Towards a conscious choice of a similarity measure: a qualitative point of view

Bernadette Bouchon-Meunier, Giulianella Coletti, Marie-Jeanne Lesot, Maria Rifqi

Towards a conscious choice of a similarity measure: a qualitative point of view

Bernadette Bouchon-Meunier[1], Giulianella Coletti[2], Marie-Jeanne Lesot[1], and Maria Rifqi[1]

[1] Université Pierre et Marie Curie - Paris 6, CNRS UMR 7606, LIP6,
104 avenue du Président Kennedy, F-75016 Paris, France
{Bernadette.Bouchon-Meunier,Marie-Jeanne.Lesot,Maria.Rifqi}@lip6.fr
[2] Dipartimento Matematica e Informatica, Università di Perugia,
Via Vanvitelli 1, 06123 Perugia, Italy
coletti@dipmat.unipg.it

Abstract. In many applications, such as case based reasoning, data mining or analogical reasoning, the choice of a particular measure of similarity is crucial. In this paper, we propose to study similarity measures from the point of view of the ordering relation they induce on object pairs. Using a classic method in measurement theory, introduced by Tversky, we establish necessary and sufficient conditions for the existence of a specific numerical measure, or a class of measures, to represent a given ordering relation, depending on the axioms this relation satisfies. The interest is particularly focused on different conditions of independence.

Key words: similarity, comparison measure, ordering relation, representability, weak independence conditions

1 Introduction

Similarity is a key concept in artificial intelligence [10] and similarity measures have been extensively studied; Bouchon-Meunier et al. [3] or Lesot et al. [6] for instance propose overviews of various approaches of similarity measures used in data mining. Following the links to the concept of similarity in cognitive science, especially in the process of categorization, the seminal work proposed by Tversky [12] has often been considered as a reference for the description of a general framework: it embeds numerous similarity measures and enables the user to make an appropriate choice of a specific similarity measure when facing a particular problem to solve. In [2, 9] we have proposed a general form for comparison measures, including similarity measures, compatible with Tversky's model, i.e. such that the proposed classes of measures satisfy the basic axioms introduced by Tversky. Nevertheless, there is still a need for a consensus about the choice of similarity measures; we study the converse approach in this paper, starting from Tversky's requirements and producing general classes of similarity measures, that accept the classical measures as particular cases. We especially focus on the independence axiom, introducing also relaxed variants.

We follow the idea of Tversky [12] of studying similarity in the environment of the theory of measurements [5, 11]. This point of view has also been used more recently in [4] and [1] to study particular classes of dissimilarity and similarity indices used in Descriptive Statistics for the comparison of frequency distributions. In this framework, and considering that object ranking is a frequent reason to use similarity measures, we introduce a binary relation on a set of pairs of objects, expressing a *comparative degree of similarity*. We study the representability of this comparative similarity by means of different numerical similarity measures: we establish axioms stating necessary and sufficient conditions under which a given comparative similarity is represented by a specific class of similarity measures. In other words, given the set of properties possessed by a given comparative degree of similarity, we characterise the form of the similarity measures that can represent it.

Thus we obtain two kinds of equivalence classes of similarity measures: the first one is given by measures representing the same ordering relation, as introduced in [7, 8]. The second, rougher, definition of equivalence is given by the measures representing orders that are not exactly identical but that possess the same properties and satisfy the same axioms. This definition permits to point out the actual rules we accept when we choose one particular measure of similarity and to make explicit underlying requirements on the induced order.

The paper is organized as follows. In Section 2, we consider as a starting point the numerical similarity measures: after recalling the classic notion of equivalence, we introduce basic axioms that are satisfied by comparative similarities induced from given classes of numerical similarity measures. We then turn to the reciprocal point of view, to relate given comparative similarities satisfying specific axioms to classes of numerical similarity measures. In Section 3, we introduce the independence axioms that are required to establish, in Section 4, these necessary and sufficient conditions for the existence of a class of measures to represent a given comparative degree of similarity.

2 From numerical similarity to comparative similarity

In this section, after introducing the notations used throughout the paper, we discuss the classic definition of equivalence between numerical similarity measures and establish basic axioms satisfied by comparative similarities induced from given classes of numerical similarities, following the ideas of Tversky to study similarity using the framework of the measurement theory [12].

For simplicity we consider the case of data described by a set of characteristics $\mathcal{A}$ that can be only present or absent in any object (so data are crisp and correspond to subsets of $\mathcal{A}$). We note that our approach can be easily extended to the case where objects are described by fuzzy subsets of $\mathcal{A}$.

2.1 Preliminaries

We consider that each object is described by p binary attributes, that is by the set of present characteristics from the predefined list $\mathcal{A}$. The data set is noted

$\mathcal{X} = \{0,1\}^p$: for every $X \in \mathcal{X}$, $X = \{x_1, ..., x_p\}$, $x_i \in \{0,1\}$. The particular object with $x_i = 0$ for every i is denoted $\underline{0}$. We note x_i^c the value $1 - x_i$, $X_k^c = \{x_1, ..., x_k^c, ..., x_p\}$ and $X^c = \{x_1^c, ..., x_p^c\}$. Finally, for any $X \in \mathcal{X}$, we note $I_X = \{i : x_i = 1\}$ and $|X|$ the cardinality of I_X.

