PeerJ Computer Science

On component-wise dissimilarity measures and metric properties in pattern recognition

Enrico De Santis[1], Alessio Martino[2] and Antonello Rizzi[1]

[1] Department of Information Engineering, Electronics and Telecommunications, University of Roma “La Sapienza”, Rome, Italy
[2] Department of Business and Management, LUISS University, Rome, Italy

ABSTRACT

In many real-world applications concerning pattern recognition techniques, it is of utmost importance the automatic learning of the most appropriate dissimilarity measure to be used in object comparison. Real-world objects are often complex entities and need a specific representation grounded on a composition of different heterogeneous features, leading to a non-metric starting space where Machine Learning algorithms operate. However, in the so-called unconventional spaces a family of dissimilarity measures can be still exploited, that is, the set of component-wise dissimilarity measures, in which each component is treated with a specific sub-dissimilarity that depends on the nature of the data at hand. These dissimilarities are likely to be non-Euclidean, hence the underlying dissimilarity matrix is not isometrically embeddable in a standard Euclidean space because it may not be structurally rich enough. On the other hand, in many metric learning problems, a component-wise dissimilarity measure can be defined as a weighted linear convex combination and weights can be suitably learned. This article, after introducing some hints on the relation between distances and the metric learning paradigm, provides a discussion along with some experiments on how weights, intended as mathematical operators, interact with the Euclidean behavior of dissimilarity matrices.

Subjects Algorithms and Analysis of Algorithms, Artificial Intelligence, Data Mining and Machine Learning, Data Science, Theory and Formal Methods
Keywords Metric learning, Pattern recognition, Dissimilarity space, Euclidean embedding, Kernel methods, Pseudo-Euclidean embedding

Submitted 6 June 2022
Accepted 26 August 2022
Published 10 October 2022

Corresponding author
Alessio Martino, amartino@luiss.it

Academic editor
Muhammad Aleem

Additional Information and Declarations can be found on page 22

DOI **10.7717/peerj-cs.1106**

INTRODUCTION

In the past few decades, the discipline of pattern recognition (PR), aiming to automatically discover regularities in data, focused most efforts in frameworks conceived to learn from examples, thus from observations. These frameworks exploit several machine learning techniques grounding on the data-driven approach (*Bishop, 2006*). Therefore, in these specific cases, the goal of a PR system is to find regularities in data aiming to reach good *generalization* capabilities by building a model from known observations (*Hart, Stork & Duda, 2000*). Thereby, at the basis of an automated PR pipeline there are the observations, that can be any type of measurements on real-world objects. Observations can be collected by hand or automatically by sensors. Furthermore, observations can be labeled and labels allow to distinguish the class or the category in which the object falls (*Martino, Giuliani &*

How to cite this article De Santis E, Martino A, Rizzi A. 2022. On component-wise dissimilarity measures and metric properties in pattern recognition. **PeerJ Comput. Sci. 8:e1106 DOI 10.7717/peerj-cs.1106**

Rizzi, 2018). Moving away from the "philosophical" problem among the differences between objects that live in the real world—a discussion that deserves a systematic and really interesting discussion—it can be stated that it is really difficult to enumerate all differences between two real-world objects, at least, at raw (atomic) level. So what we can reveal are the differences between the (physical or virtual) properties of two objects and tell whether they can be considered different. This task takes part to the PR process and deserves a thorough discussion. In the PR jargon, the problem is known as finding a good *representation* of objects, *e.g.*, if the weight is important as a property defining the objects, it should be taken into account, otherwise the system should not consider the "weight" feature. A really interesting theoretical treatment, within the context of cognitive science, on how *natural properties* arise from generic objects in building a suitable representation is provided by Peter Gärdenfors with the theory of conceptual spaces (Gärdenfors, 2004).

A *representation* can exist in several forms, such as numbers, strings, graphs, images, spectra, time series, densities and similarities (Duin & Pękalska, 2012). Robert P. W. Duin states that (Pękalska & Duin, 2012): (i) "every real-world difference between objects that may play a role in the *human judgment* of their similarity should make a difference in the representation" and ii) "the representation of a real world object, *i.e.*, the mapping from the object to its representation, should be continuous". Hence, these prescriptions indicate that the representation should consider real-world properties judged as important and, furthermore, two similar objects should be similar in their representations too. On the top of a good representation, it is possible to train a myriad of learning algorithms capable to generate a model from data objects and, finally, to generalize towards previously unseen data. In fully supervised learning, the generalization process (classification) needs labeled examples, while unlabeled examples are used in unsupervised learning schemes (Jain, Murty & Flynn, 1999; Jain, Duin & Mao, 2000). As we will see, alongside the classical learning algorithms adopted in machine learning, it can be useful to learn a dissimilarity function tailored to the data at hand. This particular task belongs to the metric learning (ML) paradigm, a florid research field in PR (Lu et al., 2018; Bengio, Courville & Vincent, 2013).

As anticipated, many real-world objects in PR cannot be simply described by a set of measurements collected in real-valued vectors. In other words, the representation of objects may not easily start from a vectorial space and in this case the dissimilarity measure cannot be simply defined as a plain Minkowski distance, for example. In this case, a data structure, known as *dissimilarity matrix*, becomes clearly important. Thereby, in many cases, the core of a PR system is a custom-based dissimilarity measure, that is a way to measure the dissimilarity between samples of a given complex process that are described by a set of measurements that can (even simultaneously) involve real numbers, integers, vectors, categorical variables, graphs, spectra, histograms, unevenly objects/events sequences, time series *etc.* This happens when real-world objects possess a complex description arising from different intrinsic characteristics, each one caught by a suitable data structure. Thence, the overall dissimilarity can be chosen within the family of Euclidean distances, or within the general class of Minkowski distances. However, the structure of the given distance needs to take into account the different data structures. In technical literature, distances involving complex and possibly heterogeneous data structures are known as *component-wise* or *element-wise* distances (*Jimenez, Gonzalez & Gelbukh, 2016*) where, for each component, it is used a specific difference operator for the data structure at hand and, once collected, all of them are synthesized in a "template" distance that may have the Euclidean or Minkowski general form. In other words, distances are a function of the additive combination of the contributions of their components (*Beals, Krantz & Tversky, 1968*). A further generalization can be derived from the weighted Euclidean distance (WED) where a weight is associated to each component. In ML tasks, these weights can be suitably learned automatically, usually through an optimization procedure.

The WED is widely applied in PR problems such as in bioinformatics and personalized medicine (*Hu & Yan, 2007*; *Martino et al., 2020*; *Di Noia et al., 2020*), speech synthesis (*Lei, Ling & Dai, 2010*) or in the industrial field (*Rao, 2012*). For example, WED is used in clustering application dealing with side information (*Xing et al., 2002*). In fact, if a clustering algorithm, such as k-means, initially fails to find a meaningful solution for the problem at hand from the user point of view, the user is forced to manually tweak the metric until sufficiently good clusters are found. In *Schultz & Joachims (2003)*, the authors present a method for learning a distance metric starting from relative comparison such as "A is closer to B than A is to C". A similar application can be found in (*Kumar & Kummamuru, 2008*), where a local metric is learned.

Moreover, in many real-world problems dealing with complex systems, the starting space is not a vectorial space, being also often non-metric (*e.g.*, in life sciences (*Münch et al., 2020*; *Martino, Giuliani & Rizzi, 2018*), engineering applications (*D'urso & Massari, 2019*; *De Santis, Arnò & Rizzi, 2022*; *Kim, Lee & Kim, 2018*) or cybersecurity (*Granato et al., 2020*, *2022*)). Consequently, only the dissimilarity representation is available through the dissimilarity matrix, as stated above. Hence, in such cases, the dissimilarity matrix is a primitive data structure compared to the data matrix. As we will see in the following, a dissimilarity matrix $\mathscr{D}$ is said to be "Euclidean" if it is perfectly (isometrically) embeddable in an Euclidean vector space in which the distances calculated in the latter are identical to the ones belonging to the entries of $\mathscr{D}$ (*De Santis, Rizzi & Sadeghian, 2018*). Several standard classifiers are designed to work effectively on Euclidean vector spaces. Operating with a non-Euclidean (or even non-metric) dissimilarity matrix may cause some problems. As an example, a non-Euclidean distance matrix leads to a non-positive definite kernel and the quadratic optimization procedure used to train a support vector machine (*Vapnik, 1998*; *Schölkopf, Burges & Smola, 1999*) may thereby fail, not being fulfilled the Mercer conditions (*Mercer, 1909*; *Duin, Pękalska & Loog, 2013*; *Pękalska & Duin, 2005*). However, in order to train standard classifiers on this kind of data, some solutions can be found. The two main solutions are based either on considering the dissimilarity matrix as the starting vector space endowed with the standard Euclidean distance (dissimilarity space representation) or by adopting a suitable transformation of the dissimilarity matrix, leading to the Pseudo-Euclidean (PE) space (*Pękalska & Duin, 2005*; *De Santis et al., 2018*; *De Santis, Rizzi & Sadeghian, 2018*). In the current study, we will consider the last case. For the first case, the interested reader can be referred to *Pękalska & Duin (2005)*.

It is well known that from dissimilarity data collected in form of a dissimilarity matrix $\mathscr{D}$ it can be "reconstructed" the starting Euclidean space where the original data points lie (*De Santis, Rizzi & Sadeghian, 2018*). The reconstruction process (known as *embedding*) tries to generate the original vector space such that the distances are preserved as well as possible. Classical multi-dimensional scaling is an example of such embedding procedure (*Borg & Groenen, 2005*). For an Euclidean space all the distances are preserved and thus an Euclidean distance matrix can be embedded isometrically in an Euclidean space. For non-Euclidean distance matrices the Euclidean space is not "large enough" to embed the dissimilarity data even if they can be still embedded in the so called PE space (*Goldfarb, 1984*). The embedding procedure involves the eigendecomposition of the kernel matrix $\mathbf{G} = \mathbf{X}\mathbf{X}^T$, where $\mathbf{X}$ is the configuration matrix with data points organized as rows, also known as the Gram matrix (*Horn & Johnson, 2013*). The latter is a similarity matrix, obtainable through a suitable linear transformation of the dissimilarity matrix $\mathscr{D}$.

In this work, we consider a class of PR problems involving a dissimilarity matrix $\mathscr{D}$ deriving from a custom-based component-wise dissimilarity measure $d(x, y; \mathbf{w}) : \mathscr{F} \times \mathscr{F} \rightarrow \mathbb{R}^+$. The following study is based on the characterization of d as a composite dissimilarity matrix of the form: $\sqrt{\bar{\mathbf{d}}_c^T \mathbf{W}^T \mathbf{W} \bar{\mathbf{d}}_c}$ computed as the ℓ_2 norm of the vector $\bar{\mathbf{d}}_c$ that collects the component wise (sub)-dissimilarity measures, of which the functional form is related to the specific features (*i.e.*, data structure) within a suitable structured non-metric feature space $\mathscr{F}$.

Within this framework, in this article we provide two characterizations. The first one tries facing the claim according to which the behavior of a general dissimilarity measure depends on the behavior of the component-wise (sub)-dissimilarities. Specifically, d generates an Euclidean dissimilarity matrix if the (sub)-dissimilarities $d_{\mathscr{F}_j}$ are Euclidean. Therefore, the features $\mathscr{F}_j$ over which it is induced a particular dissimilarity measure, *i.e.*, a structural dissimilarity in the sense of *Duin & Pękalska (2010)*, can influence the nature of the mathematical space where the learning algorithm works.

As concerns the second characterization, it is really interesting to arrange a mathematical interpretation of the weights pertaining the custom-based dissimilarity matrix, in particular wondering what is the influence of a weighting matrix $\mathbf{W}$ on the eigenspectrum of the underlying Gram matrix $\mathbf{G}_w$, that is the Gram matrix obtained from the weighted version of the dissimilarity matrix $\mathscr{D}$. Unfortunately the relationship between the eigenvalues of $\mathbf{G}$ and $\mathbf{G}_w$ in a general case is not straightforward, being an open problem of mathematics (*Zhang & Zhang, 2006*; *Fulton, 2000*). It is approachable in particular cases of commuting matrices (in the case of matrices sharing a complete set of eigenvectors, *i.e.*, *normal* matrices) or when one of the two is a scalar matrix, *i.e.*, a matrix of the form $\mathbf{W} = k\mathbf{I}$. We will trace some results in the latter case.

Although this article aims at addressing these characterizations *via* a theoretical and mathematical viewpoint, the interested reader can find practical applications in the following articles. In *De Santis et al. (2015)*, *De Santis, Rizzi & Sadeghian (2018)* a One-Class classification approach is used in the field of predictive maintenance and in the real-time recognition of faults in a real-world power grid, by processing heterogeneous information coming from smart sensors related to the power grid equipment and to the surrounding environment. The system exploits a clustering-genetic algorithm (GA) (*Goldberg, 1989*) approach where the weights of a custom based Euclidean dissimilarity measure are learned solving a suitable optimization problem. In *De Santis et al. (2018)*, we addressed the problem of finding suitable representative elements in the dissimilarity space[1] in order to classify protein contact networks according to their enzymatic properties and in *De Santis et al. (2022)*, the dissimilarity space embedding has been used to recognize signals pertaining to malfunctioning states of pressurization systems for high-speed railway trains. Finally, in *Martino et al. (2020)* the same problem of classifying protein contact networks according to their enzymatic properties has been solved by an hybridization of dissimilarity spaces and multiple kernel learning.

