A relevance-based learning model of fuzzy similarity measures

Hoel Le Capitaine

▶ **To cite this version:**

Hoel Le Capitaine. A relevance-based learning model of fuzzy similarity measures. IEEE Transactions on Fuzzy Systems, 2012, 20 (1), pp.57-68. hal-00627673

HAL Id: hal-00627673
https://hal.science/hal-00627673

Submitted on 4 Oct 2012

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A relevance-based learning model of fuzzy similarity measures

Hoel Le Capitaine
LINA (UMR CNRS 6241),
École Polytechnique de Nantes, France.
and
MIA (EA 3165),
Université de La Rochelle, France
Email: hoel.lecapitaine@univ-nantes.fr

***Abstract*—Matching pairs of objects is a fundamental operation in data analysis. However, it requires to define a similarity measure between objects 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 T-equalities derived from residual implication functions is proposed. Then a model allowing 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.**

I. Introduction

Whilst the similarity is an essential concept in human reasoning, and plays a fundamental role in theories of knowledge, there is not an unique and general-purposed definition of similarity. The reason for this lack of 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], [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 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 derived from fuzzy residual implications. It does not suffer from the drawbacks of the conventional metric approaches, and allows to obtain concave or convex iso-similarity contours.

However, the number of similarity measures induced by the proposed framework is infinite, due to 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 then are ranked according to their pairwise similarity.

Additionally, an online learning setting is adopted, where, in contrast to batch methods, samples are considered one at a time. This enable to treat large data sets 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, conclusion 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 $\mathcal{A} : [0,1]^n \to [0,1]$ satisfying boundary conditions and monotonicity. In the literature, one finds many aggregation operators, *e.g.*: t-norms , OWA (*Ordered Weighted Averaging*) operators, γ-operators, or fuzzy integrals. They belong to several categories, depending on the way the values are aggregated: conjunctives, disjunctives, averaging, and mixed operators. The interested reader can refer to [6], [7] for large, yet comprehensive, surveys on aggregation operators.

A t-norm is an increasing, associative and commutative mapping $\top : [0,1]^2 \to [0,1]$ 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 $\top_P(x,y) = x\,y$ and the Łukasiewicz t-norm $\top_L(x,y) = \max(x+y-1,0)$. Various parametric families involving a real value λ lying in a specified domain have been introduced. The parametric t-norms that are used in the sequel are given in Table I.

TABLE I
PARAMETRIC T-NORMS

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

A general problem in fuzzy logic is to handle conditional statements *if x, then y* where x and y are fuzzy predicates. A widely used method consists in managing them by using functions $I : [0,1]\times[0,1] \to [0,1]$ 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 non-increasing in the first variable, non-decreasing 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

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

In the sequel, the following notation is adopted:

- $X = \{x_1, \cdots, x_n\}$ is the (supposed finite) universe of discourse,
- $\mathcal{C}(X)$ and $\mathcal{F}(X)$ are the sets of all crisps and fuzzy sets in X, respectively,
- $f_A(x)$, $\forall x \in X$, 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 which is specified for membership functions of fuzzy sets. The second category involves set-theoretic operations for fuzzy sets (fuzzy intersection, union, cardinality) [10], [11]. Finally, a third way of defining a similarity measure consists in using logical concepts of fuzzy implication, as first suggested in [12], following the seminal paper of [13]. Note that a fourth approach, 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 $\mathcal{S} : \mathcal{F}(X)\times\mathcal{F}(X) \to [0,1]$ is called a similarity measure if it satisfies*

(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) $\forall A,B,C \in \mathcal{F}(X)$, *if* $A \subseteq B \subseteq C$, *then* $\mathcal{S}(A,C) \leq \mathcal{S}(A,B) \wedge \mathcal{S}(B,C)$
or, equivalently
$\forall A,B,C,D \in \mathcal{F}(X)$, *if* $A \subseteq B \subseteq C \subseteq D$, *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 emphasize 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 focus 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. The property (P2) is also discussed, since in some cases, the similarity of a point to itself is not 1. The 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 defined for $r \geq 1$ is usually taken:

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

This metric induces well known distance functions; Hamming (or Manhattan) for $r = 1$, Euclidean for $r = 2$, 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 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 non increasing function g such that $\mathcal{S}(A,B) = g(d(A,B))$. 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 in [15]. Another function, coming from the work of Shepard [16], consists in taking $g(x) = exp(-\alpha x)$. The use of the exponential law comes from the underlying observation that there is a non linear relationship between the distance and the similarity which is concave upward. However, obtaining 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 non-metric 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, defined by the supremum over X of $A \cap B$. The interested reader can refer to [17] for more details on set-theoretic comparison measures.

