Classification with reject options in a logical framework: a fuzzy residual implication approach

Hoel Le Capitaine Carl Frélicot

MIA Laboratory, University of La Rochelle
La Rochelle, France
Emails: {hoel.le_capitaine, carl.frelicot}@univ-lr.fr

***Abstract*— *In many classification problems, overlapping classes and outliers can significantly decrease a classifier performance. In this paper, we introduce the possibility of a given classifier to reject patterns either for ambiguity or for distance. From a set of typicality degrees for a pattern to be classified, we propose to use fuzzy implications to quantify the similarity of the degrees. A class-selective scheme based on this new family is presented, and experimental results showing the efficiency of the proposed algorithm are given.*

***Keywords*—** Ambiguity measures, Dombi and Hamacher implications, fuzzy r-implications, pattern classification, reject options

1 Introduction

Usual classification of objects, or patterns, consists in assigning incoming samples to one class belonging to a set of known ones. Since classes may overlap in the feature space and all of them do not appear during the learning step of most real classification problems, some samples may have to be not classified but rejected. In practice, two situations can lead to rejection, either the incoming sample could belong to unknown class(es) (*distance rejection*) or it should be assigned to several classes (*ambiguity rejection*) [1]. In this paper, we focus our attention on this latter situation. Initially introduced by Chow [2], it consists in rejecting the most unreliable patterns that have a low posterior probability, or more generally a low membership degree value, to its closest class. Such a reject classification rule is optimal if posterior probabilities are estimated without error, unfortunately this is not the case in practice.

We investigate some new measures that aim at quantifying to what extent such a pattern is ambiguous from its membership degrees, or more generally from its multi-class classifier soft outputs, using fuzzy implications. These measures assess the relationships between membership values with various combination of triangular norm operators (t-norms for short, see [3] for an overview).

The remainder of this paper is organized as follows. In section 2, we start by presenting distance-based class models that we will use in the sequel, and then discuss about standard approaches to pattern rejection. Next, we give a brief recall on some basic definitions of triangular norms and fuzzy residual implications, as well as some new parametrical implications (section 3). Section 4 is devoted to the proposal of using implications as ambiguity measures. In section 5, an experimental evaluation and comparisons to well known class-selective pattern recognition schemes are provided. We finally conclude and give some perspectives in section 6.

2 Classification with reject options

2.1 Classification

Let $\mathbf{x}$ be a pattern described by a p (real) features and let $\Omega = \{\omega_1, \cdots, \omega_c\}$ be a set of classes of cardinality c. The objective of classification is to assign any pattern $\mathbf{x} \in \mathbb{R}^p$ to one of the c classes of Ω. It generally consists of two steps L (*labeling*) and H (*hardening*):

- $L : \mathbf{x} \mapsto \mathbf{u}(\mathbf{x}) = {}^t(u_1(\mathbf{x}), \ldots, u_c(\mathbf{x})) \in \mathcal{L}_{\bullet c}$
- $H : \mathbf{u}(\mathbf{x}) \mapsto \mathbf{h}(\mathbf{x}) = {}^t(h_1(\mathbf{x}), \ldots, h_c(\mathbf{x})) \in \mathcal{L}_{hc}$.

where $\mathcal{L}_{hc} = \{\mathbf{h}(\mathbf{x}) \in \mathcal{L}_{fc} : h_i(\mathbf{x}) \in \{0, 1\}\}$ and $\mathcal{L}_{\bullet c}$ depends on the mathematical framework the classifier relies on, e.g. $\mathcal{L}_{pc} = [0, 1]^c$ for degrees of typicality or $\mathcal{L}_{fc} = \{\mathbf{u}(\mathbf{x}) \in \mathcal{L}_{pc} : \sum_{i=1}^{c} u_i(\mathbf{x}) = 1\}$ for posterior probabilities and membership degrees. Posterior probabilities can be obtained either from (known) class-conditional densities whose parameters are estimated using a learning set X of patterns, i.e. patterns for which the assignment is known, or from class-density estimates using the classes of their neighbors in X. However, for high dimensional feature spaces, many patterns are required to perform good class-density estimates (curse of dimensionality). When the learning set is small, distances to prototypes of each class are used, e.g.

$$d^2(\mathbf{x}, \mathbf{v}_i) = (\mathbf{x} - \mathbf{v}_i)^t \Sigma_i^{-1} (\mathbf{x} - \mathbf{v}_i) \tag{1}$$

where $\mathbf{v}_i$ and Σ_i are the mean vector and covariance matrix of the class ω_i estimated from X. In order to have a common scaling, label values are often obtained in the unit interval, e.g.

$$u_i(\mathbf{x}) = \frac{\alpha_i}{\alpha_i + d^2(\mathbf{x}, \mathbf{v}_i)} \tag{2}$$

where α_i are user-defined parameters.

