A family of measures for best top-n class-selective decision rules

Hoel Le Capitaine, Carl Frelicot

A family of measures for best top-n class-selective decision rules[☆]

Hoel Le Capitaine[a,*], Carl Frélicot[a]

[a] *Mathematics, Image and Applications Laboratory, University of La Rochelle, Av. M. Crépeau, 17042 La Rochelle, FRANCE*

Abstract

When classes strongly overlap in the feature space, or when some classes are not known in advance, the performance of a classifier heavily decreases. To overcome this problem, the reject option has been introduced. It simply consists in withdrawing the decision, and let another classifier, or an expert, take the decision whenever exclusively classifying is not reliable enough. The classification problem is then a matter of class-selection, from none to all classes. In this paper, we propose a family of measures suitable to define such decision rules. It is based on a new family of operators that are able to detect blocks of similar values within a set of numbers in the unit interval, the soft labels of an incoming pattern to be classified, using a single threshold. Experiments on synthetic and real data sets available in the public domain show the efficiency of our approach.

Keywords: Reject options, Class-selective decision rules, Fuzzy aggregation operators.

1. Introduction

Supervised classifier design aims at defining rules that allow to associate an incoming object with one of a set of known classes, based on a training set of labeled objects, (see [2]). Such exclusive classification rules must be sufficiently reliable for real-world problems, *e.g.* in medical diagnosis or nuclear plant monitoring. In many applications, some objects to be classified fit several classes (inliers), are distributed around the classes or even arise from an unknown class (outliers), so the classifier performance can significantly be affected. It is more convenient to reject such samples, *i.e.* withhold making a decision and call for a exceptional handling (use of a different rule, different classifier, human inspection) than making a wrong assignment. Pattern rejection has been first formalized in the context of statistical pattern recognition under the minimum misclassification risk theory. Chow modified the Bayes decision rule in order to reject an object if its highest posterior probability is less than a threshold [3, 4]. Ha extended it to inliers that can be associated with a subset of the known classes [5, 6] leading to a *class-selective rejection rule* as opposed to Chow's *rejection rule*. In order to avoid unnatural decisions due to posterior probabilities normalization, Horiuchi proposed to test their differences [7] while Frélicot & Dubuisson had previously used their ratios [8]. This latter idea has been introduced again several times [9, 10]. In order to overcome difficulties due to errors in probability estimations, class-dependent rejection based on multiple thresholds can be introduced [11]. Tax & Duin have recently proposed two rescaling heuristics allowing to adjust the thresholds and combine one-class models that are based either on class-densities or on distances to prototypes [12]. Mascarilla et. al have defined a class of operators based on triangular norms that combine soft class-labels suitable to select several classes [13]. As said above, it may be required to reject an incoming object from all the known classes. Dubuisson & Masson [14] first proposed to threshold the mixture density in order to *distance reject* outliers and opposed this approach to the *ambiguity rejection* dedicated to inliers as tackled by Chow and following authors (Ha, Horiuchi). The use of class-dependent thresholds to distance reject outliers can be found in [15]. Many decision rules dealing with one kind of reject or both have been proposed so far, in other theoretical frameworks, *e.g.*: neural networks [16], genetic algorithms [17], support vector machines [18, 19], fuzzy inference systems [20, 21], theory of evidence [22], but we restrict to the statistical one. Let us precise the difference authors make between a *classification method* and a *decision rule*. A classification method is a two-step procedure as illustrated in Figure 1. The first step provides soft labels to classes, *e.g.* posterior probabilities for the Bayes classifier, while the second one deals with the decision, *i.e.* associates a (set of) class(es) to the incoming pattern by means of a rule applied on the soft class-labels, *e.g.* the MAP (*Maximum A Posteriori*) rule for the Bayes classifier. Including reject options consists in defining an appropriate decision rule such that an outlier (respectively an inlier) can be associated to the empty set (respectively a subset) of classes. This paper addresses this particular problem. Therefore, the family of class-selective decision rules we propose can be included in any classification method.

In order to illustrate the different class-selective schemes, Figure 2 shows how some typical decision rules among the cited ones partition the pattern space into decision regions for a $c = 3$

[☆] This article is a widely extended version of a paper presented at the 19th International Conference on Pattern Recognition by the authors [1].

[*] Corresponding author

Email addresses: hoel.le_capitaine@univ-lr.fr (Hoel Le Capitaine), carl.frelicot@univ-lr.fr (Carl Frélicot)

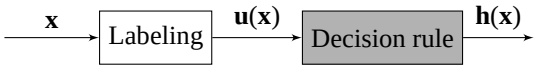
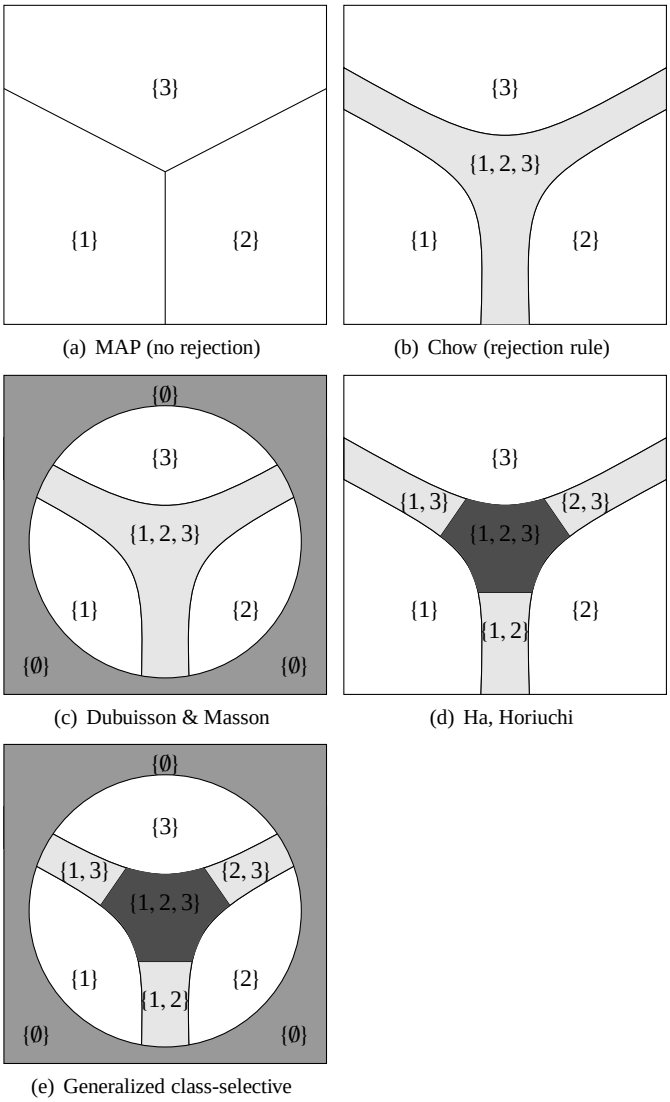


Figure 1: A classification method procedure composed of a labeling step and a decision rule.

classes problem. The MAP rule results in three regions, each of which corresponds to a single class (a). The outcomes of Chow's rejection rule are also singletons augmented by the entire set of classes $\{1, 2, 3\}$ which corresponds to the *total* ambiguity reject region (b), to which Dubuisson & Masson add the one of no class assignment $\{0\}$ dedicated to distance rejection (c). A general class-selective rejection rule results in 2^c regions (e) corresponding to all possible combination of classes, including no class assignment. This latter option is not allowed by earliest rules proposed by Ha, Horiuchi and Frélicot & Dubuisson (d).



(a) MAP (no rejection) (b) Chow (rejection rule)

(c) Dubuisson & Masson (d) Ha, Horiuchi

(e) Generalized class-selective

Figure 2: Feature space partitioning for a three-class problem and decision rules with reject options.

A short state-of-the-art in statistical pattern recognition with reject options emphasizes the following different issues related to the design of decision rules:

- which kind of rejection should be preferred?
  → it may depend on the application, however a class-selective rule should be able to cope with both types of rejection at the same time, as shown in Figure 2-(e),
- which model should be used (density-based or distance-based)?
  → to correctly distinguish whether an object has to be distance or ambiguity rejected, distance-based models are often easier to manage because they allow to compute non-normalized soft labels whereas posterior probabilities or fuzzy membership degrees have the constraint to sum up to one,
- how many thresholds are required?
  → the less the number of thresholds, the better it is from a statistical point of view as well as for the practitioner who faces their tuning, except if it allows to increase the classifier performance according to the application.

In this paper, we propose a family of block-similarity measures of class-labels and derive a general class-selective rule which allows, in one single step, either to distance reject (none class is selected), or to classify (one selected class), or to ambiguity reject (two up to all selected classes) a pattern, given only one user-specified threshold.

In Section 2 we start by discussing the standard approaches of rejecting objects and the necessary background on aggregation operators to settle it in terms of soft labels aggregation. Next, we recall some measures suitable for class-selection based on combination of basic aggregation operators. The new family of class-selective decision rules is presented in Section 3. It is based on an operator which detects the different blocks of similar values within a tuple of values in $[0, 1]$. Thus, the block of a number (from 0 to c) similar largest labels gives the classes to which the pattern has to be assigned to. We discuss its properties and illustrate its behavior on simple examples. Experimental results on several synthetic and real data sets from the public domain that show its efficiency and concluding remarks as well as perspectives remarks are given in Section 4 and 5, respectively.

2. Preliminaries

2.1. Pattern classification and reject options

Let us consider an object described by p features, namely a vector $\mathbf{x}$ in a p-dimensional real space, to be classified in a $c-$classes problem. Let $C^+ = \{1, ..., c\}$ be the set of class indexes, a classification rule for $\mathbf{x}$ supposes soft labels to the classes $u_i(\mathbf{x})$, $i \in C^+$, to be available [23], namely a c-dimensional vector $\mathbf{u}(\mathbf{x})$. For density-based models, soft labels $u_i(\mathbf{x})$ are posterior probabilities $P(i|\mathbf{x})$ computed through the Bayes formula, using class-conditional densities and prior probabilities

estimated on a training set $\aleph$ of labeled objects. For distance-based models, one can use soft labels computed as functions of distances to class-prototypes, *e.g.* a Cauchy-type function:

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

where α_i is a parameter controlling the membership, which can be user-defined or learned from $\aleph$, d is a distance in $\mathbb{R}^p$, and $\mathbf{v}_i$ is a prototype of the i-th class. Among the possible distances, one finds the squared Mahalanobis distance $d^2(\mathbf{x}, \mathbf{v}_i) = {}^t(\mathbf{x} - \mathbf{v}_i)\Sigma_i^{-1}(\mathbf{x} - \mathbf{v}_i)$ where $\mathbf{v}_i$ and Σ_i are the i-th class mean vector and covariance matrix estimated on $\aleph$. It has been shown through empirical studies that (1) is a good model for vague concepts or classes [24]. Recall that we do not address the problem of obtaining reliable soft labels but the problem of defining decision rules based on the labels, see Figure 1. Consequently, any soft labels (constrained or not) provided by a classifier can be used in the proposed framework.

