We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

6,300

Open access books available

172,000

International authors and editors

190M

Downloads

Our authors are among the

154

Countries delivered to

TOP 1%

most cited scientists

12.2%

Contributors from top 500 universities

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Class-Selective Rejection Rules based on the Aggregation of Pattern Soft Labels

Carl Frélicot and Hoel Le Capitaine
Lab. MIA - Université de La Rochelle
France

1. Introduction

Let $\Omega = \{\omega_1, \dots, \omega_c\}$ be a set of c classes and let $\mathbf{x}$ be a pattern described by p features, namely a vector $\mathbf{x} = {}^t(x_1 \dots x_p)$ in a p -dimensional real space $\mathbb{R}^p$. Classifier design aims at defining rules that allow to associate an incoming pattern $\mathbf{x}$ with one class of Ω . Let $\mathcal{L}_{hc}$ be the set of c -dimensional binary vectors whose components sum up to one. Then, such a rule, defined as a mapping $D: \mathbb{R}^p \rightarrow \mathcal{L}_{hc}$, $\mathbf{x} \mapsto \mathbf{h}(\mathbf{x})$, is called a *crisp classifier*. In most theoretical approaches to pattern classification, it is convenient to define a classifier as a couple (L, H) where:

- L is a *labeling* function: $\mathbb{R}^p \rightarrow \mathcal{L}_{\bullet c}$, $\mathbf{x} \mapsto \mathbf{u}(\mathbf{x})$, $\mathcal{L}_{\bullet c}$ depending on the mathematical framework the classifier relies on, $\mathcal{L}_{pc} = [0, 1]^c$ for degrees of typicality or $\mathcal{L}_{fc} = \{\mathbf{u}(\mathbf{x}) \in \mathcal{L}_{pc} \mid \sum_{i=1}^c u_i(\mathbf{x}) = 1\}$ for posterior probabilities and fuzzy membership degrees or even $\mathcal{L}_{hc}$;
- H is a *hardening* function: $\mathcal{L}_{\bullet c} \rightarrow \mathcal{L}_{hc}$, $\mathbf{u}(\mathbf{x}) \mapsto \mathbf{h}(\mathbf{x})$, which often reduces to the class of maximum label selection, $\mathcal{L}_{hc} = \{\mathbf{h}(\mathbf{x}) \in \mathcal{L}_{fc} \mid h_i(\mathbf{x}) \in \{0, 1\}\}$.

Thus, the crisp classifier D is a special case of the classifier (L, H) . Whenever $\mathcal{L}_{\bullet c} \neq \mathcal{L}_{hc}$, label vectors $\mathbf{u}(\mathbf{x})$ are said to be *soft* and the resulting classifier is called a *soft classifier*. Special cases can be emphasized because of the L -function: the *possibilistic* classifier when $\mathcal{L}_{\bullet c} = \mathcal{L}_{pc}$, the *fuzzy* and the *probabilistic* classifiers when $\mathcal{L}_{\bullet c} = \mathcal{L}_{fc}$. In the probabilistic case, $u_i(\mathbf{x})$ are posterior probabilities $P(\omega_i \mid \mathbf{x})$ that can be obtained either from (known) class-conditional densities whose parameters are estimated using a learning set $\mathcal{X}$ of patterns, *i.e.* patterns for which the class-assignment is known, or from class-density estimates using the classes of their neighbors in $\mathcal{X}$. Throughout this chapter we shall use these definitions because most statistical pattern classifiers share either the L -function or the H -function, see examples in (Frélicot, 1998). Furthermore, the chapter addresses the problem of aggregating the soft labels issued from the L -function by the design of special H -functions, whatever they have been obtained. However, note that some authors consider the mapping $D: \mathbb{R}^p \rightarrow \mathcal{L}_{hc}$, $\mathcal{L}_{fc}$ or $\mathcal{L}_{pc} \setminus \mathbf{0}^1$ to define a crisp, a fuzzy or a possibilistic classifier respectively (Bezdek et al., 1999). Since the L -part is out of the scope of this chapter and there are many ways to compute labels,

¹ $\mathbf{0} = {}^t(0 \dots 0)$ is the p -dimensional null vector

we shall use the degree of typicality defined by:

$$u_i(\mathbf{x}) = \frac{\alpha_i}{\alpha_i + d^2(\mathbf{x}, \mathbf{v}_i)} \quad (1)$$

where α_i is a user-defined parameter, d is a distance in $\mathbb{R}^p$, and $\mathbf{v}_i$ is a prototype of the class ω_i . Among the possible distances, one finds the Mahalanobis distance $d^2(\mathbf{x}, \mathbf{v}_i) = (\mathbf{x} - \mathbf{v}_i)^t \Sigma_i^{-1} (\mathbf{x} - \mathbf{v}_i)$ where $\mathbf{v}_i$ and Σ_i are the mean vector and covariance matrix of ω_i estimated from a learning set $\mathcal{X}$. It has been shown through empirical studies (Zimmermann & Zysno, 1985) that (1) is a good model for membership functions that model vague concepts or classes. Parameters α_i will always all be set to 1 except if mentioned otherwise.

As defined, any H -function results in an exclusive classification rule which is not efficient in practice because it supposes that Ω is exhaustively defined (*closed-world* assumption) and that classes do not overlap (*separability* assumption). In many real applications, both assumptions are not true and such a classifier can lead to very undesired decisions. It is often more convenient to withhold making a decision and direct the pattern to an exceptional handling 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 the misclassification risk. The first one, called *distance rejection* (Dubuisson & Masson, 1993) 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 (Chow, 1970), (Ha, 1997). 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 the H -function definition such that $\mathbf{h}(\mathbf{x})$ can take values in the set of vertices $\mathcal{L}_{hc}^c = \{0, 1\}^c$ of the unit hypercube 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 $\mathbf{h}(\mathbf{x})$ is $\mathbf{0}$,
- classification when $\mathbf{h}(\mathbf{x})$ is in $\mathcal{L}_{hc}$,
- ambiguity rejection when $\mathbf{h}(\mathbf{x})$ is in $\mathcal{L}_{hc}^c \setminus \{\mathcal{L}_{hc} \cup \mathbf{0}\}$.

The resulting classification rule is then a matter of selecting, by H , the appropriate number of classes varying from zero (distance rejection) to c (total ambiguity rejection) and which class/es is/are involved, provided its soft label vector $\mathbf{u}$ is available from L . This can be obtained by aggregating the components of $\mathbf{u}$ in a suitable way.

2. Aggregation Operators for Class-Selection

In this section, we first briefly review basic aggregation operators (Calvo et al., 2002), in particular the ones issued from the fuzzy sets theory which have received more attention in the last few decades because of their ability to manage imprecise and/or incomplete data. Then we present some combinations of them that can be used to select the number of classes to which an incoming pattern $\mathbf{x}$ has to be assigned because they allow to define some ambiguity measures.

2.1 Preliminary Definitions

Let us recall basic definitions of aggregation functions or operators that will be used to combine the values of interest, *i.e.* the soft labels of a pattern to be classified. In a broader sense, aggregation functions aim at associating a typical value to a number of several numerical values which are generally defined on a finite real interval on an ordinal scale. They are used in many fields, *e.g.* decision-making and pattern recognition (Grabisch, 1992). Since soft labels $u_i(x)$ are in $[0,1]$, we restrict to functions that aggregate values from the unit interval and we define an aggregation operator as a mapping $\mathcal{A}: [0,1]^c \rightarrow [0,1]$, $\mathbf{a} = \{a_1, \dots, a_c\} \mapsto \mathcal{A}(\mathbf{a})$, satisfying the following conditions:

- (A1) $\mathcal{A}(0, \dots, 0) = 0$ and $\mathcal{A}(1, \dots, 1) = 1$ (boundaries),
- (A2) $\forall c \in \mathbb{N}, a_1 \leq b_1, \dots, a_c \leq b_c \Rightarrow \mathcal{A}(a_1, \dots, a_c) \leq \mathcal{A}(b_1, \dots, b_c)$ (monotonicity).