The last main approach consists in considering the implication degrees of the elements belonging to A over the elements of B [13], and vice-versa. The implication degree is obtained by using one of the fuzzy implication functions described in the previous section. Formally, having x an element of A, and y 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, defined by

$$bI(x,y) = \min(I(x,y), I(y,x)) \tag{3}$$

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 bi-residual implication function defining a $\top$-equality $E_\top$. For arbitrary $A, B \in \mathcal{F}(X)$, let*

$$\mathcal{S}(A,B) = \mathop{\mathcal{A}}_{i=1}^{n} E_\top \left(f_A(x_i), f_B(x_i) \right) \tag{4}$$

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,

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

(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 $x_i \in X$,

$$f_D(x_i) \geq f_C(x_i) \tag{5}$$

$$f_B(x_i) \geq f_A(x_i) \tag{6}$$

hold. By non-increasingness in the first variable and non-decreasingness 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 Eq. (5) and $I_\top(f_C(x_i), f_A(x_i)) \leq I_\top(f_C(x_i), f_B(x_i))$ by Eq. (6). Using $E_\top(x,y) = I(\max(x,y), \min(x,y))$ and (5-6), 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 also fulfill the desired properties, because for any $I_\top$, $x \leq y \Rightarrow I_\top(x,y) = 1$ (by ordering property, see [8]).*

B. Examples

Taking particular t-norms and aggregation operators enables to retrieve well-known similarity measures. For instance, taking the arithmetic mean and $\top_P$, $\top_L$, the two measures

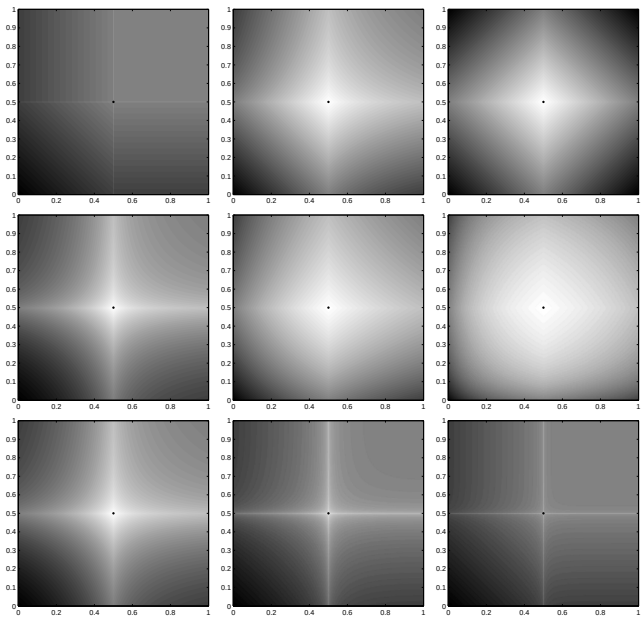

Fig. 1. Examples of iso-similarity contours between a reference set $A = [\frac{1}{2}]$ (denoted by $\cdot$) and all fuzzy sets $\mathcal{F}(X)$, $n = 2$ where high and low values correspond to white and black colors respectively. The arithmetic mean is used. First row: $I_{\top_M}$, $I_{\top_P}$ and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

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 taking 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, the minimum and maximum operators are taken as overall aggregation 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 dependencies, as in [21].*

For illustration purpose, Figure 1 shows some examples of fuzzy similarity measures as well as the influence of the parameter λ 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 two-dimensional fuzzy sets B for various $I_\top$. As one could expect, the closer to x_1 or x_2, the higher the similarity. One can also note that different $I_\top$ lead to different shapes of iso-similarity contours. According to the third line of Figure 1, one can remark that the Hamacher similarity measure gives a concave contour with a low λ, 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 modelized. 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 to obtain convex iso-similarity contours, yielding a usual metric satisfying the triangle inequality. This property is very important, since a concave contour of iso-similarities means that the triangle inequality is violated [22]. However, when dealing with fuzzy sets, the $\top$-transitivity is preserved, as discussed earlier.

TABLE II
ILLUSTRATION OF RANKING CHANGES FOR PARAMETRIC SIMILARITY MEASURES.

Fuzzy sets	λ	Hamacher similarity
A, B	2.0	**0.505**
A, C	2.0	0.475
A, B	15.0	0.650
A, C	15.0	**0.762**
Fuzzy sets	λ	**Dombi similarity**
A, B	2.0	**0.271**
A, C	2.0	0.206
A, B	0.25	0.686
A, C	0.25	**0.818**

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

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

They can be visually inspected in Figure 2. Even for a human, the ranking of $\mathcal{S}(A, B)$ and $\mathcal{S}(A, C)$ is not an easy task. In Table II, the amount of similarity between A, B and C with various λ 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 λ_1 such that $\mathcal{S}(A, B) > \mathcal{S}(A, C)$ and a value λ_2 such that $\mathcal{S}(A, B) < \mathcal{S}(A, C)$. Consequently, the three objects A, B and C are ranked differently depending on λ. If one cannot find two values of λ 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. Afterwards, 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

