Similarity measures for binary and numerical data: a survey

M-J. Lesot* and M. Rifqi*

UPMC Univ Paris 06,
UMR 7606, LIP6,
104, avenue du Président Kennedy,
75016 Paris, France
E-mail: marie-jeanne.lesot@lip6.fr
E-mail: maria.rifqi@lip6.fr
*Corresponding authors

H. Benhadda

Thales Land and Joint,
160 bd de Valmy,
BP 82, 92704 Colombes, France
E-mail: Hamid.BENHADDA@fr.thalesgroup.com

Abstract: Similarity measures aim at quantifying the extent to which objects resemble each other. Many techniques in data mining, data analysis or information retrieval require a similarity measure, and selecting an appropriate measure for a given problem is a difficult task. In this paper, the diverse forms similarity measures can take are examined, as well as their relationships and respective properties. Their semantic differences are highlighted and numerical tools to quantify these differences are proposed, considering several points of view and including global and local comparisons, order-based and value-based comparisons, and mathematical properties such as derivability. The paper studies similarity measures for two types of data: binary and numerical data, i.e., set data represented by the presence or absence of characteristics and data represented by real vectors.

Keywords: similarity measure; presence-absence data; numerical data; distance function; equivalent similarity measures; discrimination power.

Reference to this paper should be made as follows: Lesot, M-J., Rifqi, M. and Benhadda, H. (2009) 'Similarity measures for binary and numerical data: a survey', *Int. J. Knowledge Engineering and Soft Data Paradigms*, Vol. 1, No. 1, pp.63–84.

Biographical notes: Marie-Jeanne Lesot received her PhD in Computer Science from the University of Paris 6 and is currently an Associate Professor in the Department of Computer Science of Paris 6 (LIP6). Her research interests include data mining, clustering, fuzzy logic and similarity measures.

Maria Rifqi received her PhD in Computer Science (Artificial Intelligence) from the University of Paris 6. She is an Assistant Professor in the Department of Computer Science of Paris 6 (LIP6). Her research interests include fuzzy logic, machine learning and particularly similarity measures in data mining.

Hamid Benhadda received his PhD in Mathematical Statistics from the University of Paris 6. He worked for five years as an Engineer in the European Center for Applied Mathematics (ECAM) at IBM France, in the fields of data and text mining and particularly in clustering and similarity measures. He then worked for two years as a Senior Consultant in the computing decisional agency of Transiciel. He is currently a Senior Engineer Specialist in information analysis and data processing within the CeNTAI team at Thales Communications.

1 Introduction

Most machine learning tasks rely on the use of a similarity measure, or a distance function, to compare objects one with another. In a classification task for instance, they make it possible to determine whether the compared objects can be assigned to the same class or not; likewise in clustering tasks, the similarity values directly influence the obtained data subgroups. For a given learning task, the choice of a similarity measure can be as important as the choice of the data representation or the feature selection step. In some cases, e.g., in the case of kernel-based learning methods, it actually even replaces the data representation step (Schölkopf and Smola, 2002).

Informally, similarity measures are functions that quantify the extent to which objects resemble one another. They can be distinguished according to the type of data they apply to. Different measures are used for binary, numerical or structured data. In this paper, we consider the first two data types and propose an overview of the existing similarity measures. We examine the diverse forms they can take, discussing their respective properties and their relationships with one another, highlighting the choice levels that must be addressed to select a measure for a given problem. To that aim, we consider their differences from a semantic point of view, underlining the specific configuration of data that produce major differences, and we also propose numerical tools to quantify objectively their differences, from several points of view.

The paper is organised as follows. In Section 2, we first recall a general definition of similarity measures and introduce the notations used throughout the paper. In Section 3, we then consider the case of binary data, also called set data, i.e., data represented by the presence or absence of characteristics. In Section 4, we consider the case of numerical data, i.e. data represented as vectors of real numbers. For each data type, we examine how the similarity measures can be compared, before concluding in Section 5.

We do not examine the case of structured data, e.g., sequences, trees or graphs that are currently studied in the context of kernel functions and build a specific domain on their own. The interested reader is referred to Schölkopf and Smola (2002) and Shawe-Taylor and Cristianini (2004).

2 General definition of similarity measures

Informally, similarity measures are functions that quantify the extent to which objects resemble one another. They take as arguments object pairs and return numerical values that are all the higher as the objects are alike. As similarity measures are widely used in

many different domains, their terminology varies (they are also named e.g., coefficients of association, resemblance or matching), and a measure may have been proposed by several authors independently, with different names.

More formally and despite the confusion in literature, some frequent requirements for a similarity measure can be exhibited. Denoting $\mathcal{X}$ the data space, also called the universe, a similarity measure is usually defined as follows:

Definition 1: A *similarity measure S* is a function $\mathcal{X} \times \mathcal{X} \to \mathbb{R}$ that verifies the following properties:

- Positivity: $\forall x, y \in \mathcal{X}, S(x,y) \geq 0$.
- Symmetry: $\forall x, y \in \mathcal{X}, S(x,y) = S(y,x)$.
- Maximality: $\forall x, y \in \mathcal{X}, S(x,x) \geq S(x,y)$.

Other properties can be required, as for instance normalisation constraints that impose the measure to take values in the interval [0, 1]. Normalised measures can be deduced from general similarity measures through a normalisation transformation. In the following, we consider the normalised framework.

On the contrary, some of the above mentioned properties are sometimes relaxed, leading to more general definitions. In particular, Tversky (1977), going against metric approaches of similarity, proposed to reject the symmetry constraint. He argues on the directional nature of the similarity relation of the form '*x* is like *y*'. In such comparisons indeed, the two objects do not play the same role; y is considered as a referent whom the other object, x, the subject, is compared to. To model this directionality, non-symmetric measures are necessary. Such measures can either insist on what y has more than x, or on what y has less than x; the former are called satisfiability measures, and the latter inclusion measures. The interested reader is referred to Bouchon-Meunier et al. (1996) for this general framework of object comparison.

In the following, we consider the classic definition of similarity measure as stated in Definition 1 and examine how it can be instantiated for specific cases of $\mathcal{X}$, considering binary data in Section 3 and numerical data in Section 4.

3 Similarity measures for binary data

The first case we consider is $\mathcal{X} = \{0,1\}^p$, i.e., data described by *p* binary attributes. Such data can encode set data, also called presence/absence data, i.e. data described by the set of present characteristics, from a predefined list. For each characteristic, an attribute is defined, that takes value 1 if the object possesses the characteristic, and 0 otherwise. Intuitively, the data comparison will then rely on the number of characteristics shared by the objects, and the number of distinctive characteristics.

Binary data also correspond to a possible representation for nominal attributes, i.e., attributes whose possible values are elements of a predefined list and not from a continuous domain. Indeed, it is possible to replace a nominal attribute taking *k* different values by *k* binary attributes, that all take value 0 except the one corresponding to the appropriate nominal value, that takes value 1.

It is to be underlined that the notion of set data can be extended to the notion of fuzzy data. Instead of being defined as a subset of the characteristics universe, an object is then defined as a fuzzy set (Zadeh, 1965) on this universe. The similarity measures presented in this section can be extended to such fuzzy data (Rifqi, 1996).

