On Unifying Probabilistic / Fuzzy and Possibilistic Rejection-Based Classifiers

Carl Frélicot

Université de La Rochelle, Laboratoire d'Informatique et d'Imagerie Industrielle,
17042 La Rochelle Cedex 1, FRANCE
Phone: +33 546 458 234, Fax: +33 546 458 242, e-mail: cfrelico@gi.univ-lr.fr

Abstract. This paper aims at unifying the presentation of two-fold rejection-based pattern classifiers. We propose to define such a classifier as a couple of labelling and hardening functions which are independent in some way. Within this framework, crisp and probabilistic / fuzzy rejection-based classifiers are shown to be particular cases of possibilistic ones. Classifiers with no reject option remains particular cases of rejection-based ones. Examples of so-defined classifiers are presented and their ability to deal with the reject problem is shown on artificial and real data sets.

1 Pattern Classification

The pattern classification problem can be defined as follows. Let $x = \left(x_1, x_2, ..., x_p\right)^t$ be a pattern described by p features, and $\omega = \{\omega_1, \omega_2, ..., \omega_c\}$ be a set of c classes. A classifier performs a mapping: $\Re^p \rightarrow L_{.c}$ a class labels space, $x \mapsto l(x) = \left(l_1(x), l_2(x), ..., l_c(x)\right)$. Three sets of class label vectors in $\Re^c$ may be defined as in [1]:

1. $L_{pc} = [0,1]^c$, i.e. the unit hypercube in $\Re^c$;
   $l(x) \in L_{pc}$ is a *possibilistic* label vector,
   e.g. $l(x) = (0, 0.7, 0.5)^t$.
   Some authors exclude the origin of the unit hypercube from L_{pc}, e.g. in [1]. Since a possibilistic value represents a degree of typicality, we prefer not to do so.
2. $L_{fc} = \{l \in L_{pc} : \sum_{i=1}^{c} l_i(x) = 1\}$;
   $l(x) \in L_{fc}$ may bo either a *probabilistic* or a *fuzzy* label vector (depending on the way it has been generated),
   e.g. $l(x) = (0.1, 0.6, 0.3)^t$.
3. $L_{hc} = \{l \in L_{fc} : l_i(x) \in \{0,1\}, \forall i = 1, c\}$;
   $l(x) \in L_{hc}$ is a *hard* (or *crisp*) label vector,
   e.g. $l(x) = (0, 1, 0)^t$.

Depending of the nature of $l(x)$, any function: $x \mapsto l(x)$ is a *possibilistic*, a *probabilistic*, a *fuzzy*, a *hard* classifier respectively. Note that $L_{hc} \subset L_{fc} \subset L_{pc}$.

Hard classification often is the final goal of most pattern recognition processes. That is the reason why we prefer to define a classifier as a couple of functions (L, H) defined as follows.

Definition 1. Any function L: $\Re^p \rightarrow L_{.c}$ is called a *labelling function.*

Definition 2. Any function H: $L_{.c} \rightarrow L_{hc}$ is called a *hardening function.*

Definition 3. We shall call a *classifier* any couple (L, H).

In both fuzzy and possibilistic contexts for pattern classification, labelling is obtained using membership functions. Therefore, we use the following notation: $\mu_i(x)$ for labels generated by L, $l_i(x)$ for labels resulting from H. Figure 1 summarizes such a general classification process.