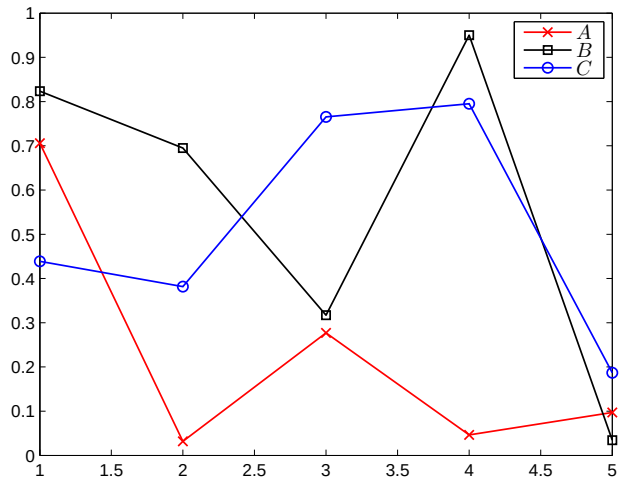

Fig. 2. Three fuzzy sets A, B and C where the ranking of similarity measures $\mathcal{S}(A,B)$ and $\mathcal{S}(A,C)$ is not obvious.

require to use all the 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 strong 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

$$\mathcal{S}_\lambda(a,b) \to \mathcal{S}_\lambda(a,c) < 1 \tag{7}$$

The $\to$ 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:

$$\ell_\lambda = \max\left\{0, \mathcal{S}_\lambda(a,b) \to \mathcal{S}_\lambda(a,c) - 1\right\} \tag{8}$$

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

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

The aim is to minimize the total loss L_λ on the learning set. The total loss is defined by summing up the individual losses over all the triplet of the learning set:

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

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 online 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, c are also uniformly 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 λ, its optimal value is searched by using an online learning algorithm based on sequential updates of λ. Depending on the similarity measure, the initial value λ_0 varies, see Section IV for comments, and subsection II for a discussion on the various similarity measures that are used in the sequel. It leads to find λ such that:

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

where $\alpha \geq 0$. In other terms, during the optimization process, λ_i is selected in order to obtain a trade-off between minimizing the loss on (a,b,c) and staying quite close to its previous value λ_{i-1}. This trade-off is controlled by the 'aggressiveness' parameter α. Naturally, if α is set to zero, then the optimal λ_i value is equal to λ_{i-1}. In contrast, setting α 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

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

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

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

Let G_i be the gradient value $\frac{\partial\left(\mathcal{S}_\lambda(a,b)-\mathcal{S}_\lambda(a,c)\right)}{\partial\lambda}$. Therefore, the optimal new value λ 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 value λ_0, which is discussed in Section IV-C. When λ is out of definition bounds of the t-norm (*e.g.* negative values for Hamacher one), the last convenient value λ_{i-1} is returned. An important case is the one of bound values. When λ reaches these bounds, then the equivalence obtained with the corresponding t-norm is used. For instance, if $\lambda = 0$ for the

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

$\mathcal{S}$	$\frac{\partial}{\partial\lambda}(\mathcal{S}_\lambda(a_k,b_k))$
$\mathcal{S}_H$	$\frac{(f_2-f_1f_2)(f_1-f_2)}{\left(f_2(\lambda_i+f_1-\lambda_i f_1)+f_1-f_2\right)^2}$
$\mathcal{S}_Y$	$-\left((1-f_2)^\lambda-(1-f_1)^\lambda\right)^{1/\lambda}\frac{1}{\lambda^2}\left(\lambda\frac{\log(1-f_2)(1-f_2)^\lambda-\log(1-f_1)(1-f_1)^\lambda}{(1-f_2)^\lambda-(1-f_1)^\lambda}-\log((1-f_2)^\lambda-(1-f_1)^\lambda)\right)$
$\mathcal{S}_D$	$\left(\left(\frac{1-f_2}{f_2}\right)^\lambda-\left(\frac{1-f_1}{f_1}\right)^\lambda\right)^{1/\lambda}\left(-\frac{1}{\lambda^2}\log\left(\left(\frac{1-f_2}{f_2}\right)^\lambda-\left(\frac{1-f_1}{f_1}\right)\right)+\frac{1}{\lambda}\frac{\log\left(\frac{1-f_2}{f_2}\right)\left(\frac{1-f_2}{f_2}\right)^\lambda-\log\left(\frac{1-f_1}{f_1}\right)\left(\frac{1-f_1}{f_1}\right)^\lambda}{\log\left(\left(\frac{1-f_2}{f_2}\right)^\lambda-\left(\frac{1-f_1}{f_1}\right)\right)}\right)$
$\mathcal{S}_F$	$\frac{1}{\log(\lambda)^2}\left(\frac{\log(\lambda)}{1+\frac{(\lambda^{f_2}-1)(\lambda-1)}{(\lambda^{f_1}-1)}}\frac{(\lambda^{f_1}-1)(\lambda^{f_2}+f_2\lambda^{f_2}-\lambda^{f_2}(f_2/\lambda)-1)-(\lambda^{f_2}-1)(\lambda-1)\lambda^{f_1}f_1}{(\lambda^{f_1}-1)^2}-\frac{1+\frac{(\lambda^{f_2}-1)(\lambda-1)}{(\lambda^{f_1}-1)}}{\lambda}\right)$
$\mathcal{S}_{SS}$	$\left(1+f_2^\lambda-f_1^\lambda\right)^{1/\lambda}\left(-\frac{1}{\lambda^2}\log(1+f_2^\lambda-f_1^\lambda)+\frac{1}{\lambda}\frac{\log(f_2)f_2^\lambda-\log(f_1)f_1^\lambda}{1+f_2^\lambda-f_1^\lambda}\right)$