Adding properties like idempotency, continuity, associativity lead to others definitions but this one is strong enough for our discourse. In the literature, one finds many families of aggregation operators, *e.g.*: triangular norms (Menger, 1942), OWA (Ordered Weighted Averaging) operators (Yager, 1988), γ -operators (Zimmerman & Zysno, 1980), or fuzzy integrals (Sugeno, 1974). They are classified either by some mathematical properties they share or by the way the values are aggregated. An aggregation operator $\mathcal{A}$ is said to be:

- (A3) conjunctive if $\mathcal{A}(\mathbf{a}) \leq \min\{a_1, a_2, \dots, a_c\}$,
- (A3') disjunctive if $\mathcal{A}(\mathbf{a}) \geq \max\{a_1, a_2, \dots, a_c\}$,
- (A3'') compensatory if $\min\{a_1, a_2, \dots, a_c\} \leq \mathcal{A}(\mathbf{a}) \leq \max\{a_1, a_2, \dots, a_c\}$,

refer to (Calvo et al., 2002; Grabisch et al., 2009) for a large survey on aggregation operators.

2.2 Basic Aggregation Operators

Beyond these operators, we choose to use the triangular norms and co-norms because of their ability to generalize the logical AND and OR crisp operators to fuzzy sets, see (Klement & Mesiar, 2005) for a survey. A fuzzy negation (or complement) is defined as a continuous, non increasing function $N: [0,1] \rightarrow [0,1]$ satisfying:

- (N1) $N(0) = 1$ and $N(1) = 0$ (boundaries),
- (N2) $N(N(a)) = a$ (involution).

A triangular norm (or t-norm) is a binary operator $\top: [0,1]^2 \rightarrow [0,1]$ satisfying the following four axioms: $\forall a, b, c \in [0,1]$

- (T1) $a \top b = b \top a$ (symmetry),
- (T2) if $b \leq c \Rightarrow a \top b \leq a \top c$ (monotonicity),
- (T3) $a \top (b \top c) = (a \top b) \top c$ (associativity),
- (T4) $a \top 1 = a$ (neutral element).

It is easy to see that these axioms make $\top$ satisfy (A1) and (A2), so any t-norm is an aggregation operator.

Given a fuzzy negation N , *e.g.* the strict negation $N(a) = 1 - a$, a triangular conorm (or t-conorm) is the dual operation $\perp, [0,1]^2 \rightarrow [0,1]$, defined as:

$$a \perp b = N(N(a) \top N(b)) \quad (2)$$

Therefore, a t-conorm satisfies axioms (T1), (T2), (T3), so any t-conorm is an aggregation operator in the sense of (A1) and (A2), and satisfies:

(S4) $a \perp 0 = 0$ (neutral element).

Axioms (T2), (T4) and (S4) imply additional axioms: $\forall a, b \in [0, 1]$

(T5) $a \top b \leq a$,

(S5) $a \leq a \perp b$,

and it ensues:

(T6) $a \top 0 = 0$ (absorbing element for any t-norm),

(T7) $a \top b \leq \min(a, b)$ ($\min$ is the largest t-norm),

(S6) $a \perp 1 = 1$ (absorbing element for any t-conorm),

(S7) $\max(a, b) \leq a \perp b$ ($\max$ is the smallest t-conorm).

Typical examples of dual couples $(\top, \perp)^2$ that will be used in the sequel are given in Table 1, including two parametric ones which generalize other t-norms and t-conorms depending of the parameter value, e.g.:

- $(\top_{H_1}, \perp_{H_1}) = (\top_A, \perp_A)$, the *product* and *probabilistic sum*
- $(\top_{D_1}, \perp_{D_1}) = (\top_{H_0}, \perp_{H_0})$, the *Hamacher's product and sum*,
- $(\top_{D_{+\infty}}, \perp_{D_{+\infty}}) = (\top_S, \perp_S)$, the *min* and *max* operators.

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 ($\gamma \in [0, +\infty[$)	$a \top_{H_\gamma} b = \frac{a b}{\gamma + (1-\gamma)(a+b-a b)}$ $a \perp_{H_\gamma} b = \frac{a+b+(\gamma-2) a b}{1+(\gamma-1) a b}$
Dombi ($\gamma \in]0, +\infty[$)	$a \top_{D_\gamma} b = \left(1 + \left(\left(\frac{1-a}{a}\right)^\gamma + \left(\frac{1-b}{b}\right)^\gamma\right)^{1/\gamma}\right)^{-1}$ $a \perp_{D_\gamma} b = 1 - \left(1 + \left(\left(\frac{a}{1-a}\right)^\gamma + \left(\frac{b}{1-b}\right)^\gamma\right)^{1/\gamma}\right)^{-1}$

Table 1. Some triangular norm dual couples.

² ones also refers to triples $(\top, N, \perp)$

2.3 Ambiguity Measures based on Combination of Basic Operators

An *ambiguity measure* is any aggregation function $\Phi: \mathcal{L}_{\bullet c} \rightarrow [0, 1]$, $\mathbf{u}(\mathbf{x}) \mapsto \Phi(\mathbf{u})$ that can reveal if an incoming pattern $\mathbf{x}$ could be associated with several classes, hence that can be used to define the H -function of a class-selective rejection classifier. We present hereafter some combinations of basic aggregation operators and the derived ambiguity measures.

Let $\mathcal{P}$ be the powerset of $C = \{1, 2, \dots, c\}$ and $\mathcal{P}_l = \{A \in \mathcal{P} : \text{card}(A) = l\}$. The *fuzzy l -order OR operator* (fOR- l for short) is an aggregation function, as defined in subsection 2.1:

$[0, 1]^c \rightarrow [0, 1]$, $\mathbf{u} \mapsto \perp^l(\mathbf{u})$, where

$$\perp^l(\mathbf{u}) = \bigwedge_{i=1, \dots, c}^l u_i = \bigwedge_{A \in \mathcal{P}_{l-1}} \left(\bigwedge_{j \in C \setminus A} u_j \right) \quad (3)$$

Some properties of fOR- l result from those of $\bigvee$ and $\bigwedge$, others have been proved in (Mascarilla et. al, 2008). Among these properties, let us recall those that are useful for the context we are interested in:

(L1) $\perp^l(\mathbf{0}) = \mathbf{0}$ and $\perp^l(\mathbf{1}) = \mathbf{1}^3$ (boundaries)

(L2) for $\mathbf{u}$ and $\mathbf{v}$ such as $u_i \leq v_i, \forall i \in C$, $\perp^l(\mathbf{u}) \leq \perp^l(\mathbf{v})$ (monotonicity)

(L3) for any permutation σ of C , $\bigwedge_{i=1, \dots, c}^l u_{\sigma(i)} = \bigwedge_{i=1, \dots, c}^l u_i$ (symmetry)

(L4) $\perp^1(\mathbf{u}) = \bigwedge(\mathbf{u})$ and $\perp^c(\mathbf{u}) = \bigvee(\mathbf{u})$, whatever c and $(\bigvee, \bigwedge)$,

(L5) if the standard norms are taken, then $\perp_S^l(\mathbf{u}) = u_{(l)}$, the l -th highest value⁴ in $\mathbf{u}$; for instance, let us take $C = \{1, 2, 3\}$ and $l = 2$, then $\mathcal{P}_{l-1} = \{\{1\}, \{2\}, \{3\}\}$ and we have

$\perp_S^2(\mathbf{u}) = \min(\max(u_2, u_3), \max(u_1, u_3), \max(u_1, u_2))$, so that if $u_2 < u_1 < u_3$, then $\perp_S^2(\mathbf{u}) = u_1$.

