Things to know about a (dis)similarity measure

Lluís A. Belanche and Jorge Orozco
`belanche@lsi.upc.edu, jorozco@lsi.upc.edu`

Soft Computing Research Group
Computer Science School
Technical University of Catalonia Barcelona, Spain

15th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems

12, 13 & 14 September 2011, Kaiserslautern, Germany

Contents

Similarity and Dissimilarity

Intuitive notion?

1. Human beings use the notion of *similarity* and *dissimilarity* for problem solving, inductive reasoning, ...

2. Computer Science: related to Artificial Intelligence, Case Based Reasoning, Data Mining, Information Retrieval, Pattern Matching, Neural Networks, ...

Similarity and Dissimilarity

Some particular thoughts

- Metric dissimilarities have been deeply studied but they are tied to a particular transitivity
- Particularly, Euclidean distances are used due to our natural understanding of Euclidean spaces.
- Not all metrics are Euclidean and many interesting dissimilarities are non-metric.
- What about similarities?

Similarity and Dissimilarity

The two sides of the same coin

- We argue that every property of a similarity should have a correspondence with one property of a dissimilarity and vice versa.

- The present work intends to make a further effort in their unification:

  1. definition of similarity and dissimilarity and a set of fundamental properties and transformations

  2. how these transformations alter the properties

Similarity and Dissimilarity

Definition

Definition 1 *A* similarity measure *is an upper bounded, exhaustive and total function* $s : X \times X \rightarrow I_s \subset \mathbb{R}$ *with* $|I_s| > 1$ *(therefore* I_s *is upper bounded and* $\sup I_s$ *exists).*

Definition 2 *A* dissimilarity measure *is a lower bounded, exhaustive and total function* $d : X \times X \rightarrow I_d \subset \mathbb{R}$ *with* $|I_d| > 1$ *(therefore* I_d *is lower bounded and* $\text{ínf}\, I_d$ *exists).*

Similarity and Dissimilarity

Properties

Reflexivity: $s(x, x) = \sup I_s \in I_s$ and $d(x, x) = \inf I_d \in I_d$.

Strong Reflexivity: $s(x, y) = \sup I_s \Leftrightarrow x = y$ and $d(x, y) = \inf I_d \Leftrightarrow x = y$.

Symmetry: $s(x, y) = s(y, x)$ and $d(x, y) = d(y, x)$.

Boundedness: A similarity s is *lower* bounded when $\inf I_s$ exists. Conversely, a dissimilarity d is *upper* bounded when$\sup I_d$ exists.

Closedness: A lower bounded s is closed if $\inf I_s \in I_s$. An upper bounded d, is closed if $\sup I_d \in I_d$.

Complementarity

Transitivity

Similarity and Dissimilarity

Transitivity (I)

Definition 3 *(Transitivity operator). Let I be a non-empty subset of $\mathbb{R}$, and let e be a fixed element of I. A transitivity operator is a function $\tau : I \times I \to I$ satisfying, for all $x, y, z \in I$:*

1. $\tau(x, e) = x$ *(null element)*

2. $y \leq z \Rightarrow \tau(x, y) \leq \tau(x, z)$ *(non-decreasing monotonicity)*

3. $\tau(x, y) = \tau(y, x)$ *(symmetry)*

4. $\tau(x, \tau(y, z)) = \tau(\tau(x, y), z)$ *(associativity)*

Similarity and Dissimilarity

Transitivity (II)

- For similarity functions, $e = \sup I$ (and then I is I_s)

- For dissimilarity functions, $e = \text{ínf}\, I$ (and then I is I_d).

This definition reduces to that of uninorms when $I = [0, 1]$.

Similarity and Dissimilarity

Transitivity (and III)

Transitivity: A similarity s defined on X is called τ_s-transitive if there is a transitivity operator τ_s such that the following inequality holds:

$$s(x, y) \geq \tau_s(s(x, z), s(z, y)) \ \forall x, y, z \in X$$

A dissimilarity d defined on X is called τ_d-transitive if there is a transitivity operator τ_d such that the following inequality holds:

$$d(x, y) \leq \tau_d(d(x, z), d(z, y)) \ \forall x, y, z \in X$$

Similarity and Dissimilarity

Equivalences (I)

1. Consider the set of all ordered pairs of elements of X and denote it $X \times X$.

2. Every similarity s induces a preorder relation in $X \times X$.

3. This preorder is "to belong to a class of equivalence with less or equal similarity value".

Formally, given X, s, we consider the preorder $\preceq$ given by

$$(x,y) \preceq (x',y') \Longleftrightarrow s(x,y) \leq s(x',y'),\ \forall (x,y),(x',y') \in X \times X$$

(analogously for dissimilarities)

Similarity and Dissimilarity

Equivalences (II)

Definition 4 *Two similarities (or two dissimilarities) defined in the same reference set X are* **equivalent** *if they induce the same preorder.*

(note equivalence is an equivalence relation)

An equivalence function $\bar{f}$ is such that $\bar{f} \circ s$ is a similarity equivalent to s.

(analogously, $\bar{f} \circ d$ is a dissimilarity equivalent to d).