3.1 Notations

Given two objects $x = (x_1, \ldots, x_p)$ and $y = (y_1, \ldots, y_p)$ both belonging to $\{0,1\}^p$, let $X = \{i / x_i = 1\}$ and $Y = \{i / y_i = 1\}$ be the set of attributes present in objects x and y respectively. Moreover, let $|\cdot|$ denote the cardinality of a set.

The existing similarity measures for binary data can be expressed as functions of the following four quantities associated to the object couple (x, y) (Lerman, 1970; Gower and Legendre, 1986):

- the number of attributes common to both objects: $|X \cap Y|$, denoted *a*
- the number of attributes present in *x* but not in y : $|X - Y|$, denoted *b*
- the number of attributes present in y but not in x : $|Y - X|$, denoted *c*
- the number of attributes in neither x nor y : $|\overline{X} \cap \overline{Y}|$, denoted *d.*

In the following, for any data couple (x, y), we will equivalently use the notations (x, y), (X, Y) or (a, b, c, d).

It is to be noticed that the sum $a + b + c + d$ is constant and equals the dimension of the data space, $a + b + c + d = p$. Besides, one can note that $a + b = 0 \Leftrightarrow a = b = 0$ only occurs if $x = (0, \ldots, 0)$, i.e., using the set representation, if *X* is the empty set. Likewise $a + c = 0$ corresponds to the case where $Y = \emptyset$.

Existing similarity measures can then be divided into two groups. The first one contains the measures that do not take into account the number of characteristics possessed by none of the objects (i.e., *d*) and only depend on the characteristics present in x or y ; these measures are called type 1 similarity measures and are studied in Section 3.2. The measures that, on the contrary, take *d* into account are called type 2 similarity measures and are examined in Section 3.3.

3.2 Type 1 similarity measures

3.2.1 Measure definitions

Type 1 similarity measures are those that only depend on characteristics present either in x or in y (possibly in both), but are independent of the attributes absent of both objects. Table 1 recalls the definitions of the main type 1 similarity measures. They are discussed in more details hereafter. All these similarity measures are increasing functions of *a*, and decreasing functions of *b* and *c*, modelling the fact that the similarity is all the higher as the objects share common characteristics and as they have few distinctive attributes.

Table 1 Type 1 similarity measures (names can vary according to authors)

Similarity measure	*Notation*	*Definition*
Jaccard (1908)	S_{Jac}	$\frac{a}{a+b+c}$
Dice (1945)	S_{Dic}	$\frac{2a}{2a+b+c}$
Sorensen (1948)	S_{Sor}	$\frac{4a}{4a+b+c}$
Anderberg (1973)	S_{And}	$\frac{8a}{8a+b+c}$
Sneath and Sokal 2 (1973)	S_{SS2}	$\frac{a}{a+2(b+c)}$
Ochiai (1957)	S_{Och}	$\frac{a}{\sqrt{a+b}\sqrt{a+c}}$
Kulczynski 2 (1927)	S_{Ku2}	$\frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$
Rifqi et al. (2000)	S_{FD}	$\frac{F_{FD}(\phi)-F_{FD}\left(\frac{\pi}{2}\right)}{F_{FD}(0)-F_{FD}\left(\frac{\pi}{2}\right)}$

As can be seen from the table, the first five measures follow the same scheme and are defined as fractions of linear combination of *a*, *b*, *c*, more precisely as the quotient between κa and the sum $\kappa a + b + c$ where κ is a multiplicative coefficient. They only differ through this multiplicative coefficient, i.e., vary the relative importance of the common and distinctive characteristics of the two objects. These measures were generalised by the contrast model proposed by Tversky (1977).

Definition 2: Given two positive real numbers α and β, a *Tversky's similarity measure* S_{Tve} is of the form:

$$\forall (x,y) \in \mathcal{X}^2,\ S_{Tve}(x,y) = \frac{a}{a+\alpha \cdot b+\beta \cdot c}. \tag{1}$$

The contrast model defines more general similarity measures than stated in Definition 1, as it does not impose the symmetry property. Symmetry is obtained if and only if $\alpha = \beta$. The previous measures are obtained through $\alpha = \beta = 1/\kappa = 2^{-n}$ and more precisely the Jaccard measure is obtained for $n = 0$, Dice for $n = 1$, Sorensen for $n = 2$, symmetric Anderberg for $n = 3$, Sokal and Sneath 2 for $n = -1$.

As shown in Table 1, there also exist type 1 similarity measures that are not fractions of linear functions of *a*, *b* and *c* and cannot be expressed in the framework of the contrast model. It is the case of the Ochiai and Kulczynski 2 measures, that can respectively be seen as the geometric and the arithmetic means of recall, *R*, and precision, *P*, defined as $R = \frac{a}{a+c}$ and $P = \frac{a}{a+b}$. Both are asymmetric measures of the contrast model. *R* can be interpreted as measuring the extent to which x is included in y, whereas *P* reciprocally measures the extent to which y is included in x. *R* is actually a satisfiability measure and

P an inclusion measure (see p.3). Their combination through the arithmetic and geometric mean in Ochiai and Kulczynski 2 then defines similarity as the extent to which x and y are included in each other.

Besides, the Ochiai measure can be interpreted as the cosine of the angle between x and y represented in the characteristic universe. Thus, it can be seen as the transposition of a classic dot product to the binary data space.

The last similarity measure in Table 1 was introduced by Rifqi et al. (2000), and is based on the Fermi-Dirac function:

$$F_{FD}(\phi) = \frac{1}{1+\exp\left(\frac{\phi-\phi_0}{\Gamma}\right)} \quad \text{with} \quad \phi = \arctan\left(\frac{b+c}{a}\right)$$

where Γ is a positive real number and $\phi_0 \in \left[0, \frac{\pi}{2}\right]$. It was introduced to allow a better control on the similarity measure discrimination power (see p.12) and is commented in more details in the section where this notion is presented.

3.2.2 Indeterminate forms

Most of the previous similarity measures are not defined for all triples (*a*, *b*, *c*) and lead to indeterminate forms that can be solved using the following remarks.

The contrast model measures are indeterminate when $a + b + c = 0$, i.e., when $a = b = c = 0$, that is $X = \emptyset$ and $Y = \emptyset$. Now from the maximality property (see Definition 1) and as all measures are normalised, one can solve the indeterminacy by setting $S(\emptyset,\emptyset) = 1$.

The Ochiai and Kulczynski 2 measures are indeterminate when $a + b = 0$ or $a + c = 0$, i.e., when *X* or *Y* is the empty set. In this case, it is more difficult to assign a relevant value. From the set data point of view, one can consider that the empty set does not resemble any set, and thus assign a similarity value 0 to any data couple (x, y) where *X* is empty and not *Y*, and reciprocally. As previously, the maximality property imposes that the similarity equals 1 when both *X* and *Y* are empty. Thus, one can define $S(\emptyset, Y) = 0$ if $Y \neq \emptyset$ and $S(\emptyset,\emptyset) = 1$.

3.3 Type 2 similarity measures

Type 2 similarity measures are those that take into account all four quantities derived from the objects, i.e., their intersection, set differences and also the intersection of their complementary sets, the *d* component, i.e., the number of characteristics that are possessed by none of the objects to be compared. The major difference with type 1 similarity measures is that for type 2 measures, the size of the universe influences the similarity. Depending on the measures, two objects can, e.g., be more similar in a small universe than in a large one.