We focus on the $H-$step because it is in charge of the classification. This step often reduces to the class of maximum label value selection, resulting in an exclusive classification rule which is not efficient in practice because it supposes that Ω is exhaustively defined (closed-world assumption) and classes do not overlap (separability assumption). Such untrue assumptions can lead to very undesired decisions. In many real applications, it is more convenient to with-hold making a decision than making a wrong assignment, e.g. in medical diagnosis where a false negative outcome can be much more costly than a false positive. Reject options have been proposed to overcome these difficulties and to reduce misclassification risk. The first one, called *distance rejection* [1], is dedicated

to outlying patterns. If $\mathbf{x}$ is far from all the class prototypes, this option allows to assign it to no class. The second one, called *ambiguity rejection*, allows to assign inlying patterns to several or all classes [2, 4]. If $\mathbf{x}$ is close to two or more class prototypes, it is associated with the corresponding classes. Including reject options leads to partition the feature space into as many regions as subsets of classes, i.e. at most 2^c ones, to which a pattern can be assigned. Formally, it consists in modifying H such that $\mathbf{h}(\mathbf{x})$ can take values in the set of vertices of the unit hypercube $\mathcal{L}_{hc}^c = \{0,1\}^c$ instead of the exclusive subset $\mathcal{L}_{hc} \subset \mathcal{L}_{hc}^c$. Different strategies can be adopted to handle these options at hand, but they all lead to a three types decision system: distance rejection when $\mathbf{h}(\mathbf{x}) = {}^t(0,\ldots,0) = \underline{0}$, exclusive classification when $\mathbf{h}(\mathbf{x}) \in \mathcal{L}_{hc}$, ambiguity rejection when $\mathbf{h}(\mathbf{x}) \in \mathcal{L}_{hc}^c \backslash \{\mathcal{L}_{hc} \cup \underline{0}\}$. The resulting classification rule is then a matter of selecting the appropriate number of classes varying from zero (distance rejection) to c (total ambiguity rejection).

2.2 Usual reject schemes

Since the work by Chow [2], many rejection schemes have been proposed. In its general form, a class-selective procedure is defined by

$$n^\star(\mathbf{x}, t) = \min_{k \in [1,c]} \{k : \mathcal{A}_k(\mathbf{x}) \leq t\} \tag{3}$$

where $\mathcal{A}_k$ is a given ambiguity measure on membership degrees, $n^\star(\mathbf{x}, t)$ is the number of selected classes for the pattern $\mathbf{x}$ to be classified, and t is a user-defined threshold which can be class dependent (e.g. t_k). This threshold can be set conditionally to cost functions relative to error, reject and correct classification rates. Propositions from the literature mainly consist in new definitions of ambiguity measures $\mathcal{A}_k$. For Chow, $\mathcal{A}_k$ was one minus the maximum value of the membership degrees. In the paper by Ha [4], the second highest value was tested to decide wether one or several classes are selected. Since this scheme is leading to unnatural decisions, Horiuchi [5] proposed a new measure defined by the difference of the membership degrees which is actually a disambiguity measure. In [6], Frélicot proposed to use the ratio of membership degrees. As usual, we use the convention that if $\mathcal{A}_k(x) > t$ for all k, then we set $n^\star(\mathbf{x}, t) = c$. Table 1 summarizes the existing ambiguity measures used for pattern rejection, where the membership degrees are assumed to be sorted in decreasing order for writing convenience, i.e. $u_1(\mathbf{x}) \geq \cdots \geq u_c(\mathbf{x})$.

Table 1: Existing ambiguity measures related to Eq. (3).

Scheme	**Ambiguity Measure** $\mathcal{A}_k(\mathbf{x})$
Chow [2]	$1 - u_1(\mathbf{x})$
Ha [4]	$u_{k+1}(\mathbf{x})$
Horiuchi [5]	$1 - (u_k(\mathbf{x}) - u_{k+1}(\mathbf{x}))$
Frélicot [6]	$u_{k+1}(\mathbf{x})/u_k(\mathbf{x})$

