

# A Relevance-Based Learning Model of Fuzzy Similarity Measures

Hoel Le Capitaine

**Abstract**—Matching pairs of objects is a fundamental operation in data analysis. However, it requires the definition of a similarity measure between objects that are to be matched. The similarity measure may not be adapted to the various properties of each object. Consequently, designing a method to learn a measure of similarity between pairs of objects is an important generic problem in machine learning. In this paper, a general framework of fuzzy logical-based similarity measures based on  $\top$ -equalities that are derived from residual implication functions is proposed. Then, a model that allows us to learn the parametric similarity measures is introduced. This is achieved by an online learning algorithm with an efficient implication-based loss function. Experiments on real datasets show that the learned measures are efficient at a wide range of scales and achieve better results than existing fuzzy similarity measures. Moreover, the learning algorithm is fast so that it can be used in real-world applications, where computation times are a key feature when one chooses an inference system.

**Index Terms**—Fuzzy logic, metrics, online algorithms, similarity learning.

## I. INTRODUCTION

WHILE similarity is an essential concept in human reasoning and plays a fundamental role in theories of knowledge, there is no unique and general-purposed definition of similarity. The reason for this lack of a definition comes from the fact that one can find practical cases where similarity properties are not satisfied (e.g., symmetry, indiscernibility, or transitivity; see [1]). Indeed, several studies (see [2] and [3] and references therein) have shown that similarity measures do not necessarily have to be transitive, implying a contradiction with the most usual approach of comparison, based on geometrical assumptions in the feature space.

Fuzzy set theory provides a consistent basis for information processing and an elegant, mathematically well-founded, representation of the uncertainty in the data. Since the data that are to be processed are often imprecise, using fuzzy set theory or its derivatives (e.g., possibility theory or belief function theory) has become a common approach in recent years [4].

In this paper, similarity measures are defined by the use of  $\top$ -equalities that are derived from fuzzy residual implications. It does not suffer from the drawbacks of the conventional metric

approaches and allows us to obtain concave or convex isosimilarity contours.

However, the number of similarity measures that are induced by the proposed framework is infinite, because of the infinite number of triangular norms (t-norms). In practice, the user must choose the similarity measure, and this wide range of choice is problematic. Consequently, *learning* the similarity measure from the available data is proposed. To this aim, the similarity measure is defined such that relevance degrees between objects are respected when these are ranked according to their pairwise similarity.

Additionally, an online learning setting is adopted, where, in contrast with batch methods, samples are considered one at a time. This enables one to treat large datasets, while keeping satisfying performances [5].

This paper is organized as follows. Section II first reviews the main approaches dealing with similarity measurement. Then,  $\top$ -equalities are used to define new classes of (parametric) similarity measures. Section III presents the learning algorithm and its properties, as well as examples. In Section IV, the use of the learning algorithm for supervised classification is described, and some comments on performances are given. Finally, conclusions and perspectives are drawn in Section V.

## II. FUZZY SIMILARITY MEASURES

### A. Basic Material

Aggregating values plays an important role in decision-making systems. Given  $n$  values, an aggregation operator is a mapping  $\mathcal{A} : [0, 1]^n \rightarrow [0, 1]$ , satisfying boundary conditions and monotonicity. In the literature, one finds many aggregation operators, e.g., t-norms, *Ordered Weighted Averaging* (OWA) operators,  $\gamma$ -operators, or fuzzy integrals. They belong to several categories, depending on the way the values are aggregated: conjunctives, disjunctives, averaging, and mixed operators. See [6] and [7] for large, yet comprehensive, surveys on aggregation operators.

A t-norm is an increasing, associative, and commutative mapping  $\top : [0, 1]^2 \rightarrow [0, 1]$ , satisfying the boundary condition  $\top(x, 1) = x$  for all  $x \in [0, 1]$ . The most popular continuous t-norms are the minimum  $\top_M(x, y) = \min(x, y)$ , the product  $\top_P(x, y) = xy$ , and the Łukasiewicz t-norm  $\top_L(x, y) = \max(x + y - 1, 0)$ . Various parametric families involving a real value  $\lambda$  lying in a specified domain have been introduced. The parametric t-norms that are used in the sequel are given in Table I.

A general problem in fuzzy logic is to handle conditional statements *if  $x$ , then  $y$* , where  $x$  and  $y$  are fuzzy predicates.

Manuscript received November 17, 2010; revised February 12, 2011 and May 16, 2011; accepted July 19, 2011. Date of publication August 30, 2011; date of current version February 7, 2012.

The author is with the Laboratoire d'Informatique de Nantes Atlantique, CNRS, UMR 6241, University of Nantes, 44000 Nantes, France and also with the University La Rochelle, Laboratoire de Mathématiques, Image et Applications, EA3165 F-17000 La Rochelle, France (e-mail: hoel.lecapitaine@univ-nantes.fr).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/TFUZZ.2011.2166079

TABLE I  
PARAMETRIC T-NORMS

Family	t-norm
Hamacher	$\top_H(x, y) = \begin{cases} \text{Drastic t-norm} & \text{if } \lambda = \infty \\ 0 & \text{if } \lambda = x = y \\ \frac{xy}{\lambda + (1-\lambda)(x+y-xy)} & \text{if } \lambda \in [0, \infty[ \text{ and } (\lambda, x, y) \neq (0, 0, 0) \end{cases}$
Dombi	$\top_D(x, y) = \begin{cases} \text{Drastic t-norm} & \text{if } \lambda = 0 \\ \top_M(x, y) & \text{if } \lambda = \infty \\ \left( 1 + \left( \left( \frac{1-x}{x} \right)^\lambda + \left( \frac{1-y}{y} \right)^\lambda \right)^{1/\lambda} \right)^{-1} & \text{if } \lambda \in ]0, \infty[ \end{cases}$
Yager	$\top_Y(x, y) = \begin{cases} \text{Drastic t-norm} & \text{if } \lambda = 0 \\ \top_M(x, y) & \text{if } \lambda = \infty \\ \max \left( 1 - ((1-x)^\lambda + (1-y)^\lambda)^{1/\lambda}, 0 \right) & \text{if } \lambda \in ]0, \infty[ \end{cases}$
Frank	$\top_F(x, y) = \begin{cases} \top_M(x, y) & \text{if } \lambda = 0 \\ \top_P(x, y) & \text{if } \lambda = 1 \\ \top_L(x, y) & \text{if } \lambda = \infty \\ \log_\lambda \left( 1 + \frac{(\lambda^x - 1)(\lambda^y - 1)}{\lambda - 1} \right) & \text{if } \lambda \in ]0, 1[ \cup ]1, \infty[ \end{cases}$
Schweizer-Sklar	$\top_{SS}(x, y) = \begin{cases} \top_M(x, y) & \text{if } \lambda = -\infty \\ \top_P(x, y) & \text{if } \lambda = 0 \\ \text{Drastic t-norm} & \text{if } \lambda = \infty \\ \left( \max(x^\lambda + y^\lambda - 1, 0) \right)^{\frac{1}{\lambda}} & \text{if } \lambda \in ]-\infty, 0[ \cup ]0, \infty[ \end{cases}$

A widely used method consists of managing them by using functions  $I : [0, 1] \times [0, 1] \rightarrow [0, 1]$  such that the truth value of  $I$  depends on the initial propositions  $x$  and  $y$ . We generally speak about an *implication function* if  $I$  is nonincreasing in the first variable, nondecreasing in the second variable,  $I(0, 0) = I(1, 1) = 1$ , and  $I(1, 0) = 0$ ; see [8] and [9] for recent surveys on fuzzy implication functions.

In this paper, residual implication functions are considered. Given a t-norm  $\top$ , its corresponding residuum is defined by

$$I_\top(x, y) = \sup_t \{t \in [0, 1] \mid \top(x, t) \leq y\}. \quad (1)$$

In the sequel, the following notation is adopted.

P1)  $X = \{x_1, \dots, x_n\}$  is the (supposed finite) universe of discourse.

P2)  $\mathcal{C}(X)$  and  $\mathcal{F}(X)$  are the sets of all crisps and fuzzy sets in  $X$ , respectively.

P3)  $f_A(x), \forall x \in X$  is the membership function of a fuzzy set  $A$  over  $X$ .

There are several ways to compare fuzzy values or fuzzy quantities. The first one is based on a broad class of measures of equality based on a distance measure that is specified for membership functions of fuzzy sets. The second category involves set-theoretic operations for fuzzy sets (fuzzy intersection, union, and cardinality) [10], [11]. Finally, a third way of defining a similarity measure consists of using logical concepts of fuzzy implication, as first suggested in [12], following the

seminal paper of [13]. Note that a fourth approach, i.e., relying on morphological operators, has been proposed in [14].

First of all, the basic definition of a fuzzy similarity measure is recalled.

*Definition 1:* A mapping  $\mathcal{S} : \mathcal{F}(X) \times \mathcal{F}(X) \rightarrow [0, 1]$  is called a similarity measure if it satisfies the following.