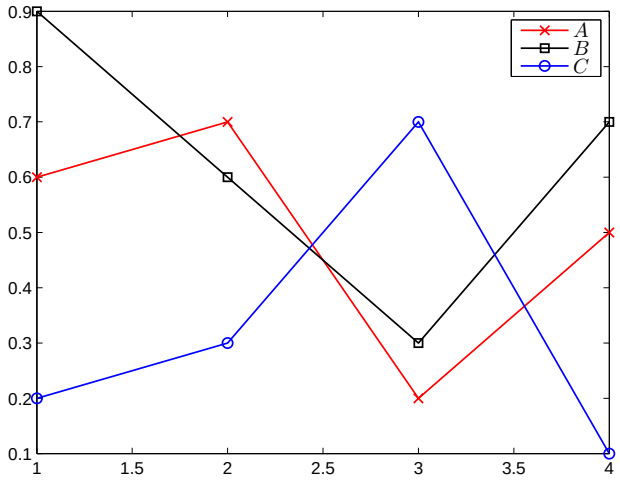

Fig. 3. Three fuzzy sets A, B and C where the similarity of A and B is clearly higher than the similarity of A and C.

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 illustration purpose, three fuzzy sets are used for which the ranking of the similarities $\mathcal{S}(A,B)$ and $\mathcal{S}(A,C)$ is natural. Such fuzzy sets are plotted in Figure 3. Here, there are only three samples, so 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 belonging to the learning data (and hopefully on the test set). The difference $\mathcal{S}(A,B)-\mathcal{S}(A,C)$ is studied as a function of the iterations. The corresponding graph is plotted in Figure 4 (dashed line). Naturally, since B is more similar to A than C, even at the beginning of the algorithm, the difference is quite large. The most interesting point is the evolution of the parameter λ that makes A and B more and more similar, and A and C more and more dissimilar as λ is updated. This difference does not exceed an upper bound reached around the 130-th update of λ.

Algorithm 1 Online Similarity Learning algorithm

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

Now, the three fuzzy sets introduced in the previous section are considered. As already mentioned, the ranking of $\mathcal{S}(A,B)$ and $\mathcal{S}(A,C)$ 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 Figure 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 $\mathcal{S}(A,C)-\mathcal{S}(A,B)$ difference curve is plotted in Figure 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 200-th iteration). In contrast, supposing that C is more similar to A than B requires much more iterations for the learning algorithm to reach its upper bound. Note that at the beginning, the difference is negative, that is to say $\mathcal{S}(A,B) > \mathcal{S}(A,C)$, and as λ 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.

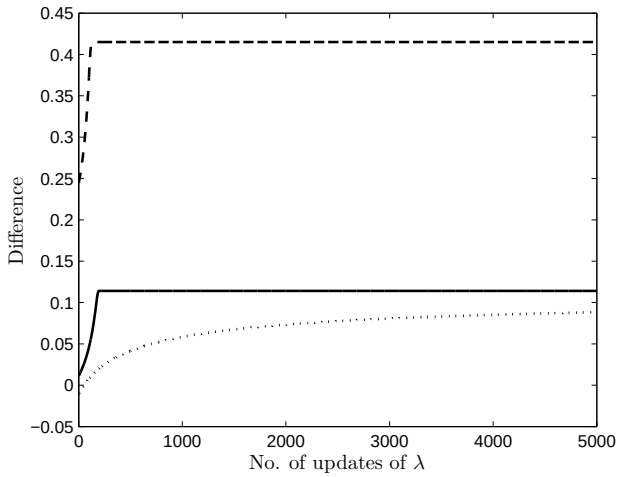

Fig. 4. Difference $\mathcal{S}(A,B)-\mathcal{S}(A,C)$ as a function of the number of update of λ, (see text for details).

A. Protocol

Let $X = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_n, y_n)\}$ 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 $y_i \in \{1, \cdots, c\}$ is known. In order to use the similarity measures 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

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

where m is the mean and σ 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 illustration purpose, we give an example for a dataset composed of a mixture of three normal 1-dimensional 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 Figure 5.*

During the learning phase, 3 samples are randomly selected in each iteration step . The first one, 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

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

procedure FUZZIFICATION(X)
  Set $p \leftarrow$ number of features
  Set $c \leftarrow$ number of classes
  for $i = 1$ to c **do**
    for $j = 1$ to p **do**
      $M(i,j) \leftarrow$ componentwise average(X_j^i)
      $S(i,j) \leftarrow$ componentwise standard dev.(X_j^i)
    end for
  end for
  for each sample $\mathbf{x}_k \in X$ **do**
    for $i = 1$ to c **do**
      for $j = 1$ to p **do**
        Set membership degrees of $\mathbf{x}_k$ using (14),
        with $M(i,j)$ and $S(i,j)$.
      end for
    end for
  end for
end procedure