Similarity and Dissimilarity

Equivalences (III)

Theorem 1 *Let s_1 be a transitive similarity and d_1 a transitive dissimilarity. Denote by τ_{s_1} and τ_{d_1} their respective transitivity operators. Let $\bar{f}$ be an equivalence function. Then:*

1. *The equivalent similarity $s_2 = \bar{f} \circ s_1$ is τ_{s_2}-transitive, where*

   $$\tau_{s_2}(a,b) = \bar{f}(\tau_{s_1}(\bar{f}^{-1}(a), \bar{f}^{-1}(b)))\ \forall a,b \in I_{s_2}$$

2. *The equivalent dissimilarity $d_2 = \bar{f} \circ d_1$ is τ_{d_2}-transitive, where*

   $$\tau_{d_2}(a,b) = \bar{f}(\tau_{d_1}(\bar{f}^{-1}(a), \bar{f}^{-1}(b)))\ \forall a,b \in I_{d_2}$$

Similarity and Dissimilarity

Transformations

Definition 5 *A* $[0,1]$*-transformation function* $\widehat{n}$ *is a decreasing bijection on [0,1] (implying that* $\widehat{n}(0) = 1, \widehat{n}(1) = 0$*, continuity and the existence of an inverse). A transformation function* $\widehat{n}$ *is involutive if* $\widehat{n}^{-1} = \widehat{n}$*.*

Definition 6 *A transformation function* $\widehat{f}$ *is the composition of two equivalence functions and a* $[0,1]$*-transformation function:*

$$\widehat{f} = \bar{f}_1^* \circ \widehat{n} \circ \bar{f}_2^{*\,-1}$$

Similarity and Dissimilarity

Duality

Definition 7 *Consider s and d and a transformation function $\widehat{f} : I_s \to I_d$. We say that s and d are* **dual** *by $\widehat{f}$ if $d = \widehat{f} \circ s$ or, equivalently, if $s = \widehat{f}^{-1} \circ d$. This relationship is written as a triple $\prec s, d, \widehat{f} \succ$.*

Theorem 2 *Given a dual triple $\prec s, d, \widehat{f} \succ$,*

1. *d is strongly reflexive if and only if s is strongly reflexive.*
2. *d is closed if and only if s is closed.*
3. *d has (unitary) complement if and only if s has (unitary) complement.*
4. *d is τ_d-transitive only if s is τ_s-transitive, where*

   $$\tau_d(x, y) = \widehat{f}(\tau_s(\widehat{f}^{-1}(x), \widehat{f}^{-1}(y)))\ \forall x, y \in I_d$$

Similarity and Dissimilarity

Example 1 (I)

Consider the function $d(x,y) = e^{|x-y|} - 1$. This is a strong reflexive and unbounded dissimilarity with codomain $I_d = [0, +\infty)$.

- It can be expressed as the composition of $\bar{f}(z) = e^z - 1$ and $d'(x,y) = |x-y|$.

- Thus, it is τ_d-transitive with $\tau_d(a,b) = ab + a + b$.

- Consequently,

$$d(x,y) = d(x,z) + d(z,y) + d(x,z) \cdot (z,y),\ \forall x,y,z \in \mathbb{R}$$

Similarity and Dissimilarity

Example 1 (II)

To see this, use that d' is d'-transitive with $d'(a,b) = a + b$ and Theorem 1:

$$\tau_d(a,b) = \bar{f}(\bar{f}^{-1}(a), \bar{f}^{-1}(b)) = e^{\ln(1+a)+ln(1+b)} - 1$$

$$= (1+a)(1+b) - 1 = ab + a + b$$

Similarity and Dissimilarity

Example 2 (I)

1. Consider a dissimilarity function between two binary trees, to measure differences between nodes but the structure of the tree.

2. Consider a simple tree coding function D that assigns a unique value for each tree. This value is first coded as a binary number of length $2^h - 1$, being h the height of the tree.

3. The reading of the code as a natural number is the tree code.

Note that D is not a bijection, since there are numbers that do not code a valid binary tree.

Similarity and Dissimilarity

Example 2 (and II)

Consider now the following dissimilarity function, where A and B are binary trees. The symbol $\oslash$ represents the empty tree with value 0.

$$d(A,B) = \begin{cases} \text{máx}\left(\frac{D(A)}{D(B)}, \frac{D(B)}{D(A)}\right) & \text{if } A \neq \oslash \text{ and } B \neq \oslash \\ 1 & \text{if } A = \oslash \text{ and } B = \oslash \\ D(A) & \text{if } A \neq \oslash \text{ and } B = \oslash \\ D(B) & \text{if } A = \oslash \text{ and } B \neq \oslash \end{cases}$$

Similarity and Dissimilarity

Example 2 (III)

- This d is strong reflexive and unbounded dissimilarity with $I_d = [1, \infty)$.

- If we impose a limit to the height of the trees, then d is also upper bounded and closed.

- It is transitive with the *product* operator:

  for any three trees A, B, C, $d(A, B) \leq d(A, C) \cdot d(C, B)$.

Similarity and Dissimilarity

Example 2 (and IV)

If we apply the (equivalence) function $\breve{f}(z) = log\ z$ to d we receive a dissimilarity $d' = \breve{f} \circ d$, where the properties of d are kept in d'.

The transitivity operator is changed using Theorem 1, to $\tau_{d'}(a,b) = a + b$.

We obtain a metric dissimilarity over trees fully equivalent to the initial choice of d.

Similarity and Dissimilarity

Conclusions

- A (new?) standard definition of similarity and dissimilarity? not quite
- Establish some operative grounds on the definition of these widely used concepts
- Keep the framework flexible, but not too much! Cannot fit all tastes
- Where do similarities and dissimilarities come from?