	t-SNE	0.350	t-SNE	t-SNE	SNN	0.615	75.7%	t-SNE	Dinc	0.536	53.2%
17. mfeat-pix	t-SNE	0.573	t-SNE	t-SNE	SNN	0.810	41.2%	t-SNE	Dinc	0.671	17.1%
18. sonar	t-SNE	0.621	t-SNE	t-SNE	SNN	0.830	33.5%	t-SNE	Dinc	0.796	28.1%
19. std-yeast	t-SNE	0.589	t-SNE	t-SNE	SNN	0.850	44.3%	t-SNE	Dinc	0.783	32.9%
20. wine	Isomap	0.946	Isomap	Eucl	Eucl	0.950	0.4%	Isomap	Dinc	0.950	0.4%

The AUC gain (in %) of the two best approaches (first and second) in the dissimilarity space over the feature space is computed according to (12).

Overall, second-order dissimilarity spaces are the best approach, and t-SNE is the best associated embedding method. Therefore, we may conclude that, although we are increasing the dimensionality of the space in the second-order dissimilarity representations, we obtain less overlap in the classes, helping the learner to identify classes better.

7 Discussion and conclusions

Novel dissimilarity representations of data based on second-order dissimilarity measures were proposed, leading to a better description of data regarding separability of classes as well as the discriminative power of features. Thus, two measures were used, namely the shared nearest neighbor (SNN) and the dissimilarity increments (Dinc). The first one is built upon the concept of “overlap” between the neighborhoods centered on each pair, while the latter is based on the increment in dissimilarity between an object and a set of representative objects, defined as an edge between a prototype object and its nearest neighbor.

The characterization of each dissimilarity representation of data is performed, by relying on measures of geometrical complexity of classification problems. Those measures give insightful information about geometry, topology, and density of the spaces. Thus, in a second-order dissimilarity representation, the datasets present less overlap between classes and have at least one feature with a high discriminant power. Furthermore, the proposed dissimilarity spaces have a non-Euclidean behavior, but not all are metric. For instance, the SNN space is metric, while the dissimilarity increments space violates the triangle inequality property.

The proposed dissimilarity representations were evaluated in two different scenarios: clustering and visualization. In the first, clustering algorithms were applied to second-order dissimilarity spaces, to the Euclidean space and to the original feature space. For the centroid-link and median-link the Dinc space performs better than any other space considered. In particular, for median-link, it allows improving the average CI in 20.7% and 12.4% regarding feature and Euclidean spaces, respectively, and for centroid-link, it improves the average CI in 11.0% and 0.9%. On the other hand, for average-link, weighted-link, single-link and complete-link the SNN space is the best overall, with an improved 17.4% and 15.4% average CI regarding feature and Euclidean spaces, respectively, when the average-link is applied, and 14.8% and 21.5% when the weighted-link is applied. Moreover, the proposed dissimilarity representations were also applied to other dataset types (time-series, categorical and graph data), showing consistently better results than the feature and first-order spaces.

For the visualization scenario, embedding techniques were applied to the same spaces and the area under the precision/recall curves were obtained. It is noteworthy that the SNN space with t-SNE performed better in all the datasets, while the Dinc space is a better representation for the principal component analysis (PCA) and locally linear embedding (LLE). Also, the application of t-SNE to the Dinc space is the second best visualization technique of all. Overall, results show that the second-order dissimilarity spaces are a better representation of data.

Acknowledgements This work was partially supported by Fundação para a Ciência e a Tecnologia (FCT) through project AIpALS, ref. PTDC/CCI-CIF/4613/2020, and the LASIGE Research Unit, ref. UIDB/00408/2020 and ref. UIDP/00408/2020.