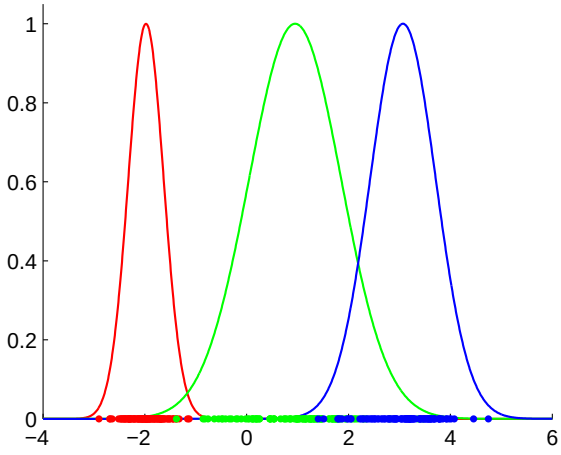

Fig. 5. Plot of the 1-dimensional dataset and the membership functions obtained by the fuzzification scheme. Samples of different classes are represented by different colored dots.

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:

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

where $\mathbf{x}_j$, $j = 1, \cdots, k$ 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:

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

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

B. Datasets

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

- the well-known IRIS dataset [24], [25]. The data contains 3 classes of 50 samples each, where each class is a type or iris plant. Each sample is described by a 4 features: sepal length and width, petal length and width. It is a classical dataset in pattern recognition literature, which is known to have one class linearly separable from the other two, and two classes which are not linearly separable.
- the Wine dataset contains 3 classes and 178 samples. Each sample is described by 13 constituents found in each wine. The classes are quite well separated, and samples' classification is not reputed to be a challenging task.
- the Pima dataset contains 2 classes of 768 samples, each described by 8 features. The class, 0 or 1, denotes the absence or presence of diabetes pathology.
- the Yeast dataset. This is a genetic dataset of phylogenetic profiles for the Yeast genome. It contains 1484 samples, described by 8 attributes. The ten classes of the problem correspond to the various localization sites of the proteins.
- the Heart dataset contains 270 samples of 13 attributes each, describing medical data of each patient. The class, 0 or 1, denotes the absence or presence of heart disease.
- the Ionosphere dataset is composed of 351 instances. For each sample, there are 17 pulse numbers, each pulse number is described by 2 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.
- The Tae dataset consists of evaluations of 151 teaching assistant. Each assistant is described by 5 attributes. The teacher may belong to three classes: low, medium or high.
- The Ecoli dataset contains 336 observations described by 7 attributes. Each sample is a sequence of protein, and the 8 classes correspond to their respective localization sites.
- the Liver Disorders dataset contains 345 samples. Each sample constitutes the record of a single male individual, and is described by 7 features such as blood tests, or number of drinks per day. The class, 0 or 1, denotes some sort of liver disorders.
- the Newthyroid dataset is composed of 215 instances. The instances are described by 5 chemical features designed to 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.
- the Vowel dataset is composed of 528 samples described by 10 features. The goal is to recognize one of the eleven vowels of British English.
- the original Breast Cancer Wisconsin dataset contains 699 samples, described by 10 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 to standard fuzzy similarity measures issued from the literature. Here, the aim is to analyze the improvement (if any) of the learning scheme compared to usual similarity measures. In all experiments, and for all similarity measures, the initial value of λ 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 average precision 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 7 out of 10 levels. The second experiment demonstrates the efficiency of the approach on the twelve aforementioned datasets. Here again, the average precision measure over the top k samples from 1 to 10 is considered, and the mean average precision is obtained by taking the means across the 10 average precision 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 comparison purpose, results obtained with a recent online similarity learning algorithm (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.

- 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.
- Although $\mathcal{S}_{mL}$ is also a point-wise measure, the average performance of this measure is better than $\mathcal{S}_M$. The reason is that the measure uses a combination of two degrees of membership, instead of one for $\mathcal{S}_M$.
- Whatever the datasets, there are at least two parametric similarity measures that perform better than commonly used similarity measures.
- The average rank of each similarity measures gives an overview of their performance compared to 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 the best results. This is the only t-norm (among the considered ones) that is equal to the four basic t-norms, depending on the λ value. Consequently, the similarity measure derived from this t-norm is more flexible than the others.

TABLE IV
AVERAGE PRECISION MEASURE (%) ON IRIS DATASET, FOR THE TOP k SIMILAR SAMPLES (k RANGES FROM 1 TO $k_{\text{MAX}} = 10$).

