Pattern Recognition 141 (2023) 109646

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Do all roads lead to Rome? Studying distance measures in the context of machine learning

Eva Blanco-Mallo[a,*], Laura Morán-Fernández[a], Beatriz Remeseiro[b,1], Verónica Bolón-Canedo[a,1]

[a] Department of Computer Science and Information Technologies, Universidade da Coruña, CITIC. Campus de Elviña s/n, 15071 A Coruña, Spain
[b] Department of Computer Science, Universidad de Oviedo. Campus de Gijón s/n, 33203 Gijón, Spain

ARTICLE INFO

Article history:
Received 14 March 2022
Revised 2 March 2023
Accepted 25 April 2023
Available online 26 April 2023

Keywords:
Distance measures
Similarity measures
Classification
Clustering
Machine learning

ABSTRACT

Many machine learning and data mining tasks are based on distance measures, so a large amount of literature addresses this aspect somehow. Due to the broad scope of the topic, this paper aims to provide an overview of the use of these measures in the most common machine learning problems, pointing out those aspects to consider to choose the most appropriate measure for a particular task. For this purpose, the most recent works addressing the subject were reviewed and seven of the most commonly used measures were analyzed, investigating in detail their main properties and applications. Different experiments were carried out to study their relationships and compare their performance. The degradation of the results in the presence of noise was also considered, as well as the execution time required by each measure.

1. Introduction

Distance and similarity measures describe how far or close two objects are. Since many real-world problems are based on finding similarities between groups of objects or populations, the list of knowledge areas that make use of them is very extensive. Some examples are biology, physics, chemistry, geography, ecology, social sciences, anthropology, algebra, statistical mathematics, engineering, and computer science [1]. In particular, within artificial intelligence, and more specifically in machine learning (ML) and data mining, many techniques rely on the use of distance and similarity measures. In fact, the knowledge areas mentioned above often make use of these techniques to tackle different issues.

Some examples of significant subareas of ML and their relationship to distance measures are as follows: (1) in some classification [2] or regression [3] problems, it is necessary to calculate the distance between new examples and those available in the training set, as well as the error in the predictions; (2) clustering algorithms require a similarity measure to group objects according to their attributes [4,5]; (3) in feature selection, the decision criterion is based on the distance between sets [6,7]; (4) in quantification problems, similarity measures are used to compare probability distributions and estimate the distribution of each class [8]; (5) in anomaly detection, the distance between the examples is measured according to a reference distribution to detect outliers [9,10]; (6) in information retrieval, objects similar to a reference object are searched [11,12]; (7) in active learning, distance measurements are used to identify the most representative samples [13]; and (8) in transfer learning, it is necessary to assess the difference between the distributions of the source and target sets [14,15].

As already mentioned, distance and similarity measures are integrated into a large number of machine learning tasks. In these scenarios, the choice of a proper distance measure is, in general, more important for success than the choice of the learning algorithm itself [16]. However, this aspect is not addressed with as much emphasis in the literature, as it depends largely on domain-specific knowledge and is not so easy to generalize. In addition, due to the extensive amount of measures available, the existing knowledge is too broad to address it completely. Therefore, it is common to find research works focused on analyzing a specific

* Corresponding author.:
E-mail addresses: eva.blanco@udc.es (E. Blanco-Mallo), laura.moranf@udc.es (L. Morán-Fernández), bremeseiro@uniovi.es (B. Remeseiro), vbolon@udc.es (V. Bolón-Canedo).
[1] These authors jointly supervised this work.

https://doi.org/10.1016/j.patcog.2023.109646

family of measures or those most used in a specific field or problem. However, it is necessary to highlight that the high theoretical content of many of them makes them hardly accessible to non-experts.

Thus, our intention is to provide a clear and intuitive background on the use of distance measures applied in machine learning problems, emphasizing various aspects or properties to be considered in order to select the most appropriate one according to the characteristics of a particular problem or application domain. For this purpose, several of the most relevant measures used in the field are discussed, providing detailed explanations and analyzing their main advantages and drawbacks. The rest of this paper is organized as follows. Section 2 introduces some historical background related to the concept of distance measures and metrics, and the main general definitions and properties that are interesting to consider in this context. Section 3 presents a summary of the main related works found in the literature in recent years. Section 4 contains the selected measures, with detailed definitions, properties, and main applications. Section 5 covers all the experimentation carried out, including a study of similarity between measures and several performance tests on different ML tasks, such as classification and clustering. Finally, the main conclusions drawn from the present work are given in Section 6.

2. Background

2.1. A brief history

The origins of the concept of a distance function involve two widely known mathematical theories: (1) Euclid's third postulate in his treatise *The Elements*, which states that a circle can be drawn with any center and any radius; and (2) the widely studied Pythagorean theorem, which states that if we know two sides of a right triangle we can solve the third unknown. When René Descartes [17] developed the Cartesian coordinate system in 1637, the theories of both math disciplines, geometry and algebra, were joined to develop the notion of distance that we use nowadays. Almost 200 years later, Cayley [18] initiated the study of n-dimensional geometry and, shortly after, Cauchy [19] was the first to define the Euclidean distance in n-dimensional space. Euclidean distance, also called Pythagorean distance, is one of the most widely used and known distance measures, and is derived directly from the two theorems mentioned above. However, it was not until 1906 when Fréchet [20] first introduced the mathematical notion of distance metric, and the first formal definition of a metric space was given by Haussdorf [21] in 1914, whose definitions are shown below.

2.2. Main definitions

A **metric** on an arbitrary set X is a **distance function** $d : X \times X \to \mathbb{R}$, such that $\forall\, x, y, z \in X$, the following properties are satisfied:

1. **Non-negativity.** The distance between x and y is always a value greater than or equal to zero: $d(x, y) \geq 0$.
2. **Identity of indiscernibles.** The distance between x and y is equal to zero if and only if x is equal to y: $d(x, y) = 0 \iff x = y$.
3. **Symmetry.** The distance between x and y is equal to the distance between y and x: $d(x, y) = d(y, x)$.
4. **Triangle inequality:** The distance between two objects is the shortest distance along any path: $d(x, y) \leq d(x, z) + d(y, z)$.

Conditions (1) and (2) together produce **positive definiteness**. A measure of distance can be seen as a measure of dissimilarity, and when the distance is in the range $[0, 1]$, the calculation of its corresponding **similarity** measure $s(x, y)$ is [1]: $s(x, y) = 1 - d(x, y)$. Many of the most used distance measures are not metrics, failing in many cases the properties of (3) symmetry or/and (4) triangle inequality. Divergences are a superset of distance functions that only require conditions (1) and (2). The lack of symmetry allows greater flexibility in formulation, which can be interesting in some problems, such as clustering [22]. Regarding triangle inequality, a direct consequence of violating it could lead to a lack of optimization or well-definiteness [23]. A **metric space** is just a set equipped with a function d that measures the distance between its elements.

Distance and similarity measures can be classified into two main groups: those based on **geometric properties** and those based on **probability distributions**. This paper focuses on the former ones, which are intended to be used in the context of Euclidean spaces, where only the positions of the coordinates of points in the related space are taken into account to calculate distances. The Minkowski distance is a generalized metric distance based on this idea, and is formulated as follows:

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

where $\mathbf{x} = \{x_1, ..., x_n\}$ and $\mathbf{y} = \{y_1, ..., y_n\}$ are two random vectors defined over a feature space $\mathcal{X}$. In the case of $p = 1$ the Manhattan distance is obtained, for $p = 2$ the Euclidean distance, and for $p = \infty$ the Chebyshev distance. If, in addition to considering the positions in space, we want to take into account the distribution of the data, there is one specific distance that stands out: the Mahalanobis distance. However, despite being a metric because it fulfills all the required properties, it is considered a special case of Bregman divergence. Bregman divergences are a family of divergences that can be defined for both general vectors and probability distributions. Being ϕ a differentiable strictly convex function, the Bregman divergence is defined as follows:

$$D_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\phi(\mathbf{y}) \rangle, \tag{2}$$

where $\nabla\phi(\mathbf{y})$ represents the gradient vector of ϕ evaluated at $\mathbf{y}$ and $\langle . \rangle$ the inner product. Depending on ϕ, different divergences can be obtained, such as squared Euclidean and Mahalanobis distances.

2.3. Relevant properties

Note that any function that fulfills at least the first two properties defined in the previous section can be considered a distance measure, so the number of functions that can be defined is infinite. In fact, there are a huge number of widely known and studied distance measures in the literature. Although they all have the same goal, their focus, and formulation can be tremendously different. Therefore, when selecting the distance measure that best solves a particular problem, it is also interesting to consider the following properties:

- **Invariance under transformations.** A distance function d is invariant under the class of transformations $\mathcal{T}$ if $d(h(x), h(y)) = d(x, y)$, $\forall h \in \mathcal{T}$.
  - Translation invariance: $d(c + x, c + y) = d(x, y)$, $\forall c \in \mathbb{R}$.
  - Rotation invariance: $d(\theta x, \theta y) = d(x, y)$, for any angle θ.
  - Scale invariance: $d(cx, cy) = d(x, y)$, $\forall c \in \mathbb{R}$.

  If we think about real-world problems, it is important that a distance or similarity measure presents certain invariance within determined transformations. Invariance to rotations and translations is convenient in many applications, such as pattern recognition [24,25]. Scale invariance is a desirable property when the absolute scale of the data is unknown or when there is a high intra-class scale variation [26].

Table 1
Summary of the main reviews of distance and similarity measures in the context of machine learning. ED: Euclidean distance. MAND: Manhattan distance. CD: cosine distance. MAHD: Mahalanobis distance. CORD: correlation distance. CAD: Canberra distance. BCD: Bray-Curtis distance. GEN: general purpose. CLA: classification. CLU: clustering. AD: anomaly detection. IR: information retrieval.