A soft label vector $\mathbf{u}(\mathbf{x})$ takes values in $\mathcal{L}_{pc} = [0,1]^c$ if it is distance-based or in $\mathcal{L}_{fc} = \left\{\mathbf{u}(\mathbf{x}) \in \mathcal{L}_{pc} | \sum_{i=1}^{c} u_i(\mathbf{x}) = 1\right\}$ if its components are posterior probabilities or fuzzy membership functions. Then, a decision rule is defined as a mapping D: $\mathcal{L}_{\bullet c} \rightarrow \mathcal{L}_{hc}$, $\mathbf{u}(\mathbf{x}) \mapsto \mathbf{h}(\mathbf{x})$ where $\mathcal{L}_{hc} = \{\mathbf{h}(\mathbf{x}) \in \mathcal{L}_{fc} | h_i(\mathbf{x}) \in \{0,1\}\}$. The MAP (*Maximum A Posteriori*) decision rule used by the probabilistic Bayes classifier is defined by $D_B : \mathcal{L}_{fc} \rightarrow \mathcal{L}_{hc}$, $\mathbf{u}(\mathbf{x}) \mapsto \mathbf{h}(\mathbf{x})$ such that

$$h_i(\mathbf{x}) = 1, \quad i = \text{argmax}_{j=1,c}\, u_j(\mathbf{x}). \tag{2}$$

Chow's rejection rule 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 trade off [4] and is defined by $D_{Ch} : \mathcal{L}_{fc} \rightarrow \{\mathcal{L}_{hc}, \mathbf{1}\}$, $\mathbf{u}(\mathbf{x}) \mapsto \mathbf{h}(\mathbf{x})$ such that

$$\begin{cases} h_i(\mathbf{x}) = 1, \quad i = \text{argmax}_{j=1,c}\, u_j(\mathbf{x}) & \text{if } u_i(\mathbf{x}) > (1-t) \\ \mathbf{h}(\mathbf{x}) = \mathbf{1} & \text{otherwise} \end{cases} \tag{3}$$

where $\mathbf{1}$ is the p-dimensional one vector ${}^t(1, ..., 1)$, meaning that $\mathbf{x}$ is ambiguity rejected between all known classes if its largest posterior probability is lower than $1 - t$, as shown in Figure 2-(b). Since $\mathbf{u}(\mathbf{x}) \in \mathcal{L}_{fc}$, it is easy to show that D_{Ch} is identical to D_B whenever $t > \frac{c-1}{c}$, *i.e.* $\mathbf{x}$ cannot be rejected [3]. Since the work by Chow, most of authors intended to avoid total ambiguity rejection whenever a number between 2 and $c - 1$ classes have to be selected. Thus, a class-selective procedure can be defined, in its general form, as the seek for the best top-n classes according to:

$$n^\star(\mathbf{x}, t) = \min_{k \in C^+} \{k : \Phi(\mathbf{u}(\mathbf{x})) \leq t\} \tag{4}$$

with the convention: if $\Phi(\mathbf{u}(\mathbf{x})) > t$ for all $k \in C^+$, then $n^\star(\mathbf{x}, t)$ is set to c. In (4), Φ is a *selection measure* on object's labels $\mathbf{u}$: $\mathcal{L}_{\bullet c} \rightarrow [0,1]$, and $n^\star(\mathbf{x}, t)$ is the number of selected classes for $\mathbf{x}$, given a user-defined threshold t. Such rules partition the feature space as shown in Figure 2-(d) and can be defined by $D_{sel} : \mathcal{L}_{\bullet c} \rightarrow \mathcal{L}_{hc}^{c,+}$, $\mathbf{u}(\mathbf{x}) \mapsto \mathbf{h}(\mathbf{x})$ such that

$$h_i(\mathbf{x}) = 1, \quad \forall i :\ 1 \leq (i) \leq n^\star(\mathbf{x}, t) \tag{5}$$

where $\mathcal{L}_{hc}^{c,+} = \{0,1\}^c \backslash \mathbf{0}$, the set of vertices of the unit hypercube $[0,1]^c$ without the origin $\mathbf{0} = {}^t(0, ..., 0)$, and (i) is the permutation of the class index i such that $u_{(i)}$ is the i-th largest value in $\mathbf{u}$. We can say that all rejection rules are the restriction of (4) to $k \in \{1, c\}$ and the measure used by Chow is $\Phi_{Ch}(\mathbf{u}) = 1 - u_{(1)}(\mathbf{x})$. Propositions from the literature mainly consist in defining new selection measures Φ suitable for class-selective rejection instead of total ambiguity rejection. In [6], the posterior probabilities are ranked in decreasing order and their values are tested up to the $(k+1)$-th to decide if k classes are selected, so the selection measure is $\Phi_{Ha}(\mathbf{u}) = u_{(k+1)}(\mathbf{x})$ with the convention $u_{(c+1)}(\mathbf{x}) = 0$. The corresponding rule D_{Ha} by Ha minimizes the error probability for a given average number of classes and the domain of t is $[0, \frac{1}{2}]$. Whenever $t > \frac{1}{2}$, $n^\star(\mathbf{x}, t) = 1$ so D_{Ha} reduces to D_B, and can increase up to c as t decreases down to zero, (see [5]). Because of the normalization constraint on $\mathbf{u}$ ($\in \mathcal{L}_{fc}$), Ha's rule can lead to counterintuitive results, so Horiuchi proposed to test the difference between successive ordered posterior probabilities [7]. Using our conventions, we can rewrite the selection measure as $\Phi_{Ho}(\mathbf{u}) = 1 - (u_{(k)}(\mathbf{x}) - u_{(k+1)}(\mathbf{x}))$ and Horiuchi's rule D_{Ho} minimizes the maximum distance between selected classes for a given average number of selected classes which is specified by $t \in [0,1]$. It is identical to D_B, so $n^\star(\mathbf{x}, t) = 1$ when $t = 0$ and increases up to c as t increases up to 1. Given $u_\star = \max_{\mathbf{y} \in \aleph} u_{(1)}(\mathbf{y})$, $\Phi_{Ds}(\mathbf{u}) = u_{(1)}(\mathbf{x})/u_\star$ is used in [25]. In [8], it has been proposed to replace posterior probabilities by soft labels, for instance as defined by (1), so $\mathbf{u} \in \mathcal{L}_{pc}$ instead of $\mathcal{L}_{fc}$, and take, given $t \in [0,1]$, $\Phi_{FD}(\mathbf{u}) = u_{(2)}(\mathbf{x})/u_{(1)}(\mathbf{x})$ as the selection measure so that (4) is restricted to $k \in \{1, c\}$. Note that the same measure has been reintroduced with posterior probabilities by several authors [9, 10], referred to as a reliability measure $\psi_b = 1 - \Phi_{FD}(\mathbf{u})$. Naturally, as defined in [8], this measure does not minimize the expected Bayesian risk anymore. The measure $\Phi_{FD}(\mathbf{u})$ has been extended to $k \in C^+$ in [26, 27] by $\Phi_{FM}(\mathbf{u}) = u_{(k+1)}(\mathbf{x})/u_{(1)}(\mathbf{x})$ and its more general form $\Phi_{FL}(\mathbf{u}) = u_{(k+1)}(\mathbf{x})/u_{(k)}(\mathbf{x})$ can be found in [28]. The resulting rule D_{FL} reduces to D_B when $t = 0$, and $n^\star(\mathbf{x}, t)$ increases from 1 up to c as t increases from 0 up to 1.

Recently, a family of operators using a combination of dual triangular norms and conorms $(\top, \perp)$ has been proposed in [13]. Denoted by $\overset{k}{\perp}(\mathbf{u})$, these operators generalize the $k - th$ largest value in $\mathbf{u}$ in the sense that it is exactly $u_{(k)}(\mathbf{x})$ when using the standard norms: $\top = \min$ the largest t-norm and $\perp = \max$ the smallest t-conorm (see section below for details). The authors showed in [28] that Ha's measure is the particular case of $\Phi_{MBF}(\mathbf{u}) = \overset{k+1}{\perp}(\mathbf{u})$ when using the standard norms. In the same paper, they introduce another selection measure $\Phi_{\underline{1}}(\mathbf{u}) = 1 - \perp(\mathbf{u})$ which generalizes Chow's one.

Earliest rules that include both distance and ambiguity reject options involve a pre-step procedure dedicated to the former case thanks to an additional measure (and a corresponding threshold), here called an *acceptation measure* Ψ. It can be either a function of $\mathbf{x}$ as by Dubuisson & Masson [14]: $\Psi(\mathbf{x}) = f(\mathbf{x})$ the mixture density for $\mathbf{x}$, or a function of $\mathbf{u}$ as in [25]: $\Psi(\mathbf{u}(\mathbf{x})) = 1 - u_{(1)}(\mathbf{x})$. Even Chow's rejection rule and class-

selective rules as defined by Ha, Horiuchi, and Frélicot & Dubuisson/Le Capitaine can be modified to include such a measure. In order to accept a pattern, the acceptation measure Ψ must be larger than a threshold. Here, we propose the measure

$$\Psi(\mathbf{u}(\mathbf{x})) = u_{(1)}(\mathbf{x}). \tag{6}$$

The pattern $\mathbf{x}$ is distance rejected if $\Psi(\mathbf{u}(\mathbf{x})) \leq t$, *i.e.* it is accepted if $\Psi(\mathbf{u}(\mathbf{x})) > t$. We claim that a class-selective rule including both options must be able to partition the feature space as shown in Figure 2-(e). In other words, it must map the entire set of vertices $\mathcal{L}^c = \{0,1\}^c$. Therefore, the selection measure requires k not to restrict in C^+ but to include the zero class-selection, *i.e.* $k \in C = \{0, C^+\}$. We express such a rule as $D_{sel} : \mathcal{L}_{\bullet c} \rightarrow \mathcal{L}^c_{hc}$, $\mathbf{u}(\mathbf{x}) \mapsto \mathbf{h}(\mathbf{x})$ such that

$$h_i(\mathbf{x}) = 1, \quad \forall i : \ 0 \leq (i) \leq n^\star(\mathbf{x}, t) \tag{7}$$

where the number $n^\star(\mathbf{x}, t)$ of classes to be selected for $\mathbf{x}$, given a threshold t, is

$$n^\star(\mathbf{x}, t) = \min_{k \in C}\{k : \Phi_{k+1}(\mathbf{u}(\mathbf{x})) \leq t\} \tag{8}$$

with the conventions: $u_{(0)}(\mathbf{x})) = 1$ and $u_{(c+1)}(\mathbf{x}) = 0$. The selection measure $\Phi_{k+1}(\mathbf{u}(\mathbf{x}))$ can be obtained by aggregating the components of $\mathbf{u}$ in a suitable way. Note that for $k = 0$ (*i.e.* distance rejection), Φ_{Ha} and Φ_{Ho} are equivalent, as given in (6).