Properties (L1) and (L2) make fOR- l be a family (parametrized by $(\bigvee, \bigwedge)$) of aggregation functions in the sense of (A1) and (A2). Axiom (T7) and property (L5) allow us to claim that the fOR- l operator measures to what extent the (generalization of the, given by the dual couple) l highest values of $\mathbf{u}$ are all high. So, if $\mathbf{u}$ is the soft label vector of a pattern $\mathbf{x}$ to be classified, it can be used as a family of ambiguity measures to reject it between the l corresponding classes as follows: given a dual couple $(\bigvee, \bigwedge)$, $\forall 2 \leq l \leq c$

$$\Phi_{l, \bigvee}(\mathbf{u}) = \perp^l(\mathbf{u}) \quad (4)$$

Let us illustrate this ability on a simple $c = 3$ classes problem. In this so-called *real line example* ($\mathbf{x} \in \mathbb{R}$), the soft labels degrees $u_i(\mathbf{x})$ are modeled by overlapping triangular functions in

³ $\mathbf{1} = {}^t(1 \dots 1)$ is the p -dimensional one vector

⁴ usually $u_{(l)}$ denotes the l -th value in ascending order but reverse order is more convenient in the context of this chapter

order to emphasize them when plotting but similar results would have been obtained using a distance model, *e.g.* by (1). Figures 1 and 2 show these degrees and the values of the ambiguity measure $\Phi_{l,\top}(\mathbf{u})$, with $l = 2$ and $l = 3$ respectively, for the different norm couples of Table 1. Of course, the ambiguity measure values depend on the couple but not in such a way that makes the remarks below not valid. However, it is obvious that some ordering appears because of the basic couples' ordering and/or the asymptotic values of the parametric ones, *e.g.* $\Phi_{l,\top_{D_\gamma}}(\mathbf{u})$ tends to $\Phi_{l,\top_S}(\mathbf{u})$ as γ increases. One can see in both Figures that $\Phi_{l,\top}(\mathbf{u}) = 0$

Fig. 1. Soft label degrees $u_{i=1,2,3}(\mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}$, and ambiguity measures $\Phi_{l=2,\top}(\mathbf{u})$ for different norm couples.

where strictly less than l degrees overlap on the real line, whatever the norm couple $(\top, \perp)$. One can reasonably expect that any pattern $\mathbf{x}$ lying outside these areas should be ambiguity rejected. For instance, if $\Phi_{l,\top}(\mathbf{u}) \geq 0.5$ in Figure 1 (respectively 0 in Figure 2), then all $\mathbf{x} \in [300, 500]$ could be associated with two (respectively three) classes, whatever $(\top, \perp)$. However, a question remains: which order ($l = 2$ or 3) induces more than the other one this ambiguity? This is a matter of selecting the appropriate number of classes which can be processed by the class-selective rule through the H -function definition. We address this problem in section 3.

The restriction to $l \geq 2$ in the definition (4) of the family of ambiguity measures $\Phi_{l,\top}(\mathbf{u})$ is motivated by an operational reason. It has been established that class-selective rejection rules that take into account the relationships between the degrees to be aggregated, e.g. the historical ones presented in section 3.1, perform better. As the fOR-1 operator reduces to $\perp$ by (L4), there are dual couples for which $\Phi_{1,\top}(\mathbf{u})$ does not depend on such relationships between the u_i values, whenever the soft label vector $\mathbf{u}$ is in $\mathcal{L}_{pc}$, in particular the standard one ($\perp = \max$). Of course, whenever $\mathbf{u}$ is in $\mathcal{L}_{fc}$, i.e. when it is a collection of posterior probabilities or membership degrees of a pattern $\mathbf{x}$ to be classified, such a relationship holds⁵ and $\Phi_{1,\top}(\mathbf{u})$ becomes useful. An alternative in the more general case consists in taking a fuzzy complement of the fOR-1 operator, e.g. the strict negation, to define a family of ambiguity measures which has never been proposed: given a t-conorm $\perp$

$$\Phi_{\bar{1},\top}(\mathbf{u}) = 1 - \perp(\mathbf{u}) = 1 - \perp(\mathbf{u}) \quad (5)$$

By definition, $\Phi_{\bar{1},\top}(\mathbf{u})$ measures to what extent the (generalization of the, given by fOR-1) highest value in $\mathbf{u}$ is not (by complement) high. Therefore, this family of ambiguity measures is suitable to define a *rejection rule* through a particular H -function, but not a *class-selective rejection rule* in a direct way because it does not enable to select the number of classes a pattern has to be associated with, contrarily to the previous family. We will show in the next section that this major difference holds for historical rules that use a single underlying ambiguity measure, so they only allow to reject patterns between at least two classes. This characteristic is well illustrated in Figure 3 where the values of the ambiguity measure $\Phi_{\bar{1},\top}(\mathbf{u})$ on the *real line example* for the different t-conorms of Table 1 are plotted.

Whatever $\top$, the values of $\Phi_{\bar{1},\top}(\mathbf{u})$ can not be used to decide whether two or three classes originate the ambiguity. Note that, once again, the ordering of the different curves is in accordance with that of the basic t-conorms and the parameter values, see for instance that $\Phi_{\bar{1},\top_{D_\gamma}}(\mathbf{u})$ tends to $\Phi_{\bar{1},\top_S}(\mathbf{u})$ as γ increases. Other remarks can be made. First of all, the highest values are obviously obtained with $\top_S$ because of property (S7). Second, even if one can find a threshold on $\Phi_{\bar{1},\top}(\mathbf{u})$ which allows to reject patterns $\mathbf{x}$ lying in areas where two degrees $u_i(\mathbf{x})$ overlap, the threshold would be so small (e.g. 0.1 for most of t-conorms) that some patterns lying outside these areas would not reasonably rejected between at least two classes as well. Look for instance all $\mathbf{x} \in [250, 350] \cup [450, 550]$ for which either $u_1(\mathbf{x})$ or $u_2(\mathbf{x})$ is greater than 0.9. Such a drawback could be somewhat avoided by normalizing the degrees. Third, we also have a high value of $\Phi_{\bar{1},\top}(\mathbf{u})$ for all $\mathbf{x} \in]-\infty, 200] \cup [600, +\infty[$ while one can reasonably expect that most of them (the farthest ones) should be distance rejected. This

⁵ if $\mathbf{u} \in \mathcal{L}_{fc}$, then $\sum_{i=1}^c u_i = 1$

Fig. 2. Soft label degrees $u_{i=1,2,3}(x), \forall x \in \mathbb{R}$, and ambiguity measures $\Phi_{l=3, \top}(u)$ for different norm couples.

unability to distinguish between both kinds of rejection (distance or ambiguity) is typical of such rejection rules, even if the degrees are normalized.

Fig. 3. Soft label degrees $u_{i=1,2,3}(\mathbf{x})$, $\forall \mathbf{x} \in \mathbb{R}$, and ambiguity measures $\Phi_{\bar{I}, \tau}(\mathbf{u})$ for different norm couples.

Given a fuzzy complement function N , the *fuzzy exclusive OR* operator (fXOR for short) is an aggregation function: $[0, 1]^c \rightarrow [0, 1]$, $\mathbf{u} \mapsto \perp(\mathbf{u})$, where (Mascarilla & Frélicot, 2001):

$$\perp(\mathbf{u}) = \bigwedge_{i=1, \dots, c} u_i = \perp(\mathbf{u}) \tau N\left(\frac{\perp(\mathbf{u})}{\perp(\mathbf{u})}\right)$$

(6)

Fig. 4. Soft label degrees $u_{i=1,2,3}(\mathbf{x}), \forall \mathbf{x} \in \mathbb{R}$, and ambiguity measures $\Phi_{\perp, \top}(\mathbf{u})$ for different norm couples.

The term on the right-hand side of $\top$ penalizes the one on the left-hand side, except if $\perp^2(\mathbf{u})$ is significantly lower than $\perp^1(\mathbf{u})$ so that the negation of the ratio becomes high and $\perp(\mathbf{u})$ tends to $\perp^1(\mathbf{u}) = \perp(\mathbf{u})$. We can say that the value of $\perp(\mathbf{u})$ is high if the (generalization of the, given

by fOR-1) highest value is large enough compared to the (generalization of, given by fOR-2) second highest one and therefore to the others. Therefore, fXOR can be used to define another family of ambiguity measures between at least two classes: given a dual triple $(\top, N, \perp)$,

$$\Phi_{\perp, \top}(\mathbf{u}) = 1 - \perp(\mathbf{u}) \quad (7)$$

By property (T5) and definition (4), we have: $\perp(\mathbf{u}) \leq \perp(\mathbf{u})$ whatever $(\top, N, \perp)$. As $\perp(\mathbf{u}) = \frac{1}{\perp(\mathbf{u})}$ by property (L4), the following property holds:

$$\Phi_{\perp, \top}(\mathbf{u}) \geq \Phi_{\top, \perp}(\mathbf{u}). \quad (8)$$

Therefore, $\Phi_{\perp, \top}(\mathbf{u})$ is expected to be less sensitive to the choice of the threshold than $\Phi_{\top, \perp}(\mathbf{u})$ in order to avoid unexpected rejection, as pointed out on the *real line example*. Figure 4 shows the values of the ambiguity measure $\Phi_{\perp, \top}(\mathbf{u})$ on the *real line example* for the different norm couples of Table 1. Once again, even if the norm triple $(\top, N, \perp)$ influences the ambiguity measure values, it does not significantly affect the results and the ordering of the different curves is in accordance with that of the basic operators and their parameter values. One can see that this family shares the same major drawbacks with the previous one. In particular, if used in a direct way, it is suitable to define a rejection rule but not a class-selective rule because this measure can not suggest that the number of classes a pattern has to be associated with is in C but in $\{1, c\}$. See, for instance, that we have $\Phi_{\perp, \top}(\mathbf{u}) \geq 0.5$ whatever the triple $(\top, N, \perp)$ for all $\mathbf{x} \in [300, 500]$, so this measure can be used to ambiguity reject such patterns as expected but without knowing which classes generate this ambiguity. However, contrarily to $\Phi_{\top, \perp}(\mathbf{u})$, the threshold to be find is not too small so all $\mathbf{x} \in [250, 350] \cup [450, 550]$ for which either $u_1(\mathbf{x})$ or $u_2(\mathbf{x})$ is greater than 0.9 will be not rejected, as one could expect.