	Field	ED	MAND	CD	MAHD	CORD	CAD	BCD
Deza and Deza [1]	GEN	√	√	√	√	-	√	√
Cha [29]	GEN	√	√	√	-	-	√	√
Choi et al. [30]	GEN	√	√	√	-	-	√	√
Todeschini et al. [31]	GEN/CLAS	√	√	√	√	√	√	√
Chomboom et al. [32]	CLAS	√	√	√	√	√	-	-
Hu et al. [33]	CLAS	√	√	√	-	√	-	-
Alfeilat et al. [34]	CLAS	√	√	√	-	-	√	√
Parmezan et al. [35]	CLAS	√	√	√	-	√	√	-
Kocher and Savoy [36]	CLAS	√	√	√	-	-	√	-
Adjabi et al. [37]	CLAS	√	√	-	-	-	-	-
Singh et al. [38]	CLUS	√	√	-	-	-	-	-
Huang [39]	CLUS	√	-	√	-	-	-	-
Shirkhorshidi et al. [4]	CLUS	√	√	√	√	-	-	-
Kumar et al. [40]	CLUS	√	√	√	√	√	√	√
Arora et al. [41]	CLUS	√	√	-	√	-	-	-
Loohach and Garg [5]	CLUS	√	√	-	-	-	-	-
Bisandu et al. [42]	CLUS	-	-	√	-	-	-	-
Chen et al. [43]	CLUS	-	-	√	-	-	-	-
Korenius et al. [44]	CLUS/IR	√	-	√	-	-	-	-
Subhashini and Kumar [45]	CLUS/IR	√	-	√	-	-	-	-
Bekhet and Ahmed [46]	IR	√	√	√	-	-	-	-
Khosla et al. [12]	IR	√	√	-	-	-	-	-
Ayyachamy et al. [47]	IR	√	√	-	√	-	√	√
Quian et al. [48]	IR	√	-	√	-	-	-	-
Vadivel et al. [11]	IR	√	√	√	-	-	-	-
Chen et al. [49]	AD	√	-	-	-	-	-	-
Weller et al. [10]	AD	√	√	√	√	-	-	-

- **Homogeneity.** A distance function d is homogeneous if it has a scaling behavior: $d(cx, cy) = c^k d(x, y)$, $\forall c, k \in \mathbb{R}$. Note that $k = 0$ is scale invariance and $k = 1$ linear homogeneity.
- **Boundedness.** A distance function d is called bounded if there exists a real number c such that $d(x, y) \leq c, \forall c \in \mathbb{R}$. This question becomes especially relevant when the magnitude of the distance or similarity is important. For example, if there is a choice between different actions depending on the degree of similarity between objects, it is necessary to set certain thresholds. If the output takes values in the range $(0, +\infty)$, determining how large is large enough becomes challenging, especially in the presence of outliers.

Additionally, when working in high-dimensional spaces, a series of problems commonly referred to as the curse of dimensionality [27] arise. As the number of attributes or dimensions increases, so does the number of examples necessary for the model to adequately generalize, specifically in an exponential way. The ideal situation is to have all possible combinations of attributes in the training set to make correct inferences for future samples, leading to data sparsity. Another consequence, especially interesting and directly related to distance functions, is that all the distances between pairs of different samples in space end up converging towards the same value as the dimensionality increases [28]. Therefore, the distance obtained in these situations is not able to capture the differences between both sets and becomes useless. Some solutions proposed in the literature to address this problem focus on reducing the dimensionality of the data in a previous step, through feature selection or feature extraction.

3. Related work

Due to the extensive use of distance and similarity measures, numerous comparisons and analyses focused on them can be found in the literature, some of which are summarized in Table 1. The selection of papers to be included in this review follows the procedure described below.

First, the relevant papers published in the last 10 years and focused on the comparative analysis of distance measures within the area of machine learning were considered. Next, the most relevant ones were selected according to two factors: (1) papers published in journals or conferences with a high impact factor and requiring peer reviews, and (2) papers with a significant number of citations (greater than 100). In addition, we included some works that, although they were published more than 10 years ago, meet the aforementioned factors and their contribution is of particular interest to the topic under discussion. Many of these works analyze several properties of the measures and the relationships between them. For example, Deza and Deza [1] offer a very extensive collection of measures in several fields of application and different implementations of the most popular distance measures. Cha [29] also presented a broad collection of distance and similarity measures between probability density functions, grouped according to their syntactic similarities. Choi et al. [30] investigate the correlations of 76 distance measures using a hierarchical clustering technique.

Focused on classification problems, there are many research works dedicated to evaluating the performance of distance measures, especially with the k-nearest neighbor algorithm (kNN). Chomboon et al. [32] studied the use of 11 distance measures, in which Euclidean, Manhattan, cosine, and Mahalanobis distances showed the best performance. The work of Hu et al. [33] is specifically focused on the medical area. For this purpose, they analyzed the behavior of distance measures in different medical datasets containing categorical, numerical, and mixed data. Among the measures analyzed, we can find the Euclidean, Manhattan, correlation, and cosine distances. Todeschini et al. [31] focused on investigating the effects of using different distance measures in classification problems, as well as the relationships between them and their properties. Based on the results obtained with the kNN algorithm on eight real datasets with 18 different measures, they concluded that the Mahalanobis distance does not seem to be useful in classification problems and that the cosine distance also

presents a weak behavior, highlighting the performance of measures such as the Euclidean or Manhattan distances. In the work of Alfeilat et al. [34], following a categorization similar to that presented by Cha [29], the performance of a large number of distance measures in kNN was analyzed and the behavior of some of them in the presence of different noise levels was studied. For their part, Parmezan et al. [35] proposed a variant of kNN and analyze the effect of using 25 distance measures in temporal data from different domains. The best-performing measures were Euclidean, Canberra, correlation, Manhattan, and cosine distances. Kocher and Savoy [36] investigated the best distance measure to solve the author profiling question using a kNN classifier. They stated that the Canberra distance achieved the best overall performance, but that the Manhattan distance had a clear advantage due to the shorter computation time. More recently, Adjabi et al. [37] evaluated various distance measures, including Euclidean and Manhattan, using kNN in a face recognition problem. They concluded that the latter yields the most reliable recognition performance.

In clustering, research related to similarity measures is also a frequent topic. Shirkhorshidi et al. [4] analyzed 12 distance measures (including Euclidean, Manhattan, cosine, and Mahalanobis) in four different clustering algorithms. In the experiments, they used 15 continuous datasets from different fields and focused on investigating performance in low and contexts. According to their results, the Mahalanobis distance is a good option for low-dimensional datasets, while the cosine distance is appropriate for high-dimensional ones. Loohach and Garg [5] discussed the effect of Euclidean and Manhattan metrics in relation to the *k*-means algorithm. Their results suggest that the Manhattan distance generally requires fewer iterations than the Euclidean distance, and therefore lower computational cost. Singh et al. [38] also investigated the use of Minkowski distances on the *k*-means algorithm on dummy data, concluding that the best result was achieved by the Euclidean distance. Arora et al. [41] evaluated the effectiveness of distance measures in the fuzzy c-means clustering algorithm, finding that Euclidean distance performed well in most of the analyzed datasets and that Manhattan distance was also equally suitable for clustering. Bisandu et al. [42] evaluated the adequacy of different measures used in text and data clustering as a function of certain properties. They claimed that, on average, the cosine distance had the highest overall performance with low memory consumption. They also concluded that choosing the best measure depends on factors such as the purpose of the research and the level of disparity of the dataset. For their part, Chen et al. [43] investigated different distance measures in clustering applied to biological data. They proposed a new measure, claiming that it is more robust than others, such as cosine distance. Finally, it is worth mentioning the work of Kumar et al. [40], which analyzes the performance of ten distance measures in different clustering techniques. They concluded that none of them was better in all cases, with large variations depending on the nature of the data and the clustering technique employed.

In the area of information retrieval, Khosla et al. [12] analyzed the performance of Euclidean and Manhattan distances in the context of content-based image retrieval (CBIR) systems. In particular, they represented images using feature vectors composed of color and texture properties, and ordered them by a distance measure to compare them to each other. Based on the experiments performed, they concluded that the Manhattan distance is more accurate than the Euclidean distance. Similarly, Vadivel et al. [11] compared the use of Euclidean, Manhattan, and cosine distances in CBIR applications. In this case, they used distance measures to compare color histograms extracted from the images. As in the previous study, the Manhattan distance showed the best performance. Ayyachamy et al. [47] discussed the performance of different distance measures in medical image retrieval, being Euclidean and Mahalanobis distances the ones that achieved better accuracy. Quian et al. [48] presented a study on the use of Euclidean distance and cosine for nearest neighbor queries in high dimensional data spaces, concluding that both measures presented very similar results. The research works of Korenius et al. [44] and Subhashini and Kumar [45] also contain analysis and comparisons of the use of Euclidean and cosine distances in retrieval information, specifically in document clustering. Subhashini and Kumar conclude that Euclidean distance is not appropriate in high dimensional sparse data environments. Bekhet and Ahmed [46] investigated the effectiveness of the most commonly used measures for video retrieval, with Manhattan and Euclidean distances being the best in terms of retrieval ability.

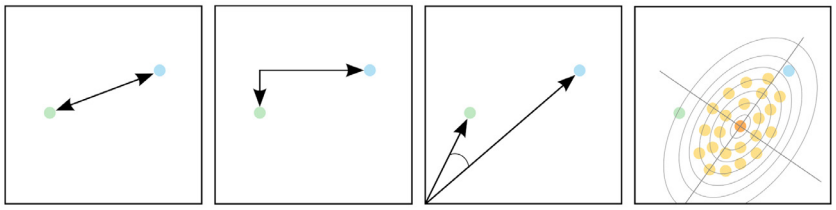
Concerning anomaly detection, some interesting projects were also carried out. Based on the work of Deza and Deza [1], Weller-Fahy et al. [10] distinguished between four categories of distance measures: power distances, distances on law distributions, correlation similarities, and other measures that use combinations of the above. The paper discusses the most commonly used measures within the network intrusion analysis detection field, by selecting 100 of the latest works in the area during the last few years. They emphasized that in most of the papers reviewed, the lack of clarity in defining the measures used is highly frequent. Nevertheless, among the most selected measures, the Euclidean and Manhattan distances stand out. Finally, Chen et al. [49] evaluated the performance of various distance measures to detect electroencephalogram (EEG) anomalies. In order to detect anomalous signals they compared the feature vectors extracted from raw EGG data using distances. The results demonstrated the poor performance of the Euclidean distance.

Finally, it should be noted that significant effort is being devoted to the study of measures in the context of fuzzy systems [50]. These measures differ from the classical ones, which are the ones discussed in the present work, but they are increasingly used in a wide range of pattern recognition, decision-making, clustering, and classification problems [51].

In summary, the Euclidean distance is always present in the comparative studies considered in all the areas addressed (Table 1). Manhattan and cosine distances are also commonly found. Both Canberra and Bray-Curtis distances are analyzed in general thematic studies, but they are not usually selected when carrying out a comparative analysis in more specific domains. Finally, the correlation distance is the least covered, usually in classification, and has a lower performance than the other measures considered. In general, in classification the most recommended measures are Euclidean and Manhattan distances, the latter also standing out in information retrieval. In clustering, there is no general consensus. The only work that includes all the measures in this study concludes that no measure stands out above the rest, since it depends on the type of dataset and algorithm used. Although the Euclidean distance offers the best response in some cases, it is also mentioned that it is not appropriate in high dimensional sparse data environments, where the cosine distance seems to perform better. Concerning anomaly detection, there is a lack of research works that discuss the performance of different measures. Since this task is usually approached using classification and clustering techniques, conclusions drawn from previous research can be applied.