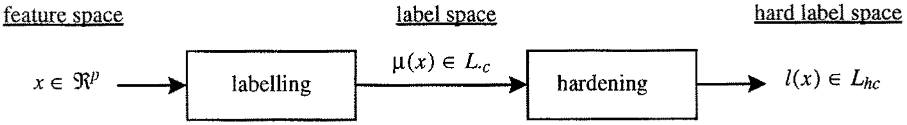

Fig. 1. Classification process

Most of classifiers developed so far either are or can be formulated as above, e.g. in [7]:

- the *Bayes rule*

Labelling Part $\quad L : \Re^p \rightarrow L_{.c} = L_{fc},\ x \mapsto \mu(x)$
given prior probabilities $P(\omega_i)$ summing up to one, compute the class conditional densities $P\left(x|\omega_j\right)$ and the mixture density $P(x) = \sum_{j=1}^{c} P\left(x|\omega_j\right) P\left(\omega_j\right)$

- $\mu_i(x) = P\left(\omega_i|x\right) = \frac{P(x|\omega_i)P(\omega_i)}{P(x)}$

Hardening Part $\quad H : L_{fc} \rightarrow L_{hc},\ \mu(x) \mapsto l(x)$

- $j = argmax_i \mu_i(x)$
- $l_i(x) = \delta_{ij}$

- the *k-NN rule*

Labelling Part $\quad L : \Re^p \rightarrow L_{.c} = L_{fc}$ or L_{hc} $(k = 1),\ x \mapsto \mu(x)$
given $N(x)$ the set of $k - NN$ of x, with respect to a distance measure d on $\Re^p$

- $\mu_i(x) = \frac{k_i}{k}$
where k_i is the number of NN issued from ω_i

Hardening Part $\quad H : L_{fc} \rightarrow L_{hc},\ \mu(x) \mapsto l(x)$

- $j = argmax_i \mu_i(x)$
- $l_i(x) = \delta_{ij}$

Another motivation for distinguishing both parts of a classifier is the kind of parameters that may be needed for the classification of a pattern whatever the classifier is: *labelling parameters* and *hardening parameters*. Most of classifiers do not need any parameter for their hardening part, but only for their labelling part, e.g. for the classifiers above: mean vectors, covariance matrices and prior probabilities for the Bayes rule, number k of neighbors for the $k-NN$ rule. Furthermore, some classifiers whose labelling part is different can share the same hardening one, e.g. the classifiers above. In some way, both parts are independent.
This formulation appears to be very useful for describing rejection-based classifiers.

2 Pattern Classification and Rejection

In the previous classification scheme, L_{hc} expresses the feature space partitioning into c mutually exclusive areas $\{\Omega_1, \Omega_2, ..., \Omega_c\} = \Omega$ associated with the classes of ω. This is clearly not efficient in many applications because: $\Omega = \Re^p$ whereas the definition of ω is rarely complete, and Ω_i areas boundaries are sharp whereas classes may partially overlap. Reject options may be used to overcome both limits and therefore reduce the misclassification risk.
Two kinds of rejection are commonly accepted: *distance rejection* which allows not to classifying a pattern in any class of ω and *ambiguity rejection* which allows to classify a pattern in several (or all) classes of ω. Classifiers including these reject options result in partitioning the feature space into $(c+2)$ areas, as shown in Figure 2: $\Omega \cup \Omega_0 \cup \Omega_a$, where Ω_0 and Ω_a are associated with a *distance reject* class ω_0 and an *ambiguity reject* class ω_a respectively.

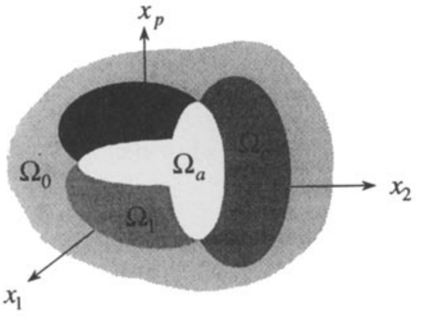

Fig. 2. Classes areas

Let us see how such classifiers can be formulated as a couple (L, H) of labelling and hardening functions as well. In order to reflect the classification areas, the hard label space must include:

1. $L_{hc} \Leftrightarrow \Omega$
2. $L_{hc}^0 = \{l \in L_{pc} \ : l_i(x) \in \{0,1\}, \sum_{i=1}^{c} l_i(x) = 0\} = \{(0,0,...,0)^t\} \Leftrightarrow \Omega_0$