Given a pair $(X,Y) \in \mathcal{X}$ we define $\mathbf{x} = |I_X \cap I_Y|$, i.e. the number of characteristics present in both objects, $\mathbf{y}^- = |I_X \setminus I_Y|$, i.e. the number of characteristics present in X but not in Y, $\mathbf{y}^+ = |I_Y \setminus I_X|$, i.e. the number of characteristics present in Y but not in X, and $\mathbf{y} = \mathbf{y}^- + \mathbf{y}^+ = |(I_X \setminus I_Y) \cup (I_Y \setminus I_X)|$. Finally we define $\mathbf{y}^* = |I_{X^c} \cap I_{Y^c}| = p - \mathbf{x} - \mathbf{y}$ that represents the number of characteristics absent of both objects.

Consider now a *comparative degree of similarity*, that is a binary relation $\preceq$ on $\mathcal{X}^2$, with the following meaning: for $X, Y, X', Y' \in \mathcal{X}$, $(X,Y) \preceq (X',Y')$ means that X is similar to Y no more than X' is similar to Y'.

The relations $\sim$ and $\prec$ are then induced by $\preceq$ as follows: $(X,Y) \sim (X',Y')$ if $(X,Y) \preceq (X',Y')$ and $(X',Y') \preceq (X,Y)$, meaning that X is similar to Y as X' is similar to Y'. Lastly $(X,Y) \prec (X',Y')$ if $(X,Y) \preceq (X',Y')$ but not $(X',Y') \preceq (X,Y)$, meaning that X is similar to Y less than X' is similar to Y'.

It is to be noticed that if $\preceq$ is complete, then $\sim$ and $\prec$ are the symmetrical and the asymmetrical parts of $\preceq$ respectively.

We now introduce the notion of representability of such a comparative degree of similarity by a numerical similarity measure:

Definition 1. *Given a comparative degree of similarity $\preceq$, a similarity measure $S : \mathcal{X}^2 \to \mathbb{R}$* represents *$\preceq$ if and only if for any $(X,Y), (X',Y') \in \mathcal{X}^2$, both following conditions hold:*

$$(X,Y) \preceq (X',Y') \Rightarrow S(X,Y) \leq S(X',Y')$$
$$(X,Y) \prec (X',Y') \Rightarrow S(X,Y) < S(X',Y')$$

We recall that if the relation $\preceq$ is complete the above conditions are equivalent to the following one: $(X,Y) \preceq (X',Y') \Leftrightarrow S(X,Y) \leq S(X',Y')$.

2.2 Similarity measure equivalence

Any similarity measure on $\mathcal{X}^2$ induces a complete comparative degree of similarity $\preceq$, defined as follows: $(X,Y) \prec (X',Y')$ if $S(X,Y) < S(X',Y')$ and $(X,Y) \sim (X',Y')$ if $S(X,Y) = S(X',Y')$.

Now the same ordering relation is induced by any similarity measure that can be expressed as an increasing transformation of S: any similarity measure $S' = \varphi(S)$, with $\varphi : \mathbb{R} \to \mathbb{R}$ strictly increasing is also a representation of $\preceq$. Moreover, no other measure $S*$ represents $\preceq$.

Thus, from a comparative point of view, all functions $\varphi(S)$ are indistinguishable. Formally speaking, the relation r defined on the set of similarity measures as SrS' if and only if S and S' induce the same comparative degree of similarity on $\mathcal{X}$ is an equivalence relation. An equivalent formulation of this concept, expressed only in terms of numerical similarity functions, is given in [7, 8].

For instance the similarity measures

$$S_\rho(X,Y) = \frac{\mathbf{x}}{\mathbf{x}+\rho\mathbf{y}} \tag{1}$$

with $\rho > 0$, are all equivalent, since each of them is an increasing transformation of any other. In particular, the Jaccard ($\rho = 1$), Dice ($\rho = 1/2$), Sorensen ($\rho = 1/4$), Anderberg ($\rho = 1/8$) and Sokal and Sneath ($\rho = 2$) measures are equivalent.

The same class also contains the function $S(X,Y) = \log(\mathbf{x}) - \log(\mathbf{y})$, which is of the kind proposed by Tversky [12] (a linear form of an increasing function): S is an increasing transformation of $S'(X,Y) = \mathbf{x}/\mathbf{y}$ which is an increasing transformation of S_1.