4. Selected measures

We selected seven distance measures that we consider relevant in the field of machine learning, specifically in areas such as classification, clustering, transfer learning, and feature selection, among others. Our selection includes some of the most referenced measures in the related work, along with their main proper-



(a) Euclidean (b) Manhattan (c) Cosine (d) Mahalanobis

Fig. 1. Graphical representation of some of the selected distance measures.

ties and applications, detailed below. In the following definitions, $\mathbf{x} = \{x_1, ..., x_n\}$ and $\mathbf{y} = \{y_1, ..., y_n\}$ are two vectors defined over a feature space $\mathcal{X}$.

4.1. Euclidean distance

The Euclidean distance (ED) is the most commonly used measure of distance between two vectors in geometric problems (see Fig. 1 a). It is often categorized as a Minkowski metric, L_2 distance, or power distance. ED is the basis of many measures of similarity and the standard in many of the most used machine learning algorithms in classification and clustering, such as k-nearest neighbor or k-means [16]. The Euclidean distance is calculated as follows:

$$ED(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} |x_i - y_i|^2} \tag{3}$$

The squared version of the ED is also a Bregman divergence, obtained by replacing $\phi = \|x\|_2$ in Eq. 2, i.e., the inner product's associated norm.

4.2. Manhattan distance

The Manhattan distance (MAND), also known as taxicab, city block, or L1 distance, was proposed by the mathematician Hermann Minkowski in the 19th century [29]. It is a special case of Minkowski's distance for $p = 1$, as mentioned in Section 2.2, and works in a different geometry than the Euclidean one, commonly known as the taxicab geometry. The MAND is calculated based on the sum of the absolute differences in all dimensions:

$$MAND(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} |x_i - y_i| \tag{4}$$

The shortest path is built through horizontal and vertical segments and, therefore, there may be more than one path to cover the shortest distance between two points (see Fig. 1 b), contrary to the Euclidean geometry.

4.3. Cosine distance

The cosine distance (CD) calculates the cosine of the angle between two projected vectors in a multidimensional space. Unlike the previous measures seen so far, this distance does not use the magnitude of the vectors to find their similarity, but only their direction (see Fig. 1 c). This is interesting when the vectors represent datasets of different sizes and we intend to determine their similarity based on their distributions, regardless of their size. It is defined as:

$$CD(\mathbf{x}, \mathbf{y}) = 1 - \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{5}$$

4.4. Mahalanobis distance

Mahalanobis distance (MAHD), first introduced in 1936 [52], was originally developed to calculate distances from a point to a center distribution (see Fig. 1 d), but it is well suited for computing distances between groups or populations using random variables. The MAHD is computed as follows:

$$MAHD(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T C^{-1} (\mathbf{x} - \mathbf{y})}, \tag{6}$$

where C is the covariance matrix of the set to which $\mathbf{x}$ and $\mathbf{y}$ belong, and T denotes the transpose operation. Note that it can be expressed as a Bregman divergence by choosing $\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T C\mathbf{x}$ in Eq. 2.

4.5. Correlation distance

The correlation distance (CORD), proposed by Székely et al. [53], is a measure of dependence between random variables in arbitrary dimensions. It is based on the Euclidean distance between its elements and is derived from other measures such as the variance distance and the covariance distance [54,55]. The empirical CORD between two random vectors is calculated as follows:

$$CORD(\mathbf{x}, \mathbf{y}) = 1 - \frac{(\mathbf{x} - \bar{\mathbf{x}}) \cdot (\mathbf{y} - \bar{\mathbf{y}})}{\|(\mathbf{x} - \bar{\mathbf{x}})\|_2 \|(\mathbf{y} - \bar{\mathbf{y}})\|_2}, \tag{7}$$

where $\cdot$ is the dot product, $\|.\|_2$ represents the Euclidean norm and $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are the mean of the elements $\mathbf{x}$ and $\mathbf{y}$, respectively.

4.6. Canberra distance

The Canberra distance (CAD) was developed in 1966 by Lance and Williams [56]. Although it was initially designed for unsigned numbers, it was later modified for signed real values. It is the weighted version of the Manhattan distance, as it calculates the absolute difference between two vectors and normalizes it by dividing it by the absolute sum of their values. The definition is as follows:

$$CAD(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \frac{|x_i - y_i|}{|x_i| + |y_i|} \tag{8}$$

4.7. Bray-Curtis distance

The Bray-Curtis distance [57] (BCD), also known as the Sorensen distance, is used to quantify the dissimilarity in the composition of two vectors based on the raw counts. Like the Canberra distance, it is a modified version of the Manhattan distance. It is calculated by dividing the absolute differences by their sum with the following formula:

$$BCD(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n} |x_i - y_i|}{\sum_{i=1}^{n} |x_i + y_i|} \tag{9}$$

Table 2
Summary of the properties satisfied by the distance measures analyzed. The check symbol (✓) indicates that the property is satisfied and the cross symbol (✗) that it is not.

	N-N	SYM	IND	TRI	TRA	SCA	ROT	HOM	BND
ED	✓ [39,58]	✓ [39,58]	✓ [39,58]	✓ [39,58]	✓ [58,59]	✗ [58,59]	✓ [58,59]	$k=1$	✗
MAND	✓ [10]	✓ [10]	✓ [10]	✓ [10]	✓ [31]	✗ [31]	✗	$k=1$	✗
CD	✓ [10,39]	✓ [10,44]	✓ [10,44]	✗ [10,44]	✗ [31,58]	✓ [58]	✓ [4]	$k=0$	✓
MAHD	✓ [10]	✓ [10]	✓ [10]	✓ [10]	✓ [31]	✓ [31,60]	✓	$k=0$	✗
CORD	✓	✓	✓	✗ [40]	✓ [40,53]	✓ [40,53]	✗	$k=0$	✓
CAD	✓ [40]	✓ [40]	✓ [40]	✓ [40]	✗	✓ [61]	✗	$k=0$	✗
BCD	✓	✓	✓	✗ [62]	✗	✓	✗	$k=0$	✓*

N-N: non-negativity. SYM: symmetry. IND: Identity of indiscernibles. TRI: triangular inequality. TRA: translation invariance. SCA: scale invariance. ROT: rotation invariance. HOM: homogeneity. BND: bounded. *: bounded only for positive vectors.

4.8. Properties summary

Table 2 contains a summary of the most popular properties in the literature related to distance measures. Specifically, the properties required to consider them as metrics and their invariance to certain transformations. For each measure, it is indicated whether it fulfills the property (✓) or not (✗), accompanied by some references (below). The cases of properties for which no reference was found (i.e., the aforementioned symbols do not have citations below them), were empirically tested by simple experiments following the definitions of the properties given in Section 2.3. The vectors and constants employed were obtained randomly.

According to the properties considered, the Mahalanobis distance (MAHD) is the most fulfilling. However, it is important to remember that taking into account the inverse of the covariance matrix, it is necessary to have more samples than features. Therefore, it may not be suitable in high-dimensional contexts. In addition, the vast majority of implemented algorithms are based on the difference between the two vectors without considering this argument, so it would require additional adjustments for its proper use. The Euclidean (ED) and correlation (CORD) distances meet all but two properties. The ED measure is not invariant to scaling, but it could be solved by normalizing or standardizing, only in case the absolute value of the variables is not relevant. It is also not bounded and its use is not recommended with high-dimensional vectors. Unlike the previous ones, CORD is bounded, but it is neither a true metric nor rotation invariant. However, it has great advantages in addition to those reflected in the table: it can apply variables of any dimension, detect nonlinear associations, and works well in high dimensions. Therefore, it is an excellent candidate when samples with several features are used and their association is a key aspect. Also worth mentioning is the cosine distance (CD), which, like the Manhattan and Canberra distances, satisfies all but three of the properties studied. The CD is not invariant to translations nor does it satisfy the triangular inequality. However, its invariance to scaling and rotations could make it a good choice for clustering. Also, it is important to remember that it can be calculated with vectors of different sizes and its output is bounded.

5. Experimental results

The experiments presented below were designed to analyze the correlation between the seven distance measures considered and evaluate their performance in different machine learning problems. These experiments are:

- **Similarity analysis** to assess the correlation between the results of the distance measures applied to the same data.
- **Discrimination ability between samples** by evaluating the performance of the measures in classification and clustering, two common tasks in areas such as pattern recognition, information retrieval, or anomaly detection. The performance degradation under different noise levels was addressed and the required execution time was assessed with respect to the performance achieved.

For experimentation purposes, we use six real datasets (see Table 3) publicly available in the UCI Machine Learning Repository [63]. Regarding the implementations of the distance measures considered, they are also available in Python libraries such as *scipy* and R libraries such as *philentropy* [64].

5.1. Similarity analysis

In order to explore the relationship between the different distance measures considered, cluster analysis was conducted following the procedure described by Cha [29]. The first step was to randomly generate 1000 random samples, $\mathbf{S} = \{\mathbf{s}_1, ..., \mathbf{s}_{1000}\}$, and a query random sample, $\mathbf{q}$. For this purpose, 100-dimensional vectors were obtained by generating random numbers in the range [0,100]. Then, the distance values $d(\mathbf{s}_i, \mathbf{q}), \forall i = 1, ..., 1000$ were calculated for each of the studied measures, $\mathbf{D} = \{d_1, ..., d_7\}$. The Pearson correlation coefficient (PCC) [65], a statistical test that computes the strength of the relationship between two variables (distance values) and their association with each other, was computed with the results obtained for each distance. PCC returns values in the range $[-1, 1]$, where 0 indicates that there is no correlation between the variables, whereas $+1$ and -1 indicate a strong positive or negative correlation, respectively. The results obtained are shown in Fig. 2(a). Finally, a hierarchical grouping of the correlation results was performed through single linkage clustering (see Fig. 2(b)) using the following formula:

$$d_{HC} = 1 - |correlation(d_j, d_k)|, \forall d_j, d_k \in \mathbf{D} \tag{10}$$

As can be seen in Fig. 2(a), there is a strong correlation ($\geq$ 0.5) between all the measures. As expected, the measures of the Minkowski family, Euclidean (ED) and Manhattan (MAND) distances, are highly correlated (0.95). In this same group are also the Mahalanobis distance (MAHD) and the cosine distance (CD). As previously mentioned in Section 4.4, MAHD is equivalent to ED when the variables are not related and it can be defined as the Euclidean norm of the standardized principal component scores [60].