3.3.1 Classic measure definitions

Table 2 provides the definition of well-known similarity measures of type 2. Note that some of them normally take values in the interval $[-1,1]$. They were normalised to $[0,1]$ using the classic transformation through the function $f(S) = (S+1)/2$.

Table 2 Type 2 similarity measures (names can vary according to authors)

Similarity measure	*Notation*	*Definition*
Rogers and Tanimoto (1960)	S_{RT}	$\frac{a+d}{a+2(b+c)+d}$
Sokal and Michener (1958)	S_{SM}	$\frac{a+d}{a+b+c+d}$
Sokal and Sneath 1 (1963)	S_{SS1}	$\frac{a+d}{a+\frac{1}{2}(b+c)+d}$
Russel and Rao (1940)	S_{RR}	$\frac{a}{a+b+c+d}$
Yule and Kendall (1950)	S_{YuQ}	$\frac{ad}{ad+bc}$
	S_{YuY}	$\frac{\sqrt{ad}}{\sqrt{ad}+\sqrt{bc}}$

Thus, the formulation of the Yule Q and Yule Y measures may differ from the classic notation. Moreover, through this transformation, the Hamman measure (Hamman, 1961) becomes identical to the simple matching measure, and thus does not appear in the table.

Like type 1 measures, type 2 measures are increasing with a and decreasing with b and c. The role of d varies. For all measures except Russel and Rao, a and d play the same role, i.e., 2 objects are all the more similar as they have the same present characteristics and the same absent characteristics. This implies that the absence of the same characteristics increases the similarity between two objects. These measures then differ by the relative weight of the $(a + d)$ and $(b + c)$ components.

The Russel and Rao measure is more severe. Its numerator reduces to a, which implies that it only bases the resemblance on the shared characteristics. It can be noticed that, as $a + b + c + d = p$ is a constant equal to the size of the universe, Russel and Rao can be simplified to a/p. This implies that it only takes $p + 1$ different values, as a only takes all values from 0 to p.

It can also be noticed that the simple matching measure S_{SM} corresponds to the classic Hamming distance between binary data.

Yule Q and Yule Y are not fractions of linear combinations of the components a, b, c and d. It must be underlined that, as opposed to the other measures, they have a specific behaviour when comparing two objects such that $X \subseteq Y$ (or reciprocally $Y \subseteq X$). Indeed, for such data, Yule Q and Yule Y lead a total similarity, $S = 1$, whereas all other similarity measures (including type 1 measures) provide a result that depends on the cardinal of Y, and that is all the smaller as Y has a low cardinal.

3.3.2 Baulieu's dissimilarity

There exist other type 2 similarity measures, derived from the general framework of Baulieu's presence/absence based dissimilarity coefficient (Baulieu, 1989) that recalls Tversky's definition of type 2 similarity (see Definition 2). Baulieu's dissimilarity coefficients are defined as rational functions:

$$D(x,y) = \frac{b+c}{\alpha' \cdot a + b + c + \beta' \cdot d}$$

with α' and β' two real numbers. Similarity measures can then be derived as $1 - D$.

More precisely, the relationship with Tversky's model is the following: as Baulieu considers dissimilarity rather than similarity, the numerator is defined as $b + c$ instead of a in the contrast model. Regarding the parameter values, the contrast model corresponds to the specific case where $\beta' = 0$, as it does not take into account the d component. Baulieu's α' parameter corresponds to the two parameters α and β of the contrast model, allowing only the special case where the coefficients assigned to b and c are equal. Indeed, Baulieu only allows symmetric measures, whereas the contrast model also encompasses non-symmetric similarity measures.

3.3.3 Indeterminate forms

Among type 2 measures, indeterminate forms can only occur for the Yule Q and the Yule Y measures: the denominator of the other measures contain the d component and cannot equal 0. For Yule Q and Yule Y, indeterminate forms occur when $ad = bc = 0$, which corresponds to different cases, examined in turn in the following:

- When $a = b = c = 0$, i.e., when $X = Y = \emptyset$: this comparison of the empty object with itself can be solved using the maximality property, setting $S(\emptyset,\emptyset) = 1$.
- When $a = b = 0$ or $a = c = 0$, i.e., when only one of the two compared objects is empty: as for Kulczynski 2 and Ochiai, one can consider that the empty set does not resemble any set except itself. Thus we set $S(\emptyset,Y) = 0$ if $Y \neq \emptyset$.
- When $d = b = 0$ or $d = c = 0$: $d = 0$ implies that one of the objects is the point possessing all characteristics, represented as (1,...,1). Thus, the other object is included in it. As $YuY(x,y) = YuQ(x,y) = 1$ if $X \subseteq Y$ or $Y \subseteq X$, by continuity, one can define $S((1,\ldots,1),y) = 1$ if $Y \neq \emptyset$. Regarding $S((1,\ldots,1),(0,\ldots,0))$ indeed, i.e., in the very particular case where the point possessing none of the characteristics is compared with the point possessing all characteristics, it seems more relevant to apply the empty set rule, and to consider that the similarity is zero. These choices imply a discontinuity when going from non-empty sets to the empty one, which is due to the specific form of these similarity measures.

3.4 Criteria for the comparison of binary data similarity measures

In the previous section, similarity measures were categorised into two groups, depending on whether they take into account the characteristics absent of both objects to be

compared or not. They were compared based on their semantics, exhibiting their respective properties in particular cases. These particularities can help selecting an appropriate measure from each type. Regarding the choice between types 1 and 2 measures, the selection strongly depends on the data: Sokal and Sneath (1963) have put forward the situations where type 2 similarity measures are more suitable than type 1. It can be added that all type 2 measures, except Russel and Rao, converge to 1 as *d* grows. Now this is a frequent situations for instance in text mining when a document is represented in the space of words.

In this section, we consider other characterisations of similarity measures, through several properties and relationships existing between the measures. Two types of comparisons are presented, depending on the characteristics of the considered problem. In tasks such as document retrieval, the similarity values obtained when comparing documents one with another are not important as such, only the induced order matters. Indeed, a user is only interested in the documents most similar to her request, the actual similarity values are of no importance. In such cases, the comparison should only deal with the order the measures induce, whereas in a more general case, the similarity values themselves must be taken into account. Section 3.4.1 deals with order-based comparisons, Section 3.4.2 with value-based comparisons.

3.4.1 Order-based comparisons

In many applications such as image retrieval, the user is interested in the list of objects most similar to his request, ignoring the similarity score of each object: as pointed out by Santini and Jain (1999), Payne et al. (1999) and Omhover et al. (2006), the relevant information is contained in the ranking induced by the similarity values, not in the values themselves. Similarly, monotone equivariant cluster analysis (Janowitz, 1979) constitutes a clustering approach where only the similarity rankings matter, and not their numerical values. In such cases, the choice between two similarity measures is of little interest if they both lead to the same ordered lists. To formalise this point, several authors introduced and studied the notion of order equivalence (Lerman, 1967; Janowitz, 1979; Baulieu, 1989; Batagelj and Bren, 1995), defining two measures as equivalent if they induce the same ranking. More formally, this leads to the following definition.

Definition 3: Two similarity measures S_1 and S_2 are *equivalent* if and only if $\forall x,y,z,t \in \mathcal{X}^4, S_1(x,y) < S_1(z,t) \Leftrightarrow S_2(x,y) < S_2(z,t)$.