3 Fuzzy residual implications

3.1 Basic definitions

Let us recall basic definitions of fuzzy operators that will be used to combine the values of interest, i.e. the pattern class-degrees of typicality. Depending on properties, aggregation functions can be classified into several categories: conjunctive, disjunctive, compensatory, and so on. We restrict on conjunctive and disjunctive functions. By definition, the output of a conjunctive operator is lower or equal than the minimum value, whereas the output of a disjunctive operator is greater or equal than the maximum value. Beyond these operators, we choose to use the triangular norms because of their ability to generalize the logical AND and OR crisp operators to fuzzy sets, see [3] for a survey. Briefly, a triangular norm (or t-norm) is a binary operation on the unit interval $\top : [0,1]^2 \to [0,1]$ which is commutative, associative, non decreasing and has 1 for neutral element. Thus, a t-norm $\top$ is conjunctive and the minimum operator $\wedge$ is the greatest t-norm. Alternatively, a triangular conorm (or t-conorm) is the dual binary operation $\perp : [0,1]^2 \to [0,1]$ having the same properties except that its neutral element is 0. Thus, a t-conorm $\perp$ is disjunctive and the maximum operator $\vee$ is the lowest t-conorm. Typical examples of dual couples (t-norm, t-conorm) that will be used in the sequel are given in Table 2.

Table 2: Examples of dual couples, including parametrical ones.

Standard	$a \top_S b = \min(a,b)$ $a \perp_S b = \max(a,b)$
Algebraic	$a \top_A b = a\,b$ $a \perp_A b = a + b - a\,b$
Łukasiewicz	$a \top_L b = \max(a + b - 1, 0)$ $a \perp_L b = \min(a + b, 1)$
Hamacher	$a \top_{H_\gamma} b = \frac{a\,b}{\gamma + (1-\gamma)\,(a+b-a\,b)}$ $a \perp_{H_\gamma} b = \frac{a+b+(\gamma-2)\,a\,b}{1+(\gamma-1)\,a\,b}$
Dombi	$a \top_{D_\gamma} b = \left(1 + \left(\left(\frac{1-a}{a}\right)^\gamma + \left(\frac{1-b}{b}\right)^\gamma\right)^{1/\gamma}\right)^{-1}$ $a \perp_{D_\gamma} b = 1 - \left(1 + \left(\left(\frac{a}{1-a}\right)^\gamma + \left(\frac{b}{1-b}\right)^\gamma\right)^{1/\gamma}\right)^{-1}$

A fuzzy residual implication, denoted R-implication (or $\rightarrow$) is defined by:

$$I(a,b) = \sup_t \{t \in [0,1] : \top(a,t) \leq b\} \tag{4}$$

Note that if $\top$ is a left-continuous t-norm, the supremum operation can be substituted by maximum. If we use additive generating functions, i.e. a strictly decreasing function $f : [0,1] \to [0,+\infty]$ with $f(1) = 0$, and admitting an inverse (or pseudo-inverse) function f^{-1}, Eq. (4) can be written as

$$I(a,b) = \max(f^{-1}(f(b) - f(a)), 0) \tag{5}$$

because f is strictly monotonic. We generally speak about an implication function if I is non-increasing in the first variable, non-decreasing in the second variable and $I(0,0) = I(1,1) = 1$, and $I(1,0) = 0$, see [7] for a large survey on fuzzy implication functions. Within theses implications, the well-known *Gödel* and *Goguen* ones are respectively given by

$$I(a,b) = \begin{cases} 1 & \text{if } b \geq a \\ b & \text{if } b < a \end{cases} \tag{6}$$

and

$$I(a,b) = \begin{cases} 1 & \text{if } b \geq a \\ \frac{b}{a} & \text{if } b \leq a \end{cases} \tag{7}$$

which are obtained with the minimum and algebraic (or product) triangular norms, respectively. In the sequel, we will assume for writing convenience that the fuzzy values are sorted in decreasing order, e.g. $a \geq b$.

3.2 Parametrical implications

Proposition 1. *Let $(\top_{H_\gamma})$, $\gamma \in [0, +\infty[$, be the family of Hamacher t-norms. The residual Hamacher implication is given by*

$$I_{H_\gamma}(a,b) = \begin{cases} 1 & \text{if } b \geq a \\ \frac{b\,(\gamma + a - \gamma a)}{b\,(\gamma + a - \gamma a) + a - b} & \text{if } b \leq a \end{cases} \tag{8}$$

Proof. By definition of R-implications (4), we can write $I_{H_\gamma}(a,b) = \sup_t\{t \in [0,1] : \top_{H_\gamma}(a,t) \leq b\}$. We also have $I_{H_\gamma}(a,b) = \max_t\{t \in [0,1] : \top_{H_\gamma}(a,t) \leq b\}$ because $\top_{H_\gamma}$ is a left-continuous t-norm. Then, solving

$$\frac{a\,t}{\gamma + (1-\gamma)\,(a + t - a\,t)} \leq b \tag{9}$$

gives

$$t \leq \frac{b\,(\gamma + a - \gamma a)}{b\,(\gamma + a - \gamma a) + a - b}. \tag{10}$$

