

# A Class-Selective Rejection Scheme based on Blockwise Similarity of Typicality Degrees

Hoel Le Capitaine and Carl Frélicot  
MIA Laboratory, University of La Rochelle, France  
{hlecap01},{cfrelico}@univ-lr.fr

## Abstract

*Overlapping classes and outliers can significantly decrease a classifier performance. We address here the problem of giving a classifier the ability to reject some patterns either for ambiguity or for distance in order to improve its performance. Given a set of typicality degrees for a pattern to be classified, we use an operator based on triangular norms and a discrete Sugeno integral to quantify their blockwise similarities. We propose a new class-selective rejection scheme which uses this operator outputs. We present the resulting algorithm which allows to assign a pattern to zero, one or several classes, and show its efficiency on real data sets.*

## 1. Introduction

The problem of aggregating collections of numerical or ordinal 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 degrees related to the given classes (posterior probabilities, membership values, ...). It has been proved that the misclassification risk significantly be reduced by allowing a classifier to reject extraneous and/or ambiguous patterns [2, 3, 5]. Thus, a classifier with reject options allows to assign a pattern to zero (distance rejection), one (exclusive classification) or several (ambiguity rejection) classes. Among the possible approaches, the fuzzy one has received more attention in the last few decades because of its ability to manage imprecise and/or incomplete data [4]. In this framework, we propose a new classification scheme with reject options, based on an operator which aggregates class-degrees of typicality of a pattern to be classified.

## 2. Fuzzy Aggregation Operators

Let us recall basic definitions of fuzzy operators that will be used to combine the values of interest, i.e. the pattern class-degrees of typicality. Depending on properties, aggregation functions can be classified into several categories: conjunctive, disjunctive, compensatory, and so on. We restrict on conjunctive and disjunctive functions. By definition, the output of a conjunctive operator is lower or equal than the minimum value, whereas the output of a disjunctive operator is greater or equal than the maximum value. Beyond these operators, we choose to use the triangular norms because of their ability to generalize the logical AND and OR crisp operators to fuzzy sets, see [7] 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. Thus, a t-norm  $\top$  is conjunctive and the minimum operator  $\wedge$  is the greatest t-norm. Alternatively, a triangular conorm (or t-conorm) is the dual binary operation  $\perp : [0, 1]^2 \rightarrow [0, 1]$  having the same properties except the latter: its neutral element is 0. Thus, a t-conorm  $\perp$  is disjunctive and the maximum operator  $\vee$  is the lowest t-conorm. Typical examples of dual couples (t-norm, t-conorm) that will be used in the sequel are given in Table 1.

**Table 1. Typical triangular norm couples**

Standard	$a \top_S b = \min(a, b)$
	$a \perp_S b = \max(a, b)$
Algebraic	$a \top_A b = a b$
	$a \perp_A b = a + b - a b$
Hamacher	$a \top_H b = \frac{ab}{\gamma + (1-\gamma)(a+b-ab)}$
	$a \perp_H b = \frac{a+b+(\gamma-2)ab}{1+(\gamma-1)ab}$

We will use another fuzzy aggregation operator, the Sugeno integral in its discrete form. It computes the mean value of a function with respect to a fuzzy mea-

sure  $m$ , which is a non-additive measure of uncertainty, i.e. more general than a possibility one and therefore a probability one. The integral of a function  $\mu$  is defined by

$$\mathcal{S}_m = \bigvee_{i=1}^n \mu(x_i) \wedge m(A_{(i)}) \quad (1)$$

where  $A_{(i)} = \{x_{(i)}, \dots, x_{(n)}\}$  with respect to a permutation so that  $\mu(x_{(i)}) \leq \dots \leq \mu(x_{(n)})$ . This integral is widely used in decision making, and in particular for pattern recognition [4] because of its ability to model some kind of interaction between features describing a pattern  $x$ .

### 3. The Class-selective Rejection Scheme

#### 3.1. Classifier design

Let  $\Omega = \{\omega_1, \dots, \omega_c\}$  be a set of  $c$  classes and  $x$  an unknown pattern described by  $p$  features. Classifier design aims at defining rules that can associate  $x \in \mathbb{R}^p$  with one class of  $\Omega$ . It generally consists of two steps  $L$  (*labeling*) and  $H$  (*hardening*):

- $L : x \mapsto \mu(x) = {}^t(\mu_1(x), \dots, \mu_c(x)) \in \mathcal{L}_{\bullet c}$ , depending on the mathematical framework the classifier relies on, e.g.  $\mathcal{L}_{pc} = [0, 1]^c$  for degrees of typicality or  $\mathcal{L}_{fc} = \{\mu(x) \in \mathcal{L}_{pc} \mid \sum_{i=1}^c \mu_i(x) = 1\}$  for posterior probabilities and membership degrees.