3. Rejection Rules and Class-Selective Rejection Rules based on Soft Labels Aggregation

This section is dedicated to the use of ambiguity measures to define the H -function of either a rejection classifier or a class-selective rejection classifier. Historical rejection rules and the ones resulting from the ambiguity measures presented above are unified. This result in a single algorithm aiming at selecting, for an incoming pattern $\mathbf{x}$ to be classified, the appropriate number of classes $n^*(\mathbf{x})$, given its soft label vector. By convention, we will define mappings

$$H: \mathcal{L}_{\bullet c} \rightarrow \mathcal{L}_{hc}^c, \mathbf{u}(\mathbf{x}) \mapsto \mathbf{h}(\mathbf{x}) \text{ such as } u_{(i)}(\mathbf{x}) \mapsto h_i(\mathbf{x}), \forall i \in C.$$

3.1 Historical Rules and Underlying Ambiguity Measures

The first rejection rule (Chow, 1957) is a probabilistic one which is based on the Bayes decision rule defined by

$$H_B: \mathcal{L}_{fc} \rightarrow \mathcal{L}_{hc} \subset \mathcal{L}_{hc}^c, \mathbf{u}(\mathbf{x}) \mapsto \mathbf{h}(\mathbf{x}) = {}^t(1, 0, \dots, 0).$$

As $u_i(\mathbf{x})$ are posterior probabilities, this rule assigns the incoming pattern $\mathbf{x}$ to the most probable class and it is known to be optimal with respect to the error probability, *i.e.* no other probabilistic classifier can yield a lower error probability. Figure 5 - (left) illustrates the partition of the feature space into three regions, resulting from H_B , each region corresponding to

a single class. The rejection rule introduced by Chow minimizes the error probability for a given reject probability which is specified by a threshold $t \in [0, \frac{c-1}{c}]$, or vice-versa. Thus, this rule yields the optimum error-reject tradeoff (Chow, 1970) and it is defined by

$$H_{Ch}: \mathcal{L}_{fc} \rightarrow \{\mathcal{L}_{hc}, \underline{1}\}, \mathbf{u}(\mathbf{x}) \mapsto \mathbf{h}(\mathbf{x}) = \begin{cases} {}^t(1, 0, \dots, 0) & \text{if } u_{(1)}(\mathbf{x}) > (1-t) \\ \underline{1} & \text{otherwise} \end{cases}$$

It means that $\mathbf{x}$ is either exclusively classified or ambiguity rejected between all the classes (*total ambiguity rejection*) if its highest posterior probability $u_{(1)}(\mathbf{x})$ is lower than some given threshold. Therefore, H_{Ch} is not a *class-selective rejection rule* but a *simple rejection rule* because $\mathbf{h}(\mathbf{x})$ can not take any value in $\mathcal{L}_{hc}^c$ but only the ones in $\{\mathcal{L}_{hc}, \underline{1}\}$. Both concepts are illustrated in Figure 5 - (*right*) and - (*center*) respectively. However, Chow's rule uses the complement of the maximum value of the posterior probabilities as an ambiguity measure:

$$\Phi_{Ch}(\mathbf{u}) = 1 - u_{(1)}(\mathbf{x}) \quad (9)$$

and the number of classes $n(\mathbf{x}, t)$ to be selected is in $\{1, c\}$. Since $\mathbf{u}(\mathbf{x}) \in \mathcal{L}_{fc}$, it is easy to show that H_{Ch} is identical to H_B whenever $t > \frac{c-1}{c}$, i.e. $\mathbf{x}$ can not be rejected. Note that, as defined, $\Phi_{Ch}(\mathbf{u})$ is a special case of $\Phi_{\bar{1}, \top}(\mathbf{u})$ by definition (5) and property (L4):

$$\Phi_{Ch}(\mathbf{u}) = \Phi_{\bar{1}, \top_s}(\mathbf{u}) \quad (10)$$

Since the work by Chow, most of rejection rules that have been proposed attempt to avoid total ambiguity rejection whenever at least two classes have to be selected. Such a class-selective procedure can be defined, in its general form, as the seek for the optimal number of classes according to:

$$n^*(\mathbf{x}, t) = \min_{k \in C} \{k : \Phi(\mathbf{u}(\mathbf{x})) \leq t\} \quad (11)$$

where Φ is an ambiguity measure on the pattern soft labels, i.e. a vector of posterior probabilities or membership degrees in $\mathcal{L}_{fc}$ or even typicality degrees in $\mathcal{L}_{pc}$, $n^*(\mathbf{x}, t)$ 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-order dependent ($t_{(k)}$). This threshold can be set conditionally to cost functions relative to error, reject and correct classification rates instead of error, reject and correct classification probabilities if needed.

Propositions from the literature mainly consist in defining new ambiguity measures Φ suitable for class-selective rejection instead of total ambiguity rejection:

- Ha ranks the posterior probabilities and test their values up to the $(k+1)$ -th to decide if k classes are selected (Ha, 1997):

$$\Phi_{Ha}(\mathbf{u}) = u_{(k+1)}(\mathbf{x}) \quad (12)$$

The corresponding rule H_{Ha} minimizes the error probability for a given average number of classes (Ha, 1996) and the domain of t is $[0, \frac{1}{2}]$. Whenever $t > \frac{1}{2}$, $n^*(\mathbf{x}, t) = 1$

Fig. 5. Feature space partitioning for a three classes problem and reject options – no rejection by H_B (left), rejection rules H_{rej} (center) and class-selective rejection rules H_{sel} (right).

(H_{Ha} reduces to H_B) and can increase up to c as t decreases down to zero. Note that, as defined, $\Phi_{Ha}(\mathbf{u})$ is a special case of $\Phi_{l,\top}(\mathbf{u})$ by definition (4) and property (L4):

$$\Phi_{Ha}(\mathbf{u}) = \Phi_{k+1,\top_s}(\mathbf{u}). \quad (13)$$

- Since this rule is leading to unnatural decisions because of the normalization of $\mathbf{u}(\mathbf{x})$, Horiuchi proposes a new measure defined by the difference of the posterior probabilities (Horiuchi, 1998). This difference is actually a disambiguity measure, so we complement it to derive an ambiguity one:

$$\Phi_{Ho}(\mathbf{u}) = 1 - (u_{(k)}(\mathbf{x}) - u_{(k+1)}(\mathbf{x})) \quad (14)$$

Horiuchi's rule H_{Ho} minimizes the maximum distance between selected classes for a given average number of classes which is specified by $t \in [0, 1]$. It is identical to H_B , so $n^*(\mathbf{x}, t) = 1$ if $t = 0$ and increases up to c as t increases up to 1.

- Before these two works, Frélicot & Dubuisson proposed to use the ratio of typicality degrees (Frélicot & Dubuisson, 1992) in order to relax the summation constraint on $\mathbf{u}(\mathbf{x})$:

$$\Phi_{FD}(\mathbf{u}) = u_{(k+1)}(\mathbf{x}) / u_{(k)}(\mathbf{x}) \quad (15)$$

The induced rule H_{FD} allows to select $n^*(\mathbf{x}, t)$ classes in C , given $t \in [0, 1]$.