It is to be noted that the function $S(X,Y) = \alpha\log(\mathbf{x}) - \beta\log(\mathbf{y})$ for $\alpha, \beta > 0$ is not in the same class, but it is equivalent to all measures

$$S *_\rho (X,Y) = \frac{\mathbf{x}^\alpha}{\mathbf{x}^\alpha + \rho\mathbf{y}^\beta}$$

2.3 Basic axioms

We are now interested in a different classification of measures of similarity: instead of considering the measures that induce the same order, we consider the measures that induce orders satisfying the same class of axioms. In this section, we consider axioms that lead to preliminary results regarding relations between similarity measures and comparative degrees of similarity.

Basic properties The first two axioms we introduce describe basic properties a binary relation has to satisfy to define a comparative degree of similarity: the first one only states the relation must be a weak order.

Axiom S1 [weak order]
$\preceq$ is a weak order, i.e it is complete, reflexive and transitive.

The second axiom expresses boundary conditions: it imposes that for any X, whatever Y, X cannot be more similar to Y than it is similar to itself, and it cannot be less similar to Y than it is to its complement. Lastly, it imposes that X is similar to itself as Y is to itself: all data are equally similar to themselves.

Axiom S2 [boundary conditions] $\forall X, Y \in \mathcal{X}$,
$(X^c, X) \sim (Y^c, Y) \preceq (X,Y) \preceq (X,X) \sim (Y,Y)$ and $(X^c, X) \prec (X,X)$

The third axiom imposes a symmetry condition.

Axiom S3 [symmetry]
$\forall X, Y \in \mathcal{X}$, $(X,Y) \sim (Y,X)$

These properties lead to the following two definitions:

Definition 2. *A binary relation $\preceq$ on $\mathcal{X}^2$ is a* comparative similarity *if and only if it satisfies axioms S1 and S2.*

Definition 3. *A comparative similarity is* symmetric *if and only if it satisfies axiom S3.*

The next axiom expresses the idea that all attributes have the same role with respect to the comparative similarity: a change in one attribute is equivalent to the modification of any attribute of the same category, i.e. attributes representing characteristics present in both objects (with indices in $I_X \cap I_Y$), only in one of them (with indices in $I_X \setminus I_Y$ or in $I_Y \setminus I_X$) or absent of both (indices not in $I_X \cup I_Y$):

Axiom S4 [attribute uniformity] $\forall h, k \in \{1, ..., p\}$,
if $h, k \in I_X \cap I_Y$, or $h, k \in I_X \setminus I_Y$, or $h, k \in I_Y \setminus I_X$, or $h, k \notin I_X \cup I_Y$,
then $(X, Y_k^c) \sim (X, Y_h^c)$ and $(X_k^c, Y) \sim (X_h^c, Y)$.

It must be underlined that any comparative similarity representable by a similarity measure depending only on $\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-$ and $\mathbf{y}^*$ satisfies this axiom. Reciprocally, as $\mathcal{X}^2$ is finite, any comparative similarity satisfying axiom S4 can be represented by a function depending only on $\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-$ and $\mathbf{y}^*$.

Monotonicity axioms The following three axioms of monotonicity govern the comparative similarity among pairs differing in the presence/absence of only one attribute: 4 pairs must then be compared, depending on whether the modification is applied to both data, one or the other, or none of them. The different axioms correspond to different choices regarding the semantics of the similarity measure, as commented below.

Axiom S5 [monotonicity] $\forall X, Y \in \mathcal{X},\ X \neq Y$
$\forall k \in I_X \cap I_Y, \quad (X, Y_k^c) \sim (X_k^c, Y) \prec (X_k^c, Y_k^c) \prec (X, Y)$
if $I_X \cap I_Y = \emptyset$, $\forall k \in I_X \quad (X, Y) \sim (X_k^c, Y)$

The first condition means that if an attribute possessed by both objects is modified, the modified objects are less similar one to another than the initial object pairs were. This corresponds to a strong semantic choice: it implies that the common presence of an attribute is preferred to a common absence. Moreover, the axiom states that modifying both objects degrades the similarity to a lesser extent than changing only one of them. Lastly, if only one object is modified, there is no difference whether X or Y is concerned. Equivalently, the axiom can be written in the following three forms, describing the expected variations when other attribute types are considered:
$\forall k \in I_X \setminus I_Y$, $(X_k^c, Y_k^c) \sim (X, Y) \prec (X_k^c, Y) \prec (X, Y_k^c)$
$\forall k \in I_Y \setminus I_X$, $(X_k^c, Y_k^c) \sim (X, Y) \prec (X, Y_k^c) \prec (X_k^c, Y)$
$\forall k \notin I_X \cup I_Y$, $(X, Y_k^c) \sim (X_k^c, Y) \prec (X, Y) \prec (X_k^c, Y_k^c)$
The second condition considers the case where the intersection $I_X \cap I_Y$ is empty, i.e. when there is no common attributes: then it is indifferent whether

the attributes are absent of both objects or present in one of them. In particular, it implies that whatever X and Y such that $I_X \cap I_Y = \emptyset$, $(X, Y) \sim (X, X^c)$.