There exists many ways to compute labels, but we do not address the labelling problem in this paper and we will use the typicality measure defined as:

$$\mu_i(x) = \frac{\alpha}{\alpha + d^2(x, v_i)} \quad (2)$$

where  $\alpha$  is a user-defined parameter,  $d$  a distance, and  $v_i$  a prototype of the class  $\omega_i$  obtained from a learning set of patterns. It has been shown through empirical studies [9] that (2) is a good model for membership functions that model vague concepts or classes.

- $H : \mu(x) \mapsto h(x) = {}^t(h_1(x), \dots, h_c(x)) \in \mathcal{L}_{hc}$ , where  $\mathcal{L}_{hc} = \{h(x) \in \mathcal{L}_{fc} \mid h_i(x) \in \{0, 1\}\}$ .

We address the hardening problem because this step, which often reduces to the class of maximum label selection, is in charge of the decision making.

#### 3.2. Reject options and the proposed class-selective scheme

As defined,  $H$  is an exclusive rule which is not efficient in practice because it supposed that:

- i*)  $\Omega$  is exhaustively defined (closed-world assumption),
- ii*) classes do not overlap (separability assumption).

Such untrue assumptions can lead to very undesired decisions. In many real applications, it is more convenient to with-hold making a decision than making a wrong assignment, e.g. in medical diagnosis where a false negative outcome can be much more costly than a false positive. Reject options have been proposed to overcome these difficulties and to reduce misclassification risk. The first one, called *distance rejection* [3], is dedicated to outlying patterns. If  $x$  is far from all the class prototypes, this option allows to assign it to no class. The second one, called *ambiguity rejection*, allows to assign inlying patterns to several or all classes [2, 5]. If  $x$  is close to two or more class prototypes, it is associated with the corresponding classes. Formally, including reject options consists in modifying  $H$  such that  $h(x)$  can take values in the set of vertices of the unit hypercube  $\mathcal{L}_{hc}^c = \{0, 1\}^c$  instead of the exclusive subset  $\mathcal{L}_{hc} \subset \mathcal{L}_{hc}^c$ . Different strategies can be adopted to handle these options at hand, but they all lead to a three types decision system: distance rejection when  $h(x) = {}^t(0, \dots, 0) = \underline{0}$ , exclusive classification when  $h(x) \in \mathcal{L}_{hc}$ , ambiguity rejection when  $h(x) \in \mathcal{L}_{hc}^c \setminus \{\mathcal{L}_{hc} \cup \underline{0}\}$ .

For any pattern  $x$  to be classified, given its label vector  $\mu(x)$  from  $L$  by (2), sorted in descending order  $\mu_1(x) \geq \mu_2(x) \geq \dots \geq \mu_c(x)$ , we propose a two-steps class-selective scheme for  $H$  as follows :

- 1) test for distance rejection :  $h(x) = \underline{0}$  if  $\mu_1(x) < s$ , where  $s$  is a user-defined threshold
- 2) if  $x$  is not distance rejected, assign it to a (sub)set of selected classes of cardinality  $k \in \{1, \dots, c\}$  ; thus it is exclusively classified if  $k = 1$  or ambiguity rejected between the selected classes if  $k > 1$ .

For the (sub)set of classes selection problem, we propose to use the  $\Phi_{j,k}$  measure introduced in [8] for another purpose. Assuming  $\mu$  to be a sorted  $c$ -tuple,  $\mu_1 \geq \mu_2 \geq \dots \geq \mu_c$ , we defined an operator based on triangular norms and the Sugeno integral which quantifies the similarity of the block of values  $\{\mu_j, \dots, \mu_k\}$ :

$$\Phi_{j,k}(\mu) = \begin{cases} \frac{\bigwedge_{i=\frac{k+j}{2}}^k \mu_i \top \mathcal{N}_\lambda(i,k)}{\bigwedge_{i=\frac{k+j}{2}}^j \mu_i \top \mathcal{N}_\lambda(i,j)} & \text{if } k-j \text{ is even} \\ \frac{\bigwedge_{i=\frac{k+j+1}{2}}^k \mu_i \top \mathcal{N}_\lambda(i,k)}{\bigwedge_{i=\frac{k+j-1}{2}}^j \mu_i \top \mathcal{N}_\lambda(i,j)} & \text{if } k-j \text{ is odd} \end{cases} \quad (3)$$

where  $\mathcal{N}_\lambda(i, l)$  is a gaussian kernel defined by:

$$\mathcal{N}_\lambda(i, l) = \exp \frac{-(i-l)^2}{\lambda} \quad (4)$$

The resolution parameter  $\lambda$  controls the area of influence: when  $\lambda \rightarrow 0$ , the kernel becomes a dirac centered in  $l$ , and when  $\lambda \rightarrow \infty$ , the kernel becomes the constant value 1. Therefore, the contribution of the intermediate values  $\mu_{j+1}, \dots, \mu_{k-1}$  to  $\Phi_{j,k}(\mu)$  is small if  $\lambda$  is close to zero and increases with  $\lambda$ . This means that increasing  $\lambda$  will not make two consecutive  $\mu_i$ 's more similar but may increase the similarity of blocks of larger size.