$k=$	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
$\mathcal{S}_H(\lambda^\star = 58.53)$	93.33	94.33	93.77	**94.50**	**94.53**	**94.66**	**94.47**	**94.33**	**94.22**	**93.66**
$\mathcal{S}_Y(\lambda^\star = 1.41)$	**95.33**	94.66	94	93.33	92.80	93	92.85	92.83	92.96	92.93
$\mathcal{S}_D(\lambda^\star = 0.10)$	**95.33**	**95**	93.77	94	93.46	93.44	93.33	93.08	92.96	92.80
$\mathcal{S}_F(\lambda^\star = 10^7)$	94.66	**95**	94.22	93.66	93.60	93.88	93.61	93.33	93.25	93.26
$\mathcal{S}_{SS}(\lambda^\star = 0.5)$	94.66	94	**94.44**	93.66	93.06	93.33	93.42	93.33	93.33	93.26

TABLE V
MEAN AVERAGE PRECISION (%) 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
$\mathcal{S}_H$	**94.18**	94.55	68.39	47.01	75.39	89.64	**49.82**	76.51	57.73	95.11	**85.11**	95.97	2.75
$\mathcal{S}_Y$	93.47	95.83	67.32	46.89	**75.64**	89.66	49.44	74.77	58.37	95.32	83.05	**96.02**	3.08
$\mathcal{S}_D$	93.72	93.92	**68.65**	46.79	75.41	**92.61**	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	**95.76**	67.13	**47.12**	75.52	89.95	49.46	**76.97**	**59.03**	**95.44**	83.28	95.94	**2.41**
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	–

- When compared to OASIS, parametric equivalences behaves favorably. In particular, all similarity measures except Dombi and Frank give better results than OASIS.

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 (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

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

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

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

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 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, defined by

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

where q_α 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 $R(\{i\}) - R(\{j\})$ obtained with similarity measures $\{i\}$ and $\{j\}$, or black if the difference is not statistically significant under the Nemenyi test, (see Figure 6).

According to Figure 6, the following remarks are drawn. Two main sets of measures : $\{1, 2, 3, 4\}$ 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

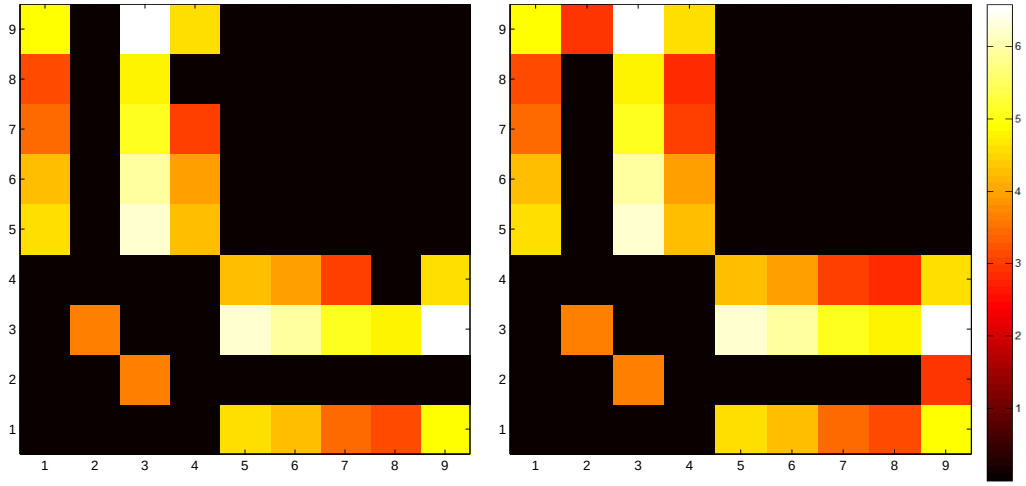

Fig. 6. Comparison of each similarity measures for 2 confidence levels $\alpha = 0.05$ (left) and $\alpha = 0.10$ (right). In each plot, entries (a, b) are the paired rank differences of similarity measures $\{i\}$ and $\{j\}$, (see text). Black entries correspond to not statistically significant differences under the Nemenyi post-hoc test.

between the measures. The second interesting measure is $\{9\}$, *i.e.* $\mathcal{S}_{SS}$, which is significantly better than the entire first set. Considering $\alpha = 0.10$, one may conclude:

- $\mathcal{S}_{SS}$ is significantly better than all commonly used similarity measures,
- none of the parametric measures is significantly better than the others,
- $\mathcal{S}_P$, $\mathcal{S}_M$ and $\mathcal{S}_{mL}$ are significantly worst than the parametric measures,
- 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 λ with respect to the performance 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 mean average precision as a function of λ_0 is plotted in Figure 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 on the performance. Note that the best mAP 94.29% is greater than the value reported in Table V. Finally, the effect of the input parameter α on the performance of the similarity measures is investigated. To this aim, the learning algorithm 1 is run with different values of α on the IRIS dataset. The mean average precisions and average losses of algorithms as a function of α varying from 0.1 to 10 are depicted in Figure 8. As can be seen from the graphs, the value α have a larger effect on the performance (up to 0.4%) of the learning algorithm than the initial value λ_0. As expected, while the loss is decreasing, the corresponding mean average precision is increasing. An interesting point is that local peaks of the loss also correspond to local decreasing of the mean average precision, *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 α is a weight applied on the loss, one can expect that