Since $a \geq b$, it is easy to show that the right part of Eq. (10) is in $[0,1]$, hence we obtain Eq.(8). □

Proposition 2. *Let $(\top_{D_\gamma})$, $\gamma \in]0, +\infty[$, be the family of Dombi t-norms. The residual Dombi implication is given by*

$$I_{D_\gamma}(a,b) = \begin{cases} 1 & \text{if } b \geq a \\ \left(1 + \left(\left(\frac{1-b}{b}\right)^\gamma - \left(\frac{1-a}{a}\right)^\gamma\right)^{1/\gamma}\right)^{-1} & \text{if } b < a \end{cases} \tag{11}$$

Proof. $I_{D_\gamma}(a,b) = \sup_t\{t \in [0,1] : \top_{H_\gamma}(a,t) \leq b\}$ by (4). Since $\top_{D_\gamma}$ is a left-continuous t-norm, we can write $I_{D_\gamma}(a,b) = \max_t\{t \in [0,1] : \top_{D_\gamma}(a,t) \leq b\}$. Then, solving

$$\left(1 + \left(\left(\frac{1-a}{a}\right)^\gamma + \left(\frac{1-t}{t}\right)^\gamma\right)^{1/\gamma}\right)^{-1} \leq b \tag{12}$$

gives

$$t \leq \left(1 + \left(\left(\frac{1-b}{b}\right)^\gamma - \left(\frac{1-a}{a}\right)^\gamma\right)^{1/\gamma}\right)^{-1}. \tag{13}$$

Since, $a \geq b$, it is easy to show that $\left(\left(\frac{1-b}{b}\right)^\gamma - \left(\frac{1-a}{a}\right)^\gamma\right)^{1/\gamma} \geq 0$, hence the right part of Eq. (13) is in $[0,1]$ and (11) is obtained. □

Note that R-implications are mostly used in fuzzy inference systems, see [8] for a large overview on the use of parametrical R-implications in fuzzy rule based systems.

4 Some parametrical measures of ambiguity

In this section, the concept of similarity measure and its relationship with ambiguity measures are described. Then we propose to use fuzzy parametrical implications as new families of ambiguity measures to be used for pattern classification with reject options. The resulting class-selection algorithm (H-step) is presented and we finally discuss the choice of the parameter for parametrical implications with the help of numerical examples.

4.1 Proposition and properties

A similarity measure $\mathcal{S}$ generally satisfies the following properties:

(P1) $\mathcal{S}(a,b) = \mathcal{S}(b,a)$, *(symmetry)*

(P2) $\mathcal{S}(a,a) \geq \mathcal{S}(a,b)$, *(minimality)*

(P3) $\mathcal{S}(a,b) = 1 \Leftrightarrow a = b$, *(identity)*

However, the symmetry property (P1) is still subject to experimental investigation: if $\mathcal{S}(a,b)$ is the answer to the question *how is a similar to b?*, then, when making comparisons, subjects focus more on the feature a than on b. This corresponds to the notion of *saliency of a and b* [9]: if b is more salient than a, then a is more similar to b than vice versa, which is experimentally confirmed. Property (P2) is also open because it can be violated experimentally, see [9] for details.

There are several ways to deal with the comparison of 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. This approach takes its roots from studies on how to measure the distance between two real functions and do not refer to any specific interpretation. The general form of a Minkowski r-metric is usually taken and leads to well known distance functions (Hamming, Euclidean, Chebyshev). A second way to compare fuzzy values comes from some basic set-theoretic considerations where union, intersection and complement are defined for fuzzy sets. Cardinal and possibility based measures belong to this category. In this paper, we focus on the third way to deal with fuzzy values comparison: the logical framework. Accordingly to [10], for a certain universe of discourse $\mathcal{D}$, the degree of equality of two fuzzy elements a and b can be defined by implications as follows:

$$(a \equiv b) = \frac{1}{2}\big((a \to b) \wedge (b \to a) + (\bar{a} \to \bar{b}) \wedge (\bar{b} \to \bar{a})\big) \tag{14}$$

where $\wedge$ stands for minimum, $\to$ is an implication and $\bar{a}$ is the strict negation $\bar{a} = 1 - a$.

Since 1 is the neutral element of t-norms, applying Eq. (4) with $a \geq b$ gives

$$(a \equiv b) = \frac{1}{2}\big((a \to b) + (\bar{b} \to \bar{a})\big). \tag{15}$$

A convenient way to define an ambiguity measure is to quantify to which extent two fuzzy membership degrees are similar, so that it is closely related to the problem of matching fuzzy quantities, or fuzzy sets similarity. So we propose to use fuzzy R-implications as generalized ambiguity measures.