As usual, we use the convention that if $\Phi(\mathbf{u}(\mathbf{x})) > t$ for all $k \in C$, then we set $n^*(\mathbf{x}, t) = c$ which will correspond to total ambiguity rejection.

Even if Chow's, Ha's and Horiuchi's rules have been defined within the probabilistic framework ($\mathbf{u}(\mathbf{x}) \in \mathcal{L}_{fc}$), one can intend to use the corresponding ambiguity measures for typicality degrees in $\mathcal{L}_{pc}$. The values of the historical ambiguity measures Φ_{Ch} , Φ_{Ha} , Φ_{Ho} and Φ_{FD} on the *real line example* are shown in Figures 6 and 7 for orders $k+1=2$ and $k+1=3$ respectively. Keeping in mind that Φ_{Ch} will induce a rejection rule and not a class-selective rule, its

values are not shown for $k + 1 = 3$, and one can see that Φ_{Ch} has a similar behaviour than $\Phi_{1,\top}(\mathbf{u})$ and $\Phi_{\perp,\top}(\mathbf{u})$, the other families of measures which do not take care of the ambiguity order. A similar discussion to the one we had for $\Phi_{l=k+1,\top}(\mathbf{u})$ in section 2.3 can be done for the other measures. Whatever the ambiguity measure, its lowest values correspond to patterns $\mathbf{x}$ lying in intervals of the real line for which one can reasonably expect that they should not be ambiguity rejected. At the opposite, the higher values correspond to intervals that contain $\mathbf{x}$ for which one can reasonably expect that they should be ambiguity rejected between two (Figure 6) or three classes (Figure 7). The selection of the most appropriate order is still a question of selecting the appropriate number of classes by (11) through the definition of the H -function.

Fig. 6. Ambiguity measures $\Phi_{Ch}(\mathbf{u})$, and for $k + 1 = 2$, $\Phi_{FD}(\mathbf{u})$, $\Phi_{Ho}(\mathbf{u})$ and $\Phi_{Ha}(\mathbf{u})$ on the real line example.

Class-selective rule	Ambiguity Measure Φ
(Ha, 1997)	$\Phi_{Ha}(\mathbf{u}) = u_{(k+1)}(\mathbf{x})$ (12) $= \Phi_{k+1,\top_s}(\mathbf{u})$
(Horiuchi, 1998)	$\Phi_{Ho}(\mathbf{u}) = 1 - (u_{(k)}(\mathbf{x}) - u_{(k+1)}(\mathbf{x}))$ (14)
(Frélicot & Dubuisson, 1992)	$\Phi_{FD}(\mathbf{u}) = u_{(k+1)}(\mathbf{x}) / u_{(k)}(\mathbf{x})$ (15)
(Mascarilla <i>et. al</i> , 2008)	$\Phi_{k,\top}(\mathbf{u}) = \perp_k(\mathbf{u})$ (4)

Table 2. Class-selective rejection rules and corresponding ambiguity measures.

Fig. 7. Ambiguity measures $\Phi_{FD}(\mathbf{u})$, $\Phi_{Ho}(\mathbf{u})$ and $\Phi_{Ha}(\mathbf{u})$ on the *real line example* for $k + 1 = 3$.

3.2 Unification of Class-selective Rejection Rules based on Ambiguity Measures

In order to derive H –functions that are able to deal with distance rejection using the ambiguity measures that allow it, namely Φ_{Ha} , Φ_{Ho} , Φ_{FD} and $\Phi_{k,\top}(\mathbf{u})$, we only need to replace (11) by:

$$n^*(\mathbf{x}, t) = \min_{k \in [0, c]} \{k : \Phi(\mathbf{u}(\mathbf{x})) \leq t\}$$

(16)

with default values $u_0(\mathbf{x}) = 1$ and $u_{c+1}(\mathbf{x}) = 0$. Therefore, the unified H_{sel} –function of a class-selective rejection rule based on any of the ambiguity measures of Table 2 is given by:

$$H_{sel}: \mathcal{L}_{\bullet c} \rightarrow \mathcal{L}_{hc}^c, \mathbf{u}(\mathbf{x}) \mapsto \mathbf{h}(\mathbf{x}) \text{ such as } h_i(\mathbf{x}) = 1 \text{ for all } i \in \{0, n^*(\mathbf{x}, t)\}.$$

and its implementation is shown in Algorithm 1.

Algorithm 1: Class-selective rejection rule (H –function implementation).