3. $L^{1,a}_{hc} = L^1_{hc} + L^a_{hc} \Leftrightarrow \Omega_a$
with $L^1_{hc} = \{l \in L_{pc} \;: l_i(x) \in \{0,1\}, \sum_{i=1}^c l_i(x) = c\} = \{(1,1,...,1)^t\}$
and $L^a_{hc} = \{l \in L_{pc} \;: l_i(x) \in \{0,1\}, 1 < \sum_{i=1}^c l_i(x) \leq c\}$

Therefore, we can define a rejection-based classifier as a couple of functions (L, H), where H is defined as follows:

Definition 4. Any function H: $L_{.c} \rightarrow L^{0,1,a}_{hc}$, where $L^{0,1,a}_{hc} = L_{hc} + L^0_{hc} + L^{1,a}_{hc}$ is the set of 2^c vertices of L_{pc} is called a rejection-based *hardening function.*

The different subsets of L_{pc} are shown in Figure 3. Superscripts ^{0}, ^{1} and ^{a} denote particular cases of rejection involved, i.e. distance rejection, total ambiguity rejection and partial ambiguity rejection respectively. We distinguish total ambiguity and partial ambiguity because some classifiers do not deal with both cases. Classifiers dealing with partial ambiguity rejection are said to be *class-selective* as in [8].

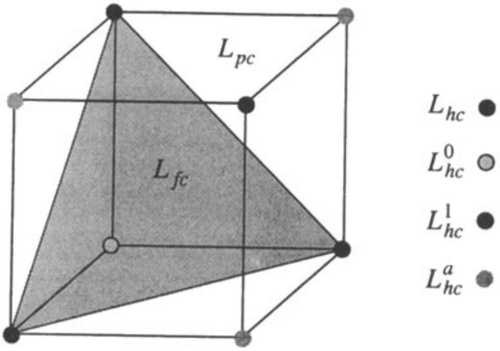

Fig. 3. Hard label spaces

In the definition above, we did not mention any condition on the label space $L_{.c}$ the label vectors $\mu_(x)$ have to be issued from. Let us discuss this point with respect to the different pattern classifier design approaches. When the L function takes values in L_{pc}, it is easy to build up a rejection-based H function taking values in $L^{0,1,a}_{hc}$ because $L^{0,1,a}_{hc} \subset L_{hc}$. The H function generally consists of two sequentially conditioned steps whose strategy is quite different:

1. *accept-first,* in which both kinds of rejection are considered independently,
2. *reject-first,* in which both kinds of rejection are considered dependently.

Whatever the strategy is, so defined rejection-based possibilistic classifiers need hardening parameters whereas non rejection-based classifiers do not. These parameters generally consist of thresholds on membership function values $\mu_i(x)$ issued from the labelling part. The number of hardening parameters can vary from 2 (one for each kind of rejection) to $2c$ (if they are class-dependant). Hardening parameters can be learned from a learning set as well as labelling ones, e.g. for a class of possibilistic classifiers in [5].

Table 1. Accept-first strategy

	Test for	**Conclude** $l(x) \in$
Step 1	•distance rejection (understood *vs* acceptation)	•L_{hc}^{0} otherwise
Step 2	if accepted, •exclusive classification (understood *vs* ambiguity rejection)	•L_{hc} else $L_{hc}^{1,a}$

Table 2. Reject-first strategy

	Test for	**Conclude** $l(x) \in$
Step 1	•test for exclusive classification (understood *vs* rejection)	•L_{hc} otherwise
Step 2	if rejected, •test for distance rejection (understood *vs* ambiguity rejection)	•L_{hc}^{0} else $L_{hc}^{1,a}$