The degree of "non-metricity" and even of non-Euclidean behavior can be measured suitably with specific indexes obtained from the PE embedding such as the *Eigen-Ratio*, the *Negative Eigen-Fraction* and the *Non-Metricity Fraction*, each of which measures the non-Euclidean behavior, *e.g*., of a given dissimilarity matrix (*Pękalska et al., 2006*). Therefore, while the second question concerns the relation between $\mathbf{G}$ and $\mathbf{G}_w$, in the first characterization we are wondering what is the influence of dissimilarity weights on the Negative Eigen-Fraction, hence we are questioning on how it is possible to tune the non-Euclidean behavior of a custom-based dissimilarity matrix.

The article is organized as follows. In "Metric Learning" it is provided a brief review of the various ML paradigms treated in the literature. "On Metric Spaces and Dissimilarity Matrices" is a concise description of metric spaces and related dissimilarity matrices that serves as background. "The Weighted Euclidean Distance" is a deepening of the Euclidean distance structure and its weighted component-wise counterpart. "Characterization of a Composite Component-wise Dissimilarity" and "On the Presence of Weights in a Component-wise Dissimilarity and the Eigenspectrum of the Gram Matrix" sketch an experimental evaluation of the proposed principal investigations and, finally, "Conclusion" concludes the article.

METRIC LEARNING

The ML problem is concerned with learning a distance function tuned to a particular task and has been shown to be useful when exploited in conjunction with techniques relying explicitly on distances or dissimilarities, such as clustering algorithms, nearest-neighbor classifiers, *etc*. For example, if the task is to asses the similarity (or dissimilarity) between two images with the aim of finding a match, *e.g*., in face recognition, we would discover a proper distance function that emphasizes appropriate features (hair color, ratios of distances between facial key-points, *etc*.). Although this task can be performed by hand, it is very useful to develop tools for learning automatically the subset of meaningful features for the problem at hand. In fact, as anticipated in "Introduction", useful representations can be also learned. However, it is unquestionable that, at least on a theoretical level, representation learning must be taken separate from classification tasks as depicted in Fig. 1 and discussed in *Bellet, Habrard & Sebban (2013)*.

[1] See *Pękalska, Duin & Paclík (2006)* for a discussion on the subject matter.

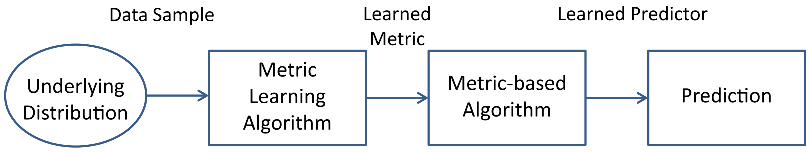

Figure 1 Scheme of the common process in Metric Learning. A metric is learned from data comingfrom a suitable distribution and plugged into a predictor (*e.g.*, a classifier, a regressor, a recommendersystem, *etc.*). The predictor fed with the learned metric hopefully performs better than a predictorinduced by a standard, non-learned, metric (*Bellet, Habrard & Sebban, 2013*).
Full-size DOI: 10.7717/peerj-cs.1106/fig-1

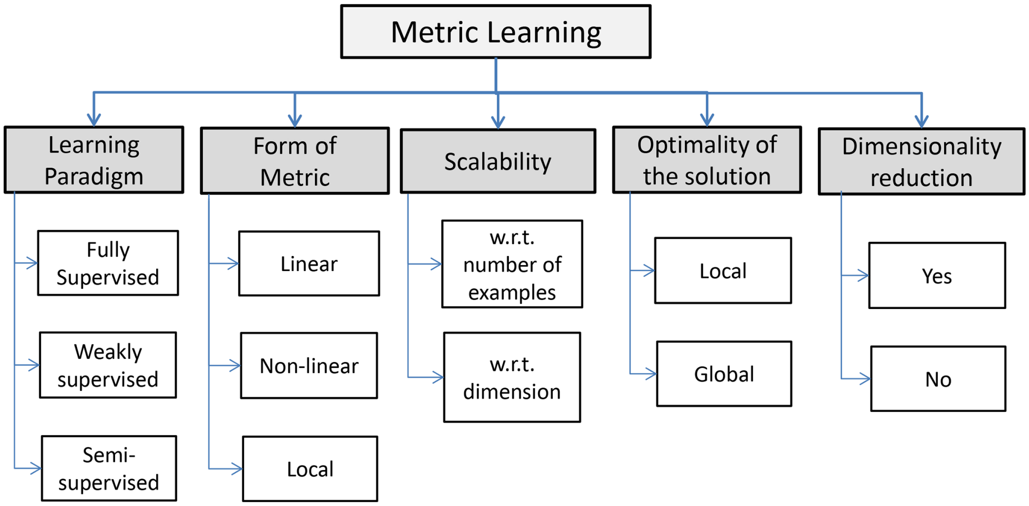

Figure 2 Five key properties of ML algorithms (*Bellet, Habrard & Sebban, 2013*).
Full-size DOI: 10.7717/peerj-cs.1106/fig-2

The ML step can be conceived as a first step in the open-loop pipeline depicted in Fig. 1, to be performed before the model synthesis stage. Moreover, both tasks can be done together in the same system, representing an *advanced* closed-loop and automatic PR system. It is the case, for example, of feature selection and feature extraction techniques. The last procedures can be done manually, but if they are automated (*i.e.*, optimized) they can be a building block of the classification system itself. There are many methodologies capable to learn a representation; some authors distinguish between neural learning, that is learning by means of deep learning techniques, and ML. Despite this distinction, in general, both approaches ground the learning procedure on optimization techniques. Neural learning is useful in finding a good feature space, while ML involves the learning of suitable manifold where data objects lie and where they can be well represented for solving the problem at hand.

Many declinations of ML are available and, according to Fig. 2, they can be resumed in three principal paradigms: *fully supervised*, *weakly supervised* and *semi supervised*. An informal formulation of the supervised ML task is as follows: given an input distance function $d(\mathbf{x},\mathbf{y})$ between objects $\mathbf{x}$ and $\mathbf{y}$ (for example, the Euclidean distance), along with supervised information regarding an ideal distance, construct a new distance function $\hat{d}(\mathbf{x},\mathbf{y})$ which is "better" than the original distance function (*Kulis, 2012*). Normally, fully supervised paradigms have access to a set of labeled training instances, whose labels are used to generate a set of constraints. In other words supervised ML is cast into pairwise constraints: the equivalence constraints where pairs of data points belong to the same classes, and inequivalence constraints where pairs of data points belong to different classes (*Bar-Hillel et al., 2003*; *Xing et al., 2002*). In weakly supervised learning algorithms we do not have to access to the label of individual training examples and learning constraints are given in a different form as side information, while semi-supervised paradigms do not use either labeled samples or side information. Some authors (*e.g.*, *Saul & Roweis (2003)*) deal with unsupervised ML paradigms, sometimes called also manifold learning, referring to the idea of learning an underlying low-dimensional manifold[2] where geometric relationships (*e.g.*, the distance) between most of the observed data are preserved. Often this paradigm coincides with the *dimensionality reduction* paradigm such as the well-known Principal Component Analysis (PCA) (*Shlens, 2014*; *Giuliani, 2017*) and the Classical Multi-Dimensional Scaling, based on linear relations. As concerns non-linear counterparts, it is worth taking note of embedding methods such as ISOMAP (*Tenenbaum, De Silva & Langford, 2000*), Locally Linear Embedding (*Roweis & Saul, 2000*) and Laplacian Eigenmap (*Belkin & Niyogi, 2003*). Other methods are based on information-theoretic relations such as the Mutual Information. Hence, the form or structure of the learned metric can be *linear*, *non-linear*, *local*. Linear ML paradigms are based on the learning of a metric in the form of a generalized Mahalanobis distance (*Mahalanobis, 1936*) between data objects, *i.e.*, $\mathscr{D}_{ij}^{\mathbf{W}} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W}^T \mathbf{W}(\mathbf{x}_i - \mathbf{x}_j)} = \sqrt{\|\mathbf{W}(\mathbf{x}_i - \mathbf{x}_j)\|_2}$, where $\mathbf{M} = \mathbf{W}^T\mathbf{W}$ is a matrix with suitable properties that has to be learned. In other words, the learning algorithm learns a linear transformation $\mathbf{x} \rightarrow \mathbf{W}\mathbf{x}$ that better represents the similarity in the target domain. Sometimes, there are some non-linear structures in the available data that linear algorithms are unable to capture. This limitation leads to a non-linear ML paradigm, that can be based on the "kernelization" of linear methods or purely non-linear mapping methods. The last cases lead, for the Euclidean distance, to a kernelized version combining the learned transformation $\phi(\mathbf{x}) : \mathbb{R}^m \rightarrow \mathbb{R}^{\bar{m}}$ with a Euclidean distance function with the capability to capture highly non-linear similarity relations, that is $\mathscr{D}_{ij}^{\phi} = \|(\phi(\mathbf{x}_i) - \phi(\mathbf{x}_j))\|_2$ (*Kedem et al., 2012*). Local metric refers to a problem where multiple local metrics are learned and often relies on heterogeneous data objects. In the last setting, algorithms learn using only local pairwise constraints. According to the scheme depicted in Fig. 2 the *scalability* of the solution is a challenging task, especially if we consider the growing of the availability of data in the Big Data era. The scalability could be important under the dataset dimension n and/or the dimensionality of data m. Finally, the intrinsic optimization task underlying the ML paradigm makes the optimality of the solution another important aspect. The latter, depends on the structure of the optimization scheme, that is, if the problem is convex or not (*Boyd & Vandenberghe, 2004*). In fact, for convex formulations it is guaranteed to reach a global maximum. On the contrary, for non-convex formulations, the solution may only be a local optimum.

[2] A manifold is a topological space that resembles Euclidean space near each point. Hence a n-dimensional manifold has a neighborhood that is homeomorphic to the Euclidean space of dimension n.

ON METRIC SPACES AND DISSIMILARITY MATRICES

Definitions

The standard Euclidean space, as vector space, is highly structured from the algebraic viewpoint. Moreover, the Euclidean distance is experienced daily by human beings. PR problems do not involve necessary spaces with such an high level structure. Basically, from the PR point of view, a finite number of objects have to possess such properties that guarantee generalization, hence learning. The principal property is the "*closeness*" that relies on the notion of *neighborhood*, that is a primitive property applicable to general topological spaces (*Deza & Deza, 2009*). Furthermore, the metric properties that enrich the structure of primitive mathematical objects can be induced not only for a space but also for a set (*e.g.*, the set of binary strings).

Definition 1 (Metric Space). Given a set X, a metric space is a pair (X, d), where d is a distance function $d : X \times X \to \mathbb{R}_+^0$[3] such that the following conditions are fulfilled for ∀ x, y, $z \in \mathrm{X}$[4]:

1. Reflexivity: $d(x, x) = 0$;
2. Symmetry: $d(x, y) = d(y, x)$;
3. Definiteness: $d(x, y) = 0 \Rightarrow x = y$;
4. Triangular inequality: $d(x, z) \leq d(x, y) + d(y, z)$.

If all conditions are fulfilled d is properly said distance function. Conversely, if some conditions are weakened the space continues to have some structure and d is better known in PR as *dissimilarity*. For example, a space (X, d) that obeys only the reflexivity condition is known as *hollow space*; a hollow space[5] that obeys the symmetry constraint is a *pre-metric space*; a pre-metric space obeying the definiteness is a *quasi-metric space*; a pre-metric space satisfying the triangle inequality is a *semi-metric space*.

Definition 2 (Metric for Dissimilarity Matrix $\mathscr{D}$ (*De Santis, Rizzi & Sadeghian, 2018*)). Let $\mathscr{D}$ be a symmetric dissimilarity matrix with positive off-diagonal elements d_{ij} built on a set of n objects $\mathscr{X} = \{o_1, o_2, \ldots, o_n\}$, where $d_{ij} = f(o_i, o_j)$ is a admissible measure of the dissimilarity between the objects o_i and o_j. $\mathscr{D}$ is metric if the triangle inequality $d_{ij} + d_{jk} \geq d_{ik}$ hold for all triplets (i, j, k).

It is worth noting that if two objects are similar in a metric sense, every other object that has a relation with one will have a similar relation with the other. This property allows for one of the given objects being eligible for becoming a prototype in learning algorithms (*Pękalska & Duin, 2005*).

Definition 3 (Euclidean behavior (*De Santis, Rizzi & Sadeghian, 2018*)) A $n \times n$ dissimilarity (distance) matrix $\mathscr{D}$ is Euclidean if it can be embedded in a Euclidean space $(\mathbb{R}^m, d_2)$, where d_2 is the standard Euclidean distance, where $n \geq m$. Hence, a configuration $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ can be determined in $\mathbb{R}^m$ such that $d_2(\mathbf{x}_i, \mathbf{x}_j) = \left\|\mathbf{x}_i - \mathbf{x}_j\right\|_2 = d_{ij}$ for all i, j.