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 0
  given any ambiguity measure  $\Phi(\mathbf{u})$  of Table 2, find  $n^*(\mathbf{x}, t)$  according to (16)
  foreach  $j = 1 : n^*(\mathbf{x}, t)$  do
    | set  $h_j(\mathbf{x}) = 1$  in decreasing order of  $u_{(j)}(\mathbf{x})$ 's
  end
end
end
```

Let us see the classification results obtained applying this common implementation of H_{sel} to some typical soft label vectors $\mathbf{u} \in \mathcal{L}_{pc}$, assumed to correspond to patterns $\mathbf{x}$, where $c = 4$:

- $\mathbf{u} = {}^t(0.70 \ 0.10 \ 0.85 \ 0.80)$, so $\mathbf{x}$ is expected to be ambiguity rejected between three classes $\{\omega_1, \omega_3, \omega_4\}$,
- $\mathbf{u} = {}^t(0.20 \ 0.10 \ 0.85 \ 0.80)$, so $\mathbf{x}$ is expected to be ambiguity rejected between two classes $\{\omega_3, \omega_4\}$,
- $\mathbf{u} = {}^t(0.20 \ 0.10 \ 0.85 \ 0.15)$, so $\mathbf{x}$ is expected to be exclusively classified in one class $\{\omega_3\}$, and
- $\mathbf{u} = {}^t(0.20 \ 0.10 \ 0.05 \ 0.15)$, so $\mathbf{x}$ is expected to be distance rejected.

The values of $\Phi(\mathbf{u}(\mathbf{x}))$ for all the values $k \in [0, c]$ are given in Table 3 as well as the interval of the threshold t leading to the (expected) correct classification above. One can see that all rules succeed by setting t to 0.5, except the ones based on $\Phi_{k, \top_H}(\mathbf{u})$ for which a smaller value is needed.

To end this presentation, we also can give the unified H_{rej} -function of a rejection rule based on any of the ambiguity measures summarized in Table 4:

$$H_{rej}: \mathcal{L}_{\bullet c} \rightarrow \{\mathcal{L}_{hc}, \underline{1}\}, \mathbf{u}(\mathbf{x}) \mapsto \mathbf{h}(\mathbf{x}) = \begin{cases} {}^t(1, 0, \dots, 0) & \text{if } \Phi(\mathbf{u}) \leq t \\ \underline{1} & \text{otherwise} \end{cases}$$

As already mentioned, such rules only allow to either exclusively classify an incoming pattern $\mathbf{x}$ or to reject it between all the c classes at hand, and do not allow to select a subset of several classes the patterns have to be associated with or even none (distance rejection), see Figure 5 (center). Table 5 reports the classification results obtained applying H_{rej} to the typical soft label vectors $\mathbf{u} \in \mathcal{L}_{pc}$ for which the (unfortunately?) expected results are:

- rejection of the patterns for which $\mathbf{u} = {}^t(0.70 \ 0.10 \ 0.85 \ 0.80)$, $\mathbf{u} = {}^t(0.20 \ 0.10 \ 0.85 \ 0.80)$ and $\mathbf{u} = {}^t(0.20 \ 0.10 \ 0.05 \ 0.15)$,
- exclusive classification of $\mathbf{x}$ in $\{\omega_3\}$ when $\mathbf{u} = {}^t(0.20 \ 0.10 \ 0.85 \ 0.15)$.

One can see that the rules based on the $\Phi_{\perp, \top}(\mathbf{u})$ family succeed in recovering the expected results, whatever the triple $(\top, N, \perp)$, for any $t > 0.5$. Not surprisingly, the rules based on the $\Phi_{\top, \top}(\mathbf{u})$ family fail most of time. The reasons have been discussed in the previous section, *e.g.* the normalization problem. For instance, if you take $\frac{\mathbf{u}}{\sum_{i=1}^c u_i(\mathbf{x})}$ instead of $\mathbf{u}$, $\Phi_{\top, \top}(\mathbf{u}) = \Phi_{Ch}(\mathbf{u})$ values are respectively: 0.65, 0.56, 0.34 and 0.60, so the expected results are obtained using the corresponding rule with $t > 0.34$.

4. Experimental Results

In this section, we report experiments carried out on both artificial and real benchmark data sets for which it is beneficial to use class-selective rejection rules or rejection rules because the classes overlap in the feature space. The different rules presented in section 3 are compared one to each other.

4.1 Data Sets and Protocol

Table 6 reports the characteristics (number n of patterns, number p of features, number c of classes, degree of overlap) of artificial and real data sets:

- synthetic datasets:
 - D contains 2000 points drawn from a mixture of $c = 2$ normal 7-dimensional distributions of 1000 points each with means $\mathbf{v}_1 = {}^t(1 \ 0 \ \dots \ 0)$ and $\mathbf{v}_2 = {}^t(-1 \ 0 \ \dots \ 0)$, and equal covariance matrices $\Sigma_1 = \Sigma_2 = I$,
 - D_H consists of two overlapping gaussian classes with different covariance matrices according to the Highleyman distribution, each composed of 800 observations in $\mathbb{R}^2$ (Highleyman, 1962), see Figure 8.
- real datasets from the UCI Machine Learning Repository (Blake & Merz, 1998).

The classification performance of the different rules are obtained by a 10-fold crossvalidation procedure on the different datasets. Each set of samples is divided at random into 10 approximately equal size and roughly balanced parts, ensuring that the classes are distributed proportionally among each of the 10 parts. The class-parameters are estimated on 90% of the samples (learning set χ) so that the degrees of typicality (1) of the remaining 10% test samples are computed and then their hard label vectors $\mathbf{h}(\mathbf{x})$ are predicted by each rule. This procedure is repeated 10 times, with each part playing the role of the test samples and the errors on all 10 parts added together to compute the overall error and therefore the performance.

In all cases, the threshold t is set to reject 10% of the data, so that 90% is the best achievable correct classification rate and then the error rate is $(90 - \text{correct})\%$. Note that there are no points in the datasets which are identified as outliers even if they lie far from the classes they are assumed to come from, so that some of them can be uncorrectly distance rejected.

4.2 Comparative Performance

Tables 7 and 8 show the classification performance obtained using the presented class-selective rejection rules H_{sel} and rejection rules H_{rej} based on ambiguity measures that aggregate pattern soft labels, namely the $\Phi_{\bar{1},\top}$ and $\Phi_{\perp,\top}$ families on one hand, and Φ_{H_0} , Φ_{FD} and the $\Phi_{k,\top}$ family on the other hand. For both types, the best score is indicated in bold and the second best score is italicized. One can immediately see that Φ_{FD} is the overall best measure, for all the considered datasets, so it seems more interesting to analyze the results by setting apart it.

Globally, class-selective rules perform better than rejection rules, so the $\Phi_{k,\top}$ family gives better results than the two other families. Also the $\Phi_{\perp,\top}$ family slightly outperforms the $\Phi_{\bar{1},\top}$ one. However, the $\Phi_{k,\top}$ family appears to be more sensitive to $(\top, \perp)$ than the two others, so its performance variation is higher, especially for some datasets having a number of classes greater than two (e.g. *Thyroid* and *Glass*). A positive consequence is its potential good performance compared to other families and the Horiuchi's rule, probably because more than two classes are overlapping so that the interaction between membership degrees is more important. A negative one is the need for the practitioner to carefully choose $(\top, \perp)$. About this choice, the particularly good results obtained using the Dombi norms with $\gamma = 2$ must be emphasized for the $\Phi_{k,\top}$ family compared to Horiuchi's rule and Ha's rule for some datasets (e.g. D , D_H and *Thyroid*), as well as compared to the other norms within the $\Phi_{\bar{1},\top}$ and $\Phi_{\perp,\top}$

Fig. 8. Highleyman data set D_H .

families. Note finally that, keeping in mind that these two families are less sensitive to $(\top, \perp)$ and the already mentioned asymptotic behaviour of Dombi norms, the Standard norms (so Chow's rule for $\Phi_{\top, \top}$) gives also quite good results for most of the datasets.

5. Conclusion