References

- Aidos H, Fred A (2012) Statistical modeling of dissimilarity increments for d -dimensional data: application in partitional clustering. *Pattern Recogn* 45(9):3061–3071
- Aidos H, Fred A (2015a) Consensus of clusterings based on high-order dissimilarities. In: *Partitional clustering algorithms*, pp 311–349. Springer
- Aidos H, Fred A (2015b) A novel data representation based on dissimilarity increments. In: *Proceedings international workshop of similarity-based pattern recognition (SIMBAD)*, pp 1–14
- Aidos H, Fred A, Duin R (2012) Classification using high order dissimilarities in non-euclidean spaces. In: *Proceedings of the international conference on pattern recognition applications and methods (ICPRAM)*, pp 306–309
- Batista D, Aidos H, Fred A, Santos J, Ferreira RC, das Neves RC (2018) Protecting the ECG signal in cloud-based user identification system: a dissimilarity representation approach. In: *Proceedings of the international joint conference on biomedical engineering systems and technologies (BIOSTEC) vol 4*, pp 78–86
- Batista L, Granger E, Sabourin R (2010) Applying dissimilarity representation to off-line signature verification. In: *International conference on pattern recognition (ICPR)*, pp 1433–1436
- Baydogan MG, Runger G (2016) Time series representation and similarity based on local autopatterns. *Data Min Knowl Disc* 30(2):476–509
- Bicego M (2005) Odor classification using similarity-based representation. *Sens Actuat B Chem* 110(2):225–230
- Calvo-Zaragoza J, Valero-Mas JJ, Rico-Juan JR (2016) Prototype generation on structural data using dissimilarity space representation. *Neural Comput Appl*, pp 1–10
- Cao H, Bernard S, Heutte L, Sabourin R (2018) Dissimilarity-based representation for radiomics applications. *arXiv preprint arXiv:1803.04460*
- Chen Y, Garcia EK, Gupta MR, Rahimi A, Cazzanti L (2009) Similarity-based classification: concepts and algorithms. *J Mach Learn Res* 10:747–776
- Cheplygina V, Tax DMJ, Loog M (2016) Dissimilarity-based ensembles for multiple instance learning. *IEEE Trans Neural Netw Learn Syst* 27(6):1379–1391
- De Santis E, Martino A, Rizzi A, Mascioli FMF (2018) Dissimilarity space representations and automatic feature selection for protein function prediction. In: *2018 International joint conference on neural networks (IJCNN)*, pp 1–8
- Demšar J (2006) Statistical comparisons of classifiers over multiple data sets. *J Mach Learn Res* 7:1–30
- Duda RO, Hart PE, Stork DG (2001) *Pattern classification*, 2nd edn. Wiley
- Duin R, Pekalska E (2010) Non-Euclidean dissimilarities: causes and informativeness. In: *Proceedings joint IAPR international workshop (SSPR/SPR) structural, syntactic, and statistical pattern recognition*, pp 324–333
- Duin RPW, Bicego M, Orozco-Alzate M, Kim S-W, Loog M (2014) Metric learning in dissimilarity space for improved nearest neighbor performance. In: *Structural, syntactic, and statistical pattern recognition—proceedings joint IAPR international workshops (SSPR/SPR)*
- Ertöz L, Steinbach M, Kumar V (2003) Finding clusters of different size, shape, and densities in noisy high dimensional data. In: *Proceedings of the SIAM international conference on data mining (SDM)*, pp 47–58
- Eskander GS, Sabourin R, Granger E (2013) Dissimilarity representation for handwritten signature verification. In: *Proceedings of the international workshop on automated forensic handwriting analysis: a satellite workshop of international conference on document analysis and recognition (AFHA)*, pp 26–30
- Fred A (2001) Finding consistent clusters in data partitions. In: *Proceedings international workshop multiple classifier systems (MCS)*, pp 309–318
- Fred A, Leitão J (2003) A new cluster isolation criterion based on dissimilarity increments. *IEEE Trans Pattern Anal Mach Intell* 25(8):944–958