P1)  $\mathcal{S}(A, B) = \mathcal{S}(B, A), \quad \forall A, B \in \mathcal{F}(X).$

P2)  $\mathcal{S}(A, A) = 1, \forall A \in \mathcal{F}(X).$

P3)  $\mathcal{S}(D, D^c) = 0, \quad \forall D \in \mathcal{C}(X).$

P4)  $\forall A, B, C \in \mathcal{F}(X), \text{ if } A \subseteq B \subseteq C, \text{ then } \mathcal{S}(A, C) \leq \mathcal{S}(A, B) \wedge \mathcal{S}(B, C)$

or, equivalently

$\forall A, B, C, D \in \mathcal{F}(X), \text{ if } A \subseteq B \subseteq C \subseteq D, \text{ then } \mathcal{S}(A, D) \leq \mathcal{S}(B, C).$

However, most of these properties, if not all, are subject to criticisms and debates. Therefore, it contributes to the emphasis on the lack of a clear definition of a general-purposed similarity measure. The symmetry property P1) is still subject to experimental investigations: If  $\mathcal{S}(x, y)$  is the answer to the question *how is  $x$  similar to  $y$ ?*, then one focusses more on the feature  $x$  than on  $y$ . This corresponds to the notion of saliency [10] of  $x$  and  $y$ : If  $y$  is more salient than  $x$ , then  $x$  is more similar to  $y$  than *vice versa*, which is experimentally confirmed by observing asymmetries in confusion matrices. Property P2) is also discussed, since in some cases, the similarity of a point to itself is not 1. Property P4) is the most debatable one. In particular, Tversky and Gati [1] present and criticize the segmental

additivity, i.e.,  $d(A, C) = d(A, B) + d(B, C)$  property of metrics. This assumption is rather intuitive and include a wide class of distance functions: all the Minkowski metrics and Riemannian curved geometries. When dealing with fuzzy sets and considering  $\top$ -equivalences and  $\top$ -equalities, the triangle inequality is replaced by the  $\top$ -transitivity.

The metric approach takes its roots from studies on how to measure the distance between two real functions. The basic concept is to consider a fuzzy set as a point in a vector space. The general form of a Minkowski  $r$ -metric that is defined for  $r \geq 1$  is usually taken:

$$d_r(A, B) = \left( \sum_{x \in X} |f_A(x) - f_B(x)|^r \right)^{1/r}. \quad (2)$$

This metric induces well-known distance functions: Hamming (or Manhattan) for  $r = 1$ , Euclidean for  $r = 2$ , and Tchebychev (or sup distance) for  $r = \infty$  (which can be written as  $\sup_{x \in X} |f_A(x) - f_B(x)|$ ). Note that for  $r < 1$ ,  $d_r$  does not define a metric, since it violates the triangle inequality. The Hamming and Euclidean distances are also denoted as  $L_1$  and  $L_2$  norms, respectively. Increasing the value  $r$  gives more weight to large differences in feature values.

Starting from a distance function, several methods have been proposed to obtain a similarity measure. The most natural way is to use a nonincreasing function  $g$  such that  $S(A, B) = g(d(A, B))$ . This idea relies on the link between a distance in a vector space and a similarity. Intuitively, when the distance increases, the similarity should decrease. Popular choices of this function  $g$  are the ‘‘Cauchy-like’’ function  $g(x) = 1/(1 + x)$  suggested by Zimmerman and Zysno [15]. Another function, which comes from the work of Shepard [16], consists of taking  $g(x) = \exp(-\alpha x)$ . The use of the exponential law comes from the underlying observation that there is a nonlinear relationship between the distance and the similarity which is concave upward. However, to obtain similarity measures with the help of distance measures implies that the similarity satisfies the metric properties, which is far from obvious in practice. This has led researchers to prefer nonmetric models of similarity.

The second way to compare fuzzy values comes from some basic set-theoretic considerations, where union, intersection, and complementation are defined for fuzzy sets. While metric-based measures can be interpreted as proximity of fuzzy sets, the set-theoretic measures can be viewed as an approximate equality. Probably, the most famous set-theoretic measure is the consistency index, which is defined by the supremum over  $X$  of  $A \cap B$ . See [17] for more details on set-theoretic comparison measures.

The last main approach consists of considering the implication degrees of the elements that belong to  $A$  over the elements of  $B$  [13], and *vice versa*. The implication degree is obtained by using one of the fuzzy implication functions that are described in the previous section. Formally, having  $x$  as an element of  $A$  and  $y$  as an element of  $B$ , their implication degree is given by  $I(x, y)$ . More details on related logical comparison measures are given in [18]. In order to obtain the implication degrees of  $A$  over  $B$  and  $B$  over  $A$ , we use the bi-implication  $bI$ , which is

defined by

$$bI(x, y) = \min(I(x, y), I(y, x)). \quad (3)$$

In [19], the authors showed that in the case of residual implications,  $bI$  is a  $\top$ -equality if and only if  $\top$  is a left-continuous t-norm. In this paper, residual implications are used, since the others (S, QL, D) do not define  $\top$ -equalities.

*Theorem 1:* Let  $bI_\top$  be a biresidual implication function defining a  $\top$ -equality  $E_\top$ . For arbitrary  $A, B \in \mathcal{F}(X)$ , let

$$\mathcal{S}(A, B) = \bigcup_{i=1}^n E_\top(f_A(x_i), f_B(x_i)) \quad (4)$$

for all  $x_i$  in  $X$ , where  $\mathcal{A}$  is an aggregation operator. Then,  $\mathcal{S}$  is a similarity measure.

*Proof:*

P1) By definition,  $E_\top(x, x) = 1$  holds for any  $x \in [0, 1]$ . By boundary conditions on  $\mathcal{A}$ , the equality  $\mathcal{S}(A, A) = 1$  is obtained.

P2) By commutativity of  $\top$ -equality

$$\begin{aligned} \mathcal{S}(A, B) &= \bigcup_{i=1}^n E_\top(f_A(x_i), f_B(x_i)) \\ &= \bigcup_{i=1}^n E_\top(f_B(x_i), f_A(x_i)) = \mathcal{S}(B, A). \end{aligned}$$

P3) By definition,  $I_\top(1, 0) = 0$  so that  $E_\top(1, 0) = 0$ . By boundary conditions on  $\mathcal{A}$ ,  $\mathcal{S}(D, D^c) = 0$ .

P4) Since  $A \subseteq B \subseteq C \subseteq D$ , for all  $x_i \in X$

$$f_D(x_i) \geq f_C(x_i) \quad (5)$$

$$f_B(x_i) \geq f_A(x_i) \quad (6)$$

hold. By nonincreasingness in the first variable and nondecreasingness in the second variable of  $I_\top$ , for all  $x_i \in X$ , we have  $I_\top(f_D(x_i), f_A(x_i)) \leq I_\top(f_C(x_i), f_A(x_i))$  by (5) and  $I_\top(f_C(x_i), f_A(x_i)) \leq I_\top(f_C(x_i), f_B(x_i))$  by (6). Using  $E_\top(x, y) = I(\max(x, y), \min(x, y))$  and (5) and (6), we obtain  $E_\top(f_D(x_i), f_A(x_i)) \leq E_\top(f_C(x_i), f_A(x_i))$  and  $E_\top(f_C(x_i), f_A(x_i)) \leq E_\top(f_C(x_i), f_B(x_i))$ . Last, monotonicity of  $\mathcal{A}$  gives  $\mathcal{S}(A, D) \leq \mathcal{S}(B, C)$  which concludes the proof. ■

*Remark 1:* The minimum operator is used in (3), but any t-norm will also fulfill the desired properties, because for any  $I_\top$ ,  $x \leq y \Rightarrow I_\top(x, y) = 1$  (by the ordering property; see [8]).

## B. Examples

Taking particular t-norms and aggregation operators enables one to retrieve well-known similarity measures. For instance, taking the arithmetic mean and  $\top_P, \top_L$ , the two measures that are proposed in [20] are retrieved. Remark that for the arithmetic mean and the Łukasiewicz implication, the exact contraposition of the  $L_1$  norm is obtained, and to take the minimum instead of the mean gives the contraposition of the Tchebychev norm. More details on fuzzy similarity measures that can be obtained with the generic framework are given in [18].

*Remark 2:* In this paper, the arithmetic mean and the minimum and maximum operators are taken as overall aggregation

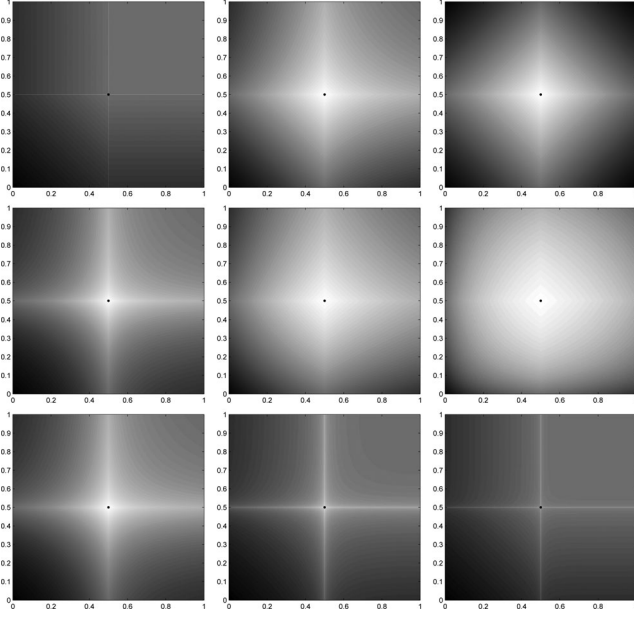


Fig. 1. Examples of isosimilarity contours between a reference set  $A = [\frac{1}{2}]$  (which is denoted by  $\cdot$ ) and all fuzzy sets  $\mathcal{F}(X)$ ,  $n = 2$ , where high and low values correspond to white and black, respectively. The arithmetic mean is used. First row:  $I_{TM}$ ,  $I_{TP}$ , and  $I_{TL}$ , respectively, from left to right. Second row: Hamacher similarity, where  $\lambda = 0, 2$ , and  $5$  from left to right. Third row: Dombi similarity, where  $\lambda = 1, 2$ , and  $5$  from left to right.

operators. However, any aggregation operator such as geometric means, OWA operators, or fuzzy integrals can be used. The use of the latter is currently under study in order to take into account feature dependences, as in [21].