A symmetric $n \times n$ matrix $\mathscr{D}$ with zero diagonal is Euclidean *iff* $\mathscr{D}_c^{*2} = \mathscr{J}\mathscr{D}^{*2}\mathscr{J}$ is negative semi-definite. The quantity $\mathscr{J} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, where $\mathbf{I}$ is the identity matrix, denotes the centering matrix. If $\mathscr{D}$ is Euclidean, it is also metric (*Gower & Legendre, 1986*).

Given a vector configuration $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ in a Euclidean space $(\mathbb{R}^m, d_2)$ equipped with the standard inner product $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$ and organized in a $n \times m$ configuration matrix[6]

[3] $\mathbb{R}_+^0 \equiv \mathbb{R}_+ \cup \{0\}$.

[4] Here it is not used the bold notation to indicate that X is a set of generic objects and not only a vector space. Hereinafter, the calligraphic notation (instead the bold one) will be used for the dissimilarity matrix $\mathscr{D}$ and for the so-called centering matrix $\mathscr{J}$.

[5] The terminology is not unified, we refer to the one adopted in (*Pękalska & Duin, 2005*).

[6] We are using the Machine Learning convention in which the n data vectors are organized as rows in the data matrix $\mathbf{X}$, hence the Gram matrix is computed as $\mathbf{X}\mathbf{X}^T$. In Linear Algebra, with data vectors organized as columns the Gram matrix is $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$ is n-times the covariance matrix, if data vectors have zero mean.

$\mathbf{X} = \left[\mathbf{x}_1^T, \mathbf{x}_2^T, \ldots, \mathbf{x}_n^T\right]^T$, the $n \times n$ Graminian (Gram) matrix $\mathbf{G}$, known in Machine Learning as *linear kernel* matrix, can be expressed by the inner product between all pairs of vectors $\mathbf{x}_i, \mathbf{x}_j$ as $\mathbf{G} = \mathbf{X}\mathbf{X}^T$. Since the squared distance d_2^2 can be expressed in terms of inner product as $d_2^2(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{x}_j \rangle = \mathbf{x}_i^T \mathbf{x}_j$, a linear relation between the Gram matrix $\mathbf{G}$ and the matrix of squared Euclidean distances $\mathscr{D}^{*2}$ can be found. The relation between $\mathbf{G}$ and $\mathscr{D}^{*2}$ is:

$$\mathbf{G} = -\frac{1}{2} \mathscr{J} \mathscr{D}^{*2} \mathscr{J}. \tag{1}$$

Conversely, the relation between $\mathscr{D}^{*2}$ and $\mathbf{G}$ is:

$$\mathscr{D}^{*2} = \mathbf{g}\mathbf{1}^T + \mathbf{1}\mathbf{g}^T - 2\mathbf{G}, \tag{2}$$

where $\mathbf{g} = diag(\mathbf{G})$.

Given a non-metric (pre-metric) or non-Euclidean symmetric dissimilarity matrix $\mathscr{D}$, the eigendecomposition of the Gram matrix $\mathbf{G}$ by the factorization $\mathbf{G} = \mathbf{Q}\Lambda\mathbf{Q}^T$, where Λ is a diagonal matrix of eigenvalues organized in descending order and $\mathbf{Q}$ is an orthogonal matrix of the correspondent eigenvectors, leads to the presence of negative eigenvalues and the indefiniteness of the corresponding Gram matrix $\mathbf{G}$. However an embedding is still possible by constructing a suitable space, *i.e.*, the PE space, with a suitable inner product and norm[7].

A generalization of the well-known Euclidean distance on a vector space $\mathbf{X} \subseteq \mathbb{R}^m$ is the Minkowski distance.

Definition 4 (Minkowski distance). Given two vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$ the Minkowski distance of order $p \in (-\infty, +\infty)$ is defined as:

$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{m} |x_i - y_i|^p\right)^{\frac{1}{p}}. \tag{3}$$

Depending on the value of the p parameter this distance generalizes the Euclidean distance ($p = 2$) or the Manhattan distance ($p = 1$). Moreover, not for all values of p the distance is metric. For $p = 2$ it is trivially metric[8] being the standard Euclidean distance. For every value $p \geq 1$ the Minkowski distance is metric, while there is a problem with the Triangular inequality for $p \in (0,1)$. In fact, if we consider a dimension $m = 2$ and three points: $A = [0, 1]^T, B = [0, 0]^T, C = [1, 0]^T$ we have $d_p(A, B) = d_p(B, C) = 1$ and $d_p(C, A) = 2^{\frac{1}{p}}$. Finally, $d_p(A, B) + d_p(B, C) = 2 < 2^{\frac{1}{p}} = d_p(C, A)$, since $p < 1$. Hence the Triangular inequality is violated and d_p is quasi-metric (*Pękalska & Duin, 2005*).

On embedding on a pseudo Euclidean space

As anticipated in *De Santis, Rizzi & Sadeghian (2018)*, a PE space $\mathbb{R}^{(p,q)}$, with signature $(p, q) \in \mathbb{N}$, can be seen as the product of a real and imaginary valued Euclidean vector space $\mathbb{R}^p \times i\mathbb{R}^q$. In other words, a PE space is a direct product space $\mathbb{R}^p \oplus \mathbb{R}^q$ with an

[7] A solution can be addressed by taking the absolute value of the negative eigenvalues, keeping in mind that the definition of the inner product generating the Gram matrix $\mathbf{G}$ changes consequently.

[8] It is noted that the Minkowski distance can be induced by a norm only for $p \geq 1$, i.e., the ℓ_p norm defined as: $||\mathbf{x}||_p = \left(\sum_{i=1}^{m} |x_i|^2\right)^{\frac{1}{p}}$.

indefinite inner product that is positive in $\mathbb{R}^p$ and negative in $\mathbb{R}^q$. Hence, given two vectors $\mathbf{x}, \mathbf{y}$ in this space the bilinear inner product can be defined as: $\langle \mathbf{x}, \mathbf{y} \rangle_{pe} \doteq \mathbf{x}^T \mathscr{J}_{pq} \mathbf{y}$, where $\mathscr{J}_{pq} = diag(\mathbf{1}_p, -\mathbf{1}_q)$. In the same way, the squared norm is defined as $\|\mathbf{x}\|^2_{pe} \doteq \langle \mathbf{x}, \mathbf{x} \rangle_{pe} = \mathbf{x}^T \mathscr{J}_{pq} \mathbf{x}$, yielding the squared distance $\|\mathbf{x} - \mathbf{y}\|^2_{pe} \doteq \langle \mathbf{x} - \mathbf{y}, \mathbf{x} - \mathbf{y} \rangle_{pe} = (\mathbf{x} - \mathbf{y})^T \mathscr{J}_{pq} (\mathbf{x} - \mathbf{y})$, that can be also negative. The Gram matrix $\mathbf{G} = -\frac{1}{2} \mathscr{J} \mathscr{D}^{*2} \mathscr{J}$ is now expressed as:

$$\mathbf{G} = \mathbf{X} \mathscr{J}_{pq} \mathbf{X}^T, \tag{4}$$

where $\mathscr{J}_{pq}$ is known as the *fundamental symmetry* in the PE space $\mathbb{R}^{(p,q)}$. The isometric embedding can be found by a proper decomposition of $\mathbf{G}$ in a PE space:

$$\mathbf{G} = \mathbf{X} \mathscr{J}_{pq} \mathbf{X}^T = \mathbf{Q} \Lambda \mathbf{Q}^T = \mathbf{Q} |\Lambda|^{\frac{1}{2}} \begin{bmatrix} \mathscr{J}_{pq} & \\ & \mathbf{0} \end{bmatrix} |\Lambda|^{\frac{1}{2}} \mathbf{Q}^T, \tag{5}$$

where $p + q = k$ and $|\Lambda|^{\frac{1}{2}}$ is a diagonal matrix whose diagonal elements are the square root of the absolute value of the eigenvalues organized in descending order, first the positive ones and after the negative ones, followed by zeros. $\mathbf{X}_k = \mathbf{Q}_k |\Lambda_k|$ is the configuration of vectors in the PE space $\mathbb{R}^k = \mathbb{R}^{(p,q)}$ where k non-zero eigenvalues corresponding to k eigenvectors in $\mathbf{Q}$ are preserved.

Finally, the estimated PE covariance matrix $\mathbf{C}$ can be found as:

$$\mathbf{C} = \frac{1}{n-1} \mathbf{X}\mathbf{X}^T \mathscr{J}_{pq} = \frac{1}{n-1} |\Lambda_k| \mathscr{J}_{pq} = \frac{1}{n-1} \Lambda_k \mathscr{J}_{pq}. \tag{6}$$

Hence $\mathbf{X}$ is an uncorrelated representation and even if $\mathbf{C}$ is not positive definite in the Euclidean sense, it is positive definite in the PE sense and $\mathbf{X}$ can be interpreted in the general context of the indefinite kernel PCA approach.

THE WEIGHTED EUCLIDEAN DISTANCE

Let be $\mathbf{X} \subseteq \mathbb{R}^{m \times n}$ a $m \times n$ data matrix with n data objects, arranged as columns, $\mathbf{x}_i = [x_{1i}, \ldots, x_{mi}]^T \in \mathbb{R}^m$, where m is the dimension of the vectorial space where data points lie. The vector space is endowed with the standard scalar product $\mathbf{x}_i \cdot \mathbf{x}_j = \langle \mathbf{x}_i, \mathbf{x}_j \rangle = \mathbf{x}_i^T \mathbf{x}_j = \sum_{k=1}^m x_{ik} x_{jk} e_{ik} e_{jk}$ while $\mathbf{e}_i$ is the i-th standard basis vector, *i.e.*, a vector of all zeros except for the entry k, which has a 1. The Euclidean distance function[9] in such space equipped with the standard inner product $\langle \cdot, \cdot \rangle$ can be expressed as:

$$\begin{aligned} d(\mathbf{x}_i, \mathbf{x}_j) &= \sqrt{\|\mathbf{x}_i - \mathbf{x}_j\|_2} = \sqrt{\langle \mathbf{x}_i - \mathbf{x}_j, \mathbf{x}_i - \mathbf{x}_j \rangle} = \\ &= \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T (\mathbf{x}_i - \mathbf{x}_j)} = \sqrt{\sum_{k=1}^m (x_{ki} - x_{kj})^2}, \end{aligned} \tag{7}$$

where the elements $\mathscr{D}_{ij} = d(\mathbf{x}_i, \mathbf{x}_j), i, j = 1, 2, \ldots, n$ form the entries of the $n \times n$ distance matrix $\mathscr{D}$ between the objects $\mathbf{x}_i$ and $\mathbf{x}_j$ in $\mathbf{X}$.

Given a symmetric positive-definite matrix $\mathbf{M}$ with real-valued entries, *i.e.*, $\mathbf{M} = \mathbf{M}^T$ and $\mathbf{x}^T \mathbf{M} \mathbf{x} \geq 0$, $\mathbf{x} \neq 0$, the entry $\mathscr{D}^{\mathbf{M}}_{ij}$ of the WED matrix $\mathscr{D}^{\mathbf{M}}$ can be expressed as:

[9] The standard Euclidean distance is an instance of a more general family of distances parametrized by the exponent p, known as Minkowski distance family. See "Characterization of a Composite Component-wise Dissimilarity" for a short introduction.

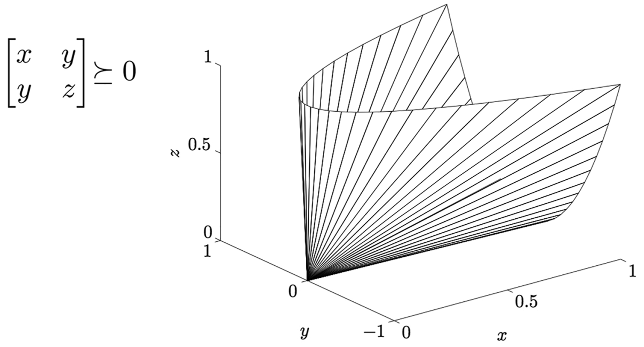

Figure 3 **The cone $\mathbb{S}_+^{\mathbf{m}}$of positive semi-definite 2 × 2 matrices.**

Full-size DOI: 10.7717/peerj-cs.1106/fig-3

$$
\begin{aligned}
\mathscr{D}_{ij}^{\mathbf{M}} = d_{\mathbf{M}}(\mathbf{x}_i, \mathbf{x}_j) \quad &= \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{M}(\mathbf{x}_i - \mathbf{x}_j)} = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{W}^T \mathbf{W}(\mathbf{x}_i - \mathbf{x}_j)} = \\
&= \sqrt{(\mathbf{W}(\mathbf{x}_i - \mathbf{x}_j))^T \mathbf{W}(\mathbf{x}_i - \mathbf{x}_j)} = \\
&= \sqrt{\langle \mathbf{W}(\mathbf{x}_i - \mathbf{x}_j), \mathbf{W}(\mathbf{x}_i - \mathbf{x}_j) \rangle} = \sqrt{\|\mathbf{W}(\mathbf{x}_i - \mathbf{x}_j)\|_2},
\end{aligned}
\tag{8}
$$

where $\mathbf{M} = \mathbf{W}^T\mathbf{W}$ is the Cholesky decomposition (*Strang, 1976*) of matrix $\mathbf{M}$, that in the Hermitian general case, is found to be the decomposition of an Hermitian matrix in the product of a lower triangular matrix and its conjugate transpose.