2.2. Aggregation operators

Aggregation operators aim at associating a typical value to a number of several numerical values which are generally defined on a finite real interval or on ordinal scales and many families of such functions are available, *e.g.*: triangular norms [29], OWA (Ordered Weighted Averaging) operators [30], γ-operators [31], fuzzy integrals [32], (see [33, 34, 35] for a survey). They are used in many fields, *e.g.* decision-making and pattern recognition. They are generally classified either by some mathematical properties they share (symmetry, associativity, monotonicity and so on) or by the way the values are aggregated (conjunctive, disjunctive, compensatory, and so on). The triangular norms are of special interest because of their ability to generalize the logical AND and OR crisp operators to fuzzy sets, (see [36] for a survey). Briefly, a triangular norm (or t-norm) is a binary operation on the unit interval $\top : [0,1]^2 \rightarrow [0,1]$ which is commutative, associative, non decreasing and has 1 for neutral element. A t-norm $\top$ is conjunctive and the minimum operator $\wedge$ is the largest one. Alternatively, a triangular conorm (or t-conorm) is the dual binary operation $\perp : [0,1]^2 \rightarrow [0,1]$ having the same properties except the latter: its neutral element is 0. A t-conorm $\perp$ is disjunctive and the maximum operator $\vee$ is the smallest one. Basic dual couples $(\top,\perp)$ that will be used in the sequel are given in Table 1. Combining norm-couples $(\top, \perp)$ allows to define operators that can be used to solve the class-selection problem, as in [13]. Let $\mathcal{P}$ be the power set of C^+ and $\mathcal{P}_k = \{A \in \mathcal{P} : card(A) = k\}$. The *fuzzy k-order OR* operator (fOR-k for short) is a family of aggregation operator parametrized by $(\top, \perp)$: $[0,1]^c \rightarrow [0,1]$, $\mathbf{u} \mapsto \overset{k}{\perp}(\mathbf{u})$, where

$$\overset{k}{\perp}(\mathbf{u}) = \bigperp_{i=1,\cdots,c}^{k} u_i = \bigtop_{A \in \mathcal{P}_{k-1}} \left(\bigperp_{j \in C^+ \setminus A} u_j \right) \tag{9}$$

Table 1: Basic triangular norm couples

Standard	$\top_S(a,b) = \min(a,b)$
	$\perp_S(a,b) = \max(a,b)$
Product	$\top_P(a,b) = a\,b$
	$\perp_P(a,b) = a + b - a\,b$
Łukasiewicz	$\top_L(a,b) = \max(a+b-1, 0)$
	$\perp_L(a,b) = \min(a+b, 1)$

Some properties of fOR-k result from those of $\top$ and $\perp$, others have been proved in [13]. Among these properties, let us recall those that are useful for the context we are interested in:

- $\overset{k}{\perp}(\mathbf{0}) = 0$ and $\overset{k}{\perp}(\mathbf{1}) = 1$ (boundaries)
- for $\mathbf{u}$ and $\mathbf{v}$ such that $u_i \leq v_i, \forall i \in C^+$, $\overset{k}{\perp}(\mathbf{u}) \leq \overset{k}{\perp}(\mathbf{v})$ (monotonicity)
- for any permutation σ of C^+, $\bigperp_{i=1,\cdots,c}^{k} u_{\sigma(i)} = \bigperp_{i=1,\cdots,c}^{k} u_i$ (symmetry)
- $\overset{1}{\perp}(\mathbf{u}) = \perp(\mathbf{u})$ and $\overset{c}{\perp}(\mathbf{u}) = \top(\mathbf{u})$, whatever c and $(\top, \perp)$,
- if the standard norms are taken, then $\overset{k}{\perp_S}(\mathbf{u}) = u_{(k)}$, the k-th largest value in $\mathbf{u}$.

Example 1. *Let us consider $C^+ = \{1,2,3,4\}$ and $k = 3$ such that $\mathcal{P}_{k-1} = \mathcal{P}_2 = \{\{1,2\},\{1,3\},\{1,4\},\{2,3\},\{2,4\},\{3,4\}\}$. For each subset in $\mathcal{P}_2$, one can select its complement with respect to C^+ and obtain, given any $\mathbf{u} \in [0,1]^c$:*

$$\begin{aligned}\overset{3}{\perp_S}(\mathbf{u}) = \min(\ &\max(u_3,u_4), \max(u_2,u_4), \max(u_2,u_3),\\ &\max(u_1,u_4), \max(u_1,u_3), \max(u_1,u_2)).\end{aligned}$$

If $\mathbf{u}$ is such that $u_1 = u_{(1)} > u_3 = u_{(2)} > u_4 = u_{(3)} > u_2 = u_{(4)}$, it gives

$$\begin{aligned}\overset{3}{\perp_S}(\mathbf{u}) &= \min(u_3, u_4, u_3, u_1, u_1, u_1)\\ &= u_4\\ &= u_{(3)}.\end{aligned}$$

Since $\top_S = min$ is the largest t-norm, this last property allows the authors to claim that their operator measures to what extent the (generalization of the) k largest values of $\mathbf{u}$ are all large. Therefore, if $\mathbf{u}$ is a vector of posterior probabilities, or a membership function of an object $\mathbf{x}$ to be classified, fOR-$(k+1)$ can be used as a family of measures to select k classes $\Phi_{MBF}(\mathbf{u}) = \overset{k+1}{\perp}(\mathbf{u})$, given a dual couple $(\top, \perp)$.

We will use another fuzzy aggregation operator, namely the Sugeno integral in its discrete form [32]. It computes the mean value of a function with respect to a fuzzy measure μ, which is

a non-additive measure of uncertainty, *i.e.* more general than a possibility one and therefore a probability one, see [35] chapter 5 for details. The Sugeno integral of a function f w.r.t μ is defined by

$$\mathcal{S}_\mu = \bigvee_{i=1}^{n} f(x_i) \wedge \mu(A_{(i)}) \tag{10}$$

where $A_{(i)} = \{x_{(i)}, \cdots, x_{(n)}\}$ with respect to a permutation so that $f(x_{(i)}) \leq \cdots \leq f(x_{(n)})$. This integral is widely used in decision making, and in particular for pattern recognition [37] because of its ability to model some kind of interaction between features describing a pattern $\mathbf{x}$.

3. The new class-selective rule

3.1. Motivation

Since $\overset{k}{\perp}(\mathbf{u})$ as defined by (9) generalizes the $k-th$ largest value in $\mathbf{u} \in \mathcal{L}_{pc}$, it seems natural to use the ratio $\overset{k+1}{\perp}(\mathbf{u})/\overset{k}{\perp}(\mathbf{u})$ to extend the selection measure $\Phi_{FL}(\mathbf{u}) = u_{(k+1)}(\mathbf{x})/u_{(k)}(\mathbf{x})$ in order to compare two consecutive (ordered) values. Another extension consists in comparing more than two successive ordered values, that is to say all the values between $u_{(i)}$ and $u_{(j>i)}$. In [38], Le Capitaine et al. have defined a *block-similarity* operator by: $[0,1]^c \to [0,1]$, $\mathbf{u} \mapsto \mathcal{B}_{(i,j)}(\mathbf{u})$ satisfying the following four properties, $\forall (i,j) \in C^+ \times C^+$, $i < j$:

- the similarity should be minimum whenever the largest and the smallest values in the block are the maximum and the minimum possible ones

$$\mathcal{B}_{(i,j)}(\mathbf{u}) = 0 \text{ whenever } u_{(i)} = 1 \text{ and } u_{(j)} = 0 \tag{11}$$

- as the opposite, it should be maximum whenever the largest and the smallest values (therefore all values) in the block are the same

$$\mathcal{B}_{(i,j)}(\mathbf{u}) = 1 \Leftrightarrow u_{(i)} = u_{(j)} \tag{12}$$

- the similarity should increase as the range of values within the block decreases
  $\forall 0 \leq \varepsilon \leq u_{(i-1)} - u_{(i)}$,

$$\mathcal{B}_{(i,j)}(u_{(1)}, ..., u_{(i)} + \varepsilon, ..., u_{(c)}) \leq \mathcal{B}_{(i,j)}(u_{(1)}, ..., u_{(i)}, ..., u_{(c)}) \tag{13}$$

- inversely, it should decrease as the range increases
  $\forall 0 \leq \varepsilon \leq u_{(j-1)} - u_{(j)}$,

$$\mathcal{B}_{(i,j)}(u_{(1)}, ..., u_{(j)} + \varepsilon, ..., u_{(c)}) \geq \mathcal{B}_{(i,j)}(u_{(1)}, ..., u_{(j)}, ..., u_{(c)}). \tag{14}$$

Unfortunately, these properties are not all satisfied by $\overset{j}{\perp}(\mathbf{u})/\overset{i}{\perp}(\mathbf{u})$ for all $(\top, \perp)$. For instance, (12) is satisfied with standard norms but not with product ones.