- García S, Derrac J, Cano JR, Herrera F (2012) Prototype selection for nearest neighbor classification: taxonomy and empirical study. *IEEE Trans Pattern Anal Mach Intell* 34(3):417–435
- Ho TK, Basu M, Law MHC (2006) Measures of geometrical complexity in classification problems. In: *Data complexity in pattern recognition*, pp 3–23. Springer
- Jain AK, Duin RPW, Mao J (2000) Statistical pattern recognition: a review. *IEEE Trans Pattern Anal Mach Intell* 22(1):4–37
- Jarvis RA, Patrick EA (1973) Clustering using a similarity measure based on shared near neighbors. *IEEE Trans Comput* 22(11):1025–1034
- Jeh G, Widom J (2002) Simrank: a measure of structural-context similarity. In: *Proceedings of the eighth ACM SIGKDD international conference on knowledge discovery and data mining*, pp 538–543
- Jin W, Tung AKH, Han J, Wang W (2006) Ranking outliers using symmetric neighborhood relationship. In: *Advances in knowledge discovery and data mining, Pacific-Asia conference (PAKDD)*, pp 577–593
- Kaski S, Nikkilä J, Ojo M, Venna J, Törönen P, Castrén E (2003) Trustworthiness and metrics in visualizing similarity of gene expression. *BMC Bioinform* 4(1):48
- Lee JA, Verleysen M (2010) Unsupervised dimensionality reduction: overview and recent advances. In: *Proceedings of the international joint conference on neural networks (IJCNN)*, pp 1–8
- Liao L, Noble WS (2003) Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *J Comput Biol* 10(6):857–868
- Livi L (2017) Designing labeled graph classifiers by exploiting the rényi entropy of the dissimilarity representation. *Entropy* 19(5):216–241
- Marques F, Carreiras C, Lourenço A, Fred A, Ferreira R (2015) ECG biometrics using a dissimilarity space representation. In: *Proceedings of the international conference on bio-inspired systems and signal processing (BIOSIGNALS)*, pp 350–359
- Martins JG, Oliveira LS, Britto AS Jr, Sabourin R (2015) Forest species recognition based on dynamic classifier selection and dissimilarity feature vector representation. *Mach Vis Appl* 26(2):279–293
- Moreno JG, Dias G, Cleuziou G (2013) Post-retrieval clustering using third-order similarity measures. In: *Proceedings of the annual meeting of the association for computational linguistics (ACL)*, pp 153–158
- Orozco-Alzate M, Duin R, Castellanos-Domínguez G (2009) A generalization of dissimilarity representations using feature lines and feature planes. *Pattern Recogn* 30(3):242–254
- Orozco-Alzate M, Castro-Cabrera PA, Bicego M, Londoño-Bonilla JM (2015) The DTW-based representation space for seismic pattern classification. *Comput Geosci*
- Pekalska E, Duin RPW (2002) Dissimilarity representations allow for building good classifiers. *Pattern Recogn Lett* 23:943–956
- Pekalska E, Duin RPW (2005) *The dissimilarity representation for pattern recognition: foundations and applications*. World Scientific Pub Co Inc
- Pelillo M (ed) (2013) *Similarity-based pattern analysis and recognition*. Springer
- Pinheiro RHW, Cavalcanti GDC, Tsang IR (2017) Combining dissimilarity spaces for text categorization. *Inf Sci* 406–407:87–101
- Plasencia-Calaña Y, Cheplygina V, Duin RPW, García-Reyes E, Orozco-Alzate M, Tax DMJ, Loog M (2013) On the informativeness of asymmetric dissimilarities. In: *Similarity-based pattern recognition - proceedings international workshop (SIMBAD)*, pp 75–89
- Pohle T, Knees P, Schedl M, Widmer G (2006) Automatically adapting the structure of audio similarity spaces. In: *Proceedings of the workshop on learning the semantics of audio signals (LSAS)*, pp 66–75
- Rossi RA, Ahmed NK (2015) The network data repository with interactive graph analytics and visualization. In: *AAAI*. <http://networkrepository.com>
- Satta R, Fumera G, Roli F (2012) Fast person re-identification based on dissimilarity representations. *Pattern Recogn Lett* 33:1838–1848
- Schleif F-M, Zhu X, Hammer B (2012) A conformal classifier for dissimilarity data. *AIAB, AIEIA, CISE, COPA, IIVC, ISQL, MHDW, and WADTMB*. In: *Artificial intelligence applications and innovations - AIAI international workshops*, pp 234–243
- Tavenard R, Faouzi J, Vandewiele G, Divo F, Androz G, Holtz C, Payne M, Yurchak R, Rußwurm M, Kolar K, Woods E (2020) Tslern, a machine learning toolkit for time series data. *J Mach Learn Res* 21(118):1–6
- Theodorakopoulos I, Kastaniotis D, Economou G, Fotopoulos S (2014) Pose-based human recognition via sparse representation in dissimilarity space. *J Vis Commun Image Represent* 25(1):12–23
- Theodoridis S, Koutroumbas K (2009) *Pattern recognition*, 4th edn. Elsevier Academic Press

Ulas A, Duin RPW, Castellani U, Loog M, Mirtuono P, Bicego M, Murino V, Bellani M, Cerruti S, Tansella M, Brambilla P (2011) Dissimilarity-based detection of schizophrenia. *Int J Imaging Syst Technol* 21(2):179–192

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.