Given a set of c truth values assumed to be sorted in decreasing order, i.e. $u_1(\mathbf{x}) \geq \cdots \geq u_c(\mathbf{x})$, with no loss of generality, let us have two predicates ($\mathbf{x}$ *is* ω_i), with the truth value $u_i(\mathbf{x})$, and ($\mathbf{x}$ *is* ω_k), with the truth value $u_k(\mathbf{x})$. The truth value of the implication *if the pattern* $\mathbf{x}$ *is* ω_i*, then the pattern* $\mathbf{x}$ *is also* ω_j*,* $\forall j$ *varying from* $i+1$ *to* k is an ambiguity measure given by $I(u_i(\mathbf{x}), u_k(\mathbf{x}))$. By convention, we assume that the assignment of $\mathbf{x}$ to ω_i is more probable than the assignment of $\mathbf{x}$ to ω_k when using this implication because $u_i(\mathbf{x}) \geq u_k(\mathbf{x})$ and obviously we have $I(u_i(\mathbf{x}), u_{i+1}(\mathbf{x})) \geq I(u_i(\mathbf{x}), u_{k>i+1}(\mathbf{x}))$.

Proposition 3. *Given* $t \in [0,1]$*, the optimum cardinality of the generalized class-selective rejection rule is given by*

$$n^\star(\mathbf{x}, t) = \min_{k \in [1,c]} \left\{ k : I\Big(u_k(\mathbf{x}), u_{k+1}(\mathbf{x})\Big) \leq t \right\} \qquad (16)$$

with $u_1(\mathbf{x}) \geq \cdots \geq u_c(\mathbf{x})$*, and the convention* $u_{c+1}(\mathbf{x}) = 0$*.*

Since $I(a,0) = 0$ if $a \neq 0$, c classes will be selected if none were previously selected.

Property 1. *If we use the Standard triangular norm* min*, we obtain the Ha class-selective rejection scheme [4].*

Property 2. *If we use the Łukasiewicz triangular norm, we obtain the Horiuchi class-selective rejection scheme [5].*

Property 3. *If we use the Algebraic triangular norm, we obtain the Frélicot class-selective rejection scheme [6].*

Proofs are straightforward and will be given in a longer forthcoming paper. Note that modifying Eq.(16) such as

$$n^\star(\mathbf{x}, t) = \min_{k \in [0,c]} \left\{ k : I\Big(u_k(\mathbf{x}), u_{k+1}(\mathbf{x})\Big) \leq t \right\} \qquad (17)$$

with the convention $u_0(\mathbf{x}) = 1$ allows to select none of the classes, i.e. to proceed to distance rejection, since $I(1,a) = a$ whatever the triangular norm.
The resulting generalized pattern classification rule with reject options ($H-$step) is presented in Algorithm 1. It can be used to compare various schemes, depending on the choice of the triangular norm.

Algorithm 1: $H-$step classification algorithm.

Data: a vector of soft class-labels $\mathbf{u}(\mathbf{x}) \in \mathcal{L}_{pc}$ and a reject threshold t
Result: a vector of class-selective assignments $\mathbf{h}(\mathbf{x}) \in \mathcal{L}_{hc}^c$
begin
  set $\mathbf{h}(\mathbf{x})$ to $\underline{0}$
  find $n^\star(\mathbf{x}, t)$ − Eq.(16) or Eq.(17)
  foreach $j = 1 : n^\star(\mathbf{x}, t)$ *in decreasing order of* $u_j(\mathbf{x})'s$ **do**
    set $h_j(\mathbf{x}) = 1$
  end
end

4.2 Discussion and examples

One of the main difficulties when using t-norms is to choose the dual couple and if needed to set the parameter value. Let us study how the choice of γ for Hamacher and Dombi implications modifies the resulting implication strength.
• *Hamacher*: increasing the value of γ will make two fuzzy values more similar because $I_{H_\gamma}(a,b) = ab/(a-b+ab)$ if $\gamma = 0$ while $\lim_{\gamma \to +\infty} I_{H_\gamma}(a,b) = 1$ whatever $(a,b) \in [0,1]^2$. Indeed, I_{H_γ} is non-decreasing with γ since

$$\frac{\partial I_{H_\gamma}}{\partial \gamma} = \frac{(b-ab)(a-b)}{(b(\gamma + a - \gamma a) + a - b)^2} \geq 0.$$

The influence of γ is much more significant for low values of a and b than for high ones because $b(\gamma + a - \gamma a)$ appears to be of order $b\,a$ (respectively $b\,(\gamma + a)$) for high (respectively low) values of a and b. It follows that if $\gamma_1 >> \gamma_2$, we have $\frac{\gamma_1}{\gamma_1+\varepsilon} >> \frac{\gamma_2}{\gamma_2+\varepsilon}$, so that large values of γ associated to low values of (a,b) will result in a high value of I_{H_γ}, see Table 3 for examples.