In ML literature the distance in Eq. (8) is known as generalized Mahalanobis distance, a family of quadratic distances parametrized by a matrix $\mathbf{M} \in \mathbb{S}_+^m$, where $\mathbb{S}_+^m$ is the cone of symmetric positive semi-definite (PSD) $m \times m$ real-valued matrices–see Fig. 3. Note that $\mathbf{M} \in \mathbb{S}_+^m$ ensures that the function $d_{\mathbf{M}}$ satisfies the properties of a pseudo-distance, *i.e.*, $\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{X}$ holds:

1. $d_{\mathbf{M}}(\mathbf{x}, \mathbf{y}) \geq 0$ (non-negativity);
2. $d_{\mathbf{M}}(\mathbf{x}, \mathbf{x}) = 0$ (identity);
3. $d_{\mathbf{M}}(\mathbf{x}, \mathbf{y}) = d_{\mathbf{M}}(\mathbf{y}, \mathbf{x})$ (symmetry);
4. $d_{\mathbf{M}}(\mathbf{x}, \mathbf{z}) \leq d_{\mathbf{M}}(\mathbf{x}, \mathbf{y}) + d_{\mathbf{M}}(\mathbf{y}, \mathbf{z})$ (triangular inequality).

The above properties hold trivially for a standard Euclidean space where $\mathbf{M} = \mathbf{I}$.

The matrix $\mathbf{W}$ can be seen as a linear operator that transforms the shape of the space where data points, *i.e.*, data vectors, lie. Specifically $\mathbf{W}$ defines a suitable transformation (endomorphism) $\mathscr{V} \rightarrow \mathscr{V}$ of the (abstract) space $\mathscr{V}$ spanned by rows vector of $\mathbf{X}$ in itself: given a vector $\mathbf{x}$ in the starting space S_1, the matrix $\mathbf{W}$ maps this vector in a new vector $\mathbf{x}_w = \mathbf{W}\mathbf{x}$ that lies in the space S_2, where S_1 and S_2 are isomorphic to $\mathscr{V}$[10]. In the new transformed space the inner product becomes the standard inner product $\langle \cdot, \cdot \rangle$, *i.e.*,

[10] If $\mathbf{M} = \mathbf{W}^T\mathbf{W}$ is strictly positive definite, $\mathbf{W}$ is a triangular matrix with no 0's entry in the principal diagonal, hence it is invertible and we have $\mathbf{W}^{-1}\mathbf{x}_w = \mathbf{x}$ and $\mathbf{W}^{-1}\mathbf{y}_w = \mathbf{y}$. Moreover, if an Hermitian matrix is positive semi-definite the Cholesky decomposition is still available, having the possibility of 0's entries on the diagonal of $\mathbf{W}$. Finally the Cholesky decomposition is unique when M is positive definite, while it is not true when it is positive semidefinite (*Golub & Van Loan, 2012*).

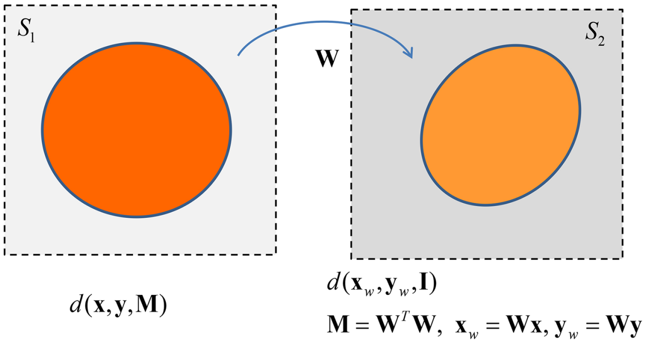
S_1
S_2
$\mathbf{W}$
$d(\mathbf{x},\mathbf{y},\mathbf{M})$
$d(\mathbf{x}_w,\mathbf{y}_w,\mathbf{I})$
$\mathbf{M}=\mathbf{W}^T\mathbf{W},\ \mathbf{x}_w=\mathbf{W}\mathbf{x},\mathbf{y}_w=\mathbf{W}\mathbf{y}$

Figure 4 **The transformation between Euclidean spaces by the linear operator W.**
Full-size DOI: 10.7717/peerj-cs.1106/fig-4

$\langle \mathbf{x}_w, \mathbf{y}_w \rangle = \langle \mathbf{W}\mathbf{x}, \mathbf{W}\mathbf{y} \rangle = \mathbf{x}_w^T\mathbf{y}_w = (\mathbf{W}\mathbf{x})^T(\mathbf{W}\mathbf{y}) = \mathbf{x}^T\mathbf{W}^T\mathbf{W}\mathbf{x}$. The arrival space is endowed with a squared norm given by $\langle \mathbf{x}_w, \mathbf{x}_w \rangle = \langle \mathbf{W}\mathbf{x}, \mathbf{W}\mathbf{x} \rangle = \|\mathbf{M}\mathbf{x}\|_2^2$, being $\mathbf{M} = \mathbf{W}^T\mathbf{W}$[11].

Observation 1. The weighted distance $d_\mathbf{M}(\mathbf{x},\mathbf{y}) = d(\mathbf{x},\mathbf{y},\mathbf{M})$ with $\mathbf{M} = \mathbf{W}^T\mathbf{W}$ equals $d_\mathbf{I}(\mathbf{x}_w,\mathbf{y}_w) = d(\mathbf{x}_w,\mathbf{y}_w,\mathbf{I})$ where $\mathbf{x}_w = \mathbf{W}\mathbf{x}$ and $\mathbf{y}_w = \mathbf{W}\mathbf{y}$ and $\mathbf{I}$ is the identity matrix.

Proof. The proof follows by the same algebraic manipulation of Eq. (8).

Let $\mathbf{M} = \mathbf{W}^T\mathbf{W}$, $\mathbf{x}_w = \mathbf{W}\mathbf{x}$ and $\mathbf{y}_w = \mathbf{W}\mathbf{y}$. It holds that:

$$\begin{aligned} d\ (\mathbf{x},\mathbf{y},\mathbf{M}) &= \sqrt{(\mathbf{x}-\mathbf{y})^T\mathbf{M}(\mathbf{x}-\mathbf{y})} = \sqrt{(\mathbf{x}-\mathbf{y})^T\mathbf{W}^T\mathbf{W}(\mathbf{x}-\mathbf{y})} = \\ &= \sqrt{(\mathbf{W}(\mathbf{x}-\mathbf{y}))^T\mathbf{W}(\mathbf{x}-\mathbf{y})} = \sqrt{(\mathbf{W}\mathbf{x}-\mathbf{W}\mathbf{y})^T(\mathbf{W}\mathbf{x}-\mathbf{W}\mathbf{y})} = \\ &= \sqrt{(\mathbf{x}_w-\mathbf{y}_w)^T(\mathbf{x}_w-\mathbf{y}_w)} = d(\mathbf{x}_w,\mathbf{y}_w,\mathbf{I}). \end{aligned} \tag{9}$$

□

The matrix **W** is an instance of an operator that defines a rotation and a scaling of the objects upon it operates. **W** maps a circle in the unweighted Euclidean space in an ellipse in the weighted Euclidean space–see Fig. 4. Hence we can state the following theorem.

Theorem 1. Applying a transformation **W** to all point of a circle of radius *r* the resulting points form an ellipse whose center is the same as the circle and length of its axes equals *r* times twice the square root of eigenvalues of $\mathbf{M} = \mathbf{W}^T\mathbf{W}$.

Proof. An interesting demonstration can be found in *Kurniawati, Jin & Shepherd (1998)*. □

The weight matrix **M** can be decomposed in its rotation and scaling components by means of the eigendecomposition operation. Specifically, by decomposing $\mathbf{M} = \mathbf{Q}\mathbf{D}\mathbf{Q}^T$ where **Q** is an orthogonal matrix with normalized column vectors, that is $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ and $\mathbf{Q}^T = \mathbf{Q}^{-1}$, and **D** is a diagonal matrix[12]. **D** contains the eigenvalues $\lambda_1,\lambda_2,\ldots,\lambda_m$ (organized in decreasing order) that are the scaling factors, while **Q** is the rotation operator matrix that leaves unchanged the (squared) norm of vectors, that is $\|\mathbf{Q}\mathbf{x}\|_2^2 = \langle \mathbf{Q}\mathbf{x}, \mathbf{Q}\mathbf{x} \rangle = \mathbf{x}^T\mathbf{Q}^T\mathbf{Q}\mathbf{x} = \mathbf{x}^T\mathbf{x} = \|\mathbf{x}\|_2^2$ (*Strang, 1976*).

[11] We can assert that each space of vectors **x** comes with its dual-space of *linear functionals* $\mathbf{w}^T$. In the scalar product $\mathbf{w}^T$ x, $\mathbf{w}^T$ acts linearly upon vectors x and y, *i.e.*, $\mathbf{w}^T(\lambda\mathbf{x}+\mu\mathbf{y}) = \lambda\mathbf{w}^T$ x+ $\mu\mathbf{w}^T$ y. At the same time **x** acts linearly upon $\mathbf{v}^T$ and $\mathbf{w}^T$, *i.e.*, $(\lambda\mathbf{v}^T+\mu\mathbf{w}^T)\mathbf{x} = \lambda\mathbf{v}^T\mathbf{x}+\mu\mathbf{w}^T\mathbf{x}$. So the linear functionals $\mathbf{w}^T$ form a vector space Dual or Conjugate to the space of vectors **x**. Each space is dual to the other, and they have the same finite dimension.

[12] The eigendecoposition results in a safe operation because **M** is a (square) real symmetric matrix, furthermore it can be demonstrated (spectral theorem (*Strang, 1976*)) that **M** is diagonalizable by the matrix of its eigenvectors, *i.e.*, from the fundamental equation about the eigendecomposition: **MQ** = **QD**, by multiplying on the left both sides by $\mathbf{Q}^T$ we have $\mathbf{Q}^T\mathbf{M}\mathbf{Q} = \mathbf{D}$. Finally the symmetry property leads to a set of real-valued eigenvalues and, being **M** a positive definite matrix, all the eigenvalues are positive.

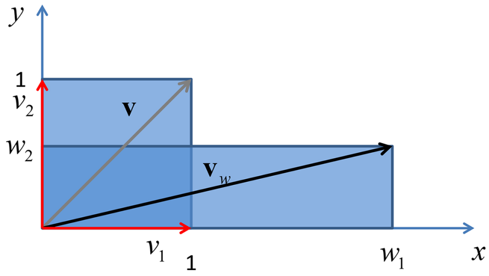

Figure 5 **Vertical shrink and horizontal stretch of a unit square through a transformation induced by adiagonal matrix with real positive entries.** Full-size DOI: 10.7717/peerj-cs.1106/fig-5

At this point it is possible to express the WED in terms of the above eigendecomposition:

$$\begin{aligned} d(\mathbf{x},\mathbf{y},\mathbf{M}) &= \sqrt{(\mathbf{x}-\mathbf{y})^T\mathbf{M}(\mathbf{x}-\mathbf{y})} = \sqrt{(\mathbf{x}-\mathbf{y})^T\mathbf{W}^T\mathbf{W}(\mathbf{x}-\mathbf{y})} = \\ &= \sqrt{(\mathbf{x}-\mathbf{y})^T\mathbf{Q}\mathbf{D}\mathbf{Q}^T(\mathbf{x}-\mathbf{y})} = \sqrt{\left(\mathbf{Q}^T\mathbf{x}-\mathbf{Q}^T\mathbf{y}\right)^T\mathbf{D}\left(\mathbf{Q}^T\mathbf{x}-\mathbf{Q}^T\mathbf{y}\right)}. \end{aligned} \tag{10}$$

From Eq. (10) it follows that $d(\mathbf{x},\mathbf{y},\mathbf{M})$ can be expressed, through the eigendecomposition of the weighting matrix $\mathbf{M}$, with another weighted distance with weights given by the eigenvalues matrix $\mathbf{D}$. This new distance takes into account new vectors: $\mathbf{Q}^T\mathbf{x} = \hat{\mathbf{x}}$ and $\mathbf{Q}^T\mathbf{y} = \hat{\mathbf{y}}$ that are the rotated counterparts of original vectors $\mathbf{x}$ and $\mathbf{y}$. In other words, the two vectors $\hat{\mathbf{x}}$ and $\hat{\mathbf{y}}$ are the rotated, but not scaled, version of $\mathbf{x}_w$ and $\mathbf{y}_w$ that originate both in space S_2. It can be demonstrated that the length of the axis of the ellipsoid in the direction of i-th eigenvalue λ_i is equal to: $\sqrt{d(\mathbf{x},\mathbf{y},\mathbf{M})/\lambda_i}$.