Since a high value of  $\Phi_{1,k}(\mu(x))$  reveals that the  $k$  highest labels have similar values, then  $x$  can be associated with the corresponding classes. We propose to use an iterative scheme in order to find  $k$ , leading to the following second part for the  $H$ -step:

- 2) for  $i$  varying from 1 to  $c$ , set  $h_i(x) = 1$  when  $\Phi_{1,i}(\mu(x)) \geq t$ , where  $t$  is a user-defined threshold

Note that  $\Phi_{1,1}(\mu(x))$  is always greater than  $t, \forall t \in [0, 1]$ , because  $\Phi_{1,1}(\mu(x)) = \frac{\mu_1(x)}{\mu_1(x)} = 1$  for any t-norm couple. This ensure that at least one class is selected, the one which corresponds to the maximum of typicality degree, i.e. the one selected by the optimum classification rule in the sense of Chow [2]. In particular, if  $t$  is set to 1, there is no ambiguity rejection. The class-selective rejection scheme presented in Algorithm 1 can be compared to the rule proposed by Ha [5].

---

**Algorithm 1:** hardening step  $H : L_{pc} \rightarrow L_{hc}^c$

---

**Data:** a sorted vector  $\mu$  of typicality degrees, a membership threshold  $s$ , an ambiguity threshold  $t$

**Result:** a vector  $h$  of class-selective assignments

**begin**

```

  if  $\mu_1(x) < s$  then
     $h_i(x) \leftarrow 0 \forall i = 1, c$ 
  if  $\sum_{i=1}^c h_i(x) > 0$  then
    for  $i \leftarrow 1$  to  $c$  do
      if  $\Phi_{1,i}(\mu(x)) \geq t$  then
         $h_i(x) \leftarrow 1$ 
      else
         $h_i(x) \leftarrow 0$ 

```

```

  return  $h(x)$ 

```

**end**

---

## 4. Experiments and Results

To validate the efficiency of the proposed class-selective scheme, we present some results obtained by a resubstitution procedure on well-known real datasets from the UCI Machine Learning Repository [1] whose characteristics (number  $p$  of features, number  $c$  of classes, degree of overlap) are summarized in Table 2. The classification performance of some usual su-

**Table 2. Datasets used in the experiments.**

data	$p$	$c$	overlap
<i>iris</i>	4	3	slight overlap, 2 classes
<i>pima</i>	8	2	medium overlap, 2 classes
<i>vowel</i>	10	11	slight overlap, by pairs
<i>glass</i>	9	6	strong overlap, up to 5 classes

pervised classifiers with no reject options are given in Table 3 for comparison purpose : the Quadratic Bayes ( $QB$ ) rule, the Nearest Neighbor ( $1-NN$ ) rule and the Maximum Classifier ( $MC$ ) based on typicality degrees in the feature space computed by (2) with  $\alpha = 1$  and  $d^2(x, v_i) = t(x - v_i)\Sigma_i^{-1}(x - v_i)$  where the covariance matrix  $\Sigma_i$  and the center  $v_i$  of the class  $\omega_i$  are estimated from the (learning) dataset. The same labeling  $L$  is used in the remaining experiments. We compare

**Table 3. Error ( $E$ ) and Correct ( $C$ ) rates of some usual classifiers with no reject options.**

data	%	QB	1-NN	MC
<i>iris</i>	$E$	2	4.67	2
	$C$	98	95.33	98
<i>pima</i>	$E$	25.39	29.53	32.55
	$C$	74.61	70.47	67.45
<i>vowel</i>	$E$	4.58	9.77	8.08
	$C$	95.42	90.23	91.92
<i>glass</i>	$E$	31.10	30.64	28.5
	$C$	68.90	69.36	71.5

the performance of the proposed scheme to two class-selective rejection ones found in the literature. In [5], Ha has proposed to set the optimum cardinality of the set of selected classes by:

$$k_{HA} = \min\{k \in \{1, \dots, c\} | \mu_{k+1}(x) \geq t\} \quad (5)$$

where  $t$  is a user-defined ambiguity threshold whose role is similar to the one in our scheme. Since this selection can lead to unnatural classification areas, Horiuchi has proposed in [6] to use:

$$k_{HO} = \min\{k \in \{1, \dots, c\} | \mu_k(x) - \mu_{k+1}(x) \geq t\} \quad (6)$$

Since these two selection schemes do not allow distance rejection, we set  $s = 0$  in the experiments so that re-

sults can be compared. Note moreover that there are no outliers in the considered datasets.