Table 3: Hamacher implications examples for $a = 0.9$, $b = 0.8$, $c = 0.1$ and $d = 0.05$.

γ	0	2	10
$I_{H_\gamma}(a,b)$	0.87	0.89	0.93
$I_{H_\gamma}(a,c)$	0.10	0.12	0.19
$I_{H_\gamma}(c,d)$	0.09	0.65	0.90

• *Dombi*: decreasing the value of γ will make two fuzzy values more similar because $\lim_{\gamma \to +\infty} I_{D_\gamma}(a,b) = b$ while $\lim_{\gamma \to 0} I_{D_\gamma}(a,b) = 1$ whatever $(a,b) \in [0,1]^2$. Analogously to Hamacher family, I_{D_γ} is non-increasing with γ since

$$\frac{\partial I_{D_\gamma}}{\partial \gamma} \leq 0.$$

Thus, decreasing γ for the Dombi family has the same impact as increasing γ for the Hamacher family, so that we expect the opposite tendency, see Table 4 for examples. The appropriate tuning of γ can be achieved using a gradient procedure, as in [11].

Table 4: Dombi implications examples for $a = 0.9$, $b = 0.8$, $c = 0.1$ and $d = 0.05$.

γ	0.5	2	10
$I_{D_\gamma}(a,b)$	0.97	0.81	0.80
$I_{D_\gamma}(a,c)$	0.12	0.10	0.10
$I_{D_\gamma}(c,d)$	0.35	0.06	0.05

The simple graphical example shown in Fig.1 illustrates how the different implications behave. The top-plot shows a one-dimensional dataset composed of two classes (□ and ×) described by a distance-based model (Eqs.(1-2)) with $\alpha_i = 1$ ($\forall i = 1, c$). In the middle, implication truth values for the standard, algebraic, and Łukasiewicz triangular norms, corresponding to Ha, Frélicot and Horiuchi class-selective schemes respectively, as well as Chow's scheme are shown. Truth values when using parametrical implications are shown in the bottom-plot. Observing the data points in the top-plot allows to obtain the bounds of intervals of x values where the classes do not overlap, therefore for which no misclassification should

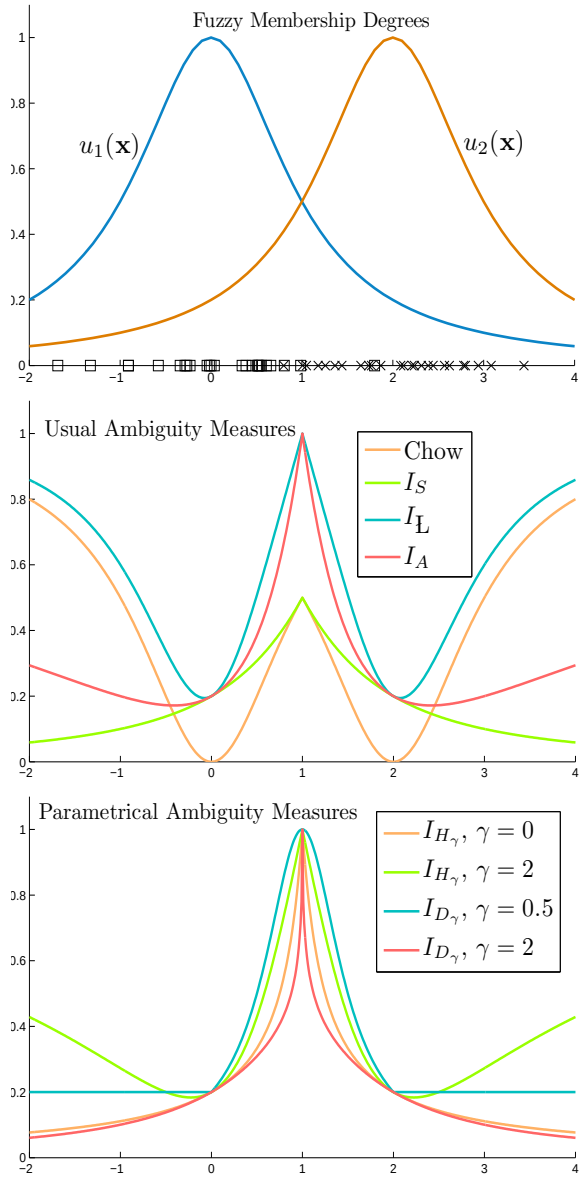

Figure 1: Original outputs using Eqs.(1-2) for two classes with normal distributions (*top*), implication truth values with usual triangular norms (*middle*) and implication truth values with Hamacher and Dombi families (*bottom*).