Finally if the weighting matrix $\mathbf{M}$ is a diagonal matrix, with real entries, the above eigendecomposition reduces to $\mathbf{M} = \mathbf{E}\mathbf{D}\mathbf{E}^T$ where $\mathbf{E}$ is the eigenvector matrix $\mathbf{E} = [\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_m]$, whose columns contains the standard basis in $\mathbb{R}^m$ with the property $\mathbf{e}_i^T\mathbf{e}_j = \delta_{ij}$, where δ_{ij} is the Kronecker delta. In this case the matrix $\mathbf{E}$ represents the identity element of the rotation operator, leaving vectors in the original place, while they are scaled by a factors given by the entries of the diagonal of $\mathbf{M}$, being the eigenvalues of a diagonal matrix the diagonal entries of the same matrix. An example of this phenomenon is given in Fig. 5.

CHARACTERIZATION OF A COMPOSITE COMPONENT-WISE DISSIMILARITY

When the PR problem at hand deals with heterogeneous measures on objects and these measures are both structurally and semantically different (graphs, time series, images, real numbers, *etc.*), a composite dissimilarity measure can be useful, for example in clustering applications. The dissimilarity measure is a combination of (sub)-dissimilarities suitably defined depending on the nature of the data.

Before constructing a toy composite dissimilarity measure, it is worth to mention the following corollary valid when a dissimilarity measure is computed by combining the dissimilarities pertaining to all of the m attributes separately. In fact, given m features, a dissimilarity measures can be computed as: $d(x, y) = \sum_{i=1}^{m} f(x_i, y_i)$, where $f(x_j, x_j) = 0$ and $f(x_j, y_j) = f(y_j, x_j) \geq 0$ for all j. The corollary states:

Corollary 1. Let $x, y \in \mathbb{R}^m$. Then $d(x, y) = \sum_{i=1}^{m} f(x_i, y_i)$ is metric *iff* f is metric on $\mathbb{R}$.

Proof. The proof can be done considering that f is non-negative, symmetric and it holds that $f(s, s) = 0$ for $s \in \mathbb{R}$, then the first three axioms about metric spaces, *i.e.*, reflexivity, symmetry and definiteness, are fulfilled. Furthermore, since d is metric $d(x, y) + d(y, z) \geq d(x, z)$ holds for all x, y, z. If we consider $x_j = c_x$, $y_j = c_y$, $z_j = z_x$, for all j and some constants c_x, c_y, c_j the Triangle inequality for d reduces to $f(x_c, y_c) + f(y_c, z_c) \geq f(x_c, z_c)$. The ⇒ proof is trivial. □

Moreover, it can be demonstrated (Gower & Legendre, 1986) that, at least for the Euclidean case ($p = 2$ in the Minkowski distance definition), if $f : X \times Y \rightarrow \mathbb{R}_{+}^{0}$ is a function, then $d(x, y) = \sum_{i=1}^{m} f(x_i, y_i)$ is metric *iff* $d^{'}(x, y) = \sum_{i=1}^{m} \left[f(x_i, y_i)^2\right]^{1/2}$ is metric.

Now we show a demonstration of the following claim valid for a composite dissimilarity measure, making use of Def. 3 that characterizes the Euclidean behavior for dissimilarity matrices and Def. 2 for metric behavior.

Claim 1. Given two general objects $x, y \in \mathscr{H}$, where $\mathscr{H}$ is a generic feature space, and a component wise custom-based dissimilarity $d(x, y) = \sqrt{(x \ominus y)^T (x \ominus y)} = \sqrt{\sum_{i=1}^{m} (x_i \ominus y_i)^2}$, then if at least one component-wise dissimilarity is not Euclidean the dissimilarity matrix that arises from d applied on object within $\mathscr{H}$, is not Euclidean.

As stated in "On Metric Spaces and Dissimilarity Matrices", the expression "non-Euclidean" means that there is no set of vectors in a vector space of any dimensionality for which the Euclidean distances between the objects are identical to the given ones (Duin & Pękalska, 2010). We show now how Claim 1 can be demonstrated with a constructive example. Let $\mathbf{x} = (x_1, x_2, \ldots x_k, x_{k+1}, \ldots x_m)$ and $\mathbf{y} = (y_1, y_2, \ldots y_k, y_{k+1}, \ldots y_m)$ be two objects in a vectorial space $\mathscr{H}_v$. We define a set of component-wise dissimilarities induced for the first k components such as $f_j^{cw}(x_j, y_j) = \left|x_j - y_j\right|, j = 1, 2, \ldots, k$ and a single component-wise dissimilarity induced for the remaining $m - k$ components such as $f^p(x^{s=k+1,\ldots,m}, y^{s=k+1,\ldots,m}) = \left(\sum_{s=k+1}^{m} |x_s - y_s|^p\right)^{\frac{1}{p}}$. In other words we divide the starting space $\mathscr{H}_v$ as the Cartesian product (Strang, 1976) between two sub-spaces, the space $\mathscr{H}_{cw}$ generated from the first k components, in which the component-wise dissimilarities are computed as f^{cw} and $\mathscr{H}_p$ where the dissimilarity is computed as the Minkowski distance f^p applied to the last $m - k$ components. Finally, the overall dissimilarity between two objects, say $\mathbf{x}, \mathbf{y}$ is induced by the ℓ_2 norm in the following way:

$$\hat{d}(\mathbf{x}, \mathbf{y}) = \sqrt{(f^{cw} \oplus f^p)^T (f^{cw} \oplus f^p)}, \tag{11}$$

where $(f^{cw} \oplus f^p)$ is the vector of dimension $k + l$ constructed by the concatenation of the two (sub)-dissimilarities $f^{cw}{}_j$, f^p, $j = 1,2,\ldots,k$.

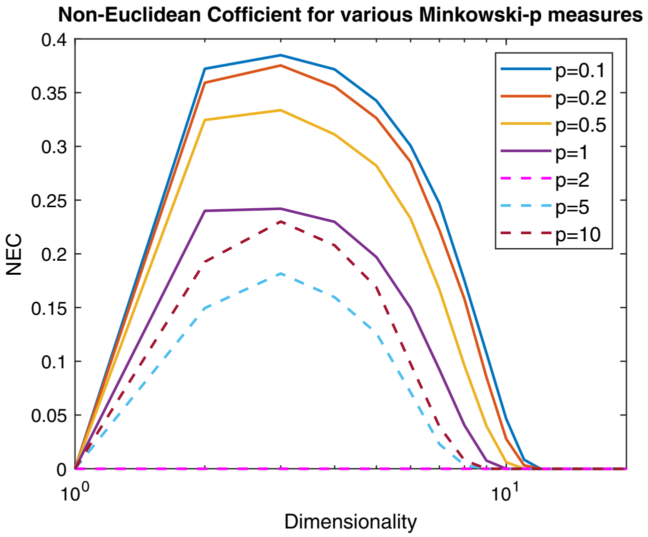

Figure 6 Non-Euclidean influence measured by the Negative Eigen-Fraction for several values of the *p* parameter of the Minkowski distance. Measures are computed starting from a 100-point multi-variate Gaussian distribution by varying the dimensionality. Full-size DOI: 10.7717/peerj-cs.1106/fig-6

To evaluate the validity of the Claim 1 the dissimilarity in Eq. (11) is computed on a sample drawn from a multi-variate Gaussian distribution with dimension m parameterized as: $m = k + l$, where k is maintained fixed without loss of generality, and l is varied. It is noted that the p parameter controls the nature of the Minkowski distance, making the (sub)-dissimilarity f^p metric or not metric (and even non-Euclidean[13]) depending on the value of p as demonstrated above, such that for $p \geq 1$ it is metric.

In order to measure the non-Euclidean behavior of the space induced by the Minkowski distance, we introduce the Negative Eigen-Fraction (NEF):

$$NEF = \frac{\sum_{j=p+1}^{p+q} |\lambda_i|}{\sum_{i=1}^{p+q} |\lambda_j|}, \tag{12}$$

where (p, q) is the signature of the PE space, and λ_i are the eigenvalues of the Gram matrix decomposition. The NEF measures the degree of the non-Euclidean influence evaluating the ratio between the sum of the negative eigenvalues and the overall set of eigenvalues. Another index that helps to commensurate the non-Euclidean influence is the Negative Eigen-Ratio (NER):

$$NER = r_1 = \frac{|\lambda_{min}|}{\lambda_{max}}, \tag{13}$$

where λ_{min} and λ_{max} are the minimum and maximum eigenvalue of the Gram matrix. In Fig. 6 are reported, following the same experimental scheme proposed in *Pękalska et al. (2006)*, several curves representing the NEF for a 100 points Gaussian sample varying the p parameter of the Minkowski distance as a function of the dimensionality. Now it is clear that the Minkowski distance is non-Euclidean for any $p \neq 2$, but for very high dimensionality values the Euclidean behavior is restored independently from p.

[13] The Euclidean property defined in Def. 3 is more restrictive than the metric property. Thereby, there are spaces that are metric but non-Euclidean. The opposite does not hold.

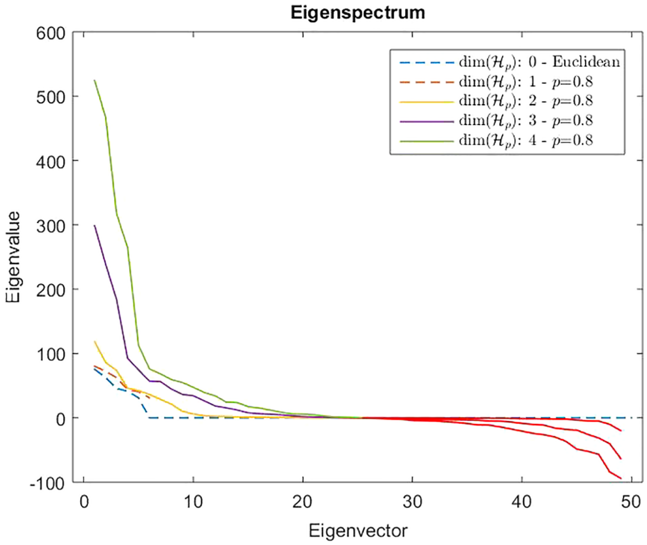

Figure 7 Eigenspectrum of the Gram matrix $\hat{\mathbf{G}}$ obtained from the dissimilarity matrix $\hat{\mathscr{D}}$ computed by means of Eq. (11) in which the parameter of the Minkowski distance in f_p is set as $p = 0.8$. Dashed lines show a positive eigenspectrum, while continuous lines show a mixed eigenspectrum.
Full-size DOI: 10.7717/peerj-cs.1106/fig-7

However, from Def. 3, we know that a $n \times n$ ($n \geq m$) dissimilarity (distance) matrix $\mathscr{D}$ is Euclidean if it can be embedded in a Euclidean space $(\mathbb{R}^m, d_2)$, where d_2 is the standard Euclidean distance. It means that the Gram matrix $\mathbf{G}$ obtained as described in "On Metric Spaces and Dissimilarity Matrices" does not contain negative eigenvalues, hence it is a positive semi-definite matrix. The remainder of the discussion is then based on the *eigenvalues spectrum* of the Gram matrix computed from the dissimilarity matrix $\hat{\mathscr{D}} = \hat{d}_{ij}$. In Fig. 7 are reported the eigenvalues spectra for the Gram matrix $\hat{\mathbf{G}}$ obtained from the dissimilarity matrix $\hat{\mathscr{D}}$ computed for a fixed $k = 5$ and varying the value for $l = 0, 1, \ldots 4$. The dashed lines are the case: $l = 0$ and $l = 1$. The first one represents the spectrum deducted from the first $k = 5$ components of $\mathscr{H}_v$ and, as we expected, it contains only positive eigenvalues, thereby the dissimilarity matrix $\hat{\mathscr{D}}$ is isometrically embeddable. The same holds for $l = 1$ because trivially we have that

$f_p(x, y) = (|x - y|^p)^{\frac{1}{p}} = |x - y|, \forall x, y \in \mathbb{R}$, thus the dissimilarity measure f_p remains metric. For $l > 1$ the several spectra contain both positive and negative eigenvalues making the Gram matrix $\hat{\mathbf{G}}$ indefinite. As counterexample in Fig. 8 are depicted the spectra of the dissimilarity in Eq. (11), where the parameter p of the Minkowski distance is set as $p = 2$. As we expect, in this case, the dissimilarity behaves in an Euclidean fashion.

ON THE PRESENCE OF WEIGHTS IN A COMPONENT-WISE DISSIMILARITY AND THE EIGENSPECTRUM OF THE GRAM MATRIX

In the discussion related to Claim 1 it is introduced a suitable component-wise custom dissimilarity that in general has the form: $d = \|\bar{\mathbf{d}}_c\|_2 = \sqrt{\bar{\mathbf{d}}_c^T \bar{\mathbf{d}}_c}$ that is the ℓ_2 norm of the vector[14] $\bar{\mathbf{d}}_c = [d_{\mathscr{F}_1}, d_{\mathscr{F}_2}, \ldots, d_{\mathscr{F}_k}]$ that is computed through suitable component-wise

[14] The vector $\mathbf{d}_c$ consists in the (sub)-dissimilarities that are computed on objects x that belongs to a suitable structured, even non-metric, dataset.