It has been shown (Batagelj and Bren, 1995; Omhover et al., 2006) that this definition is equivalent to the following one: S_1 and S_2 are equivalent if and only if there exists a strictly increasing function $f : Im(S_1) \to Im(S_2)$ such that $S_2 = f \circ S_1$, where $Im(S) = \{s \in [0,1] / \exists (x,y) \in \mathcal{X}^2, s = S(x,y)\}$.

Classes content

Using the previous definition, several classes of equivalent similarity measures can be exhibited (Omhover et al., 2006), i.e., groups of measures that always lead to the same object ranking when comparing a set of objects with a given reference. The similarities can be grouped in the following classes:

- {Jaccard, Dice, Sorensen, Symmetric Anderberg, Sokal and Sneath 2, Fermi-Dirac}
- {Sokal and Sneath 1, Rogers and Tanimoto, Simple Matching}
- {Yule Q, Yule Y}
- each of the remaining measures forms a class by itself.

For the particular case of Tversky's parametrised similarity measure, it was shown (Omhover et al., 2006) that two Tversky's measures with parameters (α, β) and (α', β') are equivalent if and only if $\alpha.\beta' = \alpha'.\beta$. This in particular implies that all symmetric Tversky's measures, for which $\alpha = \beta$, are equivalent; they belong to the first of the above mentioned classes.

Baulieu (1989) showed similar results for presence/absence based dissimilarity coefficients with linear numerator and denominator: $D_{\alpha,\beta}$ and $D_{\alpha',\beta'}$ are equivalent if and only if they have the same ratio $\alpha/\beta = \alpha'/\beta'$.

Thus, when only the similarity-induced ranking matters, the selection of an appropriate measure reduces to the selection of an appropriate equivalence class.

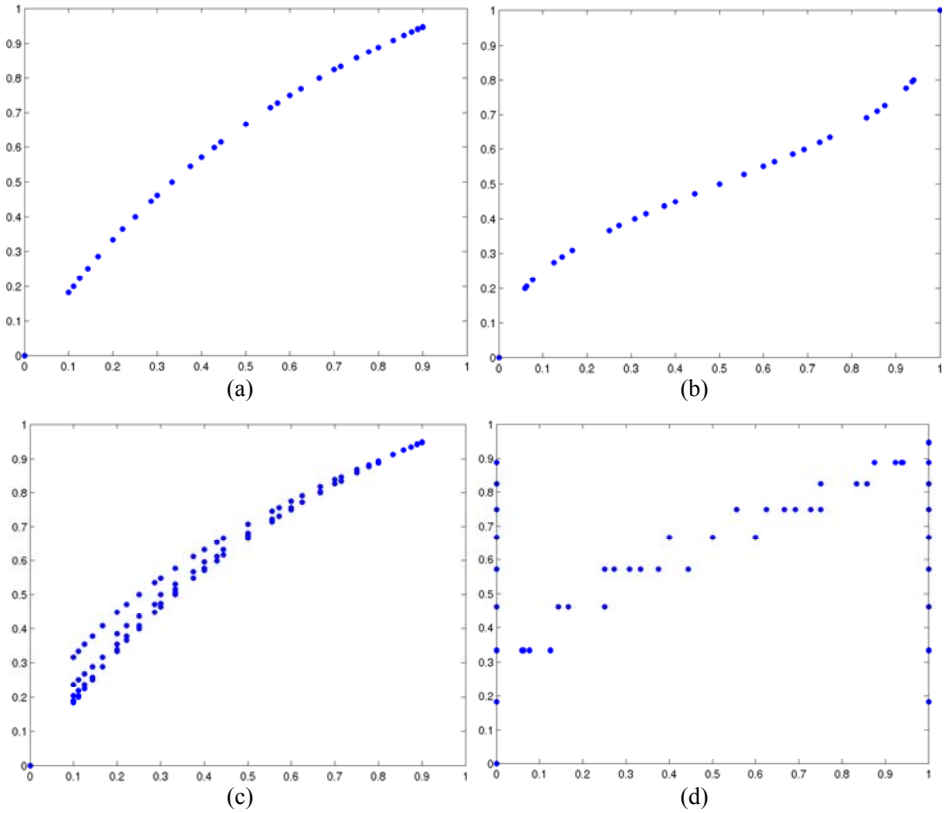
Graphical representation

The function *f* that links two equivalent similarity measures, as the definition of order equivalence states, can be graphically exhibited plotting a measure S_1 against another one S_2. Figure 1 shows such plots for four similarity measure couples. Data were uniformly sampled in $\{0,1\}^{10}$ and the similarity values were computed for all pairs of distinct data. Each point on the graph corresponds to such a couple, and has for coordinates the value S_1 and S_2 take for this data pair. It can be noticed that the plots only contain few points: the similarity measures only take few different values and many data couples lead to the same point on the plot.

From such graphs, in the case of equivalent measures [Figures 1(a) and 1(b)], we observe the function linking the two measures. An analytical study leads to $S_{Dic} = f(S_{Jac})$ with $f(x) = 2x/(1+x)$ and $S_{YuQ} = g(S_{YuQ})$ with $g(x) = x^2/(2x^2 - 2x + 1)$.

For non-equivalent measures [Figures 1(c) and 1(d)], the plot is not a curve, but a scattering of points. For instance a particular value of the Yule Q measure corresponds to several values of the Sokal and Sneath 1 measure and reciprocally [see Figure 1(d)]; thus Yule Q (YuQ in the following) cannot be written as a function of Sokal and Sneath 1 (SS1 in the following). Furthermore, Figure 1(d) illustrates the fact that SS1 and YuQ have very different semantics, due to the large point spreading. The latter is due to two cases: first, if the two objects to be compared are such that $X \subseteq Y$ (or reciprocally), then YuQ assigns a similarity value of 1, whereas the SS1 result depends on the relative sizes of *X* and *Y* and can take small values if the difference is high. Second, when one of the objects to be compared is the empty set, YuQ yields a similarity value of 0, whereas SS1 result depends on the size of the second object, and can take high value if the latter is almost empty.

Figure 1 Graphical pairwise comparisons of some similarity measures (a) S_{Jac} vs S_{Dic} (b) $S_{Y\ uQ}$ vs $S_{Y\ uY}$ (c) S_{Jac} vs S_{Och} and (d) $S_{Y\ uQ}$ vs S_{SS1} (see online version for colours)



The case of such non-equivalent measures has been studied by Rifqi et al. (2008). They propose to quantify the extent to which two measures are equivalent, and compute a degree of equivalence, based on the number and the positions of rank inversions the two measures lead to.

3.4.2 Value-based comparisons

In some machine learning tasks, not only the order induced by the similarity measures matters, the numerical values also have to be taken into account. Other comparison measures between the similarity measures are then to be considered, depending on whether the comparison is made locally, on certain parts of the universe, or globally, on the whole universe.

Global comparison: degree of severity

The characterisation of a measure in terms of degree of severity models the fact that, given two objects and a similarity level, some measures require the objects to have more common characteristics than others to reach the given level.