It is easy to prove that any comparative similarity $\preceq$ representable by a similarity measure S defined as

$$S(X, Y) = \frac{f(\mathbf{x})}{f(\mathbf{x}) + \rho g(\mathbf{y})} \tag{2}$$

with $\rho > 0$ and f and g non negative increasing functions (or any strictly increasing transformation of this measure) satisfies Axiom S5. Thus in particular, it is verified by the measures belonging to the S_ρ class defined in Equation (1), in which f and g coincide with the identity function.

Axiom S6 relaxes the conditions required by S5, insofar as it does not impose conditions on the comparison of (X, Y_k^c) and (X_k^c, Y), whereas they are equivalent in Axiom S5:

Axiom S6 [weak monotonicity] $\forall X, Y \in \mathcal{X}$, $X \neq Y$, $\forall k \in I_X \cap I_Y$,
$(X, Y_k^c) \prec (X_k^c, Y_k^c) \preceq (X, Y)$ and $(X_k^c, Y) \prec (X_k^c, Y_k^c) \preceq (X, Y)$
if $I_X \cap I_Y = \emptyset$, $\forall k \in I_X \quad (X, Y) \sim (X_k^c, Y)$

It is easy to prove that any comparative similarity $\preceq$ representable by a similarity measure S defined as

$$S(X, Y) = \frac{f(\mathbf{x})}{g(\mathbf{x} + \mathbf{y}^-)h(\mathbf{x} + \mathbf{y}^+)} \tag{3}$$

with f, g, h non negative increasing functions such that $\forall x$, $f(x) = g(x)h(x)$ (ensuring that $S(X, Y) = 1$ when $y^- = y^+ = 0$) satisfies Axiom S6 (as well as any strictly increasing transformation of this measure). A particular case is the Ochiai measure, where f is the identity function and $g(.) = h(.) = \sqrt{.}$.

Axiom S6 is also satisfied by any comparative similarity $\preceq$, representable by a similarity measure S defined as

$$S(X, Y) = \frac{f(\mathbf{x})}{g(\mathbf{x} + \mathbf{y}^-)} + \frac{f(\mathbf{x})}{h(\mathbf{x} + \mathbf{y}^+)} \tag{4}$$

with f, g, h non negative increasing functions such that $\forall x$, $f(x)(h(x) + g(x)) = g(x)h(x)$ (ensuring that $S(X, Y) = 1$ when $y^- = y^+ = 0$). In particular, it holds for the Kulczynski measure, where g and h are the identity function and $f(x) = x/2$.

Lastly Axiom S7 resembles Axiom S5 but considers the case where (X, Y) and (X_k^c, Y_k^c) are equivalent, i.e. characteristics present in both objects or absent of both objects play the same role. Besides, as Axiom S5, it requires that (X, Y_k^c) and (X_k^c, Y) are equivalent, i.e. the modification is symmetrical.

Axiom S7 [monotonicity 2] $\forall X, Y \in \mathcal{X}$, $X \neq Y$
$\forall k \in I_X \cap I_Y, \quad (X, Y_k^c) \sim (X_k^c, Y) \prec (X_k^c, Y_k^c) \sim (X, Y)$
if $I_X \cap I_Y = \emptyset$, $\forall k \in I_X \quad (X, Y) \sim (X_k^c, Y)$

This is equivalent to saying that the same property holds for any k not in $I_X \cup I_Y$ or to saying that, $\forall k \in I_X \setminus I_Y$ and $\forall k \in I_Y \setminus I_X$, $(X_k^c, Y_k^c) \sim (X, Y) \prec (X, Y_k^c) \sim (X_k^c, Y)$.

It is easy to prove that any comparative similarity $\preceq$ representable by a similarity measure S defined as

$$S(X,Y) = \frac{f(\mathbf{x}+\mathbf{y}^*)}{f(\mathbf{x}+\mathbf{y}^*)+\rho g(\mathbf{y})} \tag{5}$$

with f, g increasing functions, $\rho > 0$ satisfies Axiom S7. This corresponds to so-called type II similarity measures [6]. In particular it holds for the Rogers and Tanimoto, Sokal and Michener and Sokal and Sneath measures, for which f and g are the identity function, α takes values 2, 1 and 1/2 respectively.

3 Independence conditions

The objective is then to consider the reciprocal point of view: given the set of properties possessed by a given comparative degree of similarity, to characterise the form of the similarity measures that can represent it. To that aim, other conditions must be imposed to the comparative similarities: we take into account the basic properties considered by Tversky as fundamental for similarities in [12], focusing on the independence axiom he introduced. Starting from his classic definition, we extend it to weaker forms that will determine the class of measures a comparative similarity can be represented by, as will be shown in Section 4.