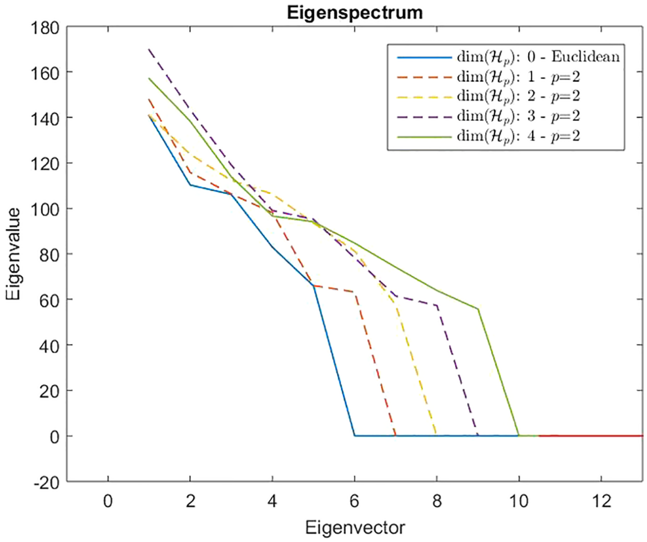

Figure 8 Eigenspectrum of the Gram matrix $\hat{\mathbf{G}}$ obtained from the dissimilarity matrix $\hat{\mathscr{D}}$ computed by means of Eq. (11) in which the parameter of the Minkowski distance is set as p = 2, hence the standard Euclidean distance. All spectra are positive, hence the custom-based dissimilarity in Eq. (11) is Euclidean. Full-size DOI: 10.7717/peerj-cs.1106/fig-8

(sub)-dissimilarities, each one induced for a specific feature type $\mathscr{F}_j$. Specifically, we had two groups: the first (sub)-dissimilarities act on a vectorial subspace and they were computed as the component-wise ℓ_1 norm: $|x_i - y_i|$, while in the second group a unique (sub)-dissimilarity is computed as the Minkowski distance (p = 0.8, hence neither metric, nor Euclidean). Now we will discuss the case in which the same family of custom-based dissimilarities are weighted, hence they have the form described in "The Weighted Euclidean Distance" for the WED. In other words, given a pair of objects x_i and y_j, the dissimilarity measure under analysis has the following form:

$$d_{\mathbf{w}}(x_i, y_j) = \left\|\bar{\mathbf{d}}_c\right\|_{\mathbf{w}} = \sqrt{\bar{\mathbf{d}}_c(x_i, y_j)^T \mathbf{W}^T \mathbf{W} \bar{\mathbf{d}}_c(x_i, y_j)}. \tag{14}$$

Given n objects x_i, i = 1,2,…,n, the weighted dissimilarity matrix whose entries are given by $d_{\mathbf{w}}(x_i, y_j)$–see Eq. (14)–is hereinafter referred to as $\mathscr{D}_{\mathbf{w}}$, for convenience. The latter can be decomposed according to Eq. (1) as $\mathbf{G}_{\mathbf{w}} = -\frac{1}{2}\mathscr{J}\mathscr{D}_{\mathbf{w}}^{*2}\mathscr{J}$, where $\mathbf{G}_{\mathbf{w}}$ is the Gram matrix parametrized by the weight matrix $\mathbf{W}$. As discussed in "Metric Learning", the weights act as a linear mapping $\mathscr{M} : \mathbf{x} \rightarrow \mathbf{Wx}$. Starting from the above settings, two questions arise. The first is if, in principle, it is possible to find a suitable weighting matrix $\mathbf{W}$ that makes the dissimilarity matrix $\mathscr{D}_{\mathbf{w}}$ "more Euclidean". The second question is about the behavior of the Gram matrix $\mathbf{G}_{\mathbf{w}}$ in terms of eigendecomposition. In other words, one may ask what is the relationship between the eigenvalues (and eigenvectors) of the non-weighted Gram matrix $\mathbf{G}$ and the weighted one $\mathbf{G}_{\mathbf{w}}$.

The two questions are strongly interrelated. By the way, the first is simpler than the second. To answer the first question one may conceive a simple problem in which one wants to minimize the NEF defined in Eq. (12), hence, we can consider a diagonal matrix $\mathbf{W}_{diag} = \mathrm{diag}([w_1, w_2, \ldots, w_d])$ and the task is to solve the following minimization problem:

$$\begin{aligned} &\arg\min_{\mathbf{G}_\mathbf{w}} \mathrm{NEF}(\mathbf{G}_\mathbf{w}), \\ &s.t. \;\; 0 \leq w_i \leq 1 \; i = 1, 2, \ldots, d. \end{aligned} \tag{15}$$

The NEF – see Eq. (12) – depends on the eigenvalues λ_i of the Gram matrix which, in turn, depend on the dissimilarity matrix $\mathscr{D}_\mathbf{w}$ through a non-linear operation, which in turn depends on the weighted dissimilarity measure $d_\mathbf{w}(x_i, y_j)$, which, finally, depends on the weights matrix $\mathbf{W}_{diag}$ (if diagonal). The optimization problem can be performed *via* the same setting used to discuss Claim 1. Specifically, it is a simple exercise in adopting a meta-heuristic, such as a GA, in minimizing the optimization problem in Eq. (15). The two subspaces, $\mathscr{H}_{cw}$ and $\mathscr{H}_v$ have a dimensionality equal to 3 and the Minkowski parameter of the distance acting on $\mathscr{H}_{cw}$ is set to 0.8 (hence neither metric, nor Euclidean).

Starting from a random population of 30 individuals (chromosomes) for the weights $\mathbf{w}$, the GA converges to the (sub)-optimal solution $\mathbf{w}^* = [1, 1, 0.999, 0.0001]$ with a fitness value (the NEF) equals to 2.0380e-06, hence negligible. As we expected, the GA finds a solution with higher weights for the "Euclidean" components and practically null value for the "Minkowski" component.

Although the answer to the first question is trivial, the second question about the relationship of the two spectra of $\mathbf{G}$ and $\mathbf{G}_\mathbf{w}$ is only apparently simple. Here we try to give a sketch of the problem. Suppose that $\mathscr{F}$ is a vectorial space endowed with the standard norm $\langle \cdot, \cdot \rangle$, and $\mathbf{X} \in \mathbb{R}^{m \times n}$ is a data matrix with the n data points organized as columns. The discussion can be restricted to an Euclidean space equipped by the standard Euclidean distance: $d(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_2$ for $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$. The scalar product matrix or the Gram matrix, with the data matrix organized with data vectors in columns and the variables as rows, is: $\mathbf{G} = \mathbf{X}^T\mathbf{X}$. The linear mapping $\mathscr{M} : \mathbf{x} \rightarrow \mathbf{W}\mathbf{x}$ transforms the data matrix $\mathbf{X}$ in $\mathscr{M}(\mathbf{X}) = \mathbf{W}\mathbf{X} = \mathbf{Y}$. Thereby, the Gram matrix becomes: $\mathbf{G}_\mathbf{w} = (\mathbf{W}\mathbf{X})^T(\mathbf{W}\mathbf{X}) = \mathbf{X}^T \mathbf{W}^T\mathbf{W}\mathbf{X} = \mathbf{Y}^T\mathbf{Y}$. We note that if $\mathbf{W}$ is invertible we have the inverse map $\mathscr{M}^{-1}(\mathbf{Y}) = \mathbf{W}^{-1}\mathbf{Y}$. In trying to find a relation between the eigenvalues of $\mathbf{Y}^T\mathbf{Y}$ and those of $\mathbf{X}^T\mathbf{X}$, we can make use of the relation between the Singular Value Decomposition (SVD) of a $m \times n$ matrix $\mathbf{A}$ and the eigendecomposition of the $n \times n$ matrix $\mathbf{A}^T\mathbf{A}$. In fact, any $m \times n$ matrix can be factored as $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^T$ (*Strang, 1976*), where the columns of matrix $\mathbf{U}$ $(m \times m)$ are the eigenvectors of $\mathbf{A}\mathbf{A}^T$ and the columns of $\mathbf{V}$ $(n \times n)$ are eigenvectors of $\mathbf{A}^T\mathbf{A}$; finally, the $r = \mathrm{rank}\ (\mathbf{A})$ singular values in the diagonal of Σ $(m \times n)$ are the square roots of the non-zero eigenvalues of both $\mathbf{A}^T\mathbf{A}$ and $\mathbf{A}\mathbf{A}^T$[15].

Let $\mathbf{Y} = \mathbf{U}_\mathbf{w}\Sigma_\mathbf{w}\mathbf{V}_\mathbf{w}^T$ be the SVD decomposition of $\mathbf{Y}$ and $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$ be the decomposition of $\mathbf{X}$. If we multiply on the left side for $\mathbf{W}^{-1}$ both sides of the first relation we obtain $\mathbf{W}^{-1}\mathbf{Y} = \mathbf{W}^{-1}\mathbf{U}_\mathbf{w}\Sigma_\mathbf{w}\mathbf{V}_\mathbf{w}^T$, hence $\mathbf{X} = \mathbf{W}^{-1}\mathbf{Y} = \mathbf{W}^{-1}\mathbf{U}_\mathbf{w}\Sigma_\mathbf{w}\mathbf{V}_\mathbf{w}^T$. If we compare

[15] It is easy to show the relation between the eigenvalues and the singular values: $\mathbf{A}^T\mathbf{A} = (\mathbf{U}\Sigma\mathbf{V}^T)^T(\mathbf{U}\Sigma\mathbf{V}^T) = \mathbf{V}\Sigma\mathbf{U}^T\mathbf{U}\Sigma\mathbf{V}^T = \mathbf{V}\Sigma^T\mathbf{V}\Sigma^T$, being $\mathbf{U}^T\mathbf{U} = \mathbf{I}$. In the same way $\mathbf{A}^T\mathbf{A} = \Sigma\mathbf{U}^T\Sigma\mathbf{U}^T$, being $\mathbf{V}^T\mathbf{V} = \mathbf{I}$. $\mathbf{V}$ and $\mathbf{U}$ are orthogonal matrices for a real $\mathbf{A}$ (for complex $\mathbf{A}$ they are unitary matrices). $\Sigma^T\Sigma = \Sigma\Sigma^T$ is a $n \times n$ diagonal matrix with diagonal entries the square roots of singular values of $\mathbf{A}$ that are the eigenvalues of $\mathbf{A}^T\mathbf{A}$ or $\mathbf{A}\mathbf{A}^T$.

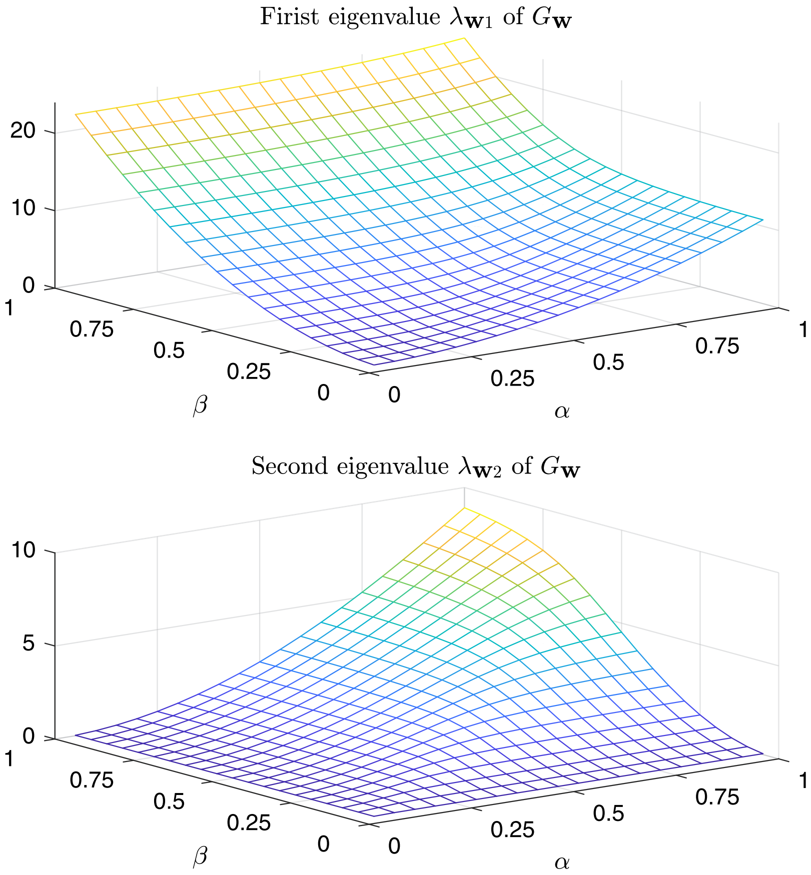

Figure 9 **Magnitude of the first and second eigenvalue of $\mathbf{G}_{\mathbf{w}}^{test}$as a function of α and β.**