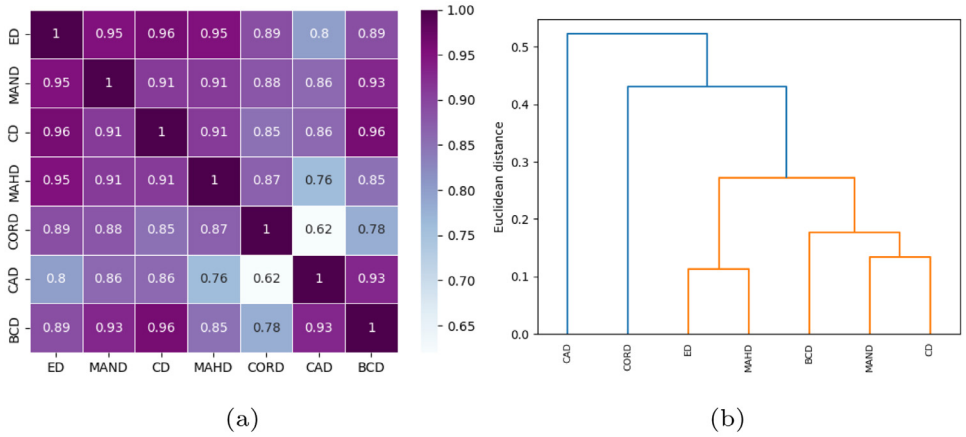


Fig. 2. Similarity between the selected distance measures: (a) Pearson correlation values and (b) dendrogram obtained from clustering.

Likewise, CD is equivalent to ED of normalized vectors. Also, the Canberra (CAD) and Bray-Curtis distances (BCD) have a high correlation with the above measures, as well as with each other. This is because both are considered modifications of the Manhattan distance. The Euclidean norm is also used as the basis in the correlation distance, so the level of correlation between it and ED is also strong, and by extension because of the aforementioned relationships, with MAND, MAHD, and CD. CAD and correlation distance (CORD) are the least alike, and in general with respect to the other measure, as shown in Fig. 2(b).

5.2. Discrimination ability between samples

Our goal is to study the behavior of the distance measures considered in two popular machine learning problems: classification and clustering. The main reason is that classification and clustering algorithms are used as the basis for numerous applications, such as pattern recognition, information retrieval, and anomaly detection, among many others. For this purpose, the six datasets listed in Table 3 were used. The degradation of the result of each distance measure in the presence of different noise levels was also evaluated. Finally, a comparative study is presented to evaluate each measure with respect to two factors: execution time and performance. The Mahalanobis distance (MAHD) makes use of the knowledge of the entire dataset to calculate the covariance matrix when performing its calculations. Therefore, this distance uses three input arguments: two vectors and the covariance matrix. The available implementations of the classification and clustering algorithms used only allow distances with two input arguments. Calculating the covariance-based only on the two input vectors instead of the whole dataset would perturb the results, so the Mahalanobis distance was not tested in the following experiments.

The procedure described by Nettleton et al. [66] was followed to generate different noise levels (Fn). The Fn values used are in the range [0,1] and the generation process is as follows:

- **Attribute noise.** For each attribute, a samples of the test set were randomly selected to be modified following a discrete uniform distribution, with $a = (Fn * b)$, where b is equal to the number of examples and Fn is the noise level. The new attribute value for each of the a samples was replaced with the value generated by a normal distribution using the mean and standard deviation that the attribute presents in the whole set.
- **Class noise.** For each sample in the training set, a reference value v generated from a continuous uniform distribution in the range [0,1] was obtained. If $v < Fn$, the class of the sample was replaced by one of the remaining classes using a discrete uniform distribution.

5.2.1. Classification

To perform the classification tests, we selected one of the best known and most used algorithms in the area, k-nearest neighbor (kNN) [67]. This algorithm works by classifying a sample according to the most common distance between its k nearest neighbors based on a distance measure. We chose this classifier because it makes few assumptions about the data and no parameter tuning is required. The datasets used are those described in Table 3, applying different noise levels. Specifically, $Fn \in [0, 0.9]$ for attribute noise and $Fn \in [0, 0.5]$ for class noise, with increments of 0.1 in both cases. The classification accuracy was used to evaluate the performance of the different measures.

Figs. 3 and 4 show the average accuracy achieved by the different measures for each noise type after 10 repetitions of the kNN algorithm with a different number of neighbors ($k \in \{1, 3, 5, 7, 9\}$). The accuracy value represented in the graph corresponds to the one reached with the best value of k, showing the standard deviation obtained when calculating the mean accuracy with all the k values. Note that, for visualization purposes, the y-axis is scaled differently in each plot. In terms of overall performance, the Canberra distance (CAD) clearly stands out from the rest of the measures. In the Wine dataset, it is the only one that achieves an accuracy higher than 0.9. This dataset includes both integer and real attributes, with particularly high integer values. As far as can be seen, CAD is the measure that is least affected by large-scale features. In this case, the Euclidean distance (ED) is the most affected. With respect to the rest of the datasets, CAD also presents a superior performance in Breast Cancer and Isolet datasets, and the difference with respect to the rest of the measures in the Iris and Lo-Res Spectrometer datasets is < 0.02 and < 0.01, respectively.

The addition of attribute noise (see Fig. 3) causes a large drop in accuracy, as expected. In the Iris, Wine, and Madelon datasets, all measures show a similar fall from the first level. In the Breast Cancer dataset, CAD is the measure that best tolerates the noise, reaching much higher accuracy values than the other measures at all levels, as in the Wine and Isolet datasets. Finally, in the Lo-Res Spectrometer dataset, all the measures show a slight drop up to $F_n = 0.4$. In this dataset, CAD again stands out with respect to the rest of the measures. Class noise is better tolerated, with a very slight decrease in accuracy up to level 0.2 in all datasets with the exception of Madelon (see Fig. 4). As can be seen, some of the previously mentioned conclusions hold. In the Iris dataset, the measures show similar behavior, as well as in Breast Cancer, where CAD stands out slightly. In the Wine dataset, CAD far outperforms

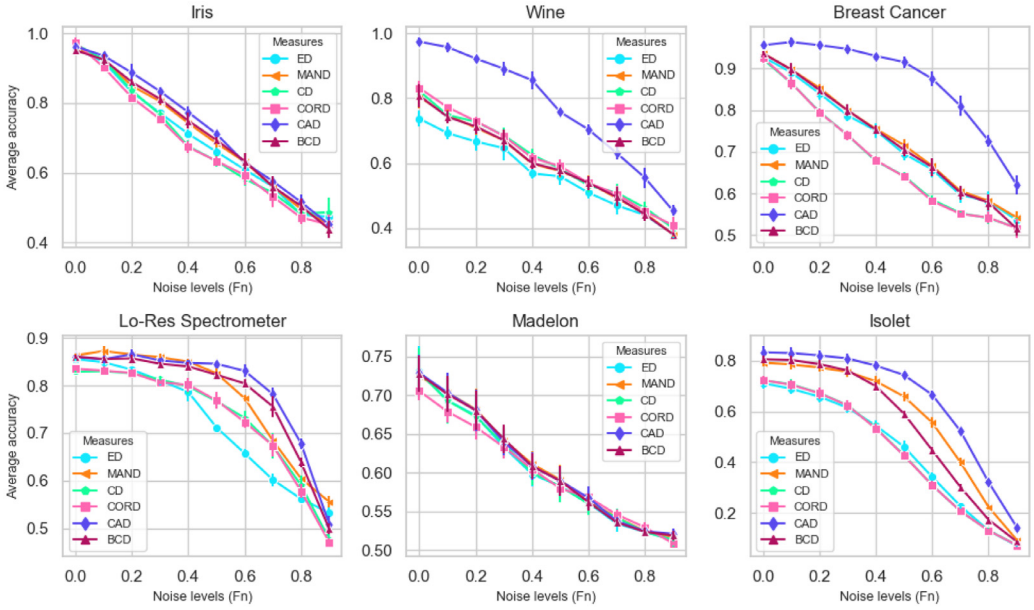


Fig. 3. Average accuracy of kNN per dataset for all the distance measures considered with attribute noise (F_n) after 10 repetitions. Note that the y-axes are scaled differently for better viewing.

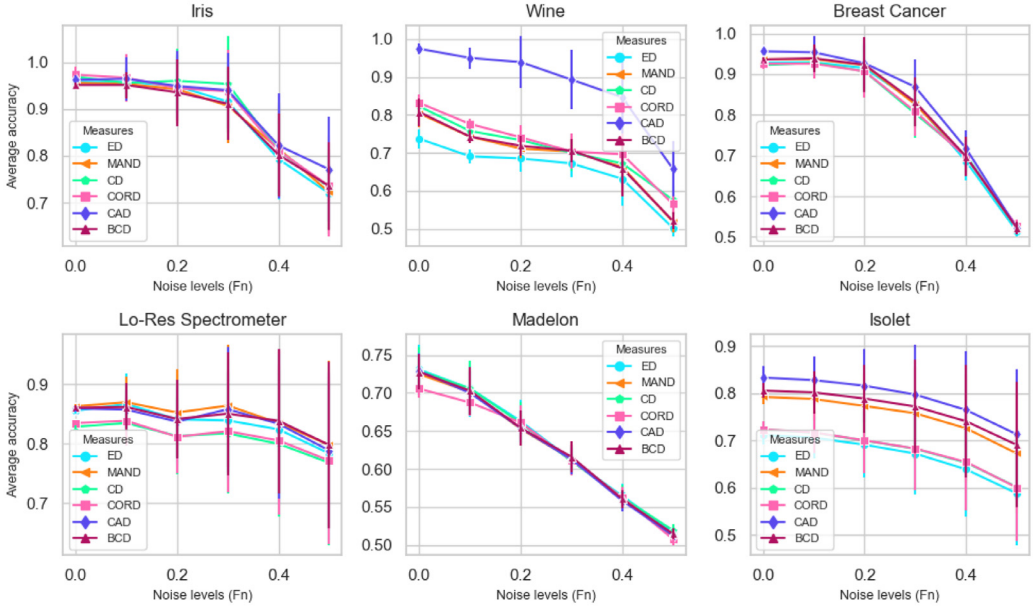


Fig. 4. Average accuracy of kNN per dataset for all the distance measures considered with class noise (F_n) after 10 repetitions. Note that the y-axes are scaled differently for better viewing.

the rest of the measures. Note that CORD and CD show akin behavior on all datasets, as do MAND and BCD.