Axiom I is the independence axiom introduced by Tversky [12]:

Axiom I [independence] For any 4-tuple (X_1, Y_1), (X_2, Y_2), (Z_1, W_1), (Z_2, W_2), if one of the following conditions holds
(i) $\mathbf{x_i} = \mathbf{z_i}$ and $\mathbf{y_i^-} = \mathbf{w_i^-}$ $(i = 1, 2)$, and $\mathbf{y_1^+} = \mathbf{y_2^+}$, $\mathbf{w_1^+} = \mathbf{w_2^+}$
(ii) $\mathbf{x_i} = \mathbf{z_i}$ and $\mathbf{y_i^+} = \mathbf{w_i^+}$ $(i = 1, 2)$, and $\mathbf{y_1^-} = \mathbf{y_2^-}$, $\mathbf{w_1^-} = \mathbf{w_2^-}$
(iii) $\mathbf{y_i^+} = \mathbf{w_i^+}$ and $\mathbf{y_i^-} = \mathbf{w_i^-}$ $(i = 1, 2)$, and $\mathbf{x_1} = \mathbf{x_2}$, $\mathbf{z_1} = \mathbf{z_2}$
then $(X_1, Y_1) \preceq (X_2, Y_2) \Leftrightarrow (Z_1, W_1) \preceq (Z_2, W_2)$.

where $z_i = |I_{Z_i} \cap I_{W_i}|$, $w_i^- = |I_{Z_i} \setminus I_{W_i}|$ and $w_i^+ = |I_{W_i} \setminus I_{Z_i}|$ (the same notations are used in the following).

Condition (i) for instance expresses that the joint effect of $\mathbf{x}$ and $\mathbf{y}^-$ is independent of the fixed component $\mathbf{y}^+$.

It must be underlined that comparative similarities representable by a similarity measure S defined by Equation (2), and in particular of the class S_h defined in Equation (1), do not satisfy this independence condition. This can be illustrated as follows in the case of the Jaccard measure, i.e. S_h with $h = 1$: considering hypothesis (i), by trivial computation, one has $(X_1, Y_1) \preceq (X_2, Y_2)$ if and only if $\mathbf{x_1}(\mathbf{y_2^-} + \mathbf{y_1^+}) \leq \mathbf{x_2}(\mathbf{y_1^-} + \mathbf{y_1^+})$ and $(Z_1, W_1) \preceq (Z_2, W_2)$ iff $\mathbf{x_1}(\mathbf{y_2^-} + \mathbf{w_1^+}) \leq \mathbf{x_2}(\mathbf{y_1^-} + \mathbf{w_1^+})$. Now the two inequalities can be independently satisfied, as can be shown using the following example: $X_1 = Z_1 = W_1 = (10000)$, $Y_1 = (11000)$, $X_2 = Z_2 = (11110)$, $Y_2 = (11101)$, and $W_2 = (11100)$.

We introduce now a weaker form of independence in which we only require that the common characteristics are independent of the totality of the characteristics present in only one element of the pair.

Axiom WI [weak independence] For any 4-tuple (X_1, Y_1), (X_2, Y_2), (Z_1, W_1), (Z_2, W_2), if one of the following conditions holds
(i) $\mathbf{x_i} = \mathbf{z_i}$ $(i = 1, 2)$, and $\mathbf{y_1} = \mathbf{y_2}$, $\mathbf{w_1} = \mathbf{w_2}$
(ii) $\mathbf{y_i} = \mathbf{w_i}$ $(i = 1, 2)$, and $\mathbf{x_1} = \mathbf{x_2}$, $\mathbf{z_1} = \mathbf{z_2}$
then $(X_1, Y_1) \preceq (X_2, Y_2) \Leftrightarrow (Z_1, W_1) \preceq (Z_2, W_2)$.

It must be underlined that the comparative similarities representable by a similarity measure S defined by Equation (2) satisfy this axiom, and thus in particular the elements of the class S_h. We prove this assertion for hypothesis (i): by trivial computation it holds that on one hand $(X_1, Y_1) \preceq (X_2, Y_2)$ iff $f(\mathbf{x_1}) \leq f(\mathbf{x_2})$, and on the other hand $(Z_1, W_1) \preceq (Z_2, W_2)$ iff $f(\mathbf{x_1}) \leq f(\mathbf{x_2})$, leading to the desired equivalence. The proof is similar for condition (ii).

Comparative similarities representable by a similarity measure S defined by Equation (3) do not satisfy the weak independence axiom WI. In particular the well known Ochiai measure does not satisfy WI, and, obviously, the independence axiom I. The same considerations hold for comparative similarities representable by a similarity measure S defined by Equation (4), and in particular for the Kulczynski measure.

We now introduce another weak kind of independence that considers as components the common characteristics and the sum of these common characteristics and the characteristics present in only one of the two objects.