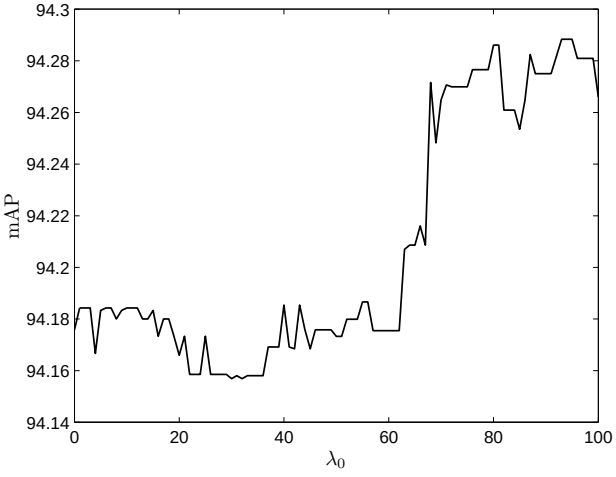

Fig. 7. Mean average precision for the Hamacher based similarity measure $\mathcal{S}_H$ as a function of the initial value λ_0.

for very small values of α, the loss will be high. On the other hand, when increasing α, the optimal λ value can be reached quickly, but noisy data influences the result.

So far, the number of iterations within the learning algorithm was fixed. The effect of α 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 mean average precision is computed using the actual λ value. Results are given in Figure 9. Naturally, a small α value ($\alpha = 0.1$) leads to a slow progress rate, since λ_i is not very different of λ_{i-1}. In contrast, when α is large ($\alpha = 100$), the precision increases faster, but at the price of a worst performance than a medium value ($\alpha = 10$) later on. Additionally, the more α, the less smooth the curve. A large value of α heavily modifies λ_i, resulting in a large difference in terms of performance. In contrast, with a small α value, λ_{i-1} and λ_i slightly differ, which results in a small variation in terms of performance.

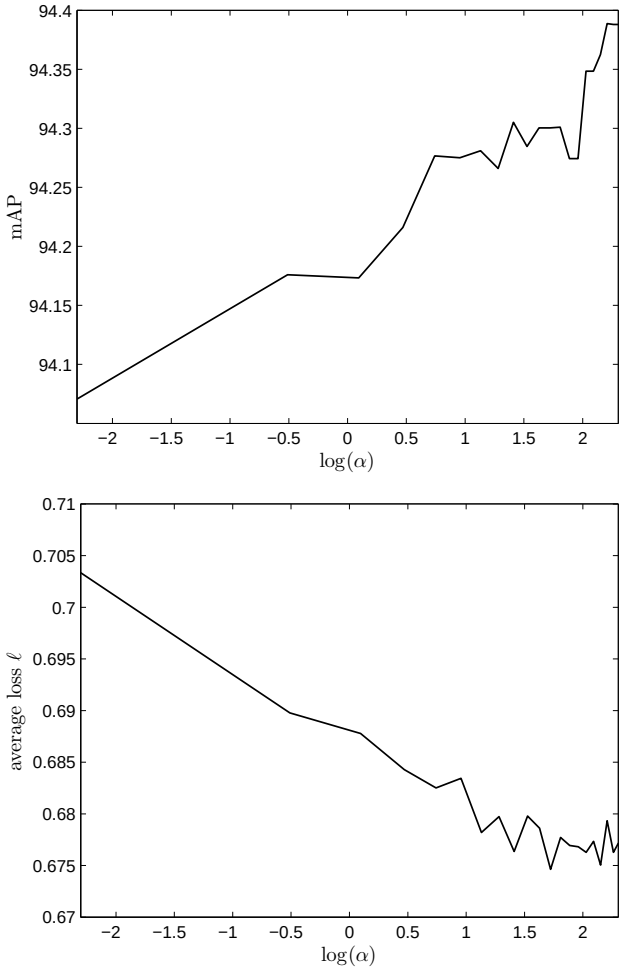

Fig. 8. Mean average precision (top) and average loss (bottom) for the Hamacher based similarity measure $\mathcal{S}_H$ as a function of $\log(\alpha)$.

V. Conclusion

The contribution of this paper lies in the development of generalized fuzzy similarity measures. First, a generic framework of designing similarity measures 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 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 developped. 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.

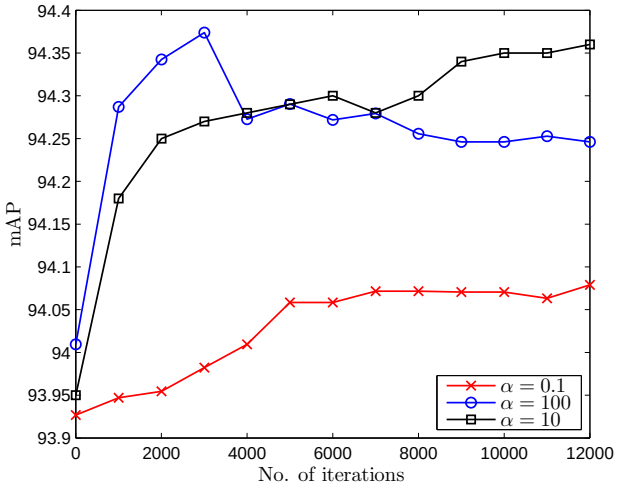

Fig. 9. Mean average precision for the Hamacher based similarity measure $\mathcal{S}_H$ as a function of the number of iterations, for different values of α.