Full-size DOI: 10.7717/peerj-cs.1106/fig-9

these two relations and multiply both sides for $\mathbf{U}^T$ on the left side and $\mathbf{V}$ on the right side and by further considering that $\mathbf{V}^T\mathbf{V} = \mathbf{I} = \mathbf{U}^T\mathbf{U}$, we come to the relation: $\mathbf{U}^T\mathbf{U}\Sigma\mathbf{V}^T\mathbf{V} = \mathbf{U}^T\mathbf{W}^{-1}\mathbf{U}_{\mathbf{w}}\Sigma_{\mathbf{w}}\mathbf{V}_{\mathbf{w}}^T\mathbf{V}$ that simplifies as:

$$\Sigma = \mathbf{U}^T\mathbf{W}^{-1}\mathbf{U}_{\mathbf{w}}\Sigma_{\mathbf{w}}\mathbf{V}_{\mathbf{w}}^T\mathbf{V}. \tag{16}$$

Equation (16) is a (complex) relation between the (diagonal) singular values matrix Σ that contains as entries the singular values of $\mathbf{X}$ and the singular values of $\mathbf{Y} = \mathbf{WX}$, placed in the diagonal of $\Sigma_{\mathbf{w}}$. Unfortunately, calculations cannot be further performed in closed form unless we make additional assumption on $\mathbf{W}$. The reason becomes clear if we think at $\mathbf{WX}$ as the product of two matrices: in fact, the original question about the relationship between the eigenvalues of the Gram matrices $\mathbf{G}$ and $\mathbf{G}_{\mathbf{w}}$ can be translated into the relation of the eigenvalues of the following matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{AB}$. However, this so-wanted relationship between the eigenvalues of the product of general matrices and its multiplicands is still an open problem of mathematics, even if in the literature there are a number of works that provide several inequalities for the matrix product and sum problem (*Zhang & Zhang, 2006*; *Fulton, 2000*; *Watkins, 1970*; *Thompson & Therianos, 1971*). If $\mathbf{W}$ is a scalar matrix of the form $\mathbf{W} = k\mathbf{I}$ the relation shown in Eq. (16) becomes simple. In

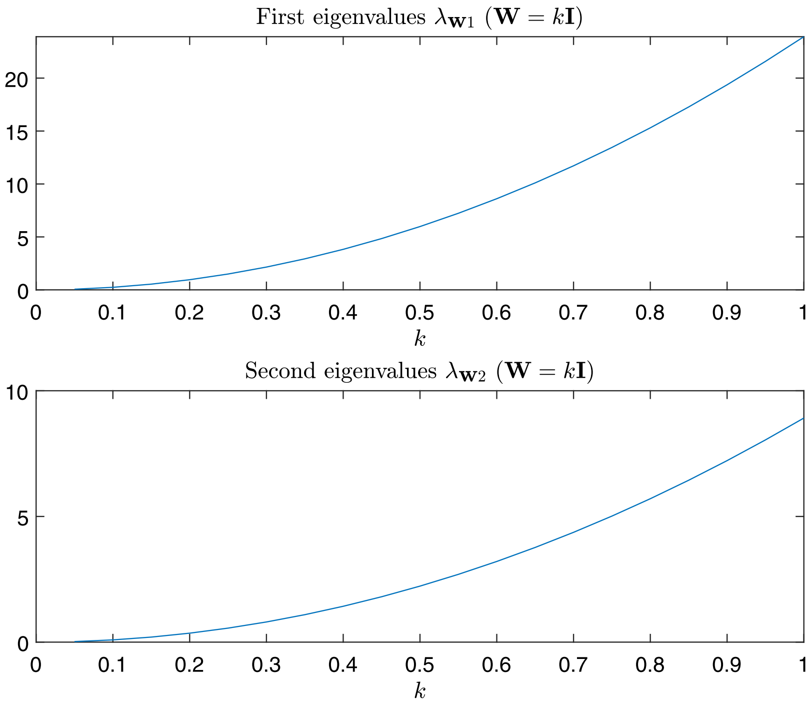

Figure 10 Magnitude of the first and second eigenvalue of $\mathbf{G}_{\mathbf{w}}^{test}$ in the case $\alpha = \beta = k$, *i.e.*, W = kI.
Full-size DOI: 10.7717/peerj-cs.1106/fig-10

fact, we can write $\mathbf{WX} = k\mathbf{IX} = \mathbf{U_w}\Sigma_\mathbf{w}\mathbf{V}_\mathbf{w}^T$, but $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$, hence $\mathbf{WX} = k\mathbf{IU}\Sigma\mathbf{V}^T = \mathbf{U_w}\Sigma_\mathbf{w}\mathbf{V}_\mathbf{w}^T$. It means that the singular vectors are the same: $\mathbf{U} = \mathbf{U_w}$ and $\mathbf{V} = \mathbf{V_w}$ and therefore Eq. (16) becomes:

$$\Sigma = k^{-1}\Sigma_\mathbf{w} \Rightarrow \Sigma_\mathbf{w} = k\Sigma, \tag{17}$$

Hence, for the spectrum of $\mathbf{G_w} = \mathbf{X}^T\mathbf{W}^T\mathbf{WX}$, we have $\Sigma^T\Sigma = k^{-1}\Sigma_\mathbf{w}^T\Sigma_\mathbf{w}$[16].

Ultimately, there are no relationships between the spectrum of the product of two generic matrices and one of the single matrices, unless in simple cases[17]. In general, two generic matrices do not share the same set of eigenvectors and this makes the analysis infeasible. In order to graphically show in a computational fashion the relationship between the eigenvalues of the Gram matrix obtained from a weighted dissimilarity matrix and those obtained from a non-weighted dissimilarity matrix, we have generated a random bi-dimensional matrix $\mathbf{X}^{test} \in \mathbb{R}^{(2\times 20)}$, hence containing 20 random 2-D vectors. Moreover, the dissimilarity matrix $\mathscr{D}_\mathbf{w}^{test}$ on $\mathbf{X}^{test}$ is computed through the standard Euclidean distance and finally the Gram matrix $\mathbf{G}_\mathbf{w}^{test}$ is extracted. The dissimilarity measure is weighted with a diagonal matrix of the form: $\mathbf{W} = \begin{bmatrix} \alpha & 0 \\ 0 & \beta \end{bmatrix}$, where $\alpha, \beta \in (0, 1]$[18]. Finally the eigendecomposition of $\mathbf{G}_\mathbf{w}^{test}$ is performed, yielding the first two eigenvalues λ_{w1} and λ_{w2} as function of $\mathbf{W}$.

In Fig. 9 are depicted the value of the first and the second eigenvalues of $\mathbf{G}_\mathbf{w}^{test}$, respectively, as a function of α and β in the predefined interval. In Fig. 10, as instead, it is

[16] A stretching or compression transformation by a scalar matrix k**I** leaves the eigenvectors unchanged, yet it modifies the eigenvalues.

[17] It is possible to demonstrate that diagonalizable matrices share the same eigenvector matrix **S** if and only if **AB** − **BA** = 0, that is, if they commute (*Strang, 1976*). The result holds also for normal matrices **N**, that is, matrices where **N** commutes with $\mathbf{N}^H$ (*Wilkinson, Wilkinson & Wilkinson, 1965*).

[18] We note that the eigenvalues of a diagonal matrix are the diagonal entries, *i.e.*, α and β, and the eigenvectors are the canonical basis in $\mathbb{R}^m$.

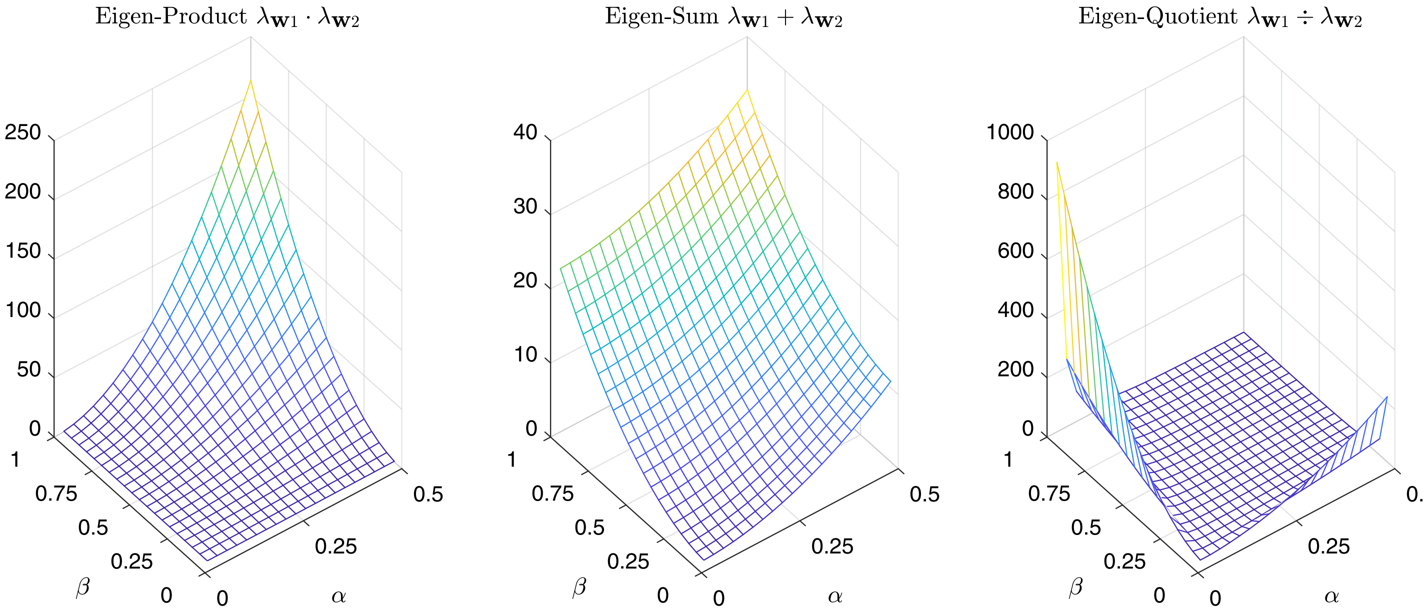

Figure 11 ***Product, Sum*** **and** ***Quotient*** **of the two eigenvalues of** $\mathbf{G}_{\mathbf{w}}^{test}$ **as a function of** α **and** β**.**
Full-size DOI: 10.7717/peerj-cs.1106/fig-11

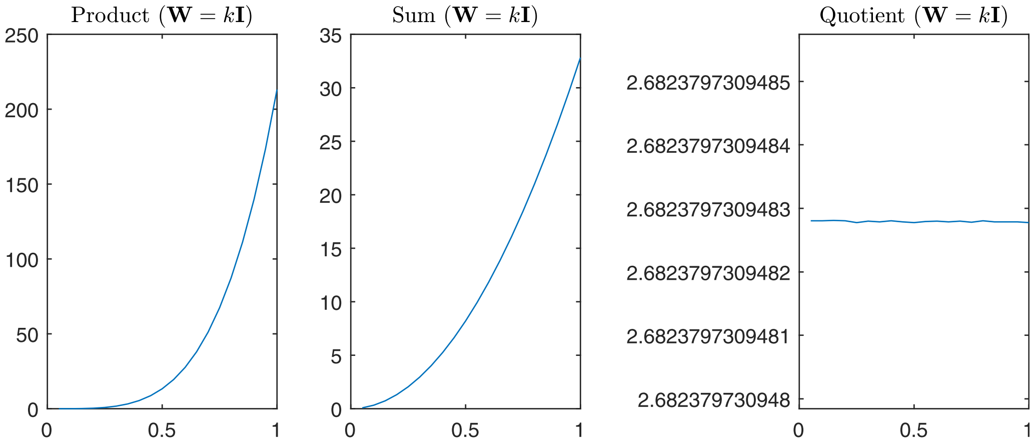

Figure 12 ***Product, Sum*** **and** ***Quotient*** **of the two eigenvalues of** $\mathbf{G}_{\mathbf{w}}^{test}$ **in the case** $\alpha = \beta = k$**,** ***i.e.,*** $\mathbf{W} = k\mathbf{I}$**.**
Full-size DOI: 10.7717/peerj-cs.1106/fig-12

reported the value of the first and second eigenvalues in the case $\alpha = \beta = k$, that is the case $\mathbf{W} = k\mathbf{I}$.

For completeness in Fig. 11 are reported the *sum*, the *product*, and the *quotient* of the first two eigenvalues of $\mathbf{G}_{\mathbf{w}}^{test}$, while in Fig. 12 we have the same operations in the case of $\alpha = \beta = k$.

CONCLUSION

In solving real-world problems in pattern recognition we may incur in a complex representation of objects with the need of a custom-based dissimilarity measure whose components are (sub)-dissimilarities tailored on the nature of the object at hand.

Moreover, the starting space can be non-metric and standard machine learning algorithms cannot operate directly due to the absence of a vectorial space endowed with some well-defined norm. The dissimilarity template can be a weighted Euclidean distance where weights are learned by exploiting a metric learning paradigm. Often, in real-world applications, the adopted custom-based dissimilarity measure leads to non-Euclidean dissimilarity matrices. The non-Euclidean behavior can be suitably measured by studying the spectrum of the related Gram matrix. The adopted framework shows how the (sub)-dissimilarity measure adopted can affect the Euclidean behavior and how a weighting scheme can suitably address this phenomenon. The weighting scheme concerns the spectra of the underlying dissimilarity, but only in some simple cases the problem can be addressed theoretically. Alongside the present work of a more theoretical nature, as regards the future directions, we have planned to evaluate the impact of the non-metricity of the dissimilarity matrices in some real-world applications (*e.g.*, predictive maintenance) and as a correction expressed directly in the objective function (in line with our theoretical discussion) of an optimization system impacts on the performance of a classification system in terms of generalization capabilities.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

The authors received no funding for this work.