This can in particular be formalised in the case of the contrast model measures when $\alpha = \beta = 2^{-n}$. Denoting M the average number of characteristics possessed by the two considered objects, $M = (|X| + |Y|)/2$, the measures can be written:

$$S(x, y) = \frac{a}{(1 - 2^{1-n})a + 2^{1-n} M}.$$

Thus, they take values higher than 0.5 if and only if

$$\alpha \geq \frac{M}{1 + 2^{n-1}}. \tag{2}$$

This implies that the lower *n* is, the higher the number of common characteristics must be to conclude to the same value of similarity, which can be interpreted in terms of degree of severity: for a low *n*, the constraints are stronger to judge the similarity of two objects than for a high *n*.

It is to be underlined that this characterisation in terms of degree of severity is complementary to the previous notion of equivalence classes and provides additional information. Indeed, inequality (2) for instance implies that the Jaccard measure is more severe than the Dice measure: even if Jaccard and Dice always lead to the same ranking, as they are equivalent, the values obtained by Jaccard are always smaller than those of Dice. This can also be observed in Figure 1(a): the curve is above the diagonal.

Local comparison: power of discrimination

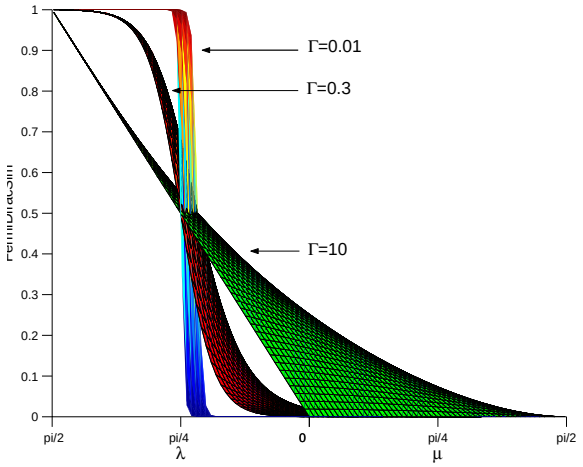
Whereas the degree of severity compares similarity measures globally, the discrimination power is a local comparison tool. It evaluates the sensitivity of the similarity measures with respect to the values of their arguments. The question is whether small variations of the input values, i.e., small variations in the configurations of the two objects to compare, lead to small differences in the similarity values or large ones. Moreover, this question is considered locally, i.e., the discrimination power studies whether such variations occur for high similarity values or for small ones.

More formally, the discrimination power of a similarity measure is a function defined as its derivative with respect to the objects. As previously described, similarity measures are generally expressed as functions of the quadruples (*a*, *b*, *c*, *d*), for which the derivative may be difficult to compute. Rifqi et al. (2003) studied in details this notion in the case of type 1 similarity measures and proposed a change of the coordinate system from (*a*, *b*, *c*) to the spherical system. This enables the reduction of the domain definition and facilitates the discrimination power computation and the comparison of measures in terms of this discrimination power.

The notion of discrimination power can be visually displayed in the case of the Fermi-Dirac measure S_{FM} (see Table 1): the Γ parameter that controls the decreased speed of the measure precisely has a direct influence on the measure discrimination power. This is illustrated in Figure 2 that represents the values of the Fermi-Dirac measure as a function of the spherical coordinates (λ, μ) for different Γ values. When Γ is small, e.g., for $\Gamma = 0.01$, small deviations between the compared objects lead to little difference in the similarity values. From a threshold that equals $\pi/4$ on the

illustrated example, the measure decreases quickly; the discrimination at this point is very high. Finally, for large deviations, the Fermi-Dirac measure again makes little difference. When Γ is high, another behaviour is observed: the resemblance varies linearly, i.e., uniformly, which means that every deviation between the compared objects is penalised in an equal manner.

Figure 2 The Fermi-Dirac measure for several values of Γ (see online version for colours)



Source: Rifqi et al. (2003)

4 Similarity measures for numerical data

The previous section considered similarity measures for binary data and presented an overview on existing measures, highlighting their semantic differences and proposing numerical tools to quantify these differences, from several points of view. In this section, we consider the case of numerical data, i.e., data represented as real vectors. Denoting as previously p the number of characteristics, the data space is thus written $\mathcal{X} = \mathbb{R}^p$. The major difference with the previous binary data is then the fact that the attributes take their values from a continuous domain, and not a discrete one.

In this case, contrary to the case of set data, the relative position of two data cannot be characterised by the four complementary components, defined as their intersection, their set differences and the intersection of their complementary sets. The information is reduced and depends on a single quantity, expressed as a distance or a scalar product between the two vectors. Thus, a normalised similarity measure can only be seen as a function

$$\begin{aligned} \mathcal{S}: \mathbb{R}^p \times \mathbb{R}^p &\rightarrow [0,1] \\ (x,y) &\mapsto S(x,y) \end{aligned}$$

that is an increasing function of a dot product or a decreasing function of a distance. Two kinds of similarity measures can be distinguished: measures derived from dissimilarity

measures and measures based on dot products. In the following, we successively consider these two categories.

4.1 Similarity measures deduced from dissimilarity

A classic definition of similarity consists in deriving a measure from a dissimilarity measure through a decreasing function; this is equivalent to deducing it from a distance function. Three choice levels must then be considered, we examine them in turn in the following. We first recall the distances that are most frequently used in the case of numerical data. We then consider normalisation processes that make it possible to derive normalised dissimilarity measures. Lastly, we study some decreasing functions that can be used to define similarity measures.

4.1.1 Distances

The distances that are most frequently used to compare numerical data are recalled in Table 3; for each of them, the set of points at distance 1 from the origin is depicted in Figure 3 in the case of 2-dimensional data. These functions are as diverse as the similarity measures presented in the previous section. The main difference is the fact that they do not depend on integer values, but on real values, and most of them are derivable functions.

Table 3 Main distances measures for numerical data, and their associated dot products

Name	*Distance*	*Dot product*
Euclidean	$d(x,y)=\sqrt{\sum_{i=1}^{p}(x_i-y_i)^2}$	$\langle x,y\rangle=\sum_{i=1}^{p}x_i,y_i=x^Ty$
Weighted Euclidean	$d(x,y)=\sqrt{\sum_{i=1}^{p}\alpha_i(x_i-y_i)^2}$	$\langle x,y\rangle=\sum_{i=1}^{p}\alpha_i x_i,y_i$
Mahalanobis	$d(x,y)=\sqrt{(x-y)^t\,\Sigma^{-1}(x-y)}$	$\langle x,y\rangle=x^T\,\Sigma^{-1}y$
Minkowski	$d_\gamma(x,y)=\left(\sum_{i=1}^{p}\lvert x_i-y_i\rvert^\gamma\right)^{\frac{1}{\gamma}}$	

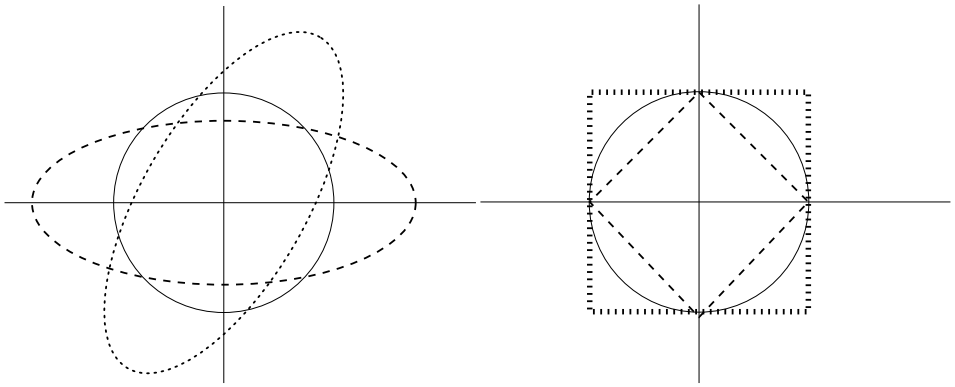
Notes: p denotes the data dimension, $x=(x_i)_{i=1..p}$ and $y=(y_i)_{i=1..p}$ two data points belonging to $\mathbb{R}^p$, $(\alpha_i)_{i=1..p}$ a vector of positive weights, Σ the data covariance matrix and γ a positive parameter.