The results obtained by a resubstitution procedure are given in Table 4 where  $HA$  and  $HO$  stand for the Ha and the Horiuchi schemes,  $\Phi_S$ ,  $\Phi_A$  and  $\Phi_H$  stand for the proposed scheme using the different triangular norms of Table 1 (with  $\gamma = 0$  for the Hamacher one). The ambiguity threshold  $t$  is (coarse) tuned so that the error rate is as much as possible equal for each rejection scheme. For the tested datasets, a very little influence of the resolution parameter  $\lambda$  setting was observed and we chose to report the results with  $\lambda = 10$ . As expected, rejecting patterns leads to decrease the error rate compared to classifiers with no reject option (Table 3). Whatever the triangular norms, the proposed class-selective rejection scheme outperforms the ones proposed by Ha and Horiuchi with respect to the correct classification rate. This efficiency is due to the fact that the ratio of membership degrees is more suited than a simple difference to ambiguity rejection: the same difference  $\varepsilon$  between two low values and two high values with Horiuchi's method will not be discriminated. On the other side, with Ha's method, the most reliable membership degree is not taken into account, which leads to unnatural decisions. We constructed the operator such that our method reap the benefits of both Ha's and Horiuchi's schemes.

**Table 4. Reject ( $R$ ), Error ( $E$ ) and Correct ( $C$ ) rates of rejection schemes.**

data	%	$HA$	$HO$	$\Phi_S$	$\Phi_A$	$\Phi_H$
<i>iris</i>	$R$	0.67	1.33	0.67	0.67	0.67
	$E$	2	1.33	1.33	1.33	1.33
	$C$	97.33	97.33	98	98	98
<i>pima</i>	$R$	5.60	4.30	3.52	4.30	4.04
	$E$	30.47	30.21	30.59	30.08	30.24
	$C$	63.93	65.49	65.89	65.62	65.72
<i>vowel</i>	$R$	15.15	9.49	8.89	8.79	8.79
	$E$	4.34	4.34	4.34	4.34	4.34
	$C$	80.51	86.16	86.77	86.87	86.87
<i>glass</i>	$R$	23.83	8.88	7.94	8.17	8.02
	$E$	22.90	22.90	22.90	22.90	22.90
	$C$	53.27	68.22	69.16	68.93	69.08

## 5. Conclusion

In this paper, a new class-selective rejection scheme is proposed. It consists of two sequential steps dealing with both reject options: distance rejection for outliers and ambiguity rejection for inliers. The latter option is based on an operator which aggregates the class-degrees of typicality of the pattern to be classified. This operator measures the blockwise similarity

of sorted degrees by combining them with triangular norms and the Sugeno integral. Experimental results we obtained on well-known real datasets show that the proposed scheme achieves better recognition accuracy than other similar class-selective rules. Due to lack of place, we did not discuss the choice of the triangular norms, for which we have some theoretical results according to the nature of the degrees (in  $\mathcal{L}_{pc}$ ,  $\mathcal{L}_{fc}$ ). We will address this problem in a forthcoming paper. Future works will concern the definition of blockwise similarity of numbers through fuzzy residual implication. We think this could generalize the concept of ambiguity for pattern recognition problems.

## References

- [1] C.L. Blake and C.J. Merz. UCI repository of machine learning databases, 1998. Dept. of Information and Computer Science, University of California, Irvine, CA, <http://archive.ics.uci.edu/ml/>.
- [2] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.
- [3] B. Dubuisson and M. Masson. A statistical decision rule with incomplete knowledge about classes. *Pattern Recognition*, 26(1):155–165, 1993.
- [4] M. Grabisch. *Pattern Recognition - From Classical to Modern Approaches*, chapter Fuzzy pattern recognition by fuzzy integrals and fuzzy rules, pages 257–280. World Scientific, 2002.
- [5] T. M. Ha. The optimum class-selective rejection rules. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(6):608–615, 1997.
- [6] T. Horiuchi. Class-selective rejection rule to minimize the maximum distance between selected classes. *Pattern Recognition*, 31(10):1579–1588, 1998.
- [7] E. P. Klement and R. Mesiar. *Logical, Algebraic, Analytic, and Probabilistic Aspects of Triangular Norms*. Elsevier, 2005.
- [8] H. Le Capitaine, T. Batard, C. Frélicot, and M. Berthier. Blockwise similarity in  $[0,1]$  via triangular norms and sugeno integrals – application to cluster validity. In *IEEE International Conference on Fuzzy Systems*, pages 835–840, 2007.
- [9] H. J. Zimmerman and P. Zysno. Quantifying vagueness in decision models. *European Journal of Operational Research*, 22(2):148–158, 1985.