For purposes of illustration, Fig. 1 shows some examples of fuzzy similarity measures, as well as the influence of the parameter  $\lambda$  for the Hamacher and the Dombi-based measures. The different plots show the similarity value of a given fuzzy set  $A = \{0.5/x_1, 0.5/x_2\}$  to all the possible 2-D fuzzy sets  $B$  for various  $I_T$ . As one could expect, the closer to  $x_1$  or  $x_2$ , the higher the similarity. One can also note that different  $I_T$  lead to different shapes of isosimilarity contours. According to the third line of Fig. 1, one can remark that the Hamacher similarity measure gives a concave contour with a low  $\lambda$ , which becomes convex when the parameter increases. Consequently, the similarity measure is much more flexible than usual metrics, since convex and concave shapes can be modeled. The interesting point is that a measure without segmental additivity can be obtained, which is consistent with the axioms of Tversky and Gati, by choosing a parameter making the measure concave. Another parameter value allows us to obtain convex isosimilarity contours, yielding a usual metric satisfying the triangle inequality. This property is very important, since a concave contour of isosimilarities means that the triangle inequality is violated [22]. However, when dealing with fuzzy sets, the  $\top$ -transitivity is preserved, as discussed earlier.

Moreover, the measures enable us to capture the idea that similarity is easier to quantify and makes more sense locally (i.e., small variations) than far away in the feature space, where comparisons and judgments of similarity are difficult. An appealing

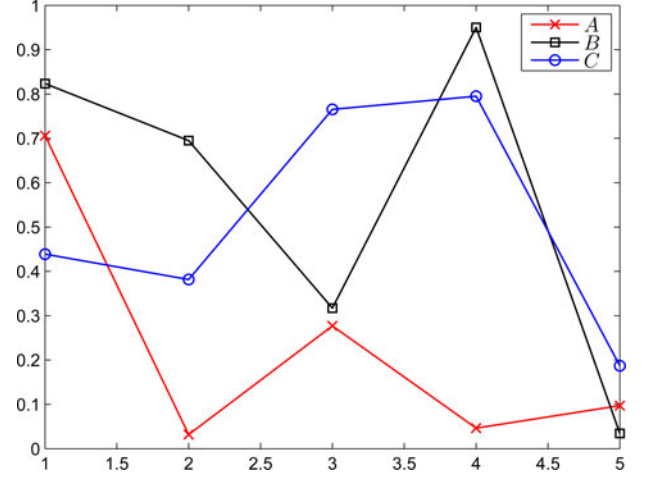


Fig. 2. Three fuzzy sets  $A$ ,  $B$ , and  $C$ , where the ranking of similarity measures  $S(A, B)$  and  $S(A, C)$  is not obvious.

TABLE II  
RANKING CHANGES FOR PARAMETRIC  
SIMILARITY MEASURES

Fuzzy sets	$\lambda$	Hamacher similarity
A, B	2.0	<b>0.505</b>
A, C	2.0	0.475
A, B	15.0	0.650
A, C	15.0	<b>0.762</b>
Fuzzy sets	$\lambda$	Dombi similarity
A, B	2.0	<b>0.271</b>
A, C	2.0	0.206
A, B	0.25	0.686
A, C	0.25	<b>0.818</b>

property of parametric similarity measures is that it allows us to rank objects in a different order depending on the parameter  $\lambda$ . Keeping the Hamacher and Dombi-based similarity measures, three fuzzy sets  $A$ ,  $B$ , and  $C$  are used in this experiment. They are defined by

- 1)  $A = \{0.7/x_1, 0.05/x_2, 0.32/x_3, 0.07/x_4, 0.10/x_5\}$ ;
- 2)  $B = \{0.82/x_1, 0.75/x_2, 0.36/x_3, 0.90/x_4, 0.04/x_5\}$ ;
- 3)  $C = \{0.45/x_1, 0.34/x_2, 0.69/x_3, 0.57/x_4, 0.16/x_5\}$ .

They can be visually inspected in Fig. 2. Even for a human, the ranking of  $S(A, B)$  and  $S(A, C)$  is not an easy task. In Table II, the amount of similarity between  $A$ ,  $B$ , and  $C$  with various  $\lambda$  values are given. In this table, maximum similarity values (with respect to the other one) are in bold font. According to this table, it can be seen that one can find a value of  $\lambda_1$  such that  $S(A, B) > S(A, C)$  and a value  $\lambda_2$  such that  $S(A, B) < S(A, C)$ . Consequently, the three objects  $A$ ,  $B$ , and  $C$  are ranked differently, depending on  $\lambda$ . If one cannot find two values of  $\lambda$  such that the ranking of three objects is different, then no matter how the measures are learned, they are all equivalent in terms of information retrieval and do not provide efficient similarity measures.



### III. LEARNING THE SIMILARITY FUNCTION

#### A. Loss Function

In this paper, an online learning procedure is adopted. In this setting, the algorithm sequentially receives samples and predicts an output. Once the output is obtained, the algorithm gets a feedback indicating its goodness. Afterward, parameters can be changed so that the probability of correct output increases in the next step. An appealing property of online algorithms is that they are relatively simple to implement and quickly (i.e., with a few number of iterations) provides good performances [5]. Moreover, the learning algorithm does not require the use of the entire learning set so that a great improvement in terms of computation is obtained when one considers large datasets.

Given information related to the relevance degree of matching between objects described by the help of fuzzy sets, the problem of learning a similarity function  $\mathcal{S}$  is addressed. A relevance degree  $\mathcal{R}$  can be understood as a pairwise function of objects  $a$  and  $b$ , which states how strongly  $a$  and  $b$  are related. For instance, in supervised classification, this could encode that both are in the same class. The relevance degree may be defined by user's input, which may be based on results of queries, or by knowledge on objects (supervised classification). More formally, there is a set  $\mathcal{C}$  of  $n$  objects and some relevance degrees  $\mathcal{R}$  between objects belonging to  $\mathcal{C}$ . The relevance degrees take their values in the unit interval, 1 meaning complete satisfaction, whereas 0 means that there is no reason at all to match the two objects. Supposing that  $\mathcal{R}(a, b) > \mathcal{R}(a, c)$ , the aim is to define and learn a parametric similarity measure  $\mathcal{S}_\lambda$  such that

$$\mathcal{S}_\lambda(a, b) \rightarrow \mathcal{S}_\lambda(a, c) < 1. \quad (7)$$

The  $\rightarrow$  operator is a fuzzy implication. In the sequel, only residual implications are considered, but this can be adapted to any fuzzy implication.

The hinge loss function for every triplet  $a, b, c$  is defined by

$$\ell_\lambda = \max\{0, \mathcal{S}_\lambda(a, b) \rightarrow \mathcal{S}_\lambda(a, c) - 1\}. \quad (8)$$

Naturally, the proposed learning framework can be built with many other fuzzy implication functions (parametric included). For simplicity, the Łukasiewicz implication is used in the sequel so that the loss becomes

$$\ell_\lambda = \max\{0, -\mathcal{S}_\lambda(a, b) + \mathcal{S}_\lambda(a, c)\}. \quad (9)$$

The aim is to minimize the total loss  $L_\lambda$  on the learning set. The total loss is defined by summing up the individual losses over all the triplet of the learning set

$$L_\lambda = \sum_{(a,b,c) \in \mathcal{C}} \ell_\lambda(a, b, c). \quad (10)$$

However, even in the case of moderately large datasets, the number of all possible triplets is very large, and exhaustive computation becomes intractable in practice. Therefore, an on-line learning scheme is adopted. Consequently, the minimum loss is not searched in the entire learning set but with randomly selected samples of the set. More precisely, the sample  $a$  is randomly selected in the learning set, and  $b$  and  $c$  are also uniformly

---

#### Algorithm 1 Online Similarity Learning algorithm

---

```

1: procedure ONLINE SIMILARITY LEARNING
2:    $\lambda_0 \leftarrow$  initial value
3:   repeat
4:     Random selection of  $a, b$  and  $c$  such that  $\mathcal{R}(a, b) >$ 
        $\mathcal{R}(a, c)$ .
5:      $G_i \leftarrow \frac{\partial(\mathcal{S}_\lambda(a, b) - \mathcal{S}_\lambda(a, c))}{\partial \lambda}$  ▷ See Table III.
6:      $\lambda \leftarrow \lambda_{i-1} + \alpha G_i$ 
7:   until convergence
8:   return  $\lambda$ 
9: end procedure

```

---

sampled such that  $a$  and  $b$  share the same class, while  $c$  belongs to another class.