happen, here $\mathbf{x}_{low} \approx 0.8$ and $\mathbf{x}_{up} \approx 1.9$. Then, the corresponding $u_1(\mathbf{x}_{low})$ and $u_2(\mathbf{x}_{up})$ ordinate values can be used to set the classification threshold t. Then, referring to measures of ambiguity in the middle - and bottom - plots, the behaviour of the different schemes can be analyzed. In such a situation, Chow's and Horiuchi's (Łukasiewicz implication) schemes would lead to reject too many points in low density areas, while other schemes do not show these drawbacks. On the other hand, Ha's scheme (standard implication) would not allow to reject simultaneously highly ambiguous and outlying patterns, whereas schemes based on algebraic or parametrical implications do (provided an appropriate value for the parameter), so a better classification performance can be expected. One can also see in the bottom-plot that tuning the parameter allows to reject patterns for ambiguity as well as for distance (a higher value of γ for the Hamacher family and a lower value for the Dombi family) as pointed out in the discussion.

5 Experimental results

In this section, we report experiments carried out on both artificial and real benchmark datasets for which it is beneficial to introduce the ambiguity reject option because the classes overlap in the feature space. The proposed class-selective rejection scheme is compared to the usual rejection schemes presented in section 2. For additional comparison, results obtained using the two very recent classification rules proposed by Tax and Duin [12] called *Outlier-norm* and *Target-norm* are given, the former being especially designed for distance rejection. Note we did not detailed these (not usual) rules here because they do not rely on an ambiguity measure, therefore they cannot be derived using fuzzy R-implications, by contrast to the previous ones. In all cases, the $L-$step was performed using Eqs.(1-2) with $\alpha_i = 1$ $(\forall i = 1, c)$.

5.1 The datasets

Two synthetic datasets are used: D which contains 2000 points drawn from two normal seven-dimensional distributions of 1000 points each with means $\mathbf{v}_1 = {}^t(1 \;\; 0 \;\cdots\; 0)$ and $\mathbf{v}_2 = {}^t(-1 \;\; 0 \;\cdots\; 0)$, and equal covariance matrices $\Sigma_1 = \Sigma_2 = I$, and DH which consists of two overlapping gaussian classes with different covariance matrices according to the Highleyman distribution, each composed of 800 observations in $\mathbb{R}^2$, see [13]. Real datasets are taken from the UCI Machine Learning Repository [14]. The characteristics (number n of patterns, number p of features, number c of classes, degree of overlap) are reported in Table 5.

Table 5: The datasets and their characteristics.

Dataset	n	p	c	**Overlap**
D	2000	7	2	slight
DH	1600	2	2	very slight
Ionosphere	351	34	2	very strong
Forest	495411	10	2	moderate
Vowel	528	10	11	very slight
Digits	10992	16	10	very slight
Thyroid	215	5	3	very slight
Pima	768	9	2	strong
Statlog	6435	36	6	slight
Glass	214	9	6	moderate

5.2 Results

Table 6 shows the classification performance of the different rejection schemes obtained by a 10-fold cross-validation procedure on the different datasets. In all cases, the threshold t was set to reject 10% of the data, so that 90% is the best achievable correct classification rate performance, then the error rate is $(90 - \text{correct})\%$. The best results are indicated in bold. Note that there are no outliers in the datasets, so that a part of the rejected points are considered as outliers whereas they belong to classes.

It appears from these results that parametrical implications generally outperform usual rejection schemes as well as non usual ones (Outlier-norm, Target-norm), e.g. $I_{H_{\gamma=2}}$ and $I_{D_{\gamma=0.5}}$. As expected in the previous section, these schemes enable to reject both ambiguous and outlying patterns. Furthermore, the tradeoff to be found between a scheme which

Table 6: Classification performance (%) on synthetic and real datasets.

Scheme	D	DH	Ionosphere	Forest	Vowel	Digits	Thyroid	Pima	Statlog	Glass
Chow	75.15	83.75	54.42	67.44	87.88	88.03	84.19	59.51	75.66	66.82
Ha	77.35	86.63	54.99	68.35	88.18	87.38	86.98	60.55	75.56	69.63
Horiuchi	77.55	84.31	56.13	69.27	89.09	88.76	86.98	62.89	77.68	68.22
Frélicot	79.65	87.12	**58.12**	69.73	89.09	89.13	**87.91**	63.28	77.48	71.03
Hamacher_0	**80.00**	87.12	55.56	69.76	88.99	88.26	**87.91**	63.15	77.26	70.09
Hamacher_2	79.75	**87.30**	**58.12**	**70.01**	**89.49**	**89.17**	**87.91**	63.28	77.78	**71.33**
$\text{Dombi}_{0.5}$	79.20	86.68	56.41	69.76	89.09	88.96	**87.91**	**63.40**	**78.12**	71.03
Dombi_2	78.85	86.38	55.56	69.65	88.89	87.91	**87.91**	62.50	76.94	70.09
Outlier-norm	76.85	83.31	57.54	67.44	87.88	86.19	84.65	61.33	75.66	69.16
Target-norm	78.80	87.19	56.83	69.69	87.27	86.39	84.19	63.15	75.86	66.82