The influence of the number of neighbors on the result is not significant without the presence of noise. In the Iris, Breast Cancer, and Lo-Res Spectrometer datasets, all the measures have a standard deviation ≤ 0.01. In the other datasets, the mean deviation value oscillates in the range $(0.01 - 0.03)$. With the introduction of noise, the influence increases considerably, especially in the presence of class noise. Overall, the best neighbor values are especially high, with $k \geq 5$ in practically all cases, rising as more noise is introduced.

Finally, in order to explore the statistical significance of our classification results, we analyzed the accuracies achieved by the different distances in each dataset without the presence of noise. For this purpose, we used a Friedman test with Nemenyi's post-hoc test [68]. Fig. 5 presents the critical difference diagrams where groups of distances that are not significantly different (at $\alpha = 0.05$) are connected. The diagrams show the mean ranks for each measure, the higher the rank (further to the right) the better the performance. If a horizontal line connects two or more measures, this means that there is no significant difference between them. As can be seen, there are only significant differences in the Wine and Isolet datasets. In the former, CAD performs significantly better than ED, CD, and MAND. In the latter, the performance of CAD is significantly better than ED, CD, and CORD. In the remaining cases, either no differences are noticed (all measures are grouped on a single horizontal line) or the data are not sufficient to detect them (the horizontal lines overlap). Although there are no significant dif-

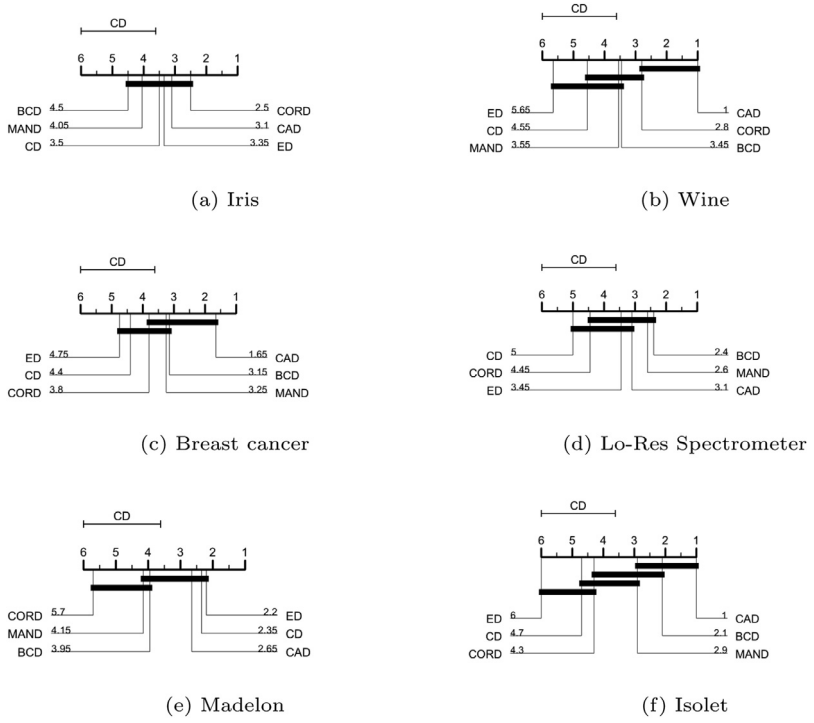


Fig. 5. Critical difference diagrams showing the accuracy obtained by the different measures on each dataset, without the presence of noise and using the kNN algorithm.

ferences, it can be seen that CAD has a much higher performance than the rest, being located further to the right of the diagram in all cases.

5.2.2. Clustering

In clustering, one of the best known and most widely adopted algorithms is *k*-means [67]. However, it is not suitable for our study, since its use with different similarity measures could lead to non-convergence problems. In *k*-means, clustering is performed by minimizing the sum of the squares of the distances between the data and the centroid of the corresponding cluster, i.e., the Euclidean distance. The centroids are selected at each iteration based on the mean of the data in each cluster, changing location step by step until no more changes occur. In order to converge, adding a new cluster center must decrease the objective function. Therefore, both the assignment step and the centroid update step must optimize the same criterion. For this reason, it cannot be used with arbitrary distance measures.

A similar clustering algorithm but that does not present this constraint is *k*-medoids [69]. Its goal is also to minimize the distance between the data relative to the centroid of the clusters, with the difference that *k*-medoids uses real representative data as centroids instead of calculating the mean in each cluster. Consequently, it not only avoids the convergence problem mentioned above but also provides more robustness to outliers. Thus, this is one of the algorithms chosen to perform the clustering experiments. The major drawback of *k*-medoids with respect to *k*-means is its complexity, and hence, its run time.

A hierarchical clustering algorithm was also chosen to evaluate the performance of different distances. This type of method focuses on building a tree, assuming that the clusters are hierarchically structured. Specifically, the complete linkage method (CLM) [70] was selected. Initially, each sample forms its own cluster. Then, the tree is built from bottom to top by merging the most similar clusters considering the maximum distance between the two farthest points in two clusters.

Since unsupervised learning algorithms were used in these experiments, only the case of attribute noise was studied. As in the previous experiment, noise levels in the range [0,0.9] were used and 10 repetitions of each experiment were conducted. The quality of the clustering result was assessed through the *purity*, a popular measure in the field [39,71] that indicates the extent to which clusters contain a single class. More specifically, it computes the coherence of the clustering result by taking into account the number of samples of the majority class per cluster. Let Ω be the set of clusters found by the algorithm and $\mathbb{C}$ the set of classes of the labeled samples, the purity is defined as follows:

$$purity(\Omega, \mathbb{C}) = \frac{1}{N} \sum_{\omega \in \Omega} \max_{c \in \mathbb{C}} |\omega \cap c| \tag{11}$$

where N is the number of labeled samples and $|\omega \cap c|$ represents the number of samples in cluster ω that belong to class c.

Figs. 6 and 7 presents the average results obtained for each dataset. Regarding the purity obtained without noise with *k*-medoids, the greatest difference is found in the Wine, Breast Cancer, and Lo-Res Spectrometer datasets, with the performance of CAD standing out above the rest of the measures. With the complete linkage algorithm, CAD again stands out together with BCD and CORD. It can also be observed that, as the number of features in the dataset increases, the performance of the measures deteriorates with both algorithms. In particular, in the Madelon and Isolet datasets, the purity obtained is very low in general. In addition to the complexity involved in dealing with a large number of features, these two data sets are the ones that contain a greater number of samples. Furthermore, both problems are multivariate and Madelon is non-linear. In terms of noise tolerance, from level 0.1 the purity plummets in all measures in both algorithms when noise is introduced into the datasets. CAD is again the measure that shows a slightly higher tolerance than the others.

Figs. 8 and 9 show the critical difference diagrams generated using the Friedman test with Nemenyi's post-hoc test ($\alpha = 0.05$) using the purity value achieved by the measures in the different datasets, without noise and using the *k*-medoids and complete

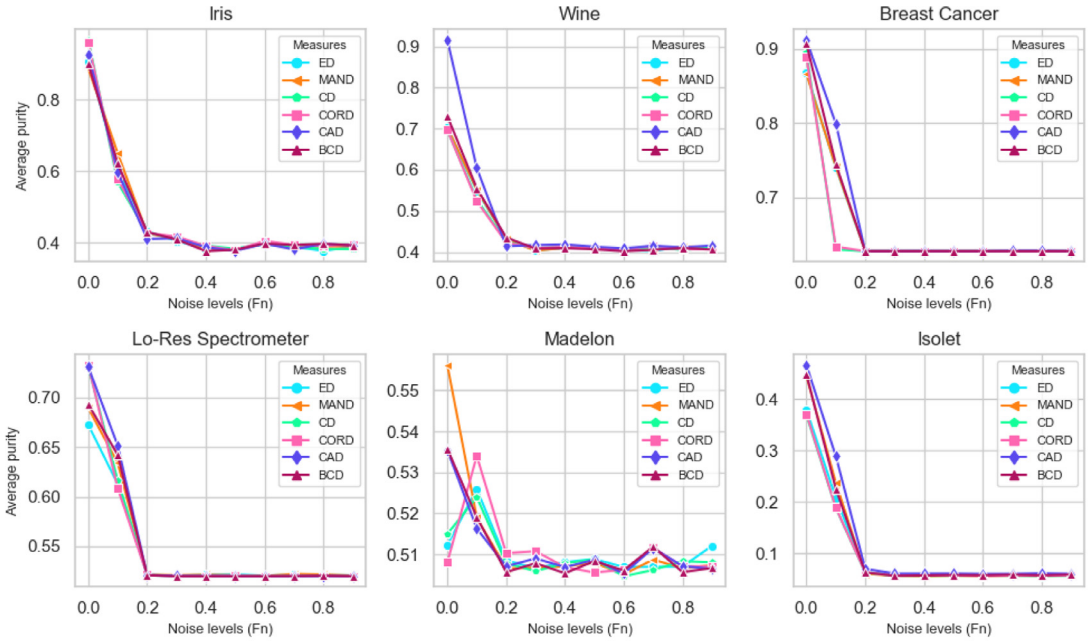


Fig. 6. Average purity of k-medoids per dataset for all the distance measures considered with attribute noise (F_n) after 10 repetitions. Note that the y-axes are scaled differently for better viewing.