#### B. Optimal Updating

Since the similarity measure is determined by its parameter value  $\lambda$ , its optimal value is searched by using an online learning algorithm based on sequential updates of  $\lambda$ . Depending on the similarity measure, the initial value  $\lambda_0$  varies; see Section IV for comments and Section II for a discussion on the various similarity measures that are used in the sequel. We then find  $\lambda$  such that

$$\lambda_i = \operatorname{argmin}_\lambda \left( \frac{1}{2} \|\lambda - \lambda_{i-1}\|^2 + \alpha \ell_\lambda(a, b, c) \right) \quad (11)$$

where  $\alpha \geq 0$ . In other words, during the optimization process,  $\lambda_i$  is selected in order to obtain a tradeoff between minimizing the loss on  $(a, b, c)$  and staying quite close to its previous value  $\lambda_{i-1}$ . This tradeoff is controlled by the “aggressiveness” parameter  $\alpha$ . Naturally, if  $\alpha$  is set to zero, then the optimal  $\lambda_i$  value is equal to  $\lambda_{i-1}$ . In contrast, setting  $\alpha$  to a high value imposes an important weight on the loss function. It is clear that when the loss  $\ell_\lambda(a, b, c)$  is equal to zero, then the optimal parameter value does not change from the previous iteration, i.e.  $\lambda_i = \lambda_{i-1}$ . Otherwise, the objective function  $\mathcal{L}$  is defined by

$$\mathcal{L}(\lambda) = \frac{1}{2} \|\lambda - \lambda_{i-1}\|^2 + \alpha (-\mathcal{S}_\lambda(a, b) + \mathcal{S}_\lambda(a, c)). \quad (12)$$

The optimal solution with respect to  $\lambda$  is such that the gradient of  $\mathcal{L}$ , i.e.,  $\partial \mathcal{L}(\lambda) / \partial \lambda$ , vanishes

$$\begin{aligned} \frac{\partial \mathcal{L}(\lambda)}{\partial \lambda} &= \lambda - \lambda_{i-1} - \alpha \frac{\partial(\mathcal{S}_\lambda(a, b) - \mathcal{S}_\lambda(a, c))}{\partial \lambda} \\ &= 0. \end{aligned} \quad (13)$$

Let  $G_i$  be the gradient value  $\frac{\partial(\mathcal{S}_\lambda(a, b) - \mathcal{S}_\lambda(a, c))}{\partial \lambda}$ . Therefore, the optimal new value  $\lambda$  is given by

$$\lambda = \lambda_{i-1} + \alpha G_i$$

The corresponding algorithm is described in Algorithm 1.

Naturally, the value  $G_i$  depends on the similarity measure that is used. The corresponding gradients for each similarity measure are given in Table III. For writing convenience,  $f_1 = \max(a_k, b_k)$  and  $f_2 = \min(a_k, b_k)$  for  $k \in N$ . The performance of the learning algorithm also depends on the initial

TABLE III  
SIMILARITY MEASURES AND THEIR CORRESPONDING RELATIVE GRADIENT WITH RESPECT TO  $\lambda$  WHEN USING THE ŁUKASIEWICZ  
IMPLICATION IN THE LOSS FUNCTION

$S$	$\frac{\partial}{\partial \lambda} (S_\lambda(a_k, b_k))$
$S_H$	$\frac{(f_2 - f_1 f_2)(f_1 - f_2)}{(f_2(\lambda_i + f_1 - \lambda_i f_1) + f_1 - f_2)^2}$
$S_Y$	$-\left((1 - f_2)^\lambda - (1 - f_1)^\lambda\right)^{1/\lambda} \frac{1}{\lambda^2} \left( \lambda \frac{\log(1 - f_2)(1 - f_2)^\lambda - \log(1 - f_1)(1 - f_1)^\lambda}{(1 - f_2)^\lambda - (1 - f_1)^\lambda} - \log((1 - f_2)^\lambda - (1 - f_1)^\lambda) \right)$
$S_D$	$\left( \left( \frac{1 - f_2}{f_2} \right)^\lambda - \left( \frac{1 - f_1}{f_1} \right)^\lambda \right)^{1/\lambda} \left( -\frac{1}{\lambda^2} \log \left( \left( \frac{1 - f_2}{f_2} \right)^\lambda - \left( \frac{1 - f_1}{f_1} \right)^\lambda \right) + \frac{1}{\lambda} \frac{\log \left( \frac{1 - f_2}{f_2} \right) \left( \frac{1 - f_2}{f_2} \right)^\lambda - \log \left( \frac{1 - f_1}{f_1} \right) \left( \frac{1 - f_1}{f_1} \right)^\lambda}{\log \left( \left( \frac{1 - f_2}{f_2} \right)^\lambda - \left( \frac{1 - f_1}{f_1} \right)^\lambda \right)} \right)$
$S_F$	$\frac{1}{\log(\lambda)^2} \left( \frac{\log(\lambda)}{1 + \frac{(\lambda^{f_2} - 1)(\lambda - 1)}{(\lambda^{f_1} - 1)}} \frac{(\lambda^{f_1} - 1)(\lambda^{f_2} + f_2 \lambda^{f_2} - \lambda^{f_2}(f_2/\lambda) - 1) - (\lambda^{f_2} - 1)(\lambda - 1)\lambda^{f_1} f_1}{(\lambda^{f_1} - 1)^2} - \frac{1 + \frac{(\lambda^{f_2} - 1)(\lambda - 1)}{(\lambda^{f_1} - 1)}}{\lambda} \right)$
$S_{SS}$	$(1 + f_2^\lambda - f_1^\lambda)^{1/\lambda} \left( -\frac{1}{\lambda^2} \log(1 + f_2^\lambda - f_1^\lambda) + \frac{1}{\lambda} \frac{\log(f_2) f_2^\lambda - \log(f_1) f_1^\lambda}{1 + f_2^\lambda - f_1^\lambda} \right)$

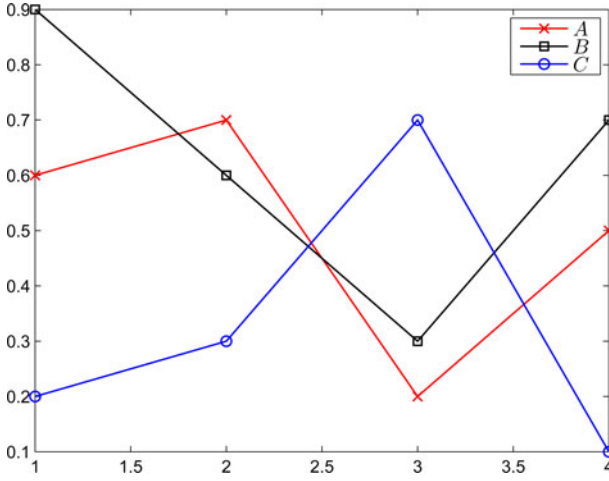


Fig. 3. Three fuzzy sets  $A$ ,  $B$ , and  $C$ , where the similarity of  $A$  and  $B$  is clearly higher than the similarity of  $A$  and  $C$ .

value  $\lambda_0$ , which is discussed in Section IV-C. When  $\lambda$  is out of definition bounds of the t-norm (e.g., negative values for Hamacher one), the last convenient value  $\lambda_{i-1}$  is returned. An important case is the one of bound values. When  $\lambda$  reaches these bounds, the equivalence that is obtained with the corresponding t-norm is used. For instance, if  $\lambda = 0$  for the SS norm, then the equivalence obtained by the product is used, thanks to the continuity around 0 of this t-norm. Note also that for a reasonable number of iterations (e.g.,  $n^2/2$ ), limit bounds such as infinity are not reached. For purposes of illustration, three fuzzy sets are used for which the ranking of the similarities  $S(A, B)$  and  $S(A, C)$  is natural. Such fuzzy sets are plotted in Fig. 3. Here, there are only three samples, and therefore, they are repeatedly selected into the learning algorithm, where  $\alpha = 1$ . If the difference between two similarities increases, then the similarity measure is more efficient for the discrimination of other objects that belong to the learning data (and hopefully on the test set). The difference  $S(A, B) - S(A, C)$  is studied as a function of the iterations. The corresponding graph is plotted in Fig. 4 (dashed line). Naturally, since  $B$  is more similar to  $A$  than  $C$ ,

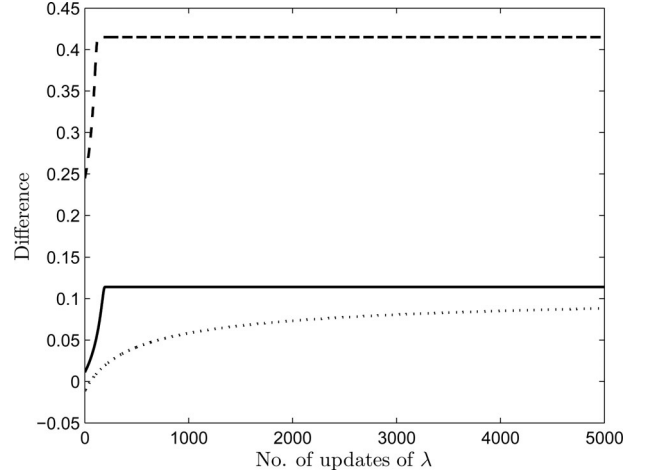


Fig. 4. Difference  $S(A, B) - S(A, C)$  as a function of the number of update of  $\lambda$  (see text for details).

even at the beginning of the algorithm, the difference is quite large. The most interesting point is the evolution of the parameter  $\lambda$  that makes  $A$  and  $B$  increasingly more similar and  $A$  and  $C$  increasingly more dissimilar as  $\lambda$  is updated. This difference does not exceed an upper bound reached around the 130th update of  $\lambda$ .