Axiom CI [cumulative independence] For any 4-tuple (X_1, Y_1), (X_2, Y_2), (Z_1, W_1), (Z_2, W_2), if one of the following conditions holds
(i) $\mathbf{x_i} = \mathbf{z_i}$ and $\mathbf{x_i} + \mathbf{y_i^-} = \mathbf{z_i} + \mathbf{w_i^-}$ $(i = 1, 2)$, and $\mathbf{x_1} + \mathbf{y_1^+} = \mathbf{x_2} + \mathbf{y_2^+}$, $\mathbf{z_1} + \mathbf{w_1^+} = \mathbf{z_2} + \mathbf{w_2^+}$
(ii) $\mathbf{x_i} = \mathbf{z_i}$ and $\mathbf{x_i} + \mathbf{y_i^+} = \mathbf{z_i} + \mathbf{w_i^+}$ $(i = 1, 2)$, and $\mathbf{x_1} + \mathbf{y_1^-} = \mathbf{x_2} + \mathbf{y_2^-}$, $\mathbf{z_1} + \mathbf{w_1^-} = \mathbf{z_2} + \mathbf{w_2^-}$
(iii) $\mathbf{x_i} + \mathbf{y_i^+} = \mathbf{z_i} + \mathbf{w_i^+}$ and $\mathbf{x_i} + \mathbf{y_i^-} = \mathbf{z_i} + \mathbf{w_i^-}$ $(i = 1, 2)$, and $\mathbf{x_1} = \mathbf{x_2}$, $\mathbf{z_1} = \mathbf{z_2}$
then $(X_1, Y_1) \preceq (X_2, Y_2) \Leftrightarrow (Z_1, W_1) \preceq (Z_2, W_2)$.

It is easy to prove that a comparative similarity representable by a similarity measure S defined by Equation (3), in particular, the Ochiai measure, satisfies the cumulative independence condition CI.

Finally we introduce another weak definition of independence that considers as components the sum of characteristics which are common and those which are absent of both objects and the sum of those present in only one object of the pair.

Axiom TWI [totally weak independence] For any 4-tuple (X_1, Y_1), (X_2, Y_2), (Z_1, W_1), (Z_2, W_2), if one of the following conditions holds

(i) $\mathbf{x_i} + \mathbf{y_i^*} = \mathbf{z_i} + \mathbf{w_i^*}$ $(i = 1, 2)$, and $\mathbf{y_1} = \mathbf{y_2}$, $\mathbf{w_1} = \mathbf{w_2}$
(ii) $\mathbf{y_i} = \mathbf{w_i}$ $(i = 1, 2)$, and $\mathbf{x_1} + \mathbf{y_1^*} = \mathbf{x_2} + \mathbf{y_2^*}$ $\mathbf{z_1} + \mathbf{w_1^*} = \mathbf{z_2} + \mathbf{w_2^*}$
then $(X_1, Y_1) \preceq (X_2, Y_2) \Leftrightarrow (Z_1, W_1) \preceq (Z_2, W_2)$.

It is easy to prove that comparative similarities representable by a similarity measure S defined by Equation (5), in particular, the Rogers and Tanimoto, the Sokal and Michener and the Sokal and Sneath measures, satisfy the axiom TWI.

4 Representation theorems

In this section we establish the theorems stating necessary and sufficient conditions for comparative similarities verifying the various independence axioms to be representable by classes of numerical measures.

Theorem 1. *Let $\preceq$ be a binary relation on $\mathcal{X}^2 \setminus \{(\underline{0}, \underline{0})\}$. The following conditions are equivalent:*

(i) $\preceq$ is a comparative similarity satisfying axioms S4 and S5 and possessing the weak independence property WI
(ii) there exist two non negative increasing functions f and g, with $f(0) = g(0) = 0$ such that the function $S : \mathcal{X}^2 \to [0, 1]$ defined by Equation (2) represents $\preceq$.

Proof: we first prove the implication (ii) $\Rightarrow$ (i) and consider $\preceq$ the ordering relation induced by a similarity measure S satisfying the conditions (ii). Then $\preceq$, representable by a function with values in $\mathbb{R}$, is a weak order, i.e. satisfies Axiom S1. Moreover, as $\forall X \in \mathcal{X}$, $S(X, X) = 1 > S(X^c, X) = S(X, X^c) = 0$ and $S(X, Y) \in [0, 1]$, $\preceq$ also satisfies Axiom S2. Thus it is a comparative similarity.

Furthermore, it satisfies the Axioms S5 and WI as already underlined in the remarks following the introduction of these axioms (see pages 5 and 8): it satisfies S5, because S is increasing with respect to $\mathbf{x}$ and decreasing with respect to $\mathbf{y}$. Besides if $I_X \cap I_Y = \emptyset$, $\mathbf{x} = 0$, thus $S(X, Y) = 0 = S(X_k^c, Y)$ for all $k \in I_X$. It satisfies WI because of the independence properties of S.

We now prove the implication (i) $\Rightarrow$ (ii) and consider a comparative similarity $\preceq$ satisfying the conditions (i). Let us indicate by $\mathbb{R}^*$ the compactification of $\mathbb{R}$, that is $\mathbb{R}^* = \mathbb{R} \cup \{-\infty, +\infty\}$. Since $\mathbb{R}^*$ is a completely ordered set containing $\mathbb{R}$, all results related to the representability of a binary relation by a function with values in $\mathbb{R}$ remain valid for functions with values in $\mathbb{R}^*$ [5].