Example 2. *Let us consider* $\mathbf{u} = {}^t(0.5 \;\; 0.2 \;\; 0.5 \;\; 0.8 \;\; 0.5)$ *and* $(i,j) = (2,4)$. *Since* $u_{(2)} = u_{(3)} = u_{(4)} = 0.5$, *we have:*

$$\begin{aligned} \overset{j}{\perp}_S(\mathbf{u})/\overset{i}{\perp}_S(\mathbf{u}) &= u_{(4)}/u_{(2)} \\ &= 1 \end{aligned}$$

but

$$\begin{aligned} \overset{j}{\perp}_P(\mathbf{u})/\overset{i}{\perp}_P(\mathbf{u}) &= 0.0558/0.7764 \\ &\neq 1. \end{aligned}$$

Let us correct this drawback by defining a generalized discrete Sugeno integral (*i.e.* using any t-norm) of $\mathbf{u}$ with respect to a fuzzy measure μ_k as:

$$\begin{aligned} \overset{k}{\mathcal{S}}_\mu(\mathbf{u}) &= \bigperp_{i=1,c} (u_{(i)} \top \mu_k(A_i)) \\ &= \left(\bigperp_{i=1,k-1} (u_{(i)} \top \mu_k(A_i)) \right) \perp \left(\bigperp_{i=k,c} (u_{(i)} \top \mu_k(A_i)) \right) \end{aligned} \tag{15}$$

where $A_i = \{j \in C^+ : u_{(j)} \geq u_{(i)}\}$. By choosing μ_k as the cardinal measure defined by

$$\mu_k(A_i) = \begin{cases} 0 & \text{if } Card(A_i) < k \\ 1 & \text{else} \end{cases}, \tag{16}$$

it is easy to show, by (15), that $\overset{k}{\mathcal{S}}_\mu(\mathbf{u})$ can be written as

$$\overset{k}{\mathcal{S}}_\mu(\mathbf{u}) = \begin{cases} \bigperp_{i=k,c} u_{(i)} & \text{if } u_{(k-1)} > u_{(k)} \\ \bigperp_{i=j,c} u_{(i)} & \text{else, } j \text{ is s.t. } u_{(j-1)} > u_{(j)} = \cdots = u_{(k)} \end{cases} \tag{17}$$

since 0 (respectively 1) is the absorbing (respectively neutral) element of any t-norm $\top$ (respectively t-conorm $\perp$).

Proposition 1. For standard and product norm couples $(\top, \perp)$, a block-similarity operator satisfying (11-14) is given by: $\forall (i,j) \in C^+ \times C^+, i < j$,

$$\mathcal{R}^{\top}_{(i,j)}(\mathbf{u}) = \overset{j}{\mathcal{S}}_\mu(\mathbf{u})/\overset{i}{\mathcal{S}}_\mu(\mathbf{u}). \tag{18}$$

For brevity, the proof of the proposition is omitted, and the reader is invited to refer to a review paper on block-similarity operators by the authors [39].

Example 3. *Let us consider* $\mathbf{u} = {}^t(0.5 \;\; 0.2 \;\; 0.5 \;\; 0.8 \;\; 0.5)$ *and* $(i,j) = (2,4)$ *as defined in Ex. 2. Since* $u_{(1)} = 0.8 > 0.5 = u_{(2)} = u_{(3)} = u_{(4)}$, *we have by (17):* $\overset{2}{\mathcal{S}}_{\top,\mu}(\mathbf{u}) = \bigperp_{k=2,c} u_{(k)}$ *and* $\overset{4}{\mathcal{S}}_{\top,\mu}(\mathbf{u}) = \bigperp_{k=2,c} u_{(k)}$, *whatever* $(\top, \perp)$. *Therefore,* $\mathcal{R}^{\top_S}_{(2,4)}(\mathbf{u}) = \mathcal{R}^{\top_P}_{(2,4)}(\mathbf{u}) = 1$.

As defined by (18), the block-similarity operator is not fully convenient for measuring the similarity between values within the block. First, it is not symmetrical because all the values between $u_{(j+1)}$ and $u_{(c)}$ are taken into account even if $u_{(j)} > u_{(j+1)}$, while values between $u_{(1)}$ and $u_{(i-1)}$ are not even if $u_{(i-1)} > u_{(i)}$. Second, the cardinal fuzzy measure (16) equally weights all values in the block whatever their position, so their relative magnitude. We propose to use a kernel function to overcome these drawbacks and make the values between $u_{(i)}$ and $u_{(j)}$, and only them, meaningful.

3.2. The symmetrical block-similarity operator

Given a symmetrical kernel function $\mathcal{K}_\lambda(l,k)$ centered on k, parametrized by a resolution parameter $\lambda \in \mathbb{R}^+$ which controls its area of influence, let us define two fuzzy Sugeno integrals:

$$\overset{k}{\mathcal{S}}_{\mathcal{K}_\lambda}(\mathbf{u}) = \bigperp_{\ell=\frac{i+j}{2}}^{k} (u_{(\ell)} \top \mathcal{K}_\lambda(\ell,k)) \qquad \text{if } (j-i) \text{ is even} \tag{19}$$

$$\overset{k^\pm}{\mathcal{S}}_{\mathcal{K}_\lambda}(\mathbf{u}) = \bigperp_{\ell=\frac{i+j\pm1}{2}}^{k} (u_{(\ell)} \top \mathcal{K}_\lambda(\ell,k)) \qquad \text{if } (j-i) \text{ is odd} \tag{20}$$

where $k^\pm$ denotes the two possible values k^+ and k^- meaning that the control variable k starts from $\frac{i+j+1}{2}$ or $\frac{i+j-1}{2}$, respectively.

Proposition 2. For strictly continuous norm couples $(\top, \perp)$, a symmetrical block-similarity operator satisfying (11-14) is given by: $\forall (i,j) \in C^+ \times C^+, i < j$,

$$\mathcal{R}^{\top,\mathcal{K}_\lambda}_{(i,j)}(\mathbf{u}) = \begin{cases} \overset{j}{\mathcal{S}}_{\mathcal{K}_\lambda}(\mathbf{u}) / \overset{i}{\mathcal{S}}_{\mathcal{K}_\lambda}(\mathbf{u}) & \text{if } (j-i) \text{ is even} \\ \overset{j^+}{\mathcal{S}}_{\mathcal{K}_\lambda}(\mathbf{u}) / \overset{i^-}{\mathcal{S}}_{\mathcal{K}_\lambda}(\mathbf{u}) & \text{if } (j-i) \text{ is odd} \end{cases} \tag{21}$$

with the convention $\mathcal{R}^{\top,\mathcal{K}_\lambda}_{(i,j)}(\mathbf{u}) = 1$ if $u_{(i)} = 0$.

Here again, the proof is omitted, (see [39] for details).

Many symmetrical functions are available, (see Table 2). Note that we impose them to be normalized so that $\mathcal{K}(l,l) = 1$ and the corresponding $u_{(l)}$ is of maximum weight in the Sugeno integrals (19-20). Figure 3-(*left*) shows different kernel functions $\mathcal{K}_\lambda(l,k)$ centered in $k = 6$.

Table 2: Examples of normalized symmetrical kernel functions $\mathcal{K}_\lambda(l,k)$ where $y = |k-l|$.

Kernel	$\mathcal{K}_\lambda$
uniform	$\mathcal{U}_\lambda(y) = \mathbb{1}_{(y \le \lambda)}$
Gaussian	$\mathcal{N}_\lambda(y) = \exp\left(-\frac{y^2}{\lambda}\right)$
exponential	$\mathcal{E}_\lambda(y) = \exp\left(-\lambda y^2\right)$
Epanechnikov	$E_\lambda(y) = \left(1-\frac{y^2}{\lambda^2}\right)\mathbb{1}_{(y \le \lambda)}$
triangular	$\mathcal{T}_\lambda(y) = \left(1-\frac{y}{\lambda}\right)\mathbb{1}_{(y \le \lambda)}$
Cauchy	$C_\lambda(y) = \frac{\lambda}{\lambda+y^2}$

With no loss of generality, let us study how $\mathcal{R}^{\top,\mathcal{K}_\lambda}_{(i,j)}(\mathbf{u})$ behaves using a Gaussian kernel $\mathcal{K}_\lambda(l,k) = \mathcal{N}_\lambda(l,k)$. Other kernels such as the exponential or Cauchy have a similar behavior since the resolution parameter can be chosen such that the shapes are roughly identical. The Gaussian kernel is defined as a function of $(\lambda > 0, k, l)$ for sake of clarity:

$$\mathcal{N}_\lambda(k,l) = \exp\Big(-\frac{(k-l)^2}{\lambda}\Big). \tag{22}$$

The resolution parameter λ controls the area of influence as follows:

- when $\lambda \to 0$, the kernel becomes a Dirac centered in l, δ_l, because the convergence is not uniform by continuity of $\mathcal{N}_\lambda$ and non continuity of δ_l,

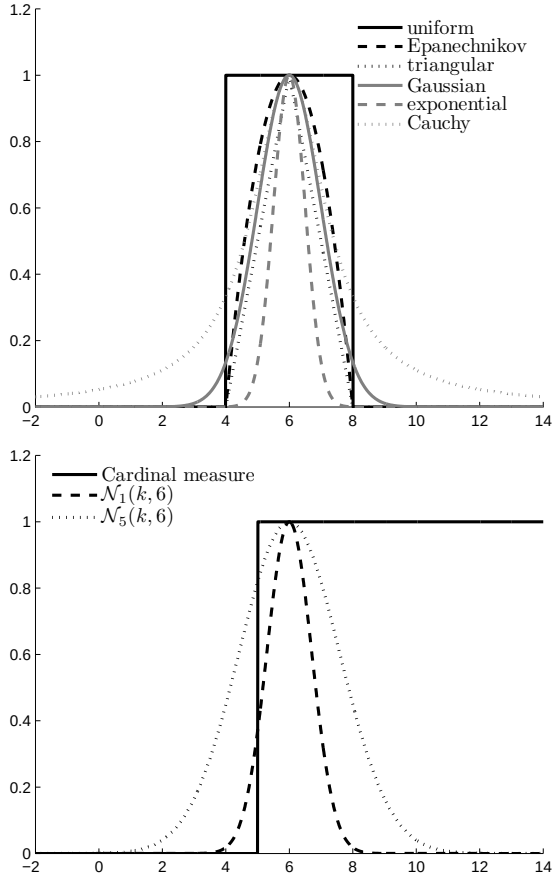


Figure 3: Examples of symmetrical kernel functions centered in $k = 6$ with $\lambda = 2$ (*left*), cardinal measure with $u_{(5)} > u_{(6)}$ and Gaussian kernels $\mathcal{N}_1(l,6)$ and $\mathcal{N}_5(l,6)$ (*right*).

- when $\lambda \to +\infty$, the kernel becomes the constant value 1.