Now, the three fuzzy sets that are introduced in the previous section are considered. As already mentioned, the ranking of  $S(A, B)$  and  $S(A, C)$  is not easy. Consequently, the learning algorithm is run a first time, with the supposition that  $B$  is more similar to  $A$  than  $C$ . The corresponding difference curve is plotted in Fig. 4 as a solid line. Then, the algorithm is run a second time with the supposition that  $C$  is more similar to  $A$  than  $B$ . The  $S(A, C) - S(A, B)$  difference curve is plotted in Fig. 4 as a dotted line. As can be seen, supposing that  $B$  is more similar to  $A$  than  $C$  leads to quickly reach the upper bound (around the 200th iteration). In contrast, supposing that  $C$  is more similar to  $A$  than  $B$  requires many more iterations for the learning algorithm to reach its upper bound. Note that at the beginning, the difference is negative, that is to say  $S(A, B) > S(A, C)$ , and

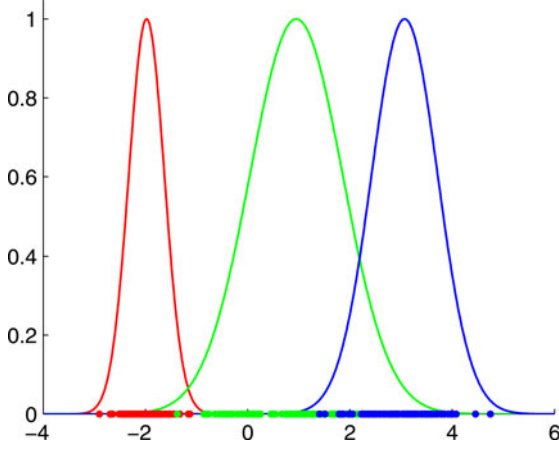


Fig. 5. Plot of the 1-D dataset and the membership functions that are obtained by the fuzzification scheme. Samples of different classes are represented by different colored dots.

as  $\lambda$  is updated,  $\mathcal{S}(A, B) < \mathcal{S}(A, C)$  is obtained, accordingly to the assumption that  $C$  is more similar to  $A$  than  $B$ . This example shows the ability of the learning algorithm to adapt the similarity measure to the data, since the ranking of the similarity between three objects can be modified according to prior knowledge.

#### IV. EXPERIMENTS

In this section, an application in supervised classification of the proposed learning model is presented.

##### A. Protocol

Let  $X = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$  be a learning set having  $c$  classes. Each sample  $\mathbf{x}_i$  is represented by a set of  $p$  features, and its discrete class label  $y_i \in \{1, \dots, c\}$  is known. In order to use the similarity measures that are previously introduced, each sample  $\mathbf{x}$  must be described by a fuzzy set. In the sequel, an automatic fuzzification scheme is proposed. Each sample is described by a discrete fuzzy set of  $c * p$  elements. The basic idea of the fuzzification is to estimate, for each feature of each class, its mean and standard deviation, giving two matrices  $M$  and  $S$  of size  $(c \times p)$ . For each sample, the fuzzy set is obtained by evaluating a membership function on each of its features with respect to all classes. For simplicity, the Gaussian membership function is used. It is defined by

$$f(x|\sigma, m) = \exp\left(-\frac{(x - m)^2}{2\sigma^2}\right) \quad (14)$$

where  $m$  is the mean, and  $\sigma$  the standard deviation. The general procedure is described in Algorithm 2, where the set  $X_j^i$  is composed of the  $j$ th feature of the class  $i$ .

*Example 1:* For purposes of illustration, we give an example for a dataset composed of a mixture of three normal 1-D distributions of 100 points each. The means are equal to  $-2, 1, 3$ , and standard deviations are equal to  $0.4, 0.9$ , and  $0.7$ , respectively. Each sample is then described by three membership degrees corresponding to the three classes. The membership functions obtained are plotted in Fig. 5.

---

#### Algorithm 2 Fuzzification of each sample $\mathbf{x}$

---

```

1: procedure FUZZIFICATION( $X$ )
2:   Set  $p \leftarrow$  number of features
3:   Set  $c \leftarrow$  number of classes
4:   for  $i = 1$  to  $c$  do
5:     for  $j = 1$  to  $p$  do
6:        $M(i, j) \leftarrow$  componentwise average( $X_j^i$ )
7:        $S(i, j) \leftarrow$  componentwise standard dev. ( $X_j^i$ )
8:     end for
9:   end for
10:  for each sample  $\mathbf{x}_k \in X$  do
11:    for  $i = 1$  to  $c$  do
12:      for  $j = 1$  to  $p$  do
13:        Set membership degrees of  $\mathbf{x}_k$  using (14),
14:        with  $M(i, j)$  and  $S(i, j)$ .
15:      end for
16:    end for
17:  end for
18: end procedure

```

---

During the learning phase, three samples are randomly selected in each iteration step. The first one, i.e.,  $a$ , is the reference one. The sample  $b$  is randomly chosen so that it belongs to the same class of  $a$ , and  $c$  is randomly selected such that its class is not the class of  $a$ . Since  $a$  and  $b$  are in the same class, and  $c$  belongs to another class,  $\mathcal{R}(a, b) > \mathcal{R}(a, c)$  holds. The performance of all similarity measures is evaluated by using standard ranking precision measures based on nearest neighbors. For each sample, all other test samples are ranked according to their similarity to the sample. The number of same-class samples among the top  $k$  similar samples is computed, giving the precision measure for this sample. When averaged over all samples, an average precision (AP) measure as a function of  $k$  is obtained as

$$\text{AP}(k) = \frac{1}{n} \sum_{i=1}^n \frac{1}{k} \sum_{j=1}^k (y(\mathbf{x}_i) = y(\mathbf{x}_j))$$

where  $\mathbf{x}_j, j = 1, \dots, k$  are the  $k$  most similar objects with respect to  $\mathbf{x}_i$ . The mean average precision (mAP) is obtained for each similarity measure by averaging the precision at level  $k$  across all  $k$  values:

$$\text{mAP} = \frac{1}{k_{\max}} \sum_{k=1}^{k_{\max}} \text{AP}(k).$$

Both AP and mAP are commonly used evaluation measures of similarity-based classifiers [5], [23].

##### B. Datasets

To validate the efficiency of the proposed learning model, a comparison of its performance on various real datasets, which is available at <http://archive.ics.uci.edu/ml/>, is conducted. The following datasets are considered.

- 1) In the well-known iris dataset [24], [25], data contains three classes of 50 samples each, where each class is a

type or iris plant. Each sample is described by four features: sepal length and width, as well as petal length and width. It is a classical dataset in the pattern recognition literature, which is known to have one class linearly separable from the other two, and two classes that are not linearly separable.

- 2) The Wine dataset contains three classes and 178 samples. Each sample is described by 13 constituents that are found in each wine. The classes are quite well separated, and samples' classification is not reputed to be a challenging task.
- 3) The Pima dataset contains two classes of 768 samples, each described by eight features. The class, i.e., 0 or 1, denotes the absence or presence of diabetes pathology.
- 4) The Yeast dataset is a genetic dataset of phylogenetic profiles for the Yeast genome. It contains 1484 samples, which are described by eight attributes. The ten classes of the problem correspond to the various localization sites of the proteins.
- 5) The Heart dataset contains 270 samples of 13 attributes each, describing medical data of each patient. The class, i.e., 0 or 1, denotes the absence or presence of heart disease.
- 6) The ionosphere dataset is composed of 351 instances. For each sample, there are 17 pulse numbers, and each pulse number is described by two attributes, giving 34 features. The goal is to discriminate radar returns from the ionosphere. The two classes correspond to radar returns showing evidence of some type of structure in the ionosphere and those that do not.
- 7) The Tae dataset consists of evaluations of 151 teaching assistants. Each assistant is described by five attributes. The teacher may belong to three classes: low, medium, or high.
- 8) The Ecoli dataset contains 336 observations described by seven attributes. Each sample is a sequence of protein, and the eight classes correspond to their respective localization sites.
- 9) The Liver Disorders dataset contains 345 samples. Each sample constitutes the record of a single male individual and is described by seven features such as blood tests or number of drinks per day. The class, i.e., 0 or 1, denotes some sort of liver disorder.
- 10) The Newthyroid dataset is composed of 215 instances. The instances are described by five chemical features that are designed for the diagnostic of the functioning of the thyroid gland. This functioning may be normal, hypo, or hyper and corresponds to the three classes of the problem.
- 11) The Vowel dataset is composed of 528 samples that are described by ten features. The goal is to recognize one of the eleven vowels of British English.
- 12) The original Breast Cancer Wisconsin dataset contains 699 samples, which are described by ten attributes. Each sample has one of two possible classes: benign or malignant.

As can be seen, the datasets that are considered present a large variety of problems, ranging from linearly separable (e.g., wine) to hard classification problems (e.g., yeast).