The most frequently used distance is the Euclidean distance. As illustrated in Figure 3(a), the set of points at distance 1 from the origin is a circle with radius 1. This distance can be modified by the use of weights to rule the relative importance of each attribute in the comparison, leading to the weighted Euclidean distance. Using the latter is equivalent to applying a linear transformation to the data before applying the Euclidean distance. Each attribute is multiplied by a factor $\sqrt{\alpha_i}$. This makes it possible to normalise the attributes, which is in particular necessary when they cover different scales: it avoids that an

attribute taking high values dominate the comparison, the other attributes having only a small influence. As indicated in Figure 3(a), the level lines are then ellipses parallel to the axes.

One can consider more general linear transformations of the data, defining, for any symmetric definite positive matrix A, $d(x,y) = \sqrt{(x-y)^t A(x-y)} = d\left(A^{1/2}x, A^{1/2}x\right)$. The weighted Euclidean distance corresponds to the case of diagonal A matrices. The Mahalanobis distance also belongs to this framework, but uses as A matrix the inverse of the data covariance matrix. Thus, the transformation applied to the data is derived from their statistical distribution. In the case where this matrix is diagonal, the Mahalanobis distance is equivalent to normalising the attributes so that they have mean 0 and standard deviation 1. In the general case, it takes into account the attribute correlations. Figure 3(a) illustrates its level lines that are ellipses.

Figure 3 Points at distance 1 from the origin for several distances, (a) Euclidean (full line), weighted Euclidean (dashes), Mahalanobis (dots) and (b) Euclidean (full line), Manhattan (dashes), Tchebychev (dots)



(a) (b)

Beyond these Euclidean distance variants, one can also use, more generally, Minkowski distances (see Table 3). The case $\gamma = 2$ corresponds to the Euclidean distance. Other classic cases are $\gamma = 1$, called Manhattan distance, and $\gamma \to \infty$, called Tchebychev distance. The latter can also be written $d_\infty(x,y) = \max_{i=1..p} |x_i - y_i|$. Their level lines are indicated in Figure 3(b) and show that they possess other properties than the Euclidean distance. For instance, the Manhattan distance is not a derivable function, but it is more robust than the Euclidean distance, insofar as it increases less rapidly and takes lower values when comparing with outliers.

4.1.2 Normalised dissimilarity measures

The previous distance functions take positive values, they must be normalised to define measures taking values in the interval [0, 1]. The definition of the normalisation process builds a second choice level that can vary the properties of the final similarity measure.

The simplest normalisation consists in applying the function

$$\eta(d) = \frac{d - d_m}{d_M - d_m} \tag{3}$$

where d_m and d_M respectively represent the minimal and maximal observed distances. This transformation guarantees that the values are between 0 and 1, these limits being obtained only in the extreme cases $d = d_m$ and $d = d_M$. The drawback of this approach comes from its sensitivity to outliers: if the maximal observed distance corresponds to an aberrant point, the scale of interesting distance values can be reduced as compared to the scale of observed distances, leading to a distorted normalisation process.

Thus, it can be interesting to consider a richer transformation that introduces saturation effects

$$\eta_Z(d) = \min\left(\max\left(\frac{d-m}{M-m}, 0\right), 1\right) \tag{4}$$

where $Z = (m, M)$ represents the function parameters. The extreme values are defined by the user and not derived from the data, which solves the problem of outlier sensitivity. Furthermore, this transformation offers a saturation property. It guarantees the value 0 for any distance $d \leq m$ and the value 1 for any distance $d \geq M$. Thus the m parameter can be interpreted as a tolerance threshold: distinct points (having a non-zero distance) can be considered as totally similar. The M parameter corresponds to the distance from which point couples are considered as totally dissimilar: the two points at maximal distance are not the only ones to have dissimilarity equals to 1.

4.1.3 Distance-based similarity measures

Similarity measures can then be derived from the previous dissimilarity measures or distance functions, through decreasing functions. One can for instance consider $S(x,y) = 1 - D(x,y)$, where d is a dissimilarity measure defined as $D = \eta_Z(d)$, d being one of the distance functions presented previously. The similarity is thus defined as the complement to 1 of dissimilarity. Indeed, when dissimilarity is 0, the similarity is total, and reciprocally. In this case, similarity is defined as a linear function of the distance.

Other, richer, functions can also be used. One can define similarity as $S(x,y) = f(d(x,y))$, where f is a decreasing function. One can for instance use the Cauchy function (denoted f_c in the following), the generalised Gaussian function (f_{gg}) or an adaptation of the Fermi-Dirac function (Rifqi et al., 2000) (see also p.6), also called sigmoid function and denoted f_{FD} in the following. They are respectively defined as

$$f_c(z) = \frac{1}{1 + \left(\frac{z}{\sigma}\right)^{\gamma}} \tag{5}$$

$$f_{gg}(z) = \exp\left(-\left(\frac{z}{\sigma}\right)^{\gamma}\right) \tag{6}$$

$$f_{FD}(z) = \frac{1}{1+\exp\left(\frac{z-\sigma}{\gamma}\right)} \tag{7}$$

where γ and σ are the parameters of the functions. These functions are represented in Figures 4–7 for different values of their parameters after normalisation by η_Z. Their global behaviour is similar, we comment hereafter the role of the γ and σ parameters.

The influence of the γ parameter can be interpreted in terms of discrimination power (Rifqi et al., 2000) (see also p.12), which measures the sensitivity of the function with respect to specific input values. For instance, for small γ values such as 0.1, the Cauchy and generalised Gaussian functions have a steep decrease for small distance values, which implies they have a high discrimination power: small differences in the input values produce major differences in the output. This implies that distances, even small, lead to an important decrease of the similarity. For γ values higher than 1 (see right part of Figures 4 and 5), the function concavity is different and the discrimination between small values is weak or zero. On the contrary, values around 0.5 are acutely distinguished. The same behaviour can be observed for the sigmoid function (see left part of Figure 6). However, the latter does not present a high discrimination for small distance values. When γ increases, the sigmoid function tends towards a linear function whose discrimination power is uniformly distributed on the interval [0, 1].

The second parameter, σ determines the position of the point with maximal discrimination power, i.e., the point where the derivative is maximal (this point is exactly located at $d = \sigma$ for the Cauchy and Fermi-Dirac functions, a dependence to γ remains in the Gaussian case for which it is located at $\sigma\left((\gamma-1)/\gamma\right)^{\gamma}$). Thus, σ determines the threshold from which the obtained values are small or zero. This is illustrated on the right part of Figures 6 and 7.