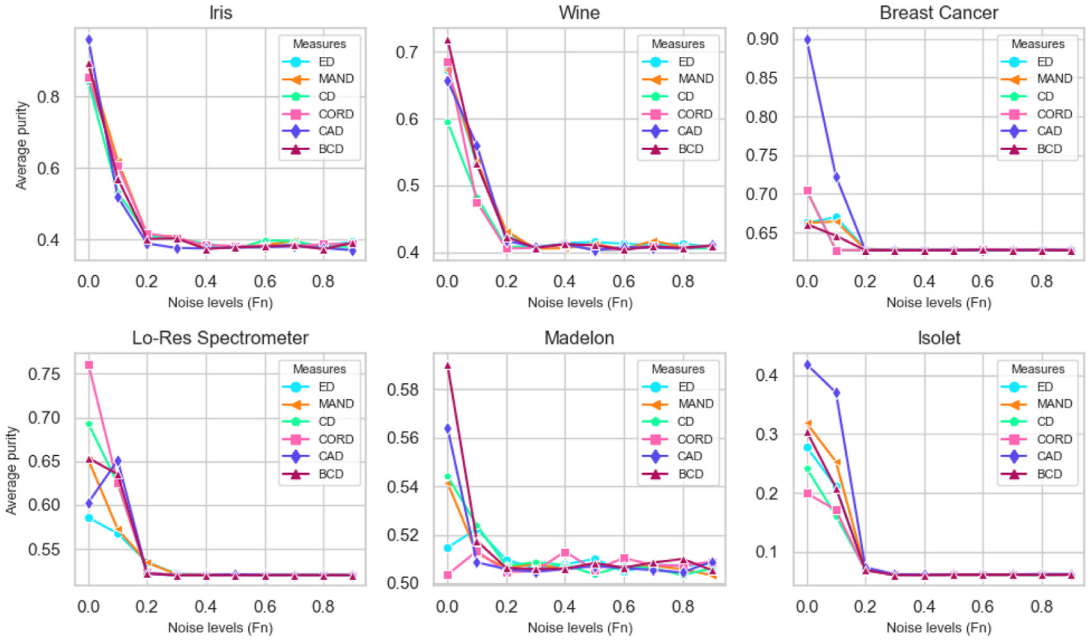


Fig. 7. Average purity of the complete linkage method per dataset for all the distance measures considered with attribute noise (F_n) after 10 repetitions. Note that the y-axes are scaled differently for better viewing.

linkage algorithms, respectively. In the k-medoids scenario, significant differences are only found in the Breast Cancer dataset, where CAD is significantly better than ED, CORD, and MAND. Note that, also in this case, CAD is on the right side of the ranking in all cases. Therefore, although there is no statistical significance over the rest of the measures, its performance is superior. In the case of the complete linkage algorithm, no significant differences between the measures are found in any dataset.

In summary, all measures suffer considerably when noise is introduced into the datasets. Without considering noise, the measure that performs best on average in both clustering methods in the considered datasets is the CAD, followed by BCD, CORD, and CD.

5.3. Performance vs. execution time

When choosing between different distance measures in a machine learning problem, the execution time plays a key role. Therefore, we evaluated each measure in terms of two factors: its performance (accuracy or purity) and its runtime. The performance results compared are those obtained without the presence of noise in the different datasets and the execution times were transformed to the range $[0, 1]$ following min-max normalization.

As can be seen, in kNN (Fig. 10 (a)) and k-medoids (Fig. 10 (b)), the fastest measures on average are Canberra (CAD), Bray-Curtis (BCD), and correlation (CORD) distances, with very similar

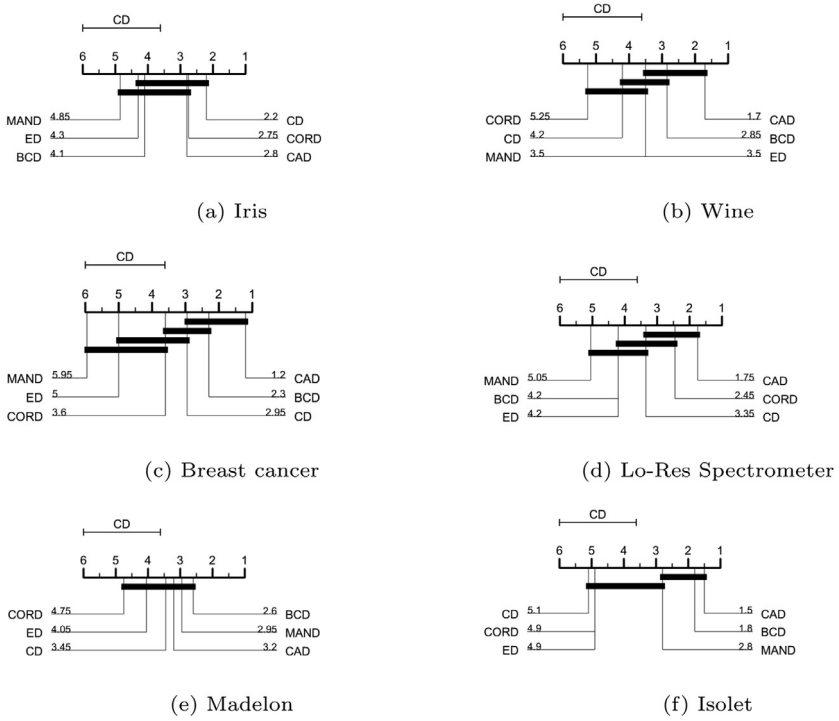


Fig. 8. Critical difference diagrams showing the purity obtained by the different measures on each dataset, without the presence of noise and using the *k*-medoids algorithm.

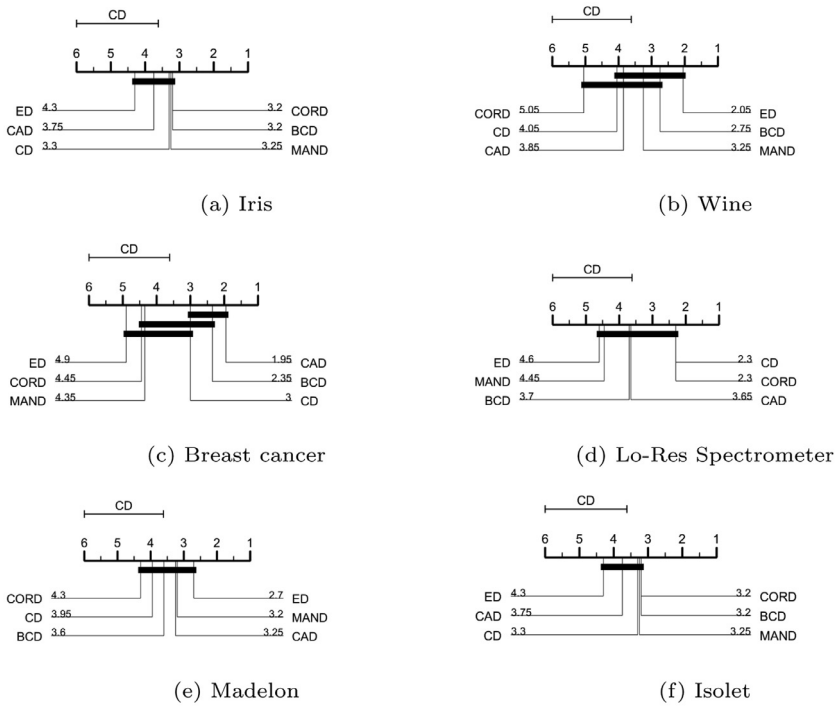


Fig. 9. Critical difference diagrams showing the purity obtained by the different measures on each dataset, without the presence of noise and using the complete linkage algorithm.

execution times. The same applies to the slowest ones, which are Euclidean (ED), Manhattan (MAND), and Cosine (CD) distances. In terms of performance/time, CAD stands out above the rest, obtaining the highest purity and accuracy values more quickly. It is followed by BCD and MAND in classification, and CORD and CD in clustering. However, in the case of CORD, an increase in execution time is noted depending on the number of samples in the dataset, which is not observed in the rest of the measures. On the contrary, ED, MAND, and CD are matching their execution time with the others as this factor increases, although their performance in comparison is lower. In the complete linkage method (Fig. 10 (c)), no clear patterns are observed as in the previous algorithms. CAD

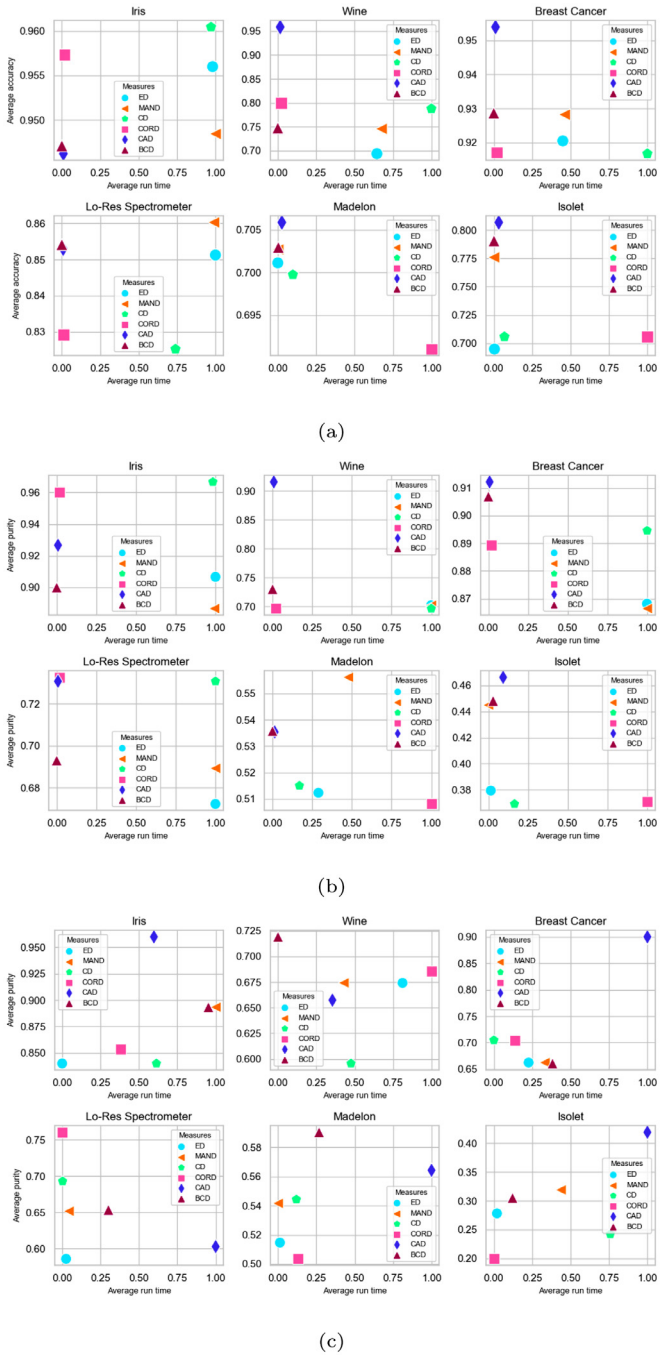


Fig. 10. Average performance *versus* average execution time over 10 repetitions of the kNN algorithm (a), 10 repetitions of the k-medoids algorithm (b), and 10 repetitions of the complete linkage method for the six real datasets considered (see Table 3). Note that the y-axes are scaled differently for better viewing.

Table 3
Properties of the real datasets used. The symbol (+) indicates that the attributes of the dataset are positive values only, while (+,–) indicates that there are both positive and negative values. Attr: Attributes, Cl: Classes, S: Samples, D.Type: Data type.