### C. Results

In this section, the learning algorithm of fuzzy similarity measure is compared with standard fuzzy similarity measures that are issued from the literature. Here, the aim is to analyze the improvement (if any) of the learning scheme compared with usual similarity measures. In all experiments, and for all similarity measures, the initial value of  $\lambda$  is set to 10. The influence of this initialization is discussed in the sequel. First, a detailed study on the IRIS dataset is provided. The number of top  $k$  samples varies from 1 to  $k_{\max} = 10$ , and the individual AP measures for each similarity measure are obtained. The corresponding performances are reported in Table IV, where the best score for each  $k$  is reported in bold font. The parametric measures are shown to be the best measures over all top  $k$  samples. The Hamacher measure performs particularly well, ranking first in seven out of ten levels. The second experiment demonstrates the efficiency of the approach on the 12 aforementioned datasets. Here again, the AP measure over the top  $k$  samples from 1 to 10 is considered, and the mAP is obtained by taking the means across the ten AP measures. The results for each dataset and each similarity measure are given in Table V, where best scores for each dataset are reported in bold font. The last column is the average rank of each similarity measure over all datasets. For purposes of comparison, results that are obtained with a recent online similarity learning algorithm Online Algorithm for Scalable Image Similarity (OASIS); see [5]) are also given in the last row of the table. In the experiments, the OASIS algorithm uses the same feature vectors as the other similarity measures. Here, the aim is to compare fuzzy similarity measures so that OASIS is not taken into account for the rank computation. According to Table V, the following remarks can be made.

- 1) Whatever the datasets,  $\mathcal{S}_M$  leads to the worst rank. This measure is a point-wise measure in the sense that it uses a single degree of membership to determine their value.
- 2) Although  $\mathcal{S}_{mL}$  is also a point-wise measure, the average performance of this measure is better than  $\mathcal{S}_M$  because the measure uses a combination of two degrees of membership instead of one for  $\mathcal{S}_M$ .
- 3) Whatever the datasets, there are at least two parametric similarity measures that perform better than commonly used similarity measures.
- 4) The average rank of each similarity measures gives an overview of their performance compared with the others. The parametric similarity measures can be ranked as follows:  $\mathcal{S}_{SS} \succ \mathcal{S}_H \succ \mathcal{S}_Y \succ \mathcal{S}_D \succ \mathcal{S}_F$ . It is not surprising to observe that the Schweizer-Sklar-based measure performs best. This is the only t-norm (among the considered ones) that is equal to the four basic t-norms, depending on the  $\lambda$  value. Consequently, the similarity measure that is derived from this t-norm is more flexible than the others.
- 5) When compared with OASIS, parametric equivalences behave favorably. In particular, all similarity measures except Dombi and Frank give better results than OASIS.



TABLE IV  
AP MEASURE (%) ON IRIS DATASET FOR THE TOP  $k$  SIMILAR SAMPLES ( $k$  RANGES FROM 1 TO  $k_{\max} = 10$ )

$k =$	1	2	3	4	5	6	7	8	9	10
$\mathcal{S}_P$	92.66	92.66	92	92.33	92.26	92.44	91.71	91.41	90.96	90.53
$\mathcal{S}_L$	94.66	94.66	93.77	92.83	92.53	92.55	92.47	92.41	92.44	92.20
$\mathcal{S}_M$	62.66	62	61.77	62	62.13	60.33	60.19	59.50	58.51	58.20
$\mathcal{S}_{mL}$	94.66	94.66	94.22	94.16	93.46	92.55	91.61	91.33	90.81	90.60
$\mathcal{S}_H(\lambda^* = 58.53)$	93.33	94.33	93.77	<b>94.50</b>	<b>94.53</b>	<b>94.66</b>	<b>94.47</b>	<b>94.33</b>	<b>94.22</b>	<b>93.66</b>
$\mathcal{S}_Y(\lambda^* = 1.41)$	<b>95.33</b>	94.66	94	93.33	92.80	93	92.85	92.83	92.96	92.93
$\mathcal{S}_D(\lambda^* = 0.10)$	<b>95.33</b>	<b>95</b>	93.77	94	93.46	93.44	93.33	93.08	92.96	92.80
$\mathcal{S}_F(\lambda^* = 10^7)$	94.66	<b>95</b>	94.22	93.66	93.60	93.88	93.61	93.33	93.25	93.26
$\mathcal{S}_{SS}(\lambda^* = 0.5)$	94.66	94	<b>94.44</b>	93.66	93.06	93.33	93.42	93.33	93.33	93.26

TABLE V  
MAP (%) FOR ALL DATASETS  $\alpha = 1$

$\mathcal{S}$	Iris	Wine	Pima	Yeast	Heart	Ionosphere	Tae	Ecoli	Liver	Newthyroid	Vowel	Breast	Avg. rank
$\mathcal{S}_P$	91.90	93.42	66.45	43.13	75.34	89	48.59	67.80	55.72	91.24	78.27	95.33	7.33
$\mathcal{S}_L$	93.05	95.65	66.39	46.26	75.56	88.97	49.31	74.64	56.05	94.95	82.98	95.87	5.25
$\mathcal{S}_M$	60.73	49.37	52.89	25.21	49.93	70.77	34.26	47.27	49.80	78.82	11.60	80.41	9
$\mathcal{S}_{mL}$	92.81	92.96	67	45.33	69.06	71.48	48.19	76.45	54.42	91.86	81.96	94.23	7.08
$\mathcal{S}_H$	<b>94.18</b>	94.55	68.39	47.01	75.39	89.64	<b>49.82</b>	76.51	57.73	95.11	<b>85.11</b>	95.97	2.75
$\mathcal{S}_Y$	93.47	95.83	67.32	46.89	<b>75.64</b>	89.66	49.44	74.77	58.37	95.32	83.05	<b>96.02</b>	3.08
$\mathcal{S}_D$	93.72	93.92	<b>68.65</b>	46.79	75.41	<b>92.61</b>	49.54	70.57	57.66	91.97	83.17	95.43	4
$\mathcal{S}_F$	93.85	95.35	68.12	46.77	75.47	89.70	49.43	75.45	55.96	93.65	84.11	95.37	4.08
$\mathcal{S}_{SS}$	93.65	<b>95.76</b>	67.13	<b>47.12</b>	75.52	89.95	49.46	<b>76.97</b>	<b>59.03</b>	<b>95.44</b>	83.28	95.94	<b>2.41</b>
OASIS	93.25	94.11	65.41	42.21	73.21	89.58	45.05	73.25	58.05	95.11	82.98	94.37	–

In order to compare multiple similarity measures over multiple datasets, a combination of a Friedman test and a Nemenyi post-hoc test is used, following the recommendations of [26]. Let  $R_j^i$  be the rank of the  $j$ th similarity measure on the  $i$ th dataset. The Friedman test compares the average ranks  $R_j$  over all datasets (see the last column of Table V). Under the null hypothesis, stating that two similarity measures are equivalent, their ranks should be equal (here,  $R_j = 5$  for all  $j$ ). The Friedman statistic is given by

$$\chi_F^2 = \frac{12N}{ns(ns+1)} \left( \sum_j R_j^2 - \frac{ns(ns+1)^2}{4} \right) \quad (15)$$

where  $N$  is the number of datasets, and  $ns$  is the number of similarity measures, which are big enough, i.e., typically  $N > 10$  and  $ns > 5$ . A derived and better statistic proposed in [27] is given by

$$F_F = \frac{(N-1)\chi_F^2}{N(ns-1) - \chi_F^2}. \quad (16)$$

The Friedman test  $\chi_F^2 = 68.95$  proves that the average ranks are significantly different from the mean rank  $R_j = 5$  expected under the null hypothesis. Moreover,  $F_F = 28.04$  is distributed according to the  $F$  distribution with  $9 - 1 = 8$  and  $9(-1) \times (12 - 1) = 88$  degrees of freedom. The  $p$ -value that is computed by using the  $F(8,88)$  distribution is almost zero so that the null hypothesis is rejected at a high level of confidence.

If the null hypothesis is rejected, the Nemenyi post-hoc test is proceeded. The performance of two similarity measures is

significantly different if the corresponding average ranks differ by at least the critical difference, which is defined by

$$CD = q_\alpha \sqrt{\frac{nc(nc+1)}{6N}} \quad (17)$$

where  $q_\alpha$  values are based on the Studentized range statistic divided by  $\sqrt{2}$  (see [26] for details). Finally, a (#similarity measures  $\times$  #similarity measures) matrix that summarizes the results is obtained. Each entry of the matrix is 1 if the difference of ranks is significant and 0 otherwise. In order to provide a more informative visualization, a new matrix is created, where each entry ( $\{i\}, \{j\}$ ) shows the difference of individual ranks  $R(\{i\}) - R(\{j\})$  that are obtained with similarity measures  $\{i\}$  and  $\{j\}$  or black if the difference is not statistically significant under the Nemenyi test (see Fig. 6).