These functions thus make it possible to transform a dissimilarity measure into a similarity measure, providing the user with parameters to precisely rule its behaviour. It can be noticed that these functions can in turn be modified to define dissimilarity measures. Defining for instance $D = 1 - f_{gg}(\eta_Z(d))$ makes it possible to control the dissimilarity discrimination power.

Figure 4 Cauchy function [see equation (5)], after normalisation through the η_Z function [see equation (4)], with $m = f_c(1)$ and $M = f_c(0)$, for several γ values, with $\sigma = 0.5$

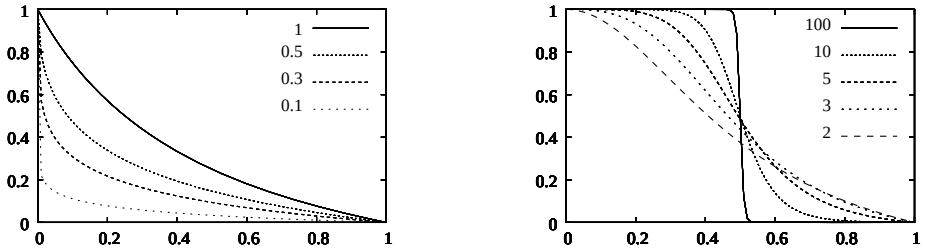


Figure 5 Generalised Gaussian function [see equation (6)], after normalisation through the η_Z function [see equation (4)], with $m = f_{gg}(1)$ and $M = f_{gg}(0)$, for several γ values, with $\sigma = 0.5$

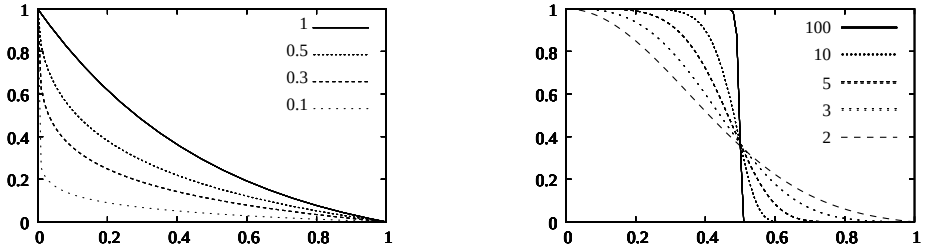


Figure 6 Fermi-Dirac function [see equation (7)], for several γ values (left) with $\sigma = 0.5$, (right) with $\sigma = 0.2$

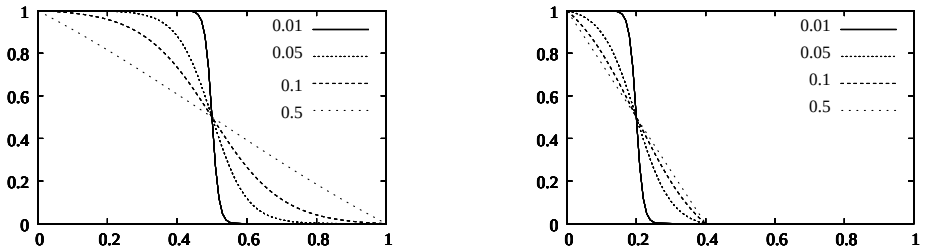
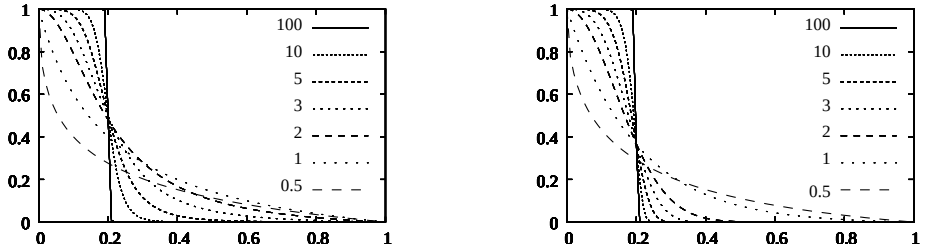


Figure 7 Cauchy function (left) and generalised Gaussian function (right) after normalisation through the η_Z function [see equation (4)], with $m = f(1)$ and $M = f(0)$ for several γ function, with $\sigma = 0.2$



4.2 Similarity measures deduced from dot products

Another definition of similarity measures for numerical data consists in basing them on dot products. Considering two data as vectors in $\mathbb{R}^p$, the scalar product depends on the angle between the two vectors and their norms:

$$\langle x,y\rangle = \|x\|\|y\|\cos(x,y).$$

It cannot be used as similarity measure directly, because it does not verify the maximality property: there exist couples (x,y) such that $\langle x,y\rangle > \langle x,x\rangle$, such as $(x,2x)$ for instance. The similarity should actually be defined as the cosine of the angle between the two vectors, that takes its maximal value when x and y are identical. It must still be stressed that such a similarity measure does not take into account the norms of the vectors, and that positively collinear vectors are considered as totally similar, e.g., $S(x,2x) = S(x,x) = 1$.

As the cosine takes values in the interval [−1, 1], it must be normalised; a simple approach is thus to define

$$S(x,y) = \frac{1}{2}\left(\frac{\langle x,y\rangle}{\|x\|\|y\|}+1\right).$$

It is to be underlined that dot products provide analytical properties, such as definite-positiveness, that can be useful for instance when the similarity measure is used in a cost function to be optimised.

4.2.1 Classic dot products

Numerous dot products can be considered: the Euclidean distances variants presented in Section 4.1.1 are associated to scalar products as indicated in the last column of Table 3.

They can be interpreted as data transformations in the same way as the corresponding distances. For instance, the Mahalanobis dot product is equivalent to applying a linear transformation that leads to attributes with mean 0 and covariance matrix identity before using the Euclidean scalar product.

4.2.2 Kernel functions

The kernel function framework, initially proposed by Vapnik (1995) (see also Schölkopf and Smola, 2002), leads to other interesting scalar products: this framework relies on a non-linear transformation of the data, performed before applying the Euclidean dot product to the transformed data. More formally, a kernel function is defined as $k(x,y) = \langle \phi(x), \phi(y)\rangle$, where ϕ denotes a non-linear transformation function. The kernel trick consists in providing a formulation of the *k* function that does not require the computation of the transformed data $\phi(x)$ but only depends on the initial representation of the data.

Consider as example the polynomial kernel, defined as $k(x,y) = (\langle x,y\rangle + l)^{\gamma}$. This kernel is equivalent to transforming the data, considering new attributes defined as

monomials of the initial attributes. In the case of 2D data for instance, $x \in \mathbb{R}^2$, $x = (x_1, x_2)$ and $y = (y_1, y_2)$, for $\gamma = 2$ et $l = 1$, one has

$$k(x,y) = \langle \phi(x), \phi(y) \rangle \text{ with } \phi(x) = \left(1, \sqrt{2}x_1, \sqrt{2}x_2, \sqrt{2}x_1x_2, x_1^2, x_2^2\right)$$

Thus, it is equivalent to transforming the data and enriching them with new attributes defined as the power 2 of the initial attributes and the product of the two coordinates, making it possible to take into account quadratic correlations between the attributes. More generally, the polynomial kernel implicitly enriches the data by components defined as monomials with degree inferior or equal to γ computed with the initial attributes. Now this enrichment remains implicit: no computation is performed in the high dimensional transformed space, the *k* function only depends on the dot product in the initial space, which has a low computational cost. Using this measure enriches the data representation without increasing the computational cost.