Proposition 3. For continuous norm couples $(\top, \perp)$, then

$$\lim_{\lambda \to 0} \mathcal{R}^{\top,\mathcal{N}_\lambda}_{(i,j)}(\mathbf{u}) = \begin{cases} 1 & \text{if } u_{(i)} = 0 \\ \frac{u_{(j)}}{u_{(i)}} & \text{else} \end{cases} \tag{23}$$

$$\lim_{\lambda \to +\infty} \mathcal{R}^{\top,\mathcal{N}_\lambda}_{(i,j)}(\mathbf{u}) = \begin{cases} \bigperp_{k=\frac{i+j}{2}}^{j} u_{(k)} / \bigperp_{k=\frac{i+j}{2}}^{i} u_{(k)} & \text{if } (j-i) \text{ is even} \\ \bigperp_{k=\frac{i+j+1}{2}}^{j} u_{(k)} / \bigperp_{k=\frac{i+j-1}{2}}^{i} u_{(k)} & \text{if } (j-i) \text{ is odd} \\ 1 & \text{if } u_{(i)} = 0 \end{cases} \tag{24}$$

Therefore, the contribution of the intermediate values $\{u_{(i+1)}, ..., u_{(j-1)}\}$ to $\mathcal{R}^{\top,\mathcal{N}_\lambda}_{(i,j)}(\mathbf{u})$ is small if λ is close to zero and increases with λ, as shown in Figure 3.
Note that $\mathcal{R}^{\top,\mathcal{K}_\lambda}_{(i,i+1)}(\mathbf{u}) = u_{(i+1)}/u_{(i)}$ does not depend on λ whatever the kernel function. This means that increasing λ does not make two successive $u_{(k)}$'s more similar but may increase the similarity of blocks of larger size (strictly greater than 2).

Figure 4 clearly shows how the values $\mathcal{R}^{\top,\mathcal{N}_\lambda}_{(i,j)}(\mathbf{u})$ depend on (i,j). Two high values are sufficient to make $\mathcal{R}^{\top,\mathcal{N}_{0.8}}_{(1,2)}(\mathbf{u})$ large

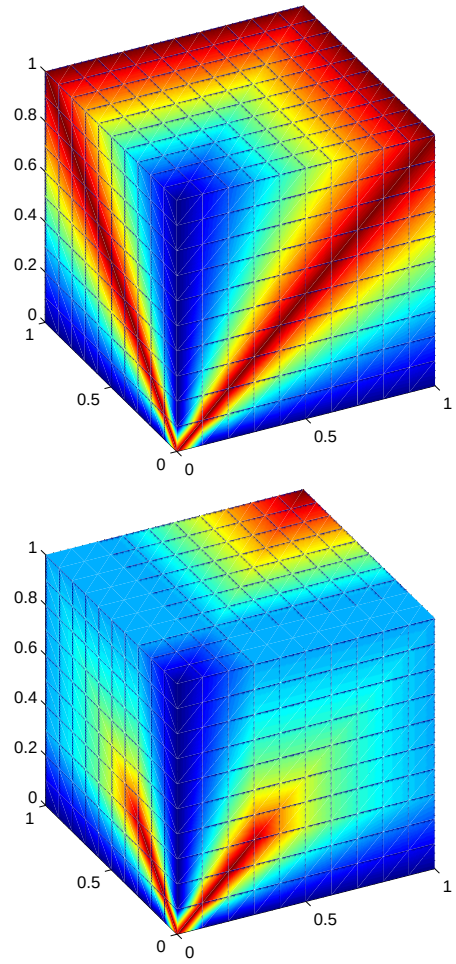


Figure 4: Values of $\mathcal{R}^{\top,\mathcal{N}_{0.8}}_{(1,2)}(\mathbf{u})$ (*left*) and $\mathcal{R}^{\top,\mathcal{N}_{0.8}}_{(1,3)}(\mathbf{u})$ (*right*) using $(\top,\perp)_S$ for all $\mathbf{u} \in [0,1]^3$. Hot (respectively cold) colors represent high (respectively low) values.

where three are necessary to make $\mathcal{R}^{\top,\mathcal{N}_{0.8}}_{(1,3)}(\mathbf{u})$ reach a similar level.

Example 4. *Let us consider the following soft label vector* $\mathbf{u} = {}^t(0.82\ \ 0.80\ \ 0.60\ \ 0.51\ \ 0.50\ \ 0.45\ \ 0.10)$ *and compute the values of* $\mathcal{R}^{\top_S,\mathcal{K}_\lambda}_{(i,j)}(\mathbf{u})$, $i = 1,6$ *and* $j = 2,7$ *for all the kernels functions of Table 2 and some resolution parameter values* λ. *One can compare the obtained values to a threshold* s, *here set to 0.95, so that if* $\mathcal{R}^{\top_S,\mathcal{K}_\lambda}_{(i,j)}(\mathbf{u}) > s$ *then the values in block* (i,j) *are detected as similar. The following blocks of similar values are detected:*

- $(1,2)$ *and* $(4,5)$, *corresponding to* $\{0.82, 0.80\}$ *and* $\{0.51, 0.50\}$, *for not too wide influence areas specified by* $\lambda < 1$ *(or* $1/\lambda$ *with* $\mathcal{E}_\lambda$*)*
- *either* $(1,2)$ *and* $(4,5)$ *exclusively or* $(1,2)$ *and* $(4,6)$, *corresponding to* $\{0.82, 0.80\}$ *and* $\{0.51, 0.50, 0.45\}$ *when increasing* λ *(or* $1/\lambda$ *with* $\mathcal{E}_\lambda$*),* e.g. *using* $\mathcal{N}_2$, $\mathcal{E}_2$, $\mathcal{T}_2$ *and* C_2.

3.3. The selection measure and the induced rule

We propose to use (18) to define a new family of class-selective decision rules including both types of reject with respect to (7-8) as: given a symmetrical kernel function $\mathcal{K}_\lambda$,

$$\Phi^{\top,\mathcal{K}_\lambda}_{k+1}(\mathbf{u}) = \mathcal{R}^{\top,\mathcal{K}_\lambda}_{(\mathbb{1}_{(k>0)},k+1)}(\mathbf{u}) \tag{25}$$

where $\mathbb{1}_{(k>0)}$ ensures that $\mathcal{R}^{\top,\mathcal{K}_\lambda}_{(1,k+1)}(\mathbf{u})$ is used whatever $k > 0$ and $\mathcal{R}^{\top,\mathcal{K}_\lambda}_{(0,1)}(\mathbf{u}) = u_{(1)}$ according to the remark following the proposition 3 and the usual convention $u_{(0)} = 1$. Note that it allows to retrieve the usual distance rejection (6) for $k = 0$, implying equivalence with Φ_{Ha}, Φ_{Ho} and Φ_{FL} for this special case. Whenever the number of classes is $c = 2$, $\Phi^{\top,\mathcal{K}_\lambda}_{k+1}$ reduces to Φ_{FL} and Φ_{FD}.

Let us see how the proposed family of rules given by (8)-(25) behaves, compared to the other class-selective rules (Ha, Horiuchi, Frélicot & Le Capitaine, referred to as literature's rules in the remaining part of the paper) on particular situations expressed by vectors of soft class-labels.

Example 5. *Let us consider the following four* $\mathbf{u}$ *vectors, each representing a particular classification case, emphasized by their expected classification hard labels* $\mathbf{h}$ *given by (7-8) assuming an appropriate threshold* t*:*

- $\mathbf{u}(\mathbf{x}) = {}^t(0.70, 0.10, 0.85, 0.80) \mapsto \mathbf{h}(\mathbf{x}) = {}^t(1,0,1,1), n^\star(\mathbf{x},t) = 3$ *;*
- $\mathbf{u}(\mathbf{y}) = {}^t(0.20, 0.10, 0.85, 0.80) \mapsto \mathbf{h}(\mathbf{y}) = {}^t(0,0,1,1), n^\star(\mathbf{y},t) = 2$ *;*
- $\mathbf{u}(\mathbf{z}) = {}^t(0.20, 0.10, 0.85, 0.15) \mapsto \mathbf{h}(\mathbf{z}) = {}^t(0,0,1,0), n^\star(\mathbf{z},t) = 1$ *;*
- $\mathbf{u}(\mathbf{v}) = {}^t(0.20, 0.10, 0.05, 0.15) \mapsto \mathbf{h}(\mathbf{v}) = {}^t(0,0,0,0), n^\star(\mathbf{v},t) = 0$.

Table 3 reports, for $k = 0,1,2,3$ and 4, the values of the selection measures $\Phi_{Ha}(\mathbf{u}) = u_{(k+1)}$, $\Phi_{Ho}(\mathbf{u}) = 1 - (u_{(k)} - u_{(k+1)})$, $\Phi_{FL}(\mathbf{u}) = u_{(k+1)}/u_{(k)}$ and $\Phi^{\top,\mathcal{K}_\lambda}_{k+1}(\mathbf{u})$ using two kernel functions and two resolution parameter values. For each $\mathbf{u}$, the range of the threshold t values leading to the correct classification is given, if there is any. The two kernel functions being considered in this example are the uniform and the Gaussian ones. The uniform kernel is used with a special purpose: emphasize its poor performances because of its equivalence with cardinal based measures such as (18). Results with other kernels such as Cauchy, exponential or triangular are roughly the same than Gaussian's ones, since a convenient resolution parameter can be tuned to obtain similar shapes, as can be seen in Figure 3.

According to Table 3, we obtain a larger range of threshold values by using the proposed selection measure than by using usual selection measures. In particular, for low values of λ, the range is much larger. The consequence is that the threshold is easier to tune with the proposed framework. Except in unusual cases, it is always possible to retrieve the right number of classes for the four considered soft label vectors. For large λ values, the Łukasiewicz t-norms are the only couple allowing

to keep a large range of threshold values. The Gaussian kernels are, as expected, more efficient than the uniform kernels. Non available (n.a.) ranges (*i.e.* there are no thresholds allowing to retrieve the right number of classes) only occur with uniform kernels. As a final remark, let us mention that in these synthetic examples, the t-norm couples can be ranked as follows: $(\top, \bot)_S \prec (\top, \bot)_P \prec (\top, \bot)_L$.

4. Experimental Results

4.1. Datasets and protocol

To validate the efficiency of the proposed class-selective rejection rules, we compare their performance on three artificial datasets and thirteen real datasets from the UCI Machine Learning Repository [40] of various characteristics in terms of: number n of objects, number p of features, number c of classes and degree of overlap, as summarized in Table 4. Corresponding degrees of overlap are roughly estimated through a three-dimensional PCA (*Principal Component Analysis*) projection. The synthetic datasets are:

- $D1$ composed of 2000 points drawn from a mixture of two normal seven-dimensional distributions of 1000 points each with means $\mathbf{v}_1 = {}^t(1, 0, ..., 0)$ and $\mathbf{v}_2 = {}^t(-1, 0, ..., 0)$, and equal covariance matrices $\Sigma_1 = \Sigma_2 = I$,
- $D2$ contains 4000 points drawn from a mixture of four normal two-dimensional distributions of 1000 points each with means $\mathbf{v}_1 = {}^t(1, 1)$, $\mathbf{v}_2 = {}^t(1, -1)$, $\mathbf{v}_3 = {}^t(-1, 1)$ and $\mathbf{v}_4 = {}^t(-1, -1)$, and equal covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma_3 = \Sigma_4 = I$, (see Figure 5-(*left*))
- *DH* consists of two overlapping Gaussian classes with different covariance matrices according to the Highleyman distribution, each composed of 800 observations in $\mathbb{R}^2$ [41], (see Figure 5-(*right*)).