According to Fig. 6, the following remarks are drawn. Two main sets of measures, i.e.,  $\{1, 2, 3, 4\}$  and  $\{5, 6, 7, 8, 9\}$ , can be identified. They correspond to the commonly used similarity measures and the new parametric similarity measures, respectively. Within the first set, one must distinguish the measure  $\{2\}$ , i.e.,  $\mathcal{S}_L$ , which is not different from the second set at significance level  $\alpha = 0.05$ . However, at  $\alpha = 0.10$ , it becomes significantly worse than the measure  $\mathcal{S}_{SS}$ . In the second set, two measures can be distinguished. The first is  $\{8\}$ , i.e.,  $\mathcal{S}_F$ . In terms of statistical significance, this is the worst parametric measure. Although its average rank is better than those of  $\{2, 4\}$ , this is not significant at level 0.05. One may argue that at significance level  $\alpha = 0.05$ , the post-hoc test is not powerful enough to detect any significant differences between the measures. The second interesting measure is  $\{9\}$ , i.e.,  $\mathcal{S}_{SS}$ , which is significantly better than the

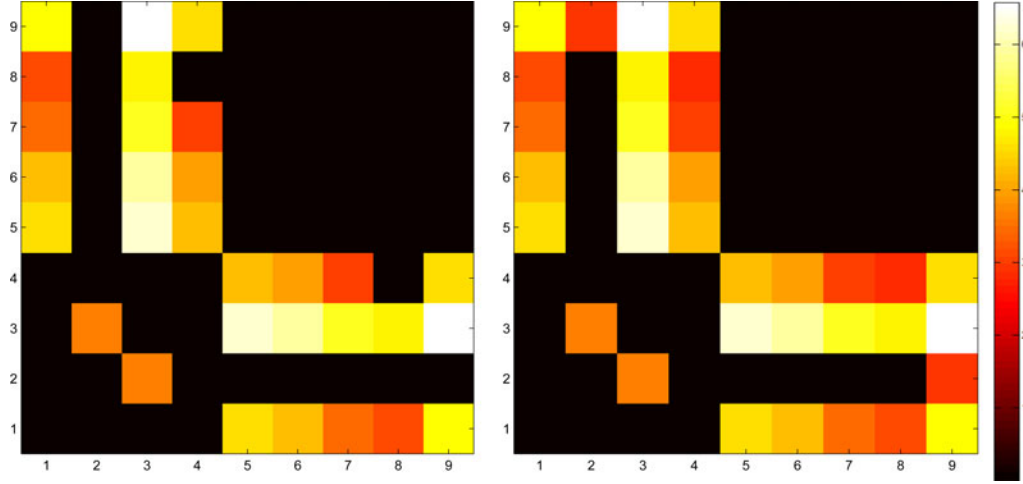


Fig. 6. Comparison of each similarity measures for two confidence levels  $\alpha = 0.05$  (left) and  $\alpha = 0.10$  (right). In each plot, entries  $(a, b)$  are the paired rank differences of similarity measures  $\{i\}$  and  $\{j\}$ . Black entries correspond to not statistically significant differences under the Nemenyi post-hoc test.

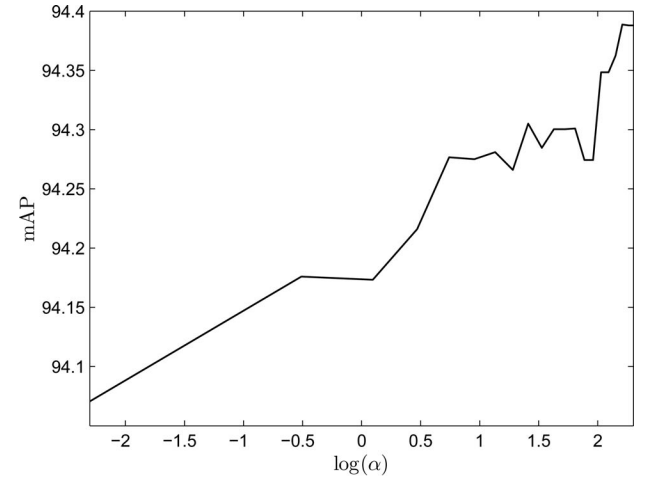
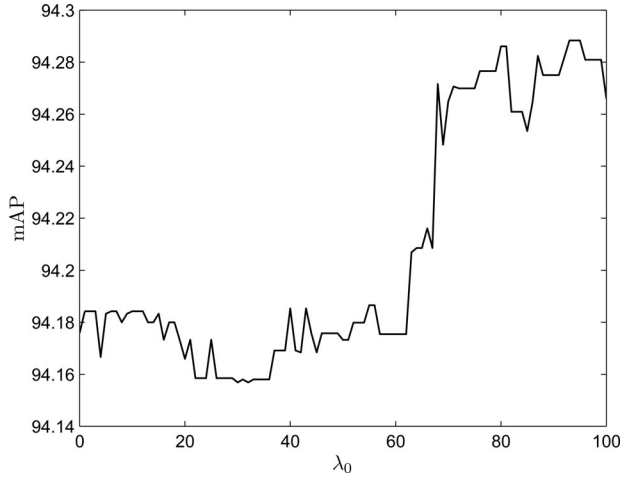


Fig. 7. mAP for the Hamacher based similarity measure  $\mathcal{S}_H$  as a function of the initial value  $\lambda_0$ .

entire first set. Considering  $\alpha = 0.10$ , one may conclude the following:

- 1)  $\mathcal{S}_{SS}$  is significantly better than all commonly used similarity measures.
- 2) None of the parametric measures is significantly better than the others.
- 3)  $\mathcal{S}_P$ ,  $\mathcal{S}_M$ , and  $\mathcal{S}_{mL}$  are significantly worst than the parametric measures.
- 4) We cannot conclude on the statistical significance of  $\mathcal{S}_L$ , except for  $\mathcal{S}_{SS}$  (worse) and  $\mathcal{S}_M$  (better).

The next experiment is the analysis of the initial value of  $\lambda$  with respect to the performance that is obtained with the learning algorithm, where  $\alpha = 1$ . In order to make the reading clear, only the results for the Hamacher-based similarity measure are reported, although similar comments are valid for the other parametric measures. The mAP as a function of  $\lambda_0$  is plotted in Fig. 7. The mAP values range from 94.16 % to 94.29 %, yielding a maximum difference of 0.13 % so that it can be concluded that the initial value does not have a large influence

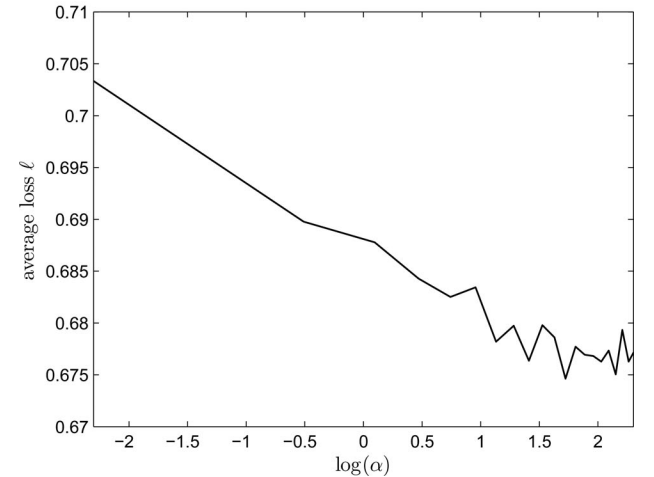


Fig. 8. (Top) mAP and (bottom) average loss for the Hamacher-based similarity measure  $\mathcal{S}_H$  as a function of  $\log(\alpha)$ .

on the performance. Note that the best mAP 94.29 % is greater than the value that is reported in Table V. Finally, the effect of the input parameter  $\alpha$  on the performance of the similarity measures is investigated. To this aim, the learning algorithm 1 is run with

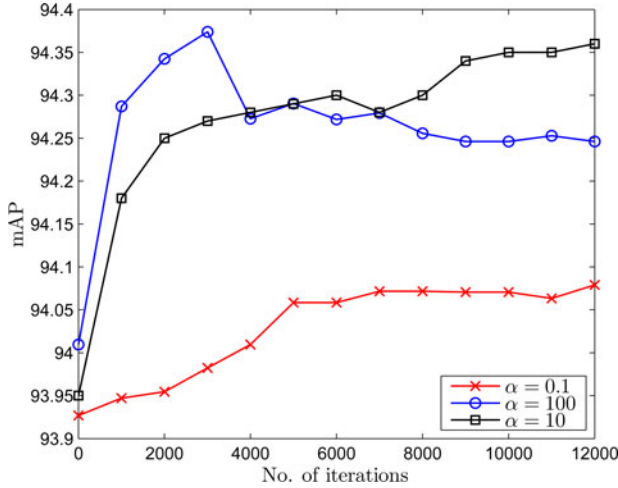


Fig. 9. mAP for the Hamacher-based similarity measure  $S_H$  as a function of the number of iterations for different values of  $\alpha$ .

different values of  $\alpha$  on the IRIS dataset. The mAPs and average losses of algorithms as a function of  $\alpha$  varying from 0.1 to 10 are depicted in Fig. 8. As can be seen from the graphs, the value  $\alpha$  have a larger effect on the performance (up to 0.4 %) of the learning algorithm than the initial value  $\lambda_0$ . As expected, while the loss is decreasing, the corresponding mAP is increasing. An interesting point is that local peaks of the loss also correspond to local decreasing of the mAP, e.g.,  $\log(\alpha) = 1.5$ , showing the ability of the loss structure to efficiently describe the problem constraints. Here again, one can note that a better performance than the one reported in Table V can be obtained, i.e., greater than 94.18 %. Since  $\alpha$  is a weight that is applied on the loss, one can expect that for very small values of  $\alpha$ , the loss will be high. On the other hand, when increasing  $\alpha$ , the optimal  $\lambda$  value can be reached quickly, but noisy data influence the result.