The problem of aggregating collections of numerical data to obtain a typical value is present in many decision systems. Aggregation operators are used to obtain an overall value for each alternative, which is exploited to establish a final decision. In the context of supervised pattern classification, such a decision consists in assigning objects (or patterns) to one class based on the aggregation of soft labels related to the given classes (posterior probabilities, fuzzy membership degrees, typicality degrees). It is well known that overlapping classes and outliers can significantly decrease a classifier performance and it has been proved that the misclassification risk can significantly be reduced by allowing a classifier to reject extraneous and/or ambiguous patterns (Dubuisson & Masson, 1993), (Tax & Duin, 2008). This results in designing classification rules that allow to assign a pattern to zero (distance rejection), one (exclusive classification) or several (ambiguity rejection) classes, in other words to select a number of classes.

This chapter addressed the problem of designing such rules that use ambiguity measures to aggregate pattern soft labels. The contribution is two-fold. An unified view of the resulting classifiers which can be either class-selective rules or simply rejection rules, depending on the ambiguity measure. Three families of ambiguity measures, based on combination of basic triangular norms and conorms as well as parametric ones, are presented. They allow to derive as many rules as many triangular norms (an infinite number!) and it is shown that they

generalize some historical rules, namely Chow's rejection rule (Chow, 1957 & 1970) and Ha's class-selective rule (Ha, 1996 & 1997). An analysis of all ambiguity measures is provided and the classification performance of all the rules on synthetic and real data sets from the public domain of various characteristics (dimensionality, number of classes, degree of overlap) is given.

From the practitioner's point of view, the choice of a particular triangular couple and/or its parameter value needs further investigation because it influences the behaviour of the ambiguity measure, hence the classification performance. For instance, the threshold on the ambiguity measure is easier to tune using a cautious couple (*e.g.* Standard or Dombi family) than a quite drastic one (*e.g.* Lukaziewicz). A future work will consist in studying more extensively how the mathematical properties of the basic norms affect the behaviour of the families of ambiguity measures and therefore the classification performances, in order to provide guidelines to choose them according to specific operational situations. Another perspective concerns the definition of a family of ambiguity measures based on an new aggregation operator which allows to measure to what extent an exact number of soft labels, but not necessarily the highest ones, are similar. A first proposition can be found in (Le Capitaine & Frélicot, 2008).

IntechOpen

$\mathbf{u} = {}^t(0.70\ 0.10\ 0.85\ 0.80)$	$k =$	0	1	2	3	4	$\forall t \in$
$\Phi_{k,\top}(\mathbf{u})$	$\Phi_{H_0}(\mathbf{u})$	0.85	0.95	0.90	0.40	0.90	$[0.40, 0.90[$
	$\Phi_{FD}(\mathbf{u})$	0.85	0.94	0.87	0.14	0	$[0.14, 0.87[$
	$\top_S \Rightarrow \Phi_{Ha}(\mathbf{u})$	0.85	0.80	0.70	0.10	0	$[0.10, 0.70[$
	$\top_A$	0.99	0.86	0.39	0.02	0	$[0.02, 0.39[$
	$\top_L$	1	1	0.65	0	0	$[0, 0.65[$
	$\top_{H_0}$	0.92	0.64	0.37	0.08	0	$[0.08, 0.37[$
	$\top_{H_2}$	0.99	0.92	0.39	0.00	0	$[0, 0.39[$
	$\top_{D_{1.5}}$	0.89	0.70	0.51	0.09	0	$[0.09, 0.51[$
	$\top_{D_2}$	0.87	0.73	0.58	0.09	0	$[0.09, 0.58[$
$\mathbf{u} = {}^t(0.20\ 0.10\ 0.85\ 0.80)$	$k =$	0	1	2	3	4	$\forall t \in$
$\Phi_{k,\top}(\mathbf{u})$	$\Phi_{H_0}(\mathbf{u})$	0.85	0.95	0.40	0.90	0.90	$[0.40, 0.85[$
	$\Phi_{FD}(\mathbf{u})$	0.85	0.94	0.25	0.50	0	$[0.25, 0.85[$
	$\top_S \Rightarrow \Phi_{Ha}(\mathbf{u})$	0.85	0.80	0.20	0.10	0	$[0.20, 0.80[$
	$\top_A$	0.97	0.70	0.10	0.00	0	$[0.10, 0.70[$
	$\top_L$	1	1	0.15	0	0	$[0.15, 1[$
	$\top_{H_0}$	0.90	0.58	0.18	0.05	0	$[0.18, 0.58[$
	$\top_{H_2}$	0.99	0.76	0.07	0.00	0	$[0.07, 0.76[$
	$\top_{D_{1.5}}$	0.88	0.67	0.21	0.07	0	$[0.21, 0.67[$
	$\top_{D_2}$	0.87	0.71	0.21	0.08	0	$[0.21, 0.71[$
$\mathbf{u} = {}^t(0.20\ 0.10\ 0.85\ 0.15)$	$k =$	0	1	2	3	4	$\forall t \in$
$\Phi_{k,\top}(\mathbf{u})$	$\Phi_{H_0}(\mathbf{u})$	0.85	0.35	0.95	0.95	0.90	$[0.35, 0.85[$
	$\Phi_{FD}(\mathbf{u})$	0.85	0.23	0.75	0.67	0	$[0.23, 0.85[$
	$\top_S \Rightarrow \Phi_{Ha}(\mathbf{u})$	0.85	0.20	0.15	0.10	0	$[0.20, 0.85[$
	$\top_A$	0.90	0.24	0.00	0.00	0	$[0.24, 0.90[$
	$\top_L$	1	0.45	0	0	0	$[0.45, 1[$
	$\top_{H_0}$	0.86	0.28	0.08	0.03	0	$[0.28, 0.86[$
	$\top_{H_2}$	0.93	0.25	0.00	0.00	0	$[0.25, 0.93[$
	$\top_{D_{1.5}}$	0.85	0.26	0.10	0.05	0	$[0.26, 0.85[$
	$\top_{D_2}$	0.85	0.24	0.11	0.07	0	$[0.24, 0.85[$
$\mathbf{u} = {}^t(0.20\ 0.10\ 0.05\ 0.15)$	$k =$	0	1	2	3	4	$\forall t \in$
$\Phi_{k,\top}(\mathbf{u})$	$\Phi_{H_0}(\mathbf{u})$	0.20	0.95	0.95	0.95	0	$[0.20, 1[$
	$\Phi_{FD}(\mathbf{u})$	0.20	0.75	0.67	0.5	0	$[0.20, 1[$
	$\top_S \Rightarrow \Phi_{Ha}(\mathbf{u})$	0.20	0.15	0.10	0.05	0	$[0.20, 1[$
	$\top_A$	0.41	0.00	0.00	0.00	0	$[0.41, 1[$
	$\top_L$	0.50	0	0	0	0	$[0.50, 1[$
	$\top_{H_0}$	0.37	0.08	0.03	0.01	0	$[0.37, 1[$
	$\top_{H_2}$	0.46	0.00	0.00	0.00	0	$[0.46, 1[$
	$\top_{D_{1.5}}$	0.28	0.10	0.05	0.02	0	$[0.28, 1[$
	$\top_{D_2}$	0.24	0.11	0.06	0.03	0	$[0.24, 1[$

Table 3. Ambiguity measures derived from class-selective rules for different $\mathbf{u}$ for which $n^*(\mathbf{x}, t) = 3, 2, 1, 0$ respectively (from top to bottom), $\forall t$ in the specified interval.

Rejection rule	Ambiguity Measure Φ
(Chow, 1970)	$\Phi_{Ch}(\mathbf{u}) = 1 - u_{(1)}(\mathbf{x}) \quad (9)$ $= \Phi_{\bar{1}, \top_S}(\mathbf{u})$
(Mascarilla & Frélicot, 2001)	$\Phi_{\perp, \top}(\mathbf{u}) = 1 - \underline{\perp}(\mathbf{u}) \quad (7)$
(Frélicot & Le Capitaine, 2009, this book)	$\Phi_{\bar{1}, \top}(\mathbf{u}) = 1 - \perp(\mathbf{u}) \quad (5)$

Table 4. Rejection rules and corresponding ambiguity measures.

$\mathbf{u} = {}^t(0.70 \ 0.10 \ 0.85 \ 0.80)$			$\mathbf{u} = {}^t(0.20 \ 0.10 \ 0.85 \ 0.80)$		
$\Phi_{\bar{1}, \top}(\mathbf{u})$	$\top_S \Rightarrow \Phi_{Ch}(\mathbf{u})$	0.15	$\Phi_{\bar{1}, \top}(\mathbf{u})$	$\top_S \Rightarrow \Phi_{Ch}(\mathbf{u})$	0.15
	$\top_A$	0.00		$\top_A$	0.02
	$\top_L$	0		$\top_L$	0
	$\top_{H_0}$	0.07		$\top_{H_0}$	0.09
	$\top_{H_2}$	0.00		$\top_{H_2}$	0.01
	$\top_{D_{1.5}}$	0.10		$\top_{D_{1.5}}$	0.11
	$\top_{D_2}$	0.12		$\top_{D_2}$	0.13
$\Phi_{\perp, \top}(\mathbf{u})$	$\top_S$	0.94	$\Phi_{\perp, \top}(\mathbf{u})$	$\top_S$	0.94
	$\top_A$	0.89		$\top_A$	0.75
	$\top_L$	1		$\top_L$	1
	$\top_{H_0}$	0.75		$\top_{H_0}$	0.70
	$\top_{H_2}$	0.94		$\top_{H_2}$	0.79
	$\top_{D_{1.5}}$	0.81		$\top_{D_{1.5}}$	0.78
	$\top_{D_2}$	0.86		$\top_{D_2}$	0.84
$\mathbf{u} = {}^t(0.20 \ 0.10 \ 0.85 \ 0.15)$			$\mathbf{u} = {}^t(0.20 \ 0.10 \ 0.05 \ 0.15)$		
$\Phi_{\bar{1}, \top}(\mathbf{u})$	$\top_S \Rightarrow \Phi_{Ch}(\mathbf{u})$	0.15	$\Phi_{\bar{1}, \top}(\mathbf{u})$	$\top_S \Rightarrow \Phi_{Ch}(\mathbf{u})$	0.80
	$\top_A$	0.09		$\top_A$	0.58
	$\top_L$	0		$\top_L$	0.5
	$\top_{H_0}$	0.13		$\top_{H_0}$	0.62
	$\top_{H_2}$	0.06		$\top_{H_2}$	0.53
	$\top_{D_{1.5}}$	0.15		$\top_{D_{1.5}}$	0.71
	$\top_{D_2}$	0.15		$\top_{D_2}$	0.75
$\Phi_{\perp, \top}(\mathbf{u})$	$\top_S$	0.24	$\Phi_{\perp, \top}(\mathbf{u})$	$\top_S$	0.80
	$\top_A$	0.37		$\top_A$	0.60
	$\top_L$	0.45		$\top_L$	0.50
	$\top_{H_0}$	0.41		$\top_{H_0}$	0.68
	$\top_{H_2}$	0.36		$\top_{H_2}$	0.55
	$\top_{D_{1.5}}$	0.34		$\top_{D_{1.5}}$	0.73
	$\top_{D_2}$	0.31		$\top_{D_2}$	0.77

Table 5. Ambiguity measures derived from rejection rules for different $\mathbf{u}$.

Data set	n	p	c	overlap
D	2000	7	2	slight
D_H	1600	2	2	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