Due to S4, $\preceq$ is representable by a function $S : \mathcal{X}^2 \to \mathbb{R}^*$, depending only on $\mathbf{x}, \mathbf{y}^+, \mathbf{y}^-$ and $\mathbf{y}^*$. Due to S5, it is strictly increasing in $\mathbf{x}$ and strictly decreasing in $\mathbf{y}^+$ and $\mathbf{y}^-$. Due to condition WI, there exists a function $f_1 : \mathbb{R} \to \mathbb{R}^*$, that moreover is strictly increasing due to S5, and two real numbers $\alpha, \beta > 0$ so that S is a strictly increasing transformation $\varphi : \mathbb{R}^* \to \mathbb{R}^*$ of a linear form of the function f_1, i.e.

$$S(X, Y) = \varphi(\alpha f_1(\mathbf{x}) - \beta f_1(\mathbf{y})) \tag{6}$$

Now from Axiom S2, necessarily $f_1(0) = -\infty$. Indeed from Axiom S2, it holds that $(X, X) \sim (Y, Y)$ and thus with $Y = X_k^c$, $(X, X) \sim (X_k^c, X_k^c)$, which implies

$S(X,X) = S(X_k^c, X_k^c)$. Applying Equation (6), $S(X,X) = \varphi(\alpha f_1(|I_X|) - \beta f_1(0))$ and $S(X_k^c, X_k^c) = \varphi(\alpha f_1(|I_X|-1) - \beta f_1(0))$. As φ is strictly increasing, the equality implies $\alpha f_1(|I_X|) - \beta f_1(0) = \alpha f_1(|I_X|-1) - \beta f_1(0)$. As f_1 is strictly increasing, $f_1(|I_X|) > f_1(|I_X|-1)$. For the equality to hold, it is necessary that $f_1(0) = -\infty$: the unique possible elements of $[-\infty, +\infty]$ which summed to two different real numbers give the same results are $-\infty$ and $+\infty$, by monotonicity of f_1, we have $f_1(0) = -\infty$.

Letting $f_2 = \exp(f_1)$, that thus satisfies $f_2(0) = 0$, and $\psi = \varphi \circ \log$, $\preceq$ is thus representable by

$$S(X,Y) = \psi\left(\frac{f_2^\alpha(\mathbf{x})}{f_2^\beta(\mathbf{y})}\right) \tag{7}$$

considering the fraction takes value $+\infty$ when $y = 0$.

Choosing as ψ_2 the increasing function $\psi_2(z) = z/(z+\rho)$, with ρ positive real number, then $\preceq$ is representable by $\psi_2(S(X,Y))$. Denoting $f(x) = f_2^\alpha(x)$ and $g(y) = f_2^\beta(y)$, the latter can be written

$$\psi_2(S(X,Y)) = \frac{f(x)}{f(x) + \rho g(y)}$$

i.e. in the form of Equation (2). Furthermore, f and g satisfy the conditions required in (ii): they are strictly increasing, and $f(0) = g(0) = 0$.

The following theorem considers the case of cumulatively independent comparative similarities:

Theorem 2. *Let $\preceq$ be a binary relation on $\mathcal{X}^2 \setminus \{(\underline{0}, \underline{0})\}$. The following conditions are equivalent:*

(i) $\preceq$ is a comparative similarity satisfying axioms S4 and S6 and possessing the cumulative independence property CI

(ii) there exists a real-valued increasing function f, with $f(0) = 0$ and $\alpha, \beta, \gamma \geq 0$, $\alpha = \beta + \gamma$, such that the function $S : \mathcal{X}^2 \to [0,1]$ defined by

$$S(X,Y) = \frac{f^\alpha(\mathbf{x})}{f^\beta(\mathbf{x}+\mathbf{y}^-) f^\gamma(\mathbf{x}+\mathbf{y}^+)} \tag{8}$$

represents $\preceq$.

Proof: The proof of implication (ii) $\Rightarrow$ (i) is direct and similar to that in Theorem 1. We prove (i) $\Rightarrow$ (ii). By the hypotheses, following the same considerations as in the previous theorem, there exists a class of functions S representing $\preceq$ that are increasing transformations of a function such as

$$S(X,Y) = \alpha f_1(\mathbf{x}) - \beta f_1(\mathbf{x}+\mathbf{y}^-) - \gamma f_1(\mathbf{x}+\mathbf{y}^+) \tag{9}$$

By the hypothesis of monotonicity S6, the function f_1 must be increasing and convex and $\alpha, \beta, \gamma > 0$. Moreover, by condition $(X,X) \sim (Y,Y)$ for every $X, Y \in$

$\mathcal{X}$, we have necessarily $\alpha = \beta + \gamma$. From condition $(X, X^c) \sim (Y^c, Y)$ of Axiom S2, with $Y = X_k^c$ and $k \in I_X$, it follows that $f_1(0) = -\infty$ using the same argument as in the proof of Theorem 1.