Dataset	#Attr	#Cl	#S	D.Type
Iris	4	3	150	Real (+)
Wine	13	3	178	Integer, Real (+)
Breast Cancer Wisconsin	30	2	569	Real (+)
Low Resolution Spectrometer	100	9	531	Integer, Real (+,–)
Madelon	500	2	2600	Real (+)
Isolet	617	26	7797	Real (+, –)

is one of the slowest measures, but it is still the best average measure in terms of purity. ED is usually fast on all datasets, but as the number of features increases its purity decreases. This is probably because ED is the measure most affected by the curse of dimensionality.

5.4. Summary results

Table 4 presents a summary of the performance of all the measures in the experiments conducted as a function of different variables. Specifically, accuracy in kNN, purity in *k*-medoids and complete linkage methods, and noise tolerance and execution time in both tasks. The performance of a measure with respect to a given variable is evaluated in the range $[1, 5]$, where a larger number of dots implies better behavior. A distance measure reaching five dots on a variable indicates that it performs well on it, i.e., it leads to high accuracy, exhibits a fast execution time or tolerates noise well, for example. In light of the results, the most suitable measure for classification is the Canberra distance (CAD), followed by the Bray-Curtis (BCD), and the Manhattan (MAND) distances. Euclidean (ED), cosine (CD), and correlation (CORD) distances show the worst behavior, proving to be the least advisable in classification. Regarding clustering, CAD again stands out favorably against the rest, followed by BCD, CORD, and CD. The measures showing the poorest performance in clustering are MAND and ED, the latter being the least preferable in general in both tasks.

In summary, the best performing measure is CAD, since it behaves equally or better than the others in all cases except for execution time in the complete linkage algorithm. Notice that it provides, on average, the highest results in terms of accuracy and purity. Additionally, it shows the best noise tolerance. The main drawback is that it exhibits a low tolerance to clustering noise, although all measures behave particularly poorly in this case.

6. Discussion and conclusion

Due to the relevant role of distance and similarity measures in a multitude of machine learning and data mining tasks, the aim of this paper is to shed light on the different types of measures used, their fundamental properties, and some relevant aspects to be taken into account depending on the needs of each particular problem. For this purpose, we summarized the most relevant publications of the last few years on the subject and selected seven outstanding measures with detailed descriptions, focusing on their main properties. Also, a similarity study between them was presented and their performance on two common ML tasks (classification and clustering) was evaluated.

Since most of the measures are highly correlated with each other, many of them achieved very similar results. Most of the analyzed measures achieved a good performance without the presence of noise, and due to their high correlation, in many cases lead to similar results. In fact, in the vast majority of critical difference diagrams analyzed, no statistical significance is found between the measures. Moreover, a multitude of variants of all of them and examples of successful applications of practically all types of problems can be found in the literature. So we could say that, in a certain sense, all roads could lead to Rome. However, if the guidelines are confusing or mistaken, things get complicated. The more noise is observed in the datasets, the more differences are observed between the performance of the distances analyzed. Although higher noise tolerance is observed in kNN, performance plummets in the clustering algorithms. In fact, none of the measures achieves good results in the Madelon and Isolet datasets in this task. These are the datasets with the largest number of samples and features, the former marked by a high level of redundancy and the latter consisting of a total of 26 classes. The more complex the problem,

Table 4
Summary of the results obtained in the experimentation.

		ED	MAND	CD	CORD	CAD	BCD
kNN	Accuracy	••	•••	••	••	••••	•••
	Attr. Noise Tolerance	•	••	•	•	•••	••
	Cl. Noise Tolerance	••	•••	••	••	••••	•••
	Execution time	•••	•••	•••	•••••	•••••	•••••
k-medoids	Purity	••	••	•••	•••	••••	•••
	Attr.Noise Tolerance	•	•	•	•	••	•
	Execution time	••••	••••	••••	•••••	•••••	•••••
C-linkage	Purity	•	••	••	•••	•••	•••
	Attr. Noise Tolerance	•	•	•	•	••	•
	Execution time	••••	•••	•••	•••	•	••

the lower the results achieved in general. This suggests that preprocessing, such as feature selection, may be of particular relevance to clustering in this type of scenario, no matter what distance is used.

Regarding kNN, although the number of neighbors does not cause a significant difference between the considered measures, the more features the dataset has, the more neighbors are needed to achieve higher accuracy, such as when introducing noise. Our hypothesis is that, the more neighbors are used, the higher the probability of having more correct samples among them, decreasing noise at the local level. Furthermore, due to the curse of dimensionality, using more neighbors implies being able to access a larger information gain, thus increasing the ability to capture the differences between the different neighborhoods. The results drawn from the experimentation reveal that the most advisable measures for classification are Canberra, Bray-Curtis, and Manhattan distances. In the case of clustering, the best results are obtained with Canberra distance, followed by Bray-Curtis, cosine, and correlation distances. In general terms, the overall performance of the Canberra distance is the most remarkable. On the contrary, the Euclidean distance, one of the most used in several applications due to its simplicity, is the one that shows the worst results on average.

The behavior of distance measures when dealing with high-dimensional datasets is a topic attracting increasing attention. Several fields such as bioinformatics, medicine, marketing, and finances, among others, make an active use of machine learning tasks based on distance measures to address their problems. A common denominator in these areas is that they work with complex data types. Therefore, future work would be focused on the integration of these measures in a distributed environment optimized for large-scale data processing, such as Spark, which would bring great improvements in terms of computational cost. In view of the results of the experimentation, the more complex the dataset, the worse the results of the measures in general. However, the concrete reasons why this occurs and how to solve it would require new experiments, which are beyond the scope of this study. Therefore, it would also be interesting to incorporate data pre-processing techniques to evaluate how they affect the behavior of different distance measures in the face of high-dimensional datasets.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The datasets used are publicly available

Acknowledgments

This work has been supported by the National Plan for Scientific and Technical Research and Innovation of the Spanish Government (Grant PID2019-109238GB, subprojects C21 and C22), by the Spanish Ministry of Science and Innovation (Grant FPI PRE2020-092608), and by the Xunta de Galicia (Grant ED431C 2022/44) with the European Union ERDF funds. CITIC, as Research Center accredited by Galician University System, is funded by "Consellería de Cultura, Educación e Universidades from Xunta de Galicia", supported in an 80% through ERDF Funds, ERDF Operational Programme Galicia 2014-2020, and the remaining 20% by "Secretaría Xeral de Universidades" (Grant ED431G 2019/01).

References

[1] M.M. Deza, E. Deza, Encyclopedia of distances, in: Encyclopedia of distances, 2009, pp. 1–583.
[2] B.K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G.R. Lanckriet, On integral probability metrics, \phi-divergences and binary classification, arXiv preprint arXiv:0901.2698 (2009).
[3] A. Botchkarev, A new typology design of performance metrics to measure errors in machine learning regression algorithms, Interdisciplinary Journal of Information, Knowledge, and Management 14 (2019) 45.
[4] A.S. Shirkhorshidi, S. Aghabozorgi, T.Y. Wah, A comparison study on similarity and dissimilarity measures in clustering continuous data, PloS One 10 (12) (2015) e0144059.
[5] R. Loohach, K. Garg, Effect of distance functions on k-means clustering algorithm, International Journal of Computer Applications 49 (6) (2012) 7–9.
[6] X. Chen, X. Chen, H. Wang, Robust feature screening for ultra-high dimensional right censored data via distance correlation, Computational Statistics & Data Analysis 119 (2018) 118–138.
[7] R. Li, W. Zhong, L. Zhu, Feature screening via distance correlation learning, Journal of the American Statistical Association 107 (499) (2012) 1129–1139.
[8] P. Pérez-Gállego, A. Castano, J.R. Quevedo, J.J. del Coz, Dynamic ensemble selection for quantification tasks, Information Fusion 45 (2019) 1–15.
[9] C.C. Phiri, C. Valle, J. Botzheim, Z. Ju, H. Liu, Fuzzy rule-based model for outlier detection in a topical negative pressure wound therapy device, ISA Transactions 117 (2021) 16–27.
[10] D.J. Weller-Fahy, B.J. Borghetti, A.A. Sodemann, A survey of distance and similarity measures used within network intrusion anomaly detection, IEEE Communications Surveys & Tutorials 17 (1) (2014) 70–91.
[11] A. Vadivel, A. Majumdar, S. Sural, Performance comparison of distance metrics in content-based image retrieval applications, in: International Conference on Information Technology, 2003, pp. 159–164.
[12] G. Khosla, N. Rajpal, J. Singh, Evaluation of euclidean and manhanttan metrics in content based image retrieval system, in: International Conference on Computing for Sustainable Global Development, 2015, pp. 12–18.
[13] Z. Wang, B. Du, W. Tu, L. Zhang, D. Tao, Incorporating distribution matching into uncertainty for multiple kernel active learning, IEEE Transactions on Knowledge and Data Engineering 33 (1) (2021) 128–142.
[14] S.J. Pan, I.W. Tsang, J.T. Kwok, Q. Yang, Domain adaptation via transfer component analysis, IEEE Transactions on Neural Networks 22 (2) (2010) 199–210.
[15] Z. Wang, B. Du, Y. Guo, Domain adaptation with neural embedding matching, IEEE Transactions on Neural Networks and Learning Systems 31 (7) (2019) 2387–2397.
[16] J. Friedman, T. Hastie, R. Tibshirani, The elements of statistical learning, volume 1, 2001.
[17] R. Descartes, The geometry of René Descartes: with a facsimile of the first edition, 2012.
[18] A. Cayley, Chapters in the analytical geometry of (n) dimensions, Cambridge Mathematical Journal 4 (1843) 119–127.