Soft labels are computed using the model defined by (1), where α_i are set to 1 for each class. Note that other decreasing functions of the distance to the class-means can be used to obtain soft labels, *e.g.* $u_i(\mathbf{x}) = \exp(-\alpha_i\, d^2(\mathbf{x}, \mathbf{v}_i))$, but since performances obtained with it are roughly the same, the results are not reported here.

There are several ways to evaluate a decision rule, depending on the type(s) or reject(s) involved (ambiguity and/or distance) and the kind of rule (rejection rule or class-selective rule). One generally uses the following quantities, as a function of the rule's parameters Θ: the correct classification $C(\Theta)$, the misclassification (or error) $E(\Theta)$ and the reject $R(\Theta)$ probabilities or rates. Chow introduced the error-reject (ER) trade off and proposed to analyze the $ER-$curve ($E(\Theta)$ *vs.* $R(\Theta)$) for all possible values of Θ (limited to $\{t\}$ for Chow's rule as well as for all other rules studied in this paper), in order to find the optimal or an operational value for t [3]. Therefore a common way to compare such rules is to plot their $ER-$curves and look for the minimum *A*rea *U*nder the *C*urve (*AUC*).

For class-selective rules, an object is generally considered as misclassified if the class it is issued from does not belong to the subset of selected ones. Therefore the $ER-$trade off is replaced by the error-average number of classes $E\overline{n}-$trade off and one analyses the $E\overline{n}-$curves as introduced by Ha [5]. The area under the $E\overline{n}-$curve [34] is given by

$$AUC(E\overline{n}) = \int_0^1 E(\overline{n}(t))dt. \tag{26}$$

Since the decision rule is evaluated for all the possible thresholds, the less the area under the $E\overline{n}-$curve is, the better the performance that is achieved [34].

We use 10-fold cross-validation, dividing the set of samples at random into 10 approximately equal-size parts. The 10 parts are roughly balanced, ensuring that the classes are distributed proportionally among each of the 10 parts. Ten-fold cross-validation works as follows: we fit the model on 90% of the samples and then predict the hard class labels $\mathbf{h}$ of the remaining 10% (the test samples). This procedure is repeated 10 times, with each part playing the role of the test samples, and the errors on all 10 parts are added together to compute the overall error.

Table 4: Datasets used in the experiments.

Dataset	n	p	c	Overlap
DH	1600	2	2	slight
*D*1	2000	7	2	moderate
*D*2	4000	2	4	strong
Forest	495411	10	2	strong
Ionosphere	351	34	2	strong
Pima	768	9	2	strong
Iris	150	4	3	slight
Thyroid	215	5	3	slight
PageBlocks	5473	10	5	moderate
Glass	214	9	6	strong
Statlog	6435	36	6	moderate
Digits	10992	16	10	slight
Yeast	1484	8	10	strong
Optical	5620	64	10	slight
Vowel	528	10	11	moderate
Letter	20000	16	26	strong

4.2. Results

The $AUC(E\overline{n})$ values obtained on the datasets of Table 4 using the new family decision rule using the three t-norm couples of Table 1 and the literature's decision rules: Ha [6], Horiuchi [7], Frélicot and Le Capitaine [28], Mascarilla *et al.* [13] referred to as $\Phi^\star_{MBF}$, Tax and Duin [12] referred to as *TD*, are given in Table 5. The $\Phi^\star_{MBF}$ rule is not directly derived from the Φ_{MBF} measure but is a rule adapted to the specific case of class selection by normalization steps in order to allow a single threshold use, as suggested by their authors (see the paper for details). For sake of brevity, we only report results obtained by $\Phi^\star_{MBF,\top_S}$, *i.e.* using the Standard t-norms. The TD rule is the

Table 3: Selection measures for $\mathbf{u}(\mathbf{x}) = {}^t(0.70, 0.10, 0.85, 0.80)$, $\mathbf{u}(\mathbf{y}) = {}^t(0.20, 0.10, 0.85, 0.80)$, $\mathbf{u}(\mathbf{z}) = {}^t(0.20, 0.10, 0.85, 0.15)$ and $\mathbf{u}(\mathbf{v}) = {}^t(0.20, 0.10, 0.05, 0.15)$ (from top to bottom) for which the right number of classes $n^\star(\mathbf{x}, t) = 3$, 2, 1 and 0 respectively is selected whatever t in the specified interval.

$\mathbf{u}$	k	Φ_{Ha}	Φ_{Ho}	Φ_{FL}	$\Phi^{\top,\mathcal{K}_\lambda}_{(k+1)}$											
					$(\top,\perp)_S$				$(\top,\perp)_P$				$(\top,\perp)_L$			
					$\mathcal{U}_\lambda$		$\mathcal{N}_\lambda$		$\mathcal{U}_\lambda$		$\mathcal{N}_\lambda$		$\mathcal{U}_\lambda$		$\mathcal{N}_\lambda$	
					$\lambda = .1$	$\lambda = 1$	$\lambda = .1$	$\lambda = 1$	$\lambda = .1$	$\lambda = 1$	$\lambda = .1$	$\lambda = 1$	$\lambda = .1$	$\lambda = 1$	$\lambda = .1$	$\lambda = 1$
u(x)	0	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
	1	0.80	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	2	0.70	0.90	0.87	0.82	0.94	0.82	0.82	0.82	0.96	0.82	0.88	0.82	1.00	0.82	0.86
	3	0.10	0.40	0.14	0.11	0.82	0.11	0.43	0.11	0.75	0.11	0.37	0.11	0.80	0.11	0.16
	4	0	0.90	0	0	0.11	0	0.11	0	0.10	0	0.05	0	0.10	0	0
	$\forall t \in$	[.1,.7[	[.4,.85[	[.14,.85[	[.11,.82[	[.82,.85[	[.11,.82[	[.43,.82[	[.11,.82[	[.75,.85[	[.11,.82[	[.37,.85[	[.11,.82[	[.8,.85[	[.11,.82[	[.16,.85[
u(y)	0	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
	1	0.80	0.95	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94	0.94
	2	0.20	0.40	0.25	0.23	0.94	0.23	0.43	0.23	0.86	0.23	0.48	0.23	1	0.23	0.36
	3	0.10	0.90	0.50	0.11	0.23	0.11	0.23	0.11	0.29	0.11	0.18	0.11	0.30	0.11	0.10
	4	0	0.90	0	0	0.11	0.01	0.11	0	0.10	0.01	0.04	0	0.10	0	0
	$\forall t \in$	[.2,.8[	[.4,0.85[	[.25,.85[	[.23,.85[	n.a.	[.23,.85[	[.43,.85[	[.23,.85[	n.a.	[.23,.85[	[.48,.85[	[.23,.85[	n.a.	[.23,.85[	[.36,.85[
u(z)	0	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85
	1	0.20	0.35	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23	0.23
	2	0.15	0.95	0.75	0.17	0.23	0.17	0.23	0.17	0.36	0.17	0.24	0.17	0.35	0.17	0.17
	3	0.10	0.95	0.67	0.11	0.17	0.11	0.17	0.11	0.26	0.11	0.17	0.11	0.25	0.11	0.11
	4	0	0.90	0	0	0.11	0.01	0.11	0	0.11	0.01	0.04	0	0.10	0	0
	$\forall t \in$	[.2,.85[	[.35,.85[	[.23,.85[	[.23,.85[	[.23,.85[	[.23,.85[	[.23,.85[	[.23,.85[	[.23,.85[	[.23,.85[	[.23,.85[	[.23,.85[	[.23,.85[	[.23,.85[	[.23,.85[
u(v)	0	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20	0.20
	1	0.15	0.95	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0.75
	2	0.10	0.95	0.67	0.50	0.75	0.50	0.75	0.50	0.73	0.50	0.61	0.50	0.71	0.50	0.50
	3	0.05	0.95	0.50	0.25	0.50	0.25	0.50	0.25	0.45	0.25	0.34	0.25	0.42	0.25	0.25
	4	0	0	0	0	0.25	0.02	0.25	0	0.15	0.01	0.08	0	0.14	0	0
	$\forall t \in$	[.2,1[	[.2,1[	[.2,1[	[.2,1[	[.2,1[	[.2,1[	[.2,1[	[.2,1[	[.2,1[	[.2,1[	[.2,1[	[.2,1[	[.2,1[	[.2,1[	[.2,1[

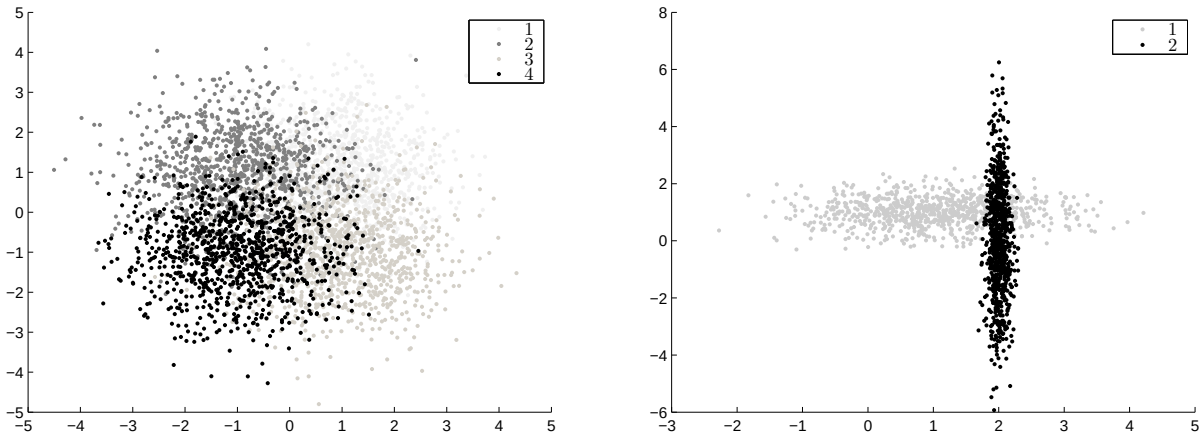


Figure 5: Synthetic datasets *D*2 (*left*) and *DH* (*right*)

one their authors call *target normalization*. Since it is not dedicated to class selection, only results obtained on two classes datasets are reported, rejection meaning that two classes are selected.