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

References

[1] A. Tversky and I. Gati, "Similarity, separability, and the triangle inequality," *Psychological review*, 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 and Systems*, vol. 133, no. 2, pp. 137–153, 2003.
[3] F. Klawonn, "Should fuzzy equality and similarity satisfy transitivity?" *Fuzzy Sets and Systems*, vol. 133, no. 2, pp. 175–180, 2003.
[4] E. Hullermeier, "Fuzzy methods in machine learning and data mining: status and prospects," *Fuzzy Sets and Systems*, 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," *Journal of Machine Learning Research*, 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, G. Mayor, Ed. 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 University Press, 2009, no. 127.
[8] M. Mas, M. Monserrat, J. Torrens, and E. Trillas, "A survey on fuzzy implication functions," *IEEE Transactions on Fuzzy Systems*, vol. 15, no. 6, pp. 1107–1121, 2007.
[9] M. Baczynski and B. Jayaram, *Fuzzy Implications*, ser. Studies in Fuzziness and Soft Computing. Springer, Berlin, 2008, vol. 231.
[10] A. Tversky, "Features of similarity," *Psychological review*, 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 and Systems*, vol. 84, no. 2, pp. 143–153, 1996.
[12] K. Hirota and W. Pedrycz, "Matching fuzzy quantities," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 6, pp. 1580–1586, 1991.
[13] W. Bandler and L. Kohout, "Fuzzy power sets and fuzzy implication operators," *Fuzzy Sets and Systems*, vol. 4, pp. 13–30, 1980.

[14] I. Bloch, "On fuzzy distances and their use in image processing," *Pattern Recognition*, vol. 32, no. 11, pp. 1873–1895, 1999.
[15] H.-J. Zimmermann and P. Zysno, "Latent connectives in human decision making," *Fuzzy Sets and Systems*, 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*. Physica-Verlag GmbH, 2002.
[18] H. Le Capitaine and C. Frélicot, "Towards a unified logical framework of fuzzy implications to compare fuzzy sets," in *13th International Fuzzy Systems Association World Congress, IFSA*, Lisboa, Portugal, 2009.
[19] B. De Baets and R. Mesiar, "Metrics and t-equalities," *Journal of Mathematical Analysis and Applications*, vol. 267, pp. 531–547, 2002.
[20] W.-J. Wang, "New similarity measures on fuzzy sets and on elements," *Fuzzy Sets and Systems*, vol. 85, pp. 305–309, 1997.
[21] S. Santini and R. Jain, "Similarity measures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, no. 9, pp. 871–883, 1999.
[22] R. N. Shepard, "Attention and the metric structure of the stimulus space," *Journal of Mathematical Psychology*, 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," *Journal of Machine Learning Research*, vol. 10, pp. 747–776, 2009.
[24] E. Anderson, "The irises of the gaspe peninsula," *Bull. Am. 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 Transactions on Fuzzy Systems*, vol. 7, no. 3, pp. 368–369, 1999.
[26] J. Demsar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
[27] R. L. Iman and J. M. Davenport, "Approximations of the critical region of the friedman statistic," *Communications in Statistics*, vol. 9, no. 6, pp. 571–595, 1980.
[28] Z. Huang, T. D. Gedeon, and M. Nikravesh, "Pattern trees induction: A new machine learning method," *IEEE Transactions on Fuzzy Systems*, vol. 16, no. 4, pp. 958–970, 2008.
[29] A. Mirzaei and M. Rahmati, "A novel hierarchical-clustering-combination scheme based on fuzzy-similarity relations," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 1, pp. 27–39, 2010.
[30] Y. Chan and J. Z. Wang, "A region-based fuzzy feature matching approach to content-based image retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1252–1267, 2002.
[31] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive-aggressive algorithms," *Journal of Machine Learning Research*, vol. 7, pp. 551–585, 2006.
[32] M. J. Lesot and M. Rifqi, "Order-based equivalence degrees for similarity and distance measures," in *LNCS 6178*, 2010, pp. 19–28.
[33] G. Beliakov, "Definition of general aggregation operators through similarity relations," *Fuzzy Sets and Systems*, vol. 114, no. 3, pp. 437–453, 2000.

Hoel Le Capitaine received the MS degree in Applied Mathematics and Computer Sciences from the University of La Rochelle, France, in 2006. In 2009, he obtained his Ph.D. degree in Signal Processing and Artificial Intelligence from the University of La Rochelle. After two years of postdoctoral research wit the Mathematics, Image and Applications Laboratory, he is now an assistant professor in the University of Nantes, with the LINA. His interests include aggregation operators for pattern recognition and machine learning, as well as image processing.