Competing Interests

Alessio Martino is an Academic Editor for PeerJ Computer Science.

Author Contributions

- Enrico De Santis conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Alessio Martino conceived and designed the experiments, performed the experiments, analyzed the data, performed the computation work, authored or reviewed drafts of the article, and approved the final draft.
- Antonello Rizzi analyzed the data, authored or reviewed drafts of the article, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

This literature review does not contain data or code.

REFERENCES

Bar-Hillel A, Hertz T, Shental N, Weinshall D. 2003. Learning distance functions using equivalence relations. *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. 11–18.

Beals R, Krantz DH, Tversky A. 1968. Foundations of multidimensional scaling. *Psychological Review* **75(2)**:127–142 DOI 10.1037/h0025470.

Belkin M, Niyogi P. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* **15(6)**:1373–1396 DOI 10.1162/089976603321780317.

Bellet A, Habrard A, Sebban M. 2013. A survey on metric learning for feature vectors and structured data. *ArXiv* DOI 10.48550/arXiv.1306.6709.

Bengio Y, Courville A, Vincent P. 2013. Representation learning: a review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **35(8)**:1798–1828 DOI 10.1109/TPAMI.2013.50.

Bishop CM. 2006. *Pattern recognition and machine learning*. First Edition. Berlin: Springer.

Borg I, Groenen PJF. 2005. *Modern multidimensional scaling: theory and applications*. Second Edition. Berlin: Springer Science & Business Media.

Boyd S, Vandenberghe L. 2004. *Convex optimization*. Cambridge: Cambridge University Press.

De Santis E, Arnò F, Martino A, Rizzi A. 2022. A statistical framework for labeling unlabelled data: a case study on anomaly detection in pressurization systems for high-speed railway trains. In: *2022 International Joint Conference on Neural Networks (IJCNN)*. In press.

De Santis E, Arnò F, Rizzi A. 2022. Estimation of fault probability in medium voltage feeders through calibration techniques in classification models. *Soft Computing* **26(15)**:7175–7193 DOI 10.1007/s00500-022-07194-6.

De Santis E, Livi L, Sadeghian A, Rizzi A. 2015. Modeling and recognition of smart grid faults by a combined approach of dissimilarity learning and one-class classification. *Neurocomputing* **170**:368–383 DOI 10.1016/j.neucom.2015.05.112.

De Santis E, Martino A, Rizzi A, Frattale Mascioli FM. 2018. Dissimilarity space representations and automatic feature selection for protein function prediction. In: *2018 International Joint Conference on Neural Networks (IJCNN)*. Piscataway: IEEE, 1–8.

De Santis E, Rizzi A, Sadeghian A. 2018. A cluster-based dissimilarity learning approach for localized fault classification in smart grids. *Swarm and Evolutionary Computation* **39(3)**:267–278 DOI 10.1016/j.swevo.2017.10.007.

Deza MM, Deza E. 2009. *Encyclopedia of distances*. First Edition. Berlin, Heidelberg: Springer.

Di Noia A, Martino A, Montanari P, Rizzi A. 2020. Supervised machine learning techniques and genetic optimization for occupational diseases risk prediction. *Soft Computing* **24(6)**:4393–4406 DOI 10.1007/s00500-019-04200-2.

Duin RPW, Pękalska E. 2010. Non-euclidean dissimilarities: causes and informativeness. In: Hancock ER, Wilson RC, Windeatt T, Ulusoy I, Escolano F, eds. *Structural, Syntactic, and Statistical Pattern Recognition*. Berlin, Heidelberg: Springer Berlin Heidelberg, 324–333.

Duin RPW, Pękalska E. 2012. The dissimilarity space: bridging structural and statistical pattern recognition. *Pattern Recognition Letters* **33(7)**:826–832 DOI 10.1016/j.patrec.2011.04.019.

Duin RPW, Pękalska E, Loog M. 2013. Non-euclidean dissimilarities: causes, embedding and informativeness. In: Pelillo M, ed. *Similarity-Based Pattern Analysis and Recognition*. London: Springer London, 13–44.

D'urso P, Massari R. 2019. Fuzzy clustering of mixed data. *Information Sciences* **505(3)**:513–534 DOI 10.1016/j.ins.2019.07.100.

Fulton W. 2000. Eigenvalues, invariant factors, highest weights, and schubert calculus. *Bulletin of the American Mathematical Society* **37(3)**:209–249 DOI 10.1090/S0273-0979-00-00865-X.

Gärdenfors P. 2004. *Conceptual spaces: the geometry of thought*. Cambridge: MIT press.

Giuliani A. 2017. The application of principal component analysis to drug discovery and biomedical data. *Drug Discovery Today* **22(7)**:1069–1076 DOI 10.1016/j.drudis.2017.01.005.

Goldberg DE. 1989. *Genetic algorithms in search, optimization and machine learning*. First Edition. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.

Goldfarb L. 1984. A unified approach to pattern recognition. *Pattern Recognition* **17(5)**:575–582 DOI 10.1016/0031-3203(84)90056-6.

Golub GH, Van Loan CF. 2012. *Matrix computations*. Vol. 3. Baltimore: JHU Press.

Gower JC, Legendre P. 1986. Metric and euclidean properties of dissimilarity coefficients. *Journal of Classification* **3(1)**:5–48 DOI 10.1007/BF01896809.

Granato G, Martino A, Baldini L, Rizzi A. 2020. Intrusion detection in wi-fi networks by modular and optimized ensemble of classifiers. In: *Proceedings of the 12th International Joint Conference on Computational Intelligence—NCTA*. INSTICC, SciTePress, 412–422.

Granato G, Martino A, Baldini L, Rizzi A. 2022. Intrusion detection in wi-fi networks by modular and optimized ensemble of classifiers: an extended analysis. *SN Computer Science* **3(4)**:310 DOI 10.1007/s42979-022-01191-0.

Hart PE, Stork DG, Duda RO. 2000. *Pattern classification*. Second Edition. Hoboken: Wiley & Sons.

Horn RA, Johnson CR. 2013. *Matrix analysis*. second Edition. Cambridge: Cambridge University Press.

Hu J, Yan C. 2007. Predicting protein subcelluar localizations using weighted euclidian distance. In: *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*. Piscataway: IEEE, 1370–1373.

Jain AK, Duin RPW, Mao J. 2000. Statistical pattern recognition: a review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22(1)**:4–37 DOI 10.1109/34.824819.

Jain AK, Murty MN, Flynn PJ. 1999. Data clustering: a review. *ACM Computing Surveys (CSUR)* **31(3)**:264–323 DOI 10.1145/331499.331504.

Jimenez S, Gonzalez FA, Gelbukh A. 2016. Mathematical properties of soft cardinality: enhancing jaccard, dice and cosine similarity measures with element-wise distance. *Information Sciences* **367(1)**:373–389 DOI 10.1016/j.ins.2016.06.012.

Kedem D, Tyree S, Sha F, Lanckriet G, Weinberger KQ. 2012. Non-linear metric learning. In: Pereira F, Burges C, Bottou L, Weinberger K, eds. *Advances in Neural Information Processing Systems*. Vol. 25. San Jose: Curran Associates, Inc.

Kim J, Lee Y, Kim H. 2018. Detection and clustering of mixed-type defect patterns in wafer bin maps. *IISE Transactions* **50(2)**:99–111 DOI 10.1080/24725854.2017.1386337.

Kulis B. 2012. Metric learning: a survey. *Foundations and Trends in Machine Learning* **5(4)**:287–364 DOI 10.1561/2200000019.

Kumar N, Kummamuru K. 2008. Semisupervised clustering with metric learning using relative comparisons. *IEEE Transactions on Knowledge and Data Engineering* **20(4)**:496–503 DOI 10.1109/TKDE.2007.190715.

Kurniawati R, Jin JS, Shepherd JA. 1998. Efficient nearest-neighbour searches using weighted euclidean metrics. In: Embury SM, Fiddian NJ, Gray WA, Jones AC, eds. *Advances in Databases*. Berlin, Heidelberg, Berlin Heidelberg: Springer, 64–76.

Lei M, Ling Z-H, Dai L-R. 2010. Minimum generation error training with weighted Euclidean distance on LSP for HMM-based speech synthesis. In: *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. Piscataway: IEEE, 4230–4233.

Lu J, Wang R, Mian A, Kumar A, Sarkar S. 2018. Distance metric learning for pattern recognition. *Pattern Recognition* **75**:1–3 DOI 10.1016/j.patcog.2017.10.032.

Mahalanobis PC. 1936. On the generalised distance in statistics. In: *Proceedings of the National Institute of Sciences of India, 1936*. 49–55.

Martino A, De Santis E, Giuliani A, Rizzi A. 2020. Modelling and recognition of protein contact networks by multiple kernel learning and dissimilarity representations. *Entropy* **22(7)**:794 DOI 10.3390/e22070794.

Martino A, Giuliani A, Rizzi A. 2018. Granular computing techniques for bioinformatics pattern recognition problems in non-metric spaces. In: Pedrycz W, Chen S-M, eds. *Computational Intelligence for Pattern Recognition*. Cham: Springer International Publishing, 53–81.

Mercer J. 1909. Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* **209**:415–446 DOI 10.1098/RSTA.1909.0016.

Münch M, Raab C, Biehl M, Schleif F-M. 2020. Data-driven supervised learning for life science data. *Frontiers in Applied Mathematics and Statistics* **6**:183 DOI 10.3389/fams.2020.553000.

Pękalska E, Duin RPW. 2005. The dissimilarity representation for pattern recognition: foundations and applications. In: *Series in Machine Perception and Artificial Intelligence*. Singapore: World Scientific.

Pękalska E, Duin RPW. 2012. Representation and generalization. *Available at http://www.37steps.com/720/representation-and-generalisation/* (accessed 26 May 2022).

Pękalska E, Duin RPW, Paclík P. 2006. Prototype selection for dissimilarity-based classifiers. *Pattern Recognition* **39(2)**:189–208 DOI 10.1016/j.patcog.2005.06.012.

Pękalska E, Harol A, Duin RPW, Spillmann B, Bunke H. 2006. Non-euclidean or non-metric measures can be informative. In: Yeung D-Y, Kwok JT, Fred A, Roli F, de Ridder D, eds. *Structural, Syntactic, and Statistical Pattern Recognition*. Berlin, Heidelberg: Springer, 871–880.

Rao R. 2012. Weighted Euclidean distance based approach as a multiple attribute decision making method for plant or facility layout design selection. *International Journal of Industrial Engineering Computations* **3(3)**:365–382 DOI 10.5267/j.ijiec.2012.01.003.

Roweis ST, Saul LK. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* **290(5500)**:2323–2326 DOI 10.1126/science.290.5500.2323.

Saul LK, Roweis ST. 2003. Think globally, fit locally: unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research* **4(Jun)**:119–155.

Schölkopf B, Burges CJC, Smola AJ. 1999. *Advances in kernel methods: support vector learning*. Cambridge, MA, USA: MIT Press.

Schultz M, Joachims T. 2003. Learning a distance metric from relative comparisons. In: Thrun S, Saul L, Schölkopf B, eds. *Advances in Neural Information Processing Systems.* Vol. 16. Cambridge: MIT Press.

Shlens J. 2014. A tutorial on principal component analysis. *ArXiv preprint*. DOI 10.48550/arXiv.1404.1100.

Strang G. 1976. *Linear algebra and its applications*. Cambridge: Academic Press.

Tenenbaum JB, De Silva V, Langford JC. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* **290(5500)**:2319–2323 DOI 10.1126/science.290.5500.2319.

Thompson R, Therianos S. 1971. The eigenvalues and singular values of matrix sums and product, viii: displacement of indices. *Aequationes Mathematicae* **7(2)**:219–242 DOI 10.1007/BF01818518.

Vapnik V. 1998. *Statistical learning theory*. New York: Wiley.

Watkins W. 1970. On the singular values of a product of matrices. *Journal of Research, National Bureau of Standards: Mathematics and mathematical physics. Section B* **74(4)**:311 DOI 10.6028/jres.074B.025.

Wilkinson JH, Wilkinson JH, Wilkinson JH. 1965. *The algebraic eigenvalue problem.* Vol. 87. Oxford: Clarendon Press Oxford.

Xing E, Jordan M, Russell SJ, Ng A. 2002. Distance metric learning with application to clustering with side-information. In: Becker S, Thrun S, Obermayer K, eds. *Advances in Neural Information Processing Systems.* Vol. 15. Cambridge: MIT Press.

Zhang F, Zhang Q. 2006. Eigenvalue inequalities for matrix product. *IEEE Transactions on Automatic Control* **51(9)**:1506–1509 DOI 10.1109/TAC.2006.880787.