A result in the special case where $\preceq$ is symmetrical can be established: if, in condition (i), $\preceq$ is also required to satisfy Axiom S3 (symmetry), then the function representing $\preceq$ is such that $\beta = \gamma = \alpha/2$.

Lastly we consider the case of totally weak independent comparative similarities:

Theorem 3. *Let $\preceq$ be a binary relation on $\mathcal{X}^2$. The following conditions are equivalent:*

(i) $\preceq$ is a comparative similarity satisfying axioms S4 and S7, and possessing the totally weak independence property TWI

(ii) there exist three real-valued increasing functions f, g, with $f(0) = 0$ and $\alpha, \beta \geq 0$, such that the function $S : \mathcal{X}^2 \to [0, 1]$ defined by Equation (5) represents $\preceq$.

The proof is very similar to that given in Theorem 1.

Again for this theorem a "symmetric version" can be proved: if, in condition (i), $\preceq$ is also required to satisfy S3, then the function representing $\preceq$ is such that $g = f$.

5 Conclusion

The approach of similarity we have presented is based on several basic hypotheses. The first one is the environment of measurement theory, stemming from Tversky's reference work, which has been considered since then by the community as a reasonable approach to model similarities managed by human beings. The second hypothesis is the importance of ranking in the management of similarities, which means that similarities are regarded as relative characteristics of families of objects, rather than intrinsic descriptions of these families. We are often interested in the comparison of similarity degrees attached to two pairs of objects, more than in the level of similarity attached to each of these pairs. We can remark that changing the measure of similarity in a model provides different values for these degrees, and it is therefore reasonable not to attach too much importance to the similarity degrees themselves, but to their relative values. The third hypothesis we make is the importance of the independence axiom among those proposed by Tversky.

We have therefore established a link between what we call comparative similarities in a qualitative approach on the one hand, and possible numerical representations of these similarities on the other hand, providing general forms of similarity measures compatible with the independence axiom and with weaker forms of this axiom. We show that such a framework embeds well-known similarity measures, and we point out classes of such measures with the same behavior with respect to independence.

It is to be hoped that this work will help users of similarities in all domains of artificial intelligence and image processing, in particular, to make an appropriate choice of a convenient measure when they have to manage resemblances. It will for instance avoid them to compare results based on the choice of several similarity measures, since results appear to be analogous when the measures belong to a same class. The choice of a similarity measure is then reduced to the choice of a class of measures.

Future works will take into account extensions of such similarity measures to graded values of attributes, in a fuzzy set based knowledge representation, replacing the binary attributes we have only considered in this paper. Such a work will meet a general framework for measures of similarity between fuzzy sets we have already proposed [2, 9], providing a qualitative view of similarities associated with such numerical evaluations of similarities.

References

1. Bertoluzza, C., Di Bacco, M., Doldi, V.: An axiomatic characterization of the measures of similarity. Sankhya **66** (2004) 474–486
2. Bouchon-Meunier, B., Rifqi, M., Bothorel, S.: Towards general measures of comparison of objects. Fuzzy Sets and Systems **84** (1996) 143–153
3. Bouchon-Meunier, B., Rifqi, M., Lesot, M.J.: Similarities in fuzzy data mining: from a cognitive view to real-world applications. In Zurada, J., Yen, G., Wang, J., eds.: Computational Intelligence: Research Frontiers. Springer, LNCS (2008)
4. Coletti, G., Di Bacco, M.: Qualitative characterization of a dissimilarity and concentration index. Metron **XLVII** (1989) 121–130
5. Krantz, D., Luce, R., Suppes, P., Tversky, A.: Foundations of measurement. Volume I. Academic Press (1971)
6. Lesot, M.J., Rifqi, M., Benhadda, H.: Similarity measures for binary and numerical data: a survey. Intern. J. of Knowledge Engineering and Soft Data Paradigms (KESDP) **1** (2009) 63–84
7. Omhover, J.F., Bouchon-Meunier, B.: Equivalence entre mesures de similarités floues: application à la recherche d'images par le contenu. In: 6eme Congrès Européen de Science des Systèmes. (2005)
8. Omhover, J.F., Detyniecki, M., Rifqi, M., Bouchon-Meunier, B.: Image retrieval using fuzzy similarity : Measure of equivalence based on invariance in ranking. In: IEEE Int. Conf. on Fuzzy Systems. (2004)
9. Rifqi, M.: Mesures de comparaison, typicalité, et classification d'objets flous: théorie et pratique. PhD thesis, University Paris 6 (1996)
10. Rissland, E.: Ai and similarity. IEEE Intelligent Systems **21** (2006) 33–49
11. Suppes, P., Krantz, D., Luce, R., Tversky, A.: Foundations of measurement. Volume II. Academic Press, New York (1989)
12. Tversky, A.: Features of similarity. Psychological Review **84** (1977) 327–352