Table 6. Datasets and their characteristics.

		(%)	D	D_H	$Digits$	$Thyroid$	$Pima$	$Statlog$	$Glass$
		$\Phi_{H_0}(\mathbf{u})$	77.95	83.88	88.76	86.98	62.89	77.68	68.22
		$\Phi_{FD}(\mathbf{u})$	79.05	86.75	89.13	87.91	63.28	77.48	71.03
$\Phi_{k,\top}(\mathbf{u})$	$\top_S \Rightarrow \Phi_{H_a}(\mathbf{u})$	78.15	86.63	87.38	86.98	60.55	75.56	69.63	
	$\top_A$	77.55	86.19	86.21	84.19	60.16	75.23	67.76	
	$\top_L$	77.55	86.50	86.60	84.38	60.16	75.29	67.76	
	$\top_{H_0}$	77.75	86.50	86.60	86.98	60.68	75.29	68.22	
	$\top_{H_2}$	77.55	86.06	86.21	84.19	60.16	75.23	67.76	
	$\top_{D_{1.5}}$	78.10	86.56	86.82	86.98	60.55	75.23	68.69	
	$\top_{D_2}$	78.20	86.63	86.96	87.44	60.55	75.15	68.69	

Table 7. Classification performance of class-selective rejection rules on datasets.

		(%)	D	D_H	$Digits$	$Thyroid$	$Pima$	$Statlog$	$Glass$
$\Phi_{\perp,\top}(\mathbf{u})$	$\top_S \Rightarrow \Phi_{Ch}(\mathbf{u})$	75.15	83.75	88.03	84.19	59.51	75.66	66.82	
	$\top_A$	75.55	83.13	87.55	85.12	60.81	74.73	65.89	
	$\top_L$	75.55	83.13	87.55	85.12	60.81	74.73	65.89	
	$\top_{H_0}$	75.55	83.13	87.55	85.12	60.81	74.73	65.89	
	$\top_{H_2}$	75.55	83.13	87.55	85.12	60.81	74.73	65.89	
	$\top_{D_{1.5}}$	75.65	83.13	87.76	85.12	61.07	74.97	66.36	
	$\top_{D_2}$	75.75	83.19	87.84	85.12	61.33	75.12	65.82	
$\Phi_{\perp,\top}(\mathbf{u})$	$\top_S$	75.90	83.44	88.04	85.12	61.59	75.66	66.82	
	$\top_A$	77.55	83.13	87.55	85.12	60.81	74.73	65.89	
	$\top_L$	77.55	83.13	87.55	85.12	60.81	74.73	65.89	
	$\top_{H_0}$	77.55	83.13	87.55	85.12	60.81	74.73	65.89	
	$\top_{H_2}$	77.55	83.13	87.55	85.12	60.81	74.73	65.89	
	$\top_{D_{1.5}}$	77.65	83.13	87.76	85.12	60.55	74.97	66.36	
	$\top_{D_2}$	77.75	83.19	87.84	85.12	61.33	75.12	66.82	

Table 8. Classification performance of rejection rules on datasets.

6. References

- Bezdek, J.C. & Keller, J.M. & Krishnapuram, R. & Pal, N.R. (1999). *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer Academic
- Blake, C. & Merz, C. (1998). UCI repository of machine learning databases, <http://www.ics.uci.edu/mllearn/MLRepository.html>
- Calvo, C. & Mayor, G. & Mesiar, R. (2002). *Aggregation Operators: New Trends and Applications*, Physica-Verlag
- Chow, C.K. (1957). An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, Vol. 6, No 4, 247-254
- Chow, C.K. (1970). On optimum error and reject tradeoff. *IEEE Transactions on Information Theory*, Vol. 16, No 1, 41-46
- Dubuisson, B. & Masson, M. (1993). A statistical decision rule with incomplete knowledge about classes. *Pattern Recognition*, Vol. 26, No 1, 155-165
- Frélicot, C. (1998). On unifying probabilistic/fuzzy and possibilistic rejection-based classifiers, In *Lecture Notes in Computer Science 1451: Advances in Pattern Recognition*, A. Amin & D. Dori & P. Pudil & H. Freeman (Eds), 736-745, Springer
- Frélicot, C. & Dubuisson, B. (1992). A multi-step predictor of membership function as an ambiguity reject solver in pattern recognition, *Proceedings of 4th International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems*, pp. 709-715, Palma de Mallorca, Spain
- Frélicot, C. & Mascarilla, L. (2002). Reject strategies driven combination of pattern classifiers. *Pattern Analysis and Applications*, Vol. 5, No 2, 234-243
- Grabisch, M. (1992). Fuzzy pattern recognition by fuzzy integrals and fuzzy rules, In: *Pattern Recognition - From Classical to Modern Approaches*, S. Pal and P. Pal (Eds), 257-280, World Scientific
- Grabisch, M. & Marichal, J. & Mesiar, R. & Pap, E. (2009). *Aggregation Functions*, Cambridge University Press, No 127 in *Encyclopedia of Mathematics and its Applications*
- Ha, T. (1996). An optimum class-selective rejection rule for pattern recognition, *Proceedings of 13th International Conference on Pattern Recognition*, Vol. 2, pp. 75-80, Vienna, Austria
- Ha, T. (1997). The optimum class-selective rejection rule. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 6, 608-615
- Highleyman, W. (1962). Linear decision functions, with application to pattern recognition, *Proceedings of the IRE*, Vol. 50, No 6, 1501-1514
- Horiuchi, T. (1998). Class-selective rejection rule to minimize the maximum distance between selected classes. *Pattern Recognition*, Vol. 31, No 10, 1579-1588
- Klement, E. & Mesiar, R. (2005). *Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms*, Elsevier
- Le Capitaine, H. & Frélicot, C. (2008). A class-selective rejection scheme based on blockwise similarity of typicality degrees, *Proceedings of 19th IAPR International Conference on Pattern Recognition*, Tampa, Florida
- Mascarilla, L. & Berthier, M. & Frélicot, C. (2008). A k-order fuzzy or operator for pattern classification with k-order ambiguity rejection. *Fuzzy Sets and Systems*, Vol. 159, No 15, 2011-2029
- Mascarilla, L. & Frélicot, C. (2001). A class of reject-first possibilistic classifiers based on dual triples, *Proceedings of 9th International Fuzzy Systems Association Worldcongress*, pp. 743-747, Vancouver, Canada

- Menger, K (1942). Statistical metrics, Proceedings of the National Academy of Science USA, Vol. 28, No 12, pp. 535-537
- Sugeno, M. (1974). Theory of fuzzy integrals and its applications, PhD thesis of the Tokyo Institute of Technology
- Tax, D. & Duin, R. (2008). Growing a multi-class classifier with a reject option. Pattern Recognition Letters, Vol. 29, No 10, 1565-1570
- Yager, R.R. (1988). Ordered weighted averaging aggregation operators in multi-criteria decision making. IEEE Transactions on Systems, Man, and Cybernetics, Vol. 18, No 1, 183-186
- Zimmermann, H.J. & Zysno, P. (1985). Quantifying vagueness in decision models. European Journal of Operational Research, Vol. 22, No 2, 148-158

IntechOpen

Pattern Recognition Recent Advances

Edited by Adam Herout

ISBN 978-953-7619-90-9

Hard cover, 524 pages

Publisher InTech

Published online 01, February, 2010

Published in print edition February, 2010

Nos aute magna at aute doloreetum erostrud eugiam zzriuscipsum dolorper iliquate velit ad magna feugiamet, quat lore dolore modolor ipsum vullutat lorper sim inci blan vent utet, vero er sequatum delit lortion sequip eliquatet ilit aliquip eui blam, vel estrud modolor irit nostinc iliquiscinit er sum vero odip eros numsandre dolessisisim dolorem volupta tionsequam, sequamet, sequis nonnulla conulla feugiam euis ad tat. Igna feugiam et ametuercil enim dolore commy numsandiam, sed te con hendit iuscidunt wis nonse volenis molorer suscip er illan essit ea feugue do dunt utetum vercili quamcon ver sequat utem zzriure modiat. Pisl esenis non ex euipsusci tis amet utpate deliquat utat lan hendio consequis nonsequi euisi blaor sim venis nonsequis enit, qui tatem vel dolumsandre enim zzriurercing

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Carl Frelicot and Hoel Le Capitaine (2010). Class-Selective Rejection Rules Based on the Aggregation of Pattern Soft Labels, Pattern Recognition Recent Advances, Adam Herout (Ed.), ISBN: 978-953-7619-90-9, InTech, Available from: <http://www.intechopen.com/books/pattern-recognition-recent-advances/class-selective-rejection-rules-based-on-the-aggregation-of-pattern-soft-labels>

INTECH
open science | open minds

InTech Europe

University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China

Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

© 2010 The Author(s). Licensee IntechOpen. This chapter is distributed under the terms of the [Creative Commons Attribution-NonCommercial-ShareAlike-3.0 License](https://creativecommons.org/licenses/by-nc-sa/3.0/), which permits use, distribution and reproduction for non-commercial purposes, provided the original is properly cited and derivative works building on this content are distributed under the same license.

IntechOpen

IntechOpen