As explained in the previous section, all the kernels of Table 2, except the uniform one, behave well provided a reasonable tuning of the resolution parameter. Therefore, we only present the results obtained with a good kernel (Gaussian $\mathcal{N}_\lambda$) and the poorest one (uniform $\mathcal{U}_\lambda$) for two values of λ (0.1,1). The results show that the rules based on Φ_{Ha} and Φ_{Ho} selection

Table 5: Area under the $E\overline{n}$−curve for class-selective decision rules. Best results are in bold.

Datasets	Φ_{Ha}	Φ_{Ho}	Φ_{FL}	$\Phi^{\star}_{MBF,\top_S}$	TD	$\Phi^{\top,\mathcal{K}_\lambda}_{(\mathbb{1}_{(k>0)},k+1)}$											
						$(\top,\perp)_S$				$(\top,\perp)_P$				$(\top,\perp)_L$			
						$\mathcal{U}_\lambda$		$\mathcal{N}_\lambda$		$\mathcal{U}_\lambda$		$\mathcal{N}_\lambda$		$\mathcal{U}_\lambda$		$\mathcal{N}_\lambda$	
						$\lambda=0.1$	$\lambda=1$	$\lambda=0.1$	$\lambda=1$	$\lambda=0.1$	$\lambda=1$	$\lambda=0.1$	$\lambda=1$	$\lambda=0.1$	$\lambda=1$	$\lambda=0.1$	$\lambda=1$
DH	0.011	0.029	**0.008**	**0.008**	0.010	**0.008**	**0.008**	**0.008**	**0.008**	**0.008**	**0.008**	**0.008**	**0.008**	**0.008**	**0.008**	**0.008**	**0.008**
*D*1	0.070	0.073	**0.046**	**0.046**	0.053	**0.046**	**0.046**	**0.046**	**0.046**	**0.046**	**0.046**	**0.046**	**0.046**	**0.046**	**0.046**	**0.046**	**0.046**
*D*2	0.223	0.408	0.250	0.259	-	**0.207**	0.259	**0.207**	0.258	**0.207**	0.257	**0.207**	0.229	**0.207**	0.253	**0.207**	**0.207**
Forest	0.146	0.151	**0.128**	**0.128**	0.138	**0.128**	**0.128**	**0.128**	**0.128**	**0.128**	**0.128**	**0.128**	**0.128**	**0.128**	**0.128**	**0.128**	**0.128**
Ionosphere	0.186	0.196	**0.163**	**0.163**	0.176	**0.163**	**0.163**	**0.163**	**0.163**	**0.163**	**0.163**	**0.163**	**0.163**	**0.163**	**0.163**	**0.163**	**0.163**
Pima	0.164	0.141	**0.115**	**0.115**	0.139	**0.115**	**0.115**	**0.115**	**0.115**	**0.115**	**0.115**	**0.115**	**0.115**	**0.115**	**0.115**	**0.115**	**0.115**
Iris	0.034	0.022	0.009	0.012	-	0.009	0.011	0.009	0.011	**0.008**	0.010	**0.008**	0.009	**0.008**	0.010	**0.008**	**0.008**
Thyroid	0.056	0.034	0.007	0.013	-	**0.006**	0.012	**0.006**	0.012	**0.006**	0.010	**0.006**	0.008	**0.006**	0.010	**0.006**	**0.006**
PageBlocks	0.215	0.146	0.086	0.087	-	0.087	0.087	0.086	0.087	0.087	0.083	0.087	**0.080**	0.087	0.083	0.087	0.087
Glass	0.526	0.537	0.125	0.166	-	**0.106**	0.165	**0.106**	0.175	**0.106**	0.144	**0.106**	0.118	**0.106**	0.143	**0.106**	**0.106**
Statlog	0.436	0.436	0.116	0.154	-	0.089	0.114	**0.086**	0.129	0.089	0.102	0.089	0.093	0.089	0.102	0.089	0.089
Digits	0.171	0.167	0.010	0.010	-	0.008	0.010	0.008	0.010	0.008	0.009	0.008	**0.007**	0.008	0.009	0.008	0.008
Yeast	1.665	2.122	1.189	0.911	-	**0.790**	0.898	**0.790**	0.954	**0.790**	0.854	**0.790**	0.807	**0.790**	0.849	**0.790**	**0.790**
Optical	0.099	0.097	0.042	0.039	-	0.004	0.006	**0.003**	0.004	**0.003**	0.005	**0.003**	0.004	**0.003**	0.005	**0.003**	**0.003**
Vowel	0.404	0.423	0.075	0.047	-	**0.035**	0.044	**0.035**	0.049	**0.035**	0.042	**0.035**	0.036	**0.035**	0.042	**0.035**	**0.035**
Letter	1.387	1.379	0.190	0.201	-	**0.058**	0.070	**0.058**	0.076	**0.058**	0.069	**0.058**	0.061	**0.058**	0.069	**0.058**	**0.058**

measures never reach the performance of the others. Let us go further into the analysis according to:

- the number of classes of the datasets and their degree of overlap between the classes,
- the parameters of the rules, namely the kernel, its resolution parameter value, and the triangular norm couple.

As expected, the rules based on Φ_{MBF}, Φ_{FL} and $\Phi^{\top,\mathcal{K}_\lambda}_{k+1}$ give the same results on two-classes datasets {*D*1, *DH*, *Forest*, *Ionosphere*, *Pima*}, whatever the kernel and the resolution parameter, and outperform the rules based on other measures, Φ_{Ha}, Φ_{Ho} and TD. The number of classes is of great interest because the proposed family of rules can take into account simultaneously from two up to c values so that the whole similarity can be exhibited while the literature's rules that only involve two successive ordered values cannot. The more the number of classes, the more the proposed family of rules outperforms the others, see datasets {*PageBlocks*, *Glass*, *Statlog*, *Digits*, *Yeast*, *Vowel*, *Optical*, *Letter*} for which $c > 3$. The way the tested rules perform with respect to the degree of overlap can be underlined by the scores ratios and differences. The *AUC* ratios are much larger for datasets presenting a slight or moderate overlap of classes {*DH*, *D*1, *Iris*, *Thyroid*, *PageBlocks*, *Statlog*, *Digits*, *Vowel*, *Optical*} than for those presenting a strong overlap {*D*2, *Ionosphere*, *Forest*, *Pima*, *Glass*, *Yeast*, *Letter*}. The reason is that for better separated classes, there are more patterns for which the largest soft label is much larger than the other values, and this situation is handled by the proposed operator whereas it is not taken into account in literature's selection measures. While the *AUC* ratios are lower for datasets presenting a strong overlap, the *AUC* differences are much larger (e.g. {*Glass*, *Statlog*, *Vowel*, *Letter* } and except for *Yeast*) for datasets presenting better separated classes. This means that the less the classes overlap, the less the benefit is, as it is for any decision rule including a reject option, *e.g.* [3, 14, 6].

As expected, using a Gaussian kernel $\mathcal{N}_\lambda$ leads to better performances than using an uniform one $\mathcal{U}_\lambda$. For almost all datasets (except *PageBlocks* and *Digits*), the best scores are obtained for small values of λ, whatever the kernel. Setting a small value of λ consists in considering only the first large and last low soft labels. This result shows that middle values of detected blocks are not critical for the class selection problem. Moreover, the Gaussian kernel is less sensitive than the uniform one to variation of their resolution parameter. For instance, changing λ from 0.1 to 1 for $\mathcal{U}_\lambda$ significantly affects the performance. As discussed in Section 3, the uniform kernel handles the similarity analogously to the cardinal measure and does not allow to make soft labels of various importances contributing. Consequently, the cardinal measure, as well as the uniform kernel, is not well adapted and convenient for the class selection problem. Moreover, as shown in Section 3, tuning λ for $\mathcal{U}_\lambda$ is much more difficult because the range of threshold values allowing to select a reasonable number of classes is smaller. What about the triangular norm couples? According to Table 5, the results are quite similar. However, considering the number of cases for which the proposed family of rules gives the best result, whatever the kernel, one can observe the following ranking: $(\top,\perp)_S \prec (\top,\perp)_P \prec (\top,\perp)_L$. This confirms the results reported in the previous section in Example 5. More generally, the product and Lukasiewicz norms induce rules that are less sensitive, in terms of performances, to changes of kernels and resolution parameter value. For final illustration purpose, some of the $(E,\overline{n})$ curves are shown in Figure 6. For sake of brevity, only six datasets presenting various amount of overlap, number of classes and number of patterns, are chosen: {*D*2,*Digits*,*Statlog*,*Yeast*,*PageBlocks*, *Vowel*}. One can see that for every average number of selected classes, the error rate obtained with the proposed family of rules is lower than the ones obtained with literature's rules.

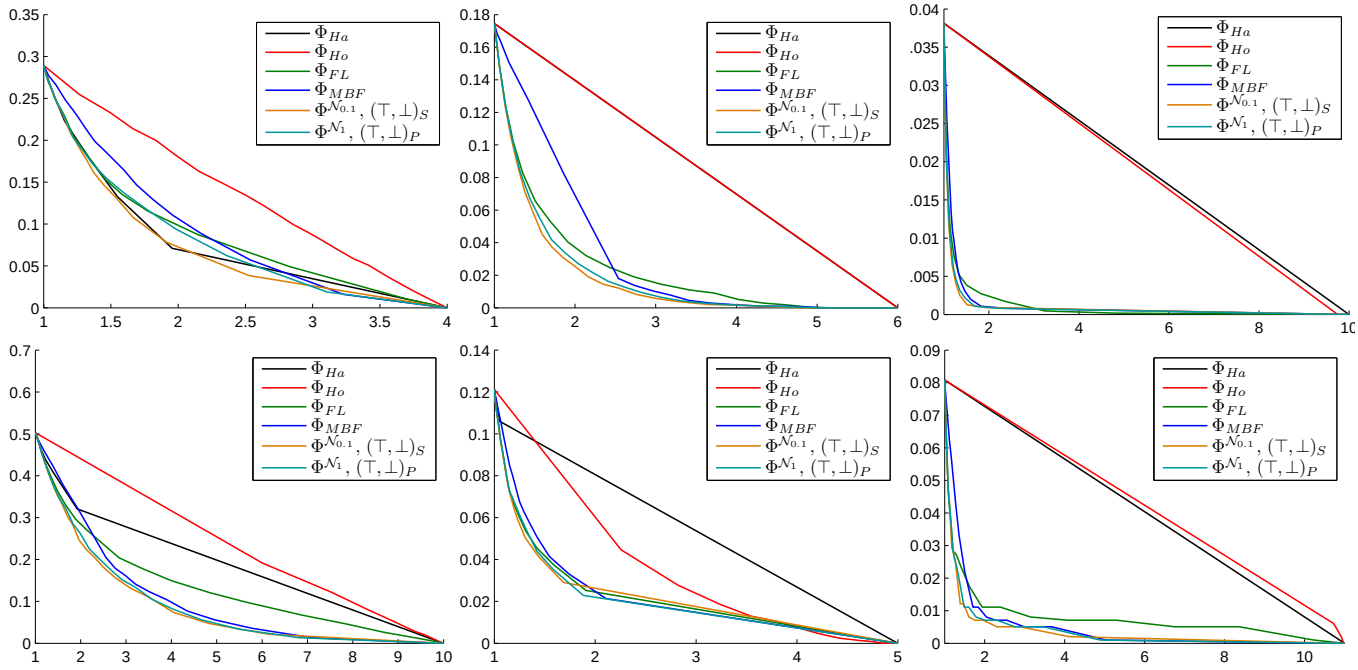


Figure 6: $(E, \overline{n})$ curves for Ha, Horiuchi, FL, MBF rules, and two examples of new rules obtained on six datasets (from left to right, up to bottom): {*D2*,*Digits*,*Statlog*,*Yeast*,*PageBlocks*, *Vowel*}.