[19] A. Cauchy, Mémoire sur les lieux analytiques, CR Acad. Sci. Paris 24 (1847) 885–887.
[20] M.M. Fréchet, Sur quelques points du calcul fonctionnel, Rendiconti del Circolo Matematico di Palermo (1884-1940) 22 (1) (1906) 1–72.
[21] F. Hausdorff, Grundzüge der mengenlehre, volume 7, 1914.
[22] M. Teboulle, P. Berkhin, I. Dhillon, Y. Guan, J. Kogan, Clustering with entropy-like k-means algorithms, in: Grouping Multidimensional Data, 2006, pp. 127–160.
[23] A. l'Hostis, Misunderstanding geographical distances: two errors and an issue in the interpretation of violations of triangle inequality, Cybergeo: European Journal of Geography (2016).
[24] P.Y. Simard, Y.A. LeCun, J.S. Denker, B. Victorri, Transformation invariance in pattern recognitiontangent distance and tangent propagation, in: Neural networks: tricks of the trade, 1998, pp. 239–274.
[25] M. Vlachos, D. Gunopulos, G. Das, Rotation invariant distance measures for trajectories, in: ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2004, pp. 707–712.
[26] E. Bernuau, D. Efimov, W. Perruquetti, Analysis of scale invariance property applying homogeneity, IFAC Proceedings Volumes 47 (3) (2014) 8235–8240.
[27] M. Verleysen, D. François, The curse of dimensionality in data mining and time series prediction, in: International Work-conference on Artificial Neural Networks, 2005, pp. 758–770.
[28] A. Kabán, On the distance concentration awareness of certain data reduction techniques, Pattern Recognition 44 (2) (2011) 265–277.
[29] S.-H. Cha, Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions, International Journal of Mathematical Models and Methods in Applied Sciences 1 (4) (2007) 300–307.
[30] S.-S. Choi, S.-H. Cha, C.C. Tappert, A survey of binary similarity and distance measures, Journal of Systemics, Cybernetics and Informatics 8 (1) (2010) 43–48.
[31] R. Todeschini, D. Ballabio, V. Consonni, Distances and other dissimilarity measures in chemometrics, Encyclopedia of Analytical Chemistry: Applications, Theory and Instrumentation (2006) 1–34.
[32] K. Chomboon, P. Chujai, P. Teerarassamee, K. Kerdprasop, N. Kerdprasop, An empirical study of distance metrics for k-nearest neighbor algorithm, in: International Conference on Industrial Application Engineering, 2015, pp. 280–285.
[33] L.-Y. Hu, M.-W. Huang, S.-W. Ke, C.-F. Tsai, The distance function effect on k-nearest neighbor classification for medical datasets, SpringerPlus 5 (1) (2016) 1–9.
[34] H.A. Abu Alfeilat, A.B. Hassanat, O. Lasassmeh, A.S. Tarawneh, M.B. Alhasanat, H.S. Eyal Salman, V.S. Prasath, Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review, Big Data 7 (4) (2019) 221–248.
[35] A.R. Parmezan, V.M. Souza, G.E. Batista, Time series prediction via similarity search: Exploring invariances, distance measures and ensemble functions, IEEE Access 10 (2022) 78022–78043.
[36] M. Kocher, J. Savoy, Distance measures in author profiling, Information Processing & Management 53 (5) (2017) 1103–1119.
[37] I. Adjabi, A. Ouahabi, A. Benzaoui, S. Jacques, Multi-block color-binarized statistical images for single-sample face recognition, Sensors 21 (3) (2021) 728.
[38] A. Singh, A. Yadav, A. Rana, K-means with three different distance metrics, International Journal of Computer Applications 67 (10) (2013).
[39] A. Huang, Similarity measures for text document clustering, in: New Zealand Computer Science Research Student Conference, volume 4, 2008, pp. 9–56.
[40] V. Kumar, J.K. Chhabra, D. Kumar, Performance evaluation of distance metrics in the clustering algorithms, INFOCOMP Journal of Computer Science 13 (1) (2014) 38–52.
[41] J. Arora, K. Khatter, M. Tushir, Fuzzy c-means clustering strategies: A review of distance measures, Software Engineering (2019) 153–162.
[42] D.B. Bisandu, R. Prasad, M.M. Liman, Data clustering using efficient similarity measures, Journal of Statistics and Management Systems 22 (5) (2019) 901–922.
[43] J. Chen, Y.K. Ng, L. Lin, X. Zhang, S. Li, On triangle inequalities of correlation-based distances for gene expression profiles, BMC Bioinformatics 24 (1) (2023) 1–16.
[44] T. Korenius, J. Laurikkala, M. Juhola, On principal component analysis, cosine and euclidean measures in information retrieval, Information Sciences 177 (22) (2007) 4893–4905.
[45] R. Subhashini, V.J.S. Kumar, Evaluating the performance of similarity measures used in document clustering and information retrieval, in: International Conference on Integrated Intelligent Computing, 2010, pp. 27–31.
[46] S. Bekhet, A. Ahmed, Evaluation of similarity measures for video retrieval, Multimedia Tools and Applications 79 (9) (2020) 6265–6278.
[47] S. Ayyachamy, V.S. Manivannan, Distance measures for medical image retrieval, International Journal of Imaging Systems and Technology 23 (1) (2013) 9–21.
[48] G. Qian, S. Sural, Y. Gu, S. Pramanik, Similarity between euclidean and cosine angle distance for nearest neighbor queries, in: ACM Symposium on Applied Computing, 2004, pp. 1232–1237.
[49] G. Chen, G. Lu, Z. Xie, W. Shang, Anomaly detection in eeg signals: A case study on similarity measure, Computational Intelligence and Neuroscience 2020 (2020) 6925107.
[50] B. Gohain, R. Chutia, P. Dutta, A distance measure for optimistic viewpoint of the information in interval-valued intuitionistic fuzzy sets and its applications, Engineering Applications of Artificial Intelligence 119 (2023) 105747.
[51] M.S. Khan, Q.D. Lohani, Topological analysis of intuitionistic fuzzy distance measures with applications in classification and clustering, Engineering Applications of Artificial Intelligence 116 (2022) 105415.
[52] P.C. Mahalanobis, On the generalized distance in statistics, National Institute of Science of India 2 (1) (1936) 49–55.
[53] G.J. Székely, M.L. Rizzo, N.K. Bakirov, Measuring and testing dependence by correlation of distances, The Annals of Statistics 35 (6) (2007) 2769–2794.
[54] C.F. Mendes, M.W. Beims, Distance correlation detecting Lyapunov instabilities, noise-induced escape times and mixing, Physica A: Statistical Mechanics and its Applications 512 (2018) 721–730.
[55] G.J. Székely, M.L. Rizzo, The distance correlation t-test of independence in high dimension, Journal of Multivariate Analysis 117 (2013) 193–213.
[56] G.N. Lance, W.T. Williams, Computer programs for hierarchical polythetic classification ("similarity analyses"), The Computer Journal 9 (1) (1966) 60–64.
[57] J.R. Bray, J.T. Curtis, An ordination of the upland forest communities of southern wisconsin, Ecological Monographs 27 (4) (1957) 326–349.
[58] A. Strehl, J. Ghosh, R. Mooney, Impact of similarity measures on web-page clustering, in: Workshop on Artificial Intelligence for Web Search, volume 58, 2000, p. 64.
[59] M.-T. Pham, O.J. Woodford, F. Perbet, A. Maki, B. Stenger, R. Cipolla, A new distance for scale-invariant 3D shape recognition and registration, in: International Conference on Computer Vision, 2011, pp. 145–152.
[60] P. Galeano, E. Joseph, R.E. Lillo, The mahalanobis distance for functional data with applications to classification, Technometrics 57 (2) (2015) 281–291.
[61] H.S. Brandi, R. Daroda, A. Olinto, The use of the canberra metrics to aggregate metrics to sustainability, Clean Technologies and Environmental Policy 16 (5) (2014) 911–920.
[62] R. Shyam, Y.N. Singh, Face recognition using augmented local binary pattern and bray curtis dissimilarity metric, in: International Conference on Signal Processing and Integrated Networks, 2015, pp. 779–784.
[63] D. Dua, C. Graff, UCI machine learning repository, 2017. http://archive.ics.uci.edu/ml.
[64] H.-G. Drost, Philentropy: information theory and distance quantification with r, Journal of Open Source Software 3 (26) (2018) 765.
[65] J. Benesty, J. Chen, Y. Huang, I. Cohen, Pearson correlation coefficient, in: Noise reduction in speech processing, 2009, pp. 1–4.
[66] D.F. Nettleton, A. Orriols-Puig, A. Fornells, A study of the effect of different types of noise on the precision of supervised learning techniques, Artificial Intelligence Review 33 (4) (2010) 275–306.
[67] X. Wu, V. Kumar, The top ten algorithms in data mining, 2009.
[68] D.G. Pereira, A. Afonso, F.M. Medeiros, Overview of friedmans test and post-hoc analysis, Communications in Statistics-Simulation and Computation 44 (10) (2015) 2636–2653.
[69] T.S. Madhulatha, Comparison between k-means and k-medoids clustering algorithms, in: International Conference on Advances in Computing and Information Technology, 2011, pp. 472–481.
[70] A.S. Hadi, A new distance between multivariate clusters of varying locations, elliptical shapes, and directions, Pattern Recognition 129 (2022) 108780.
[71] H. Kim, H. Park, Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis, Bioinformatics 23 (12) (2007) 1495–1502.

Eva Blanco-Mallo received her B.A. degree (2012) in Publicity and Public Relations from the University of Vigo (Spain) and her B.S. degree (2019) in Computer Sciences (2019) from the University of A Coruña (Spain). She is currently a Ph.D. student in the Department of Computer Science and Information Technologies of the University of A Coruña. Her research interest include machine learning, transfer learning and recommender systems.

Laura Morán-Fernández received her B.S. (2015) and Ph.D. (2020) degrees in Computer Science from the University of A Coruña (Spain). She is currently an Assistant Professor in the Department of Computer Science and Information Technologies of the University of A Coruña. She received the Frances Allen Award (2021) from the Spanish Association of Artificial Intelligence (AEPIA). Her research interests include machine learning, feature selection and big data. She has co-authored four book chapters, and more than 20 research papers in international journals and conferences.

Beatriz Remeseiro received her B.S. degree (2008), M.S. degree (2010), and Ph.D. degree "Cum Laude with International Honors" (2014) in Computer Science from the University of A Coruña (Spain). After two postdoctoral fellowships from 2015 to 2017, at the INESC TEC Institute for Systems and Computer Engineering, Technology and Science (Portugal) and the University of Barcelona (Spain), she is currently an Associate Professor at the Department of Computer Science of the University of Oviedo (Spain). Her main research interests include computer vision and deep learning. On these topics, she has co-authored 12 book chapters, and more than 70 research papers in international journals and conferences.

Verónica Bolón-Canedo received her B.S. (2009), M.S. (2010) and Ph.D. (2014) degrees in Computer Science from the University of A Coruña (Spain). After a postdoctoral fellowship in the University of Manchester, UK (2015), she is currently an Associate Professor in the Department of Computer Science and Information Technologies of the University of A Coruña. She has extensively published in the area of machine learning and feature selection. On these topics, she has co-authored two books, seven book chapters and more than 100 research papers in international conferences and journals.