So far, the number of iterations within the learning algorithm was fixed. The effect of  $\alpha$  as a function of the number of iterations is now discussed. The learning algorithm is run for  $\alpha = 0.1$ ,  $\alpha = 10$ , and  $\alpha = 100$ . At the end of each iteration, the mAP is computed using the actual  $\lambda$  value. Results are given in Fig. 9. Naturally, a small  $\alpha$  value (i.e.,  $\alpha = 0.1$ ) leads to a slow progress rate, since  $\lambda_i$  is not very different of  $\lambda_{i-1}$ . In contrast, when  $\alpha$  is large (i.e.,  $\alpha = 100$ ), the precision increases faster but at the price of a worst performance than a medium value (i.e.,  $\alpha = 10$ ) later on. Additionally, the more the  $\alpha$ , the less smooth the curve. A large value of  $\alpha$  heavily modifies  $\lambda_i$ , resulting in a large difference in terms of performance. In contrast, with a small  $\alpha$  value,  $\lambda_{i-1}$  and  $\lambda_i$  slightly differ, which results in a small variation in terms of performance.

## V. CONCLUSION

The contribution of this paper lies in the development of generalized fuzzy similarity measures. First, a generic framework of designing similarity measures that are based on the use of residual implication functions is proposed. This construction presents two main advantages: 1) Classical fuzzy similarity measures are retrieved for particular residual functions; 2) verifying

if a newly constructed similarity measure satisfies the required properties is facilitated. Then, some new families of parametric similarity measures by using parametric residual implications are presented. An algorithm that allows us to learn the parameter of each similarity measure based on relevance degrees is given. Experiments on a number of real datasets show the superiority, which is statistically significant, of the learning algorithm over commonly used similarity functions. The proposed similarity measures can be used in many classification methods, e.g., induction of pattern trees [28], hierarchical clustering [29], content-based image retrieval [30], or ranking image similarities [5].

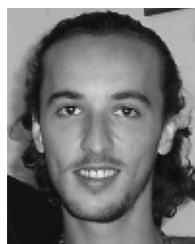
Among the potential perspectives, a more sophisticated updating scheme, using the passive-aggressive family of algorithms [31], can be developed. In each step, the weight applied on the loss is varying so that the new parameter value better reflects the data. Another interesting perspective is to analyze the fuzzy equivalences between each similarity measures as proposed in [32] in order to select a particular function within the large variety of measures.

Returning to aggregation operators, it would be interesting to consider *ideal* samples that characterize the classes and then adopt a metric that can be learned with the help of particular aggregation operators [33].

## REFERENCES

- [1] A. Tversky and I. Gati, "Similarity, separability, and the triangle inequality," *Psychol. Rev.*, vol. 89, no. 2, pp. 123–154, 1982.
- [2] M. De Cock and E. Kerre, "On (un)suitable fuzzy relations to model approximate equality," *Fuzzy Sets Syst.*, vol. 133, no. 2, pp. 137–153, 2003.
- [3] F. Klawonn, "Should fuzzy equality and similarity satisfy transitivity?" *Fuzzy Sets Syst.*, vol. 133, no. 2, pp. 175–180, 2003.
- [4] E. Hullermeier, "Fuzzy methods in machine learning and data mining: Status and prospects," *Fuzzy Sets Syst.*, vol. 156, no. 3, pp. 387–407, 2005.
- [5] G. Chechik, V. Sharma, U. Shalit, and S. Bengio, "Large scale online learning of image similarity through ranking," *J. Mach. Learn. Res.*, vol. 11, pp. 1109–1135, 2010.
- [6] T. Calvo, A. Kolesarova, M. Komornikova, and R. Mesiar, "Aggregation operators: Properties, classes and construction methods," in *Aggregation Operators: New Trends and Applications* (ser. Studies in Fuzziness and Soft Computing), R. M. T. Calvo and G. Mayor, Eds. Heidelberg, Germany: Physica-Verlag, 2002, vol. 97, pp. 1–104.
- [7] M. Grabisch, J. Marichal, R. Mesiar, and E. Pap, *Aggregation Functions* (ser. Encyclopedia of Mathematics and its Applications). Cambridge, U.K.: Cambridge Univ. Press, 2009, no. 127.
- [8] M. Mas, M. Monserrat, J. Torrens, and E. Trillas, "A survey on fuzzy implication functions," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 6, pp. 1107–1121, Dec. 2007.
- [9] M. Baczyński and B. Jayaram, *Fuzzy Implications* (ser. Studies in Fuzziness and Soft Computing). Berlin, Germany: Springer-Verlag, 2008, vol. 231.
- [10] A. Tversky, "Features of similarity," *Psychol. Rev.*, vol. 84, no. 4, pp. 327–352, 1977.
- [11] B. Bouchon-Meunier, M. Rifqi, and S. Bothorel, "Towards general measures of comparison of objects," *Fuzzy Sets Syst.*, vol. 84, no. 2, pp. 143–153, 1996.
- [12] K. Hirota and W. Pedrycz, "Matching fuzzy quantities," *IEEE Trans. Syst., Man, Cybern.*, vol. 21, no. 6, pp. 1580–1586, Nov./Dec. 1991.
- [13] W. Bandler and L. Kohout, "Fuzzy power sets and fuzzy implication operators," *Fuzzy Sets Syst.*, vol. 4, pp. 13–30, 1980.
- [14] I. Bloch, "On fuzzy distances and their use in image processing," *Pattern Recognit.*, vol. 32, no. 11, pp. 1873–1895, 1999.
- [15] H.-J. Zimmermann and P. Zysno, "Latent connectives in human decision making," *Fuzzy Sets Syst.*, vol. 4, pp. 37–51, 1980.

- [16] R. N. Shepard, "Toward a universal law of generalization for psychological science," *Science*, vol. 237, no. 4820, pp. 1317–1323, 1987.
- [17] V. V. Cross and T. A. Sudkamp, *Similarity and Compatibility in Fuzzy Set Theory: Assessment and Applications*. Heidelberg, Germany: Physica-Verlag GmbH, 2002.
- [18] H. Le Capitaine and C. Frélicot, "Towards a unified logical framework of fuzzy implications to compare fuzzy sets," presented at the 13th Int. Fuzzy Syst. Assoc. World Congr., Lisboa, Portugal, 2009.
- [19] B. De Baets and R. Mesiar, "Metrics and t-equalities," *J. Math. Anal. Appl.*, vol. 267, pp. 531–547, 2002.
- [20] W.-J. Wang, "New similarity measures on fuzzy sets and on elements," *Fuzzy Sets Syst.*, vol. 85, pp. 305–309, 1997.
- [21] S. Santini and R. Jain, "Similarity measures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 9, pp. 871–883, Sep. 1999.
- [22] R. N. Shepard, "Attention and the metric structure of the stimulus space," *J. Math. Psychol.*, vol. 1, pp. 54–87, 1964.
- [23] Y. Chen, E. K. Garcia, M. R. Gupta, A. Rahimi, and L. Cazzanti, "Similarity-based classification: Concepts and algorithms," *J. Mach. Learning Res.*, vol. 10, pp. 747–776, 2009.
- [24] E. Anderson, "The irises of the gaspe peninsula," *Bull. Amer. IRIS Soc.*, vol. 59, pp. 2–5, 1935.
- [25] J. C. Bezdek, J. M. Keller, R. Krishnapuram, L. I. Kuncheva, and N. R. Pal, "Will the iris data please stand up?" *IEEE Trans. Fuzzy Syst.*, vol. 7, no. 3, pp. 368–369, Jun. 1999.
- [26] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.
- [27] R. L. Iman and J. M. Davenport, "Approximations of the critical region of the friedman statistic," *Commun. Statist.*, vol. 9, no. 6, pp. 571–595, 1980.
- [28] Z. Huang, T. D. Gedeon, and M. Nikraves, "Pattern trees induction: A new machine learning method," *IEEE Trans. Fuzzy Syst.*, vol. 16, no. 4, pp. 958–970, Aug. 2008.
- [29] A. Mirzaei and M. Rahmati, "A novel hierarchical-clustering-combination scheme based on fuzzy-similarity relations," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 1, pp. 27–39, Feb. 2010.
- [30] Y. Chan and J. Z. Wang, "A region-based fuzzy feature matching approach to content-based image retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 9, pp. 1252–1267, Sep. 2002.
- [31] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *J. Mach. Learn. Res.*, vol. 7, pp. 551–585, 2006.
- [32] M. J. Lesot and M. Rifqi, "Order-based equivalence degrees for similarity and distance measures," in *Lecture Notes in Computer Science*. Berlin, Germany: Springer, 2010, vol. 6178, pp. 19–28.
- [33] G. Beliakov, "Definition of general aggregation operators through similarity relations," *Fuzzy Sets Syst.*, vol. 114, no. 3, pp. 437–453, 2000.



**Hoel Le Capitaine** received the M.S. degree in applied mathematics and computer sciences in 2006 and the Ph.D. degree in signal processing and artificial intelligence, in 2009, both from the University of La Rochelle, La Rochelle, France.

After two years of postdoctoral research with the Mathematics, Image, and Applications Laboratory, he is currently an Assistant Professor with the LINA, University of Nantes, Nantes, France. His interests include aggregation operators for pattern recognition and machine learning, as well as image processing.