5. Conclusion

A generic formulation of class-selective decision rules is presented. It allows to write state-of-the-art single threshold based decision rules with a single equation. A new family of measures suitable to define top-*n* class-selective decision rules is proposed as well. These selection measures rely on the block similarity detection of the soft labels to the classes of a pattern to be classified by means of a new aggregation operator. It is based on specific discrete fuzzy integrals of the ordered soft labels with respect to a symmetrical kernel function which weights the degrees according to their relative position within the block. The ratio of such two integrals is used to define the selection measures from which class-selective decision rules can be derived. Since such rules depend on only one user-defined threshold reflecting the costs of rejection and error, it is proposed to use the area under the curve of the error rate as a function of all possible threshold values as a performance measure. An extensive comparison with the usual one-threshold based class-selective decision rules on sixteen datasets is given. The results show that the new family of decision rules largely outperforms the existing ones on all datasets.

The choice of the triangular norms used in the fuzzy integrals is not trivial and remains an open problem from a theoretical point of view. It requires a further study on the mathematical properties of each t-norm, and to what extent the properties would affect the performance of the derived class-selective rules. Other future research will concern the study of a new mixed class-selective-rejective decision rule which should jointly optimize the number of selected (in the sense of class acceptation) and rejected (in the sense of elimination, see the preliminary work in [42]) classes over the user-defined threshold definition domain. We also plan to use this optimum decision rule for outliers detection, *i.e.* patterns that do not match any of the known classes so that they must be distance rejected. This problem, as well as its variant for support vector machines, will be addressed using hinge loss minimization [43], or surrogate convex loss functions [44]. It will be studied by taking into account new results on *AUC* variants [45, 46].

[1] H. Le Capitaine, C. Frélicot, A class-selective rejection scheme based on blockwise similarity of typicality degrees, in: 19th International Conference on Pattern Recognition, 2008, Tampa, USA, pp. 1–4.

[2] R. Duda, P. Hart, D. Stork, Pattern Classification, 2nd edition, Wiley Interscience, 2000.

[3] C. Chow, An optimum character recognition system using decision functions, IRE Transactions on Electronic Computers 6 (1957) 247–254.

[4] C. Chow, On optimum error and reject tradeoff, IEEE Transactions on Information Theory 16 (1970) 41–46.

[5] T. Ha, On Functional Relation between Recognition Error and Class-Selective Reject, Technical Report, Institute of Computer Science and Applied Mathematics, University of Berne, 1996.

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

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

[8] C. Frélicot, B. Dubuisson, A multi-step predictor of membership function as an ambiguity reject solver in pattern recognition, in: International Conference on Information Processing and Management of Uncertainty in Knowledge-based Systems, 1992, pp. 709–715.

[9] P. Foggia, C. Sansone, F. Tortorella, M. Vento, Multiclassification: reject criteria for the bayesian combiner, Pattern Recognition 32 (1999) 1435–1447.

[10] H. Mouchère, E. Anquetil, A unified strategy to deal with different natures of reject, in: 18th International Conference on Pattern Recognition, 2006, Hong-Kong, pp. 792–795.

[11] G. Fumera, F. Roli, G. Giacento, Reject option with multiple thresholds, Pattern Recognition 33 (2000) 2099–2101.

[12] D. Tax, R. Duin, Growing a multi-class classifier with a reject option, Pattern Recognition Letters 29 (2008) 1565–1570.

[13] L. Mascarilla, M. Berthier, C. Frélicot, A k-order fuzzy or operator for pattern classification with k-order ambiguity rejection, Fuzzy Sets and Systems 159 (2008) 2011–2029.
[14] B. Dubuisson, M. Masson, A statistical decision rule with incomplete knowledge about classes, Pattern Recognition 26 (1993) 155–165.
[15] R. Muzzolini, Y.-H. Yang, R. Pierson, Classifier design with incomplete knowledge, Pattern Recognition 31 (1998) 345–369.
[16] H. Ishibuchi, N. Nii, Neural networks for soft decision making, Fuzzy Sets and Systems 115 (2000) 121–140.
[17] L. S. Oliveira, R. Sabourin, F. Bortolozzi, C. Y. Suen, A methodology for feature selection using multiobjective genetic algorithms for handwritten digit string recognition, International Journal of Pattern Recognition and Aritificial Intelligence 17 (2003) 903–929.
[18] G. Fumera, F. Roli, Support vector machines with embedded reject option, in: 1st International Workshop on Pattern Recognition with Support Vector Machines, 2002, volume 2388, Niagara Falls, Canada, pp. 68–82.
[19] Y. Grandvalet, A. Rakotomamonjy, J. Keshet, S. Canu, Support vector machines with a reject option, in: D. Koller, D. Schuurmans, Y. Bengio, L. Bottou (Eds.), Advances in Neural Information Processing Systems 21, 2009, pp. 537–544.
[20] H. Ishibuchi, T. Nakashima, Fuzzy classification with reject options by fuzzy if-then rules, in: 7th IEEE International Conference on Fuzzy Systems, 1998, volume 3, Anchorage, Alaska, pp. 1452–1457.
[21] H. Le Capitaine, C. Frélicot, A new fuzzy 3-rules pattern classifier with reject options based on aggregation of membership degrees, in: International Conference on Information Processing and Management of Uncertainty, 2008, Malaga, Spain, pp. 473–480.
[22] T. Denoeux, A neural network classifier based on dempster-shafer theory, IEEE Transactions on Systems, Man and Cybernetics (Part A) 30 (2000) 131–150.
[23] C. Frélicot, On unifying probabilistic/fuzzy and possibilistic rejection-based classifiers, in: LNCS 1451, 1998, pp. 736–745.
[24] H.-J. Zimmerman, P. Zysno, Quantifying vagueness in decision models, European Journal of Operational Research 22 (1985) 148–158.
[25] C. D. Stefano, C. Sansone, M. Vento, To reject or not to reject: that is the question - an answer in case of neural classifiers, IEEE Transactions on Systems, Man and Cybernetics (Part C) 30 (2000) 84–94.
[26] C. Frélicot, Learning rejection thresholds for a class of fuzzy classifiers from possibilistic clustered noisy data, in: 7th International Fuzzy Systems Association World Congress, IFSA, 1997, volume 3, Prague, Czech Republic, pp. 111–116.
[27] C. Frélicot, L. Mascarilla, Reject strategies driven combination of pattern classifiers, Pattern Analysis and Applications 5 (2002) 234–243.
[28] C. Frélicot, H. Le Capitaine, Class-selective Rejection Rules based on the Aggregation of Pattern Soft Labels, Pattern Recognition: Recent Advances, Intech Publising, pp. 25–48.
[29] K. Menger, Statistical metrics, Proc. National Academy of Science USA 28 (1942) 535–537.
[30] R. Yager, Ordered weighted averaging aggregation operators in multicriteria decision making, IEEE Transactions on Systems, Man and Cybernetics 18 (1988) 183–190.
[31] H.-J. Zimmermann, P. Zysno, Latent connectives in human decision making, Fuzzy Sets and Systems 4 (1980) 37–51.
[32] M. Sugeno, Theory of fuzzy integrals and its applications, Ph.D. thesis, Tokyo Institute of Technology, 1974.
[33] C. Calvo, G. Mayor, R. Mesiar, Aggregation Operators: New Trends and Applications, Physica-Verlag, 2002.
[34] H. Le Capitaine, Aggregation operators for similarity measures. Application to ambiguity in pattern recognition, Ph.D. thesis, Univ. of La Rochelle, 2009.
[35] M. Grabisch, J. Marichal, R. Mesiar, E. Pap, Aggregation Functions, Number 127 in Encyclopedia of Mathematics and its Applications, Cambridge University Press, 2009.
[36] E. P. Klement, R. Mesiar, Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms, Elsevier, 2005.
[37] M. Grabisch, Fuzzy pattern recognition by fuzzy integrals and fuzzy rules, Pattern Recognition - From Classical to Modern Approaches, World Scientific, 2002, pp. 257–280.
[38] H. Le Capitaine, T. Batard, C. Frélicot, M. Berthier, Blockwise similarity in [0,1] via triangular norms and Sugeno integrals – application to cluster validity, in: 16th IEEE International Conference on Fuzzy Systems, 2007, London, England, pp. 835–840.
[39] H. Le Capitaine, C. Frélicot, Block-similarity of fuzzy tuples, Submitted to Fuzzy Sets and Systems (2010).
[40] A. Frank, A. Asuncion, UCI machine learning repository, 2010. University of California, Irvine, School of Information and Computer Sciences.
[41] W. Highleyman, Linear decision functions, with application to pattern recognition, Proc. of the IRE 50 (1962) 1501–1514.
[42] H. Le Capitaine, C. Frélicot, An optimum class-rejective decision rule and its evaluation, in: 20th International Conference on Pattern Recognition, 2010, Istanbul, Turkey, pp. 3312–3315.
[43] P. Bartlett, M. Wegkamp, Classification with a reject option using a hinge loss, Journal of Machine Learning Research 9 (2008).
[44] M. Yuan, B. Wegkamp, Classification methods with reject option based on convex risk minimization, Journal of Machine Learning Research 11 (2010) 111–130.
[45] S. Vanderlooy, E. Hüllermeier, A critical analysis of variants of the AUC, Machine Learning 72 (2008) 247–262.
[46] D. J. Hand, Measuring classifier performance: a coherent alternative to the area under the ROC curve, Machine Learning 77 (2009) 103–123.