As other example, one can mention the Gaussian kernel, defined as

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right). \tag{8}$$

It can be shown that in this case, the implicit function ϕ maps the data to an infinite dimensional space, providing significant data description enrichment. It is to be noticed that this function is similar to the similarity measure defined as a decreasing function of the distance, using a generalised Gaussian function [see equation (6)] with $\gamma = 2$. The kernel framework provides a complementary theoretical justification of this similarity measure definition.

Other kernel functions for numerical data are described for instance by Schölkopf and Smola (2002). It is to be noticed that this formalism also applies more generally to any type of data, and in particular to structured data.

5 Conclusions

In this paper, we proposed an overview of existing similarity measures for binary and numerical data, and of tools for comparing them and analysing their behaviours. This characterisation provides guidelines for the problem of selecting a similarity measure for a given learning task.

More precisely, as a result of this study, it first appears that the initial choice depends on the data nature itself, as different measures are to be used when the data are binary or numerical.

In the case of binary data, the selection depends on the nature of the considered problem. In the case where only the data ranking induced by similarity matters, the selection problem reduces to the choice of a similarity equivalence class instead of a measure itself. In the case where the similarity values are of importance, the choice can be guided by the notion of discrimination power. In all cases, some general considerations regarding, e.g., the expected behaviours for specific data configurations (dependence to the universe size, case of included data) can help selecting an appropriate measure.

For numerical data, the selection can be decomposed into three steps. The choice of a distance function, or of a scalar product, must depend on knowledge about the attributes, e.g., whether they are of equal importance, correlated, or sufficient (or whether an implicit transformation to a space with higher dimension is required). The second step regards a possible normalisation, which depends on the data characteristics, and for instance on the presence of outliers. Lastly, the main selection tool for a decreasing function is the discrimination power.

Acknowledgements

The authors would like to acknowledge the many helpful and precious comments and suggestions of the reviewers.

References

Anderberg, M.R. (1973) *Cluster Analysis for Applications*, Academic Press, New York.

Batagelj, V. and Bren, M. (1995) 'Comparing resemblance measures', *Journal of Classification*, Vol. 12, pp.73–90.

Baulieu, F.B. (1989) 'A classification of presence/absence based dissimilarity coefficients', *Journal of Classification*, Vol. 6, pp.233–246.

Bouchon-Meunier, B., Rifqi, M. and Bothorel, S. (1996) 'Towards general measures of comparison of objects', *Fuzzy Sets and Systems*, Vol. 84, No. 2, pp.143–153.

Dice, L.R. (1945) 'Measures of the amount of ecological association between species', *Ecology*, Vol. 26, pp.297–302.

Gower, J. and Legendre, P. (1986) 'Metric and Euclidean properties of dissimilarity coefficients', *Journal of Classification*, Vol. 3, pp.5–48.

Hamann, V. (1961) 'Merkmalbestand und Verwandtschaftsbeziehungen der Farinosae, Ein Beitrag zum System der Monokotyledonen', *Willdenowia*, Vol. 2, pp.639–768.

Jaccard, P. (1908) 'Nouvelles recherches sur la distribution florale', *Bulletin de la Société Vaudoise des Sciences Naturelles*, Vol. 44, pp.223–270.

Janowitz, M.F. (1979) 'Monotone equivariant cluster analysis', *Journal of Mathematical Psychology*, Vol. 37, pp.148–165.

Kulczynski, S. (1927) 'Classe des sciences mathématiques et naturelles', *Bulletin International de l'Académie Polonaise des Sciences et des Lettres*, pp.57–203.

Lerman, I.C. (1967) 'Indice de similarité et préordonnance associée', in *Séminaire sur les ordres totaux finis*, pp.233–243.

Lerman, I.C. (1970) *Les Bases de la Classification Automatique*, Gauthier-Villars.

Ochiai, A. (1957) 'Zoogeographic studies on the soleoid fishes found in Japan and its neighbouring regions', *Bulletin of the Japanese Society for Science and Fisheries*, Vol. 22, pp.526–530.

Omhover, J-F., Rifqi, M. and Detyniecki, M. (2006) 'Ranking invariance based on similarity measures in document retrieval', in M. Detyniecki, J.M. Jose, A. Nürnberger and C.J.K. Rijsbergen (Eds.): *Adaptive Multimedia Retrieval: User, Context and Feedback: Third Int. Workshop, AMR 2005, Revised Selected Papers*, LNCS, Springer, pp.55–64.

Payne, J.S., Hepplewhite, L. and Stonham, T.J. (1999) 'Perceptually based metrics for the evaluation of textural image retrieval methods', *Int. Conf. on Multimedia Computing and Systems*, Vol. 2, pp.793–797.

Rifqi, M. (1996) 'Mesure de comparaison, typicalité et classification d'objets flous: théorie et pratique', PhD thesis, University of Paris 6.

Rifqi, M., Berger, V. and Bouchon-Meunier, B. (2000) 'Discrimination power of measures of comparison', *Fuzzy Sets and Systems*, Vol. 110, pp.189–196.

Rifqi, M., Detyniecki, M. and Bouchon-Meunier, B. (2003) 'Discrimination power of measures of resemblance', in *Int. Fuzzy Systems Association World Congress, IFSA'03*.

Rifqi, M., Lesot, M-J. and Detyniecki, M. (2008) 'Fuzzy order-equivalence for similarity measures', in *Proc. of NAFIPS 2008*.

Rogers, D.J. and Tanimoto, T.T. (1960) 'A computer program for classifying plants', *Science*, Vol. 132, pp.1115–1118.

Russel, P.F. and Rao, T.R. (1940) 'On habitat and association of species of anopheline larvae in south-eastern madras', *Journal of Malaria India Institute*, Vol. 3, pp.153–178.

Santini, S. and Jain, R. (1999) 'Similarity measures', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 9.

Schölkopf, B. and Smola, A. (2002) *Learning with Kernels*, MIT Press.

Shawe-Taylor, J. and Cristianini, N. (2004) *Kernel Methods for Pattern Analysis, Chapter 11: Kernels for Structured Data*, Cambridge University Press.

Sneath, P.H.A. and Sokal, R.R. (1973) *Numerical Taxonomy: The Principles and Practice of Numerical Classification*, W.H. Freeman, San Francisco.

Sokal, R. and Sneath, P.H.A. (1963) *Priciples of Numerical Taxonomy,* W.H. Freeman, San Francisco.

Sokal, R.R. and Michener, C. (1958) 'A statistical method for evaluating systematic relationships', *University of Kansas Science Bulletin*, Vol. 38, pp.1409–1438.

Sorensen, T. (1948) 'A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons', *K. Dan. Vidensk. Selsk. Biol. Skr*, Vol. 5, pp.1–34.

Tversky, A. (1977) 'Features of similarity', *Psychological Review*, Vol. 84, pp.327–352.

Vapnik, V. (1995) *The Nature of Statistical Learning Theory*, Springer, New York.

Yule, G.U. and Kendall, M.G. (1950) *An Introduction to the Theory of Statistics*, 14th ed., Hafner, New York.

Zadeh, L.A. (1965) 'Fuzzy sets', *Information and Control*, Vol. 8, pp.338–353.