does not reject outliers (Ha) and others which reject too much patterns (Chow, Horuichi) appears to favour the choice of implications based on a t-norm which is lower than $\top_S$ (the highest t-norm) and greater than $\top_L$, e.g. the algebraic or the parametrical implications (provided an appropriate value for γ). More generally, rejection schemes that take into account relationship between fuzzy membership degrees (Horiuchi, Frélicot, Hamacher, Dombi) perform better than all the others schemes we tested. Looking at the degrees of overlap, $I_{D_{\gamma=0.5}}$ (respectively $I_{H_{\gamma=2}}$) is more efficient for datasets presenting a slight/very slight (respectively strong/very strong) overlap.

6 Conclusion and perspectives

In this paper, a generalized class-selective rejection scheme based on a logical approach to pattern assignation is presented allowing to either reject only ambiguous patterns or ambiguous and outlying patterns. For this purpose, we propose to design a family of ambiguity measures based on fuzzy residual implication functions, which indicate to which extent a pattern should be assigned to $n^\star$ classes depending on its membership degrees from the truth value of the implication. These measures assess the relationships between membership values with various combination of triangular norm operators, including parametrical families. It is shown that the proposed scheme generalizes well-known ones of the literature on pattern classification with reject options. It appears from experiments on both synthetic and real datasets that using, as the basis for residual implication computation, triangular norms greater than the Łukasiewicz triangular norm and lower than the standard one, gives better classification performance. Furthermore, measures taking in consideration several membership degrees, so that interactions between classes are taken into account, also give better results.

A future work will consist in defining new class-selective rejection schemes based on other parametrical triangular families (Sklar, Frank, Yager and so on) and compare their classification performance. We also think that, depending on the context of the pattern recognition problem, other implications functions than the residual implication ones would be suitable. We plan to make an extensive study on the behavior of *S-implications* which are an immediate generalization of the usual boolean implication, *QL-implications* coming from quantum mechanic logic, and *D-implications* which are the contraposition with respect to a negation of QL-implications, see [7] for definitions.

References

[1] B. Dubuisson and M. Masson. A statistical decision rule with incomplete knowledge about classes. *Pattern Recognition*, 26:155–165, 1993.

[2] C.K. Chow. On optimum error and reject tradeoff. *IEEE Transactions on Information Theory*, 16:41–46, 1970.

[3] E.P Klement and R. Mesiar. *Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms*. Elsevier, 2005.

[4] T. Ha. The optimum class-selective rejection rule. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):608–615, 1997.

[5] T. Horiuchi. Class-selective rejection rule to minimize the maximum distance between selected classes. *Pattern Recognition*, 31:1579–1588, 1998.

[6] C. Frélicot and B. Dubuisson. A multi-step predictor of membership function as an ambiguity reject solver in pattern recognition. In *4th Int. Conf. on Information Processing and Management of Uncertainty in Knowledge-based Systems*, pages 709–715. IPMU'92, July 1992.

[7] M. Mas, M. Monserrat, J. Torrens, and E. Trillas. A survey on fuzzy implication functions. *IEEE Transactions on Fuzzy Systems*, 15(6):1107–1121, 2007.

[8] Th. Whalen. Parameterized r-implications. *Fuzzy Sets and Systems*, 134(2):231–281, 2003.

[9] A. Tversky. Features of similarity. *Psychological review*, 84(4):327–352, 1977.

[10] K. Hirota and W. Pedrycz. Matching fuzzy quantities. *IEEE Transactions on Systems, Man, and Cybernetics*, 21(6):1580–1586, 1991.

[11] L. Rutkowski and K. Cpalka. Flexible neuro-fuzzy systems. *IEEE Transactions on Neural Networks*, 14(3):554–574, 2003.

[12] D.M.J. Tax and R.P.W. Duin. Growing a multi-class classifier with a reject option. *Pattern Recognition Letters*, 29(10):1565–1570, 2008.

[13] W. Highleyman. Linear decision functions, with application to pattern recognition. *Proc. IRE 50*, 50(6):1501–1514, 1962.

[14] C. Blake and C. Merz. Uci repository of machine learning databases, 1998.