If the L function takes values in L_{fc}, it is not so easy to design fuzzy / probabilistic classifiers that include both kinds of rejection because L_{fc} do not contain $L_{hc}^{0,1,a}$ but only L_{hc}. The use of a possibilistic rejection-based classifier with fuzzy / probabilistic labels unfortunately leads to undesirable results such as confusion of Ω_0 and Ω_a or Ω_0 and Ω, depending on the degree of separability of the classes and on the hardening parameters values. If a classification problem needs the design of a classifier which include both kinds of rejection (e.g. overlapping classes and possibility of outliers), and if the available labels are in L_{fc}, it is more appropriate to follow one of the listed approaches:

1. relaxing the normalization condition $\sum_{i=1}^{c} \mu_i(x) = 1$ and therefore assuming a L function taking values in L_{pc}
2. not directly basing *Step 1* on $\mu(x) \in L_{fc}$, whatever the strategy is, but on a function $f(x)$ taking values in $\Re^+$
3. designing a classifier which either exclusively classify a pattern or simply reject it without distinguishing between distance rejection and ambiguity rejection ;
   this approach corresponds to a reject-first strategy without *Step 2*
4. considering a unique type of rejection ;
   this approach might corresponds to a accept-first strategy without *Step 1* or to a reject-first strategy without *Step 2* depending on the involved reject option.

3 Examples of classifiers

Every classifier including both reject options can be formulated as a couple of functions (L, H) with the mentioned recommendations.
A label vector $l(x) \in L_{hc}^{0,1,a}$ being a boolean vector, it is convenient to express its computation as the result of a relational operation. The notation below assumes that if both operands are vectors of same dimension, the operation is performed component by component and the result is 0 for false or 1 for true. We allow one or both operands to be scalars and implicitly transform them into constant vectors whose dimension agrees with $l(x)$. Consequently, we allow $l(x)$ to be a logical expression in a test, assuming that $l(x)$ is true if it is not a zero vector.

3.1 A Possibilistic Classifier

The characteristics of the first classifier we p‥sent are: possibilistic labelling, reject-first strategy, class-selective ambiguity rejection. It is a modified version of a classifier we proposed [6].

Labelling Part $\quad L : \Re^p \to L_{pc},\ x \mapsto \mu(x)$

given class prototypes (or estimated on a learning set), e.g. class centers m_i and covariance matrices Σ_i,

- $\mu_i(x) = \frac{1}{1+d(x,\omega_i)}$, with $d^2(x, \omega_i) = (x - m_i)^t\, \Sigma_i^{-1} (x - m_i)$

Note that m_i and Σ_i being labelling parameters, they have to be known or estimated on a learning set. The used labelling function could be replaced by another one without changing the hardening function, which is:

Hardening Part $\quad H : L_{fc} \to L_{hc}^{0,1,a},\ \mu(x) \mapsto l(x)$

given a reject vector $\mu_r = \left(\mu_{r,1}, \mu_{r,2}, ..., \mu_{r,c}\right)^t$ and a distance vector $\mu_0 = \left(\mu_{0,1}, \mu_{0,2}, ..., \mu_{0,c}\right)^t$,

- $M = (M_1, M_2, ..., M_c)^t$, where $M_i = \max_{j \neq i} \mu_j(x)$
- $l(x) = \left(\mu(x) \geq \mu_r + M\right)$

```
1.  if $l(x)$, then                                  (*Step 1* ; $l(x) \in L_{hc}$)
2.          else $l(x) = \left(\mu(x) > \mu_0\right)$           (*Step 2* ; $l(x) \in L^0_{hc}$ or $L^{1,a}_{hc}$)
    endif
```

It is worthy of note that the hardening parameters $\mu_{r,i}$ and $\mu_{0,i}$ can be class-independent ; they are in [0, 1]. The more $\mu_{r,i}$ is, the less exclusively classified and the more rejected the patterns are. Given $\mu_{r,i}$ parameters, the more $\mu_{0,i}$ are, the less ambiguity rejected and the more distance rejected the patterns are. The classifier reduces to an exclusive one, i.e. with no reject options, when the hardening parameters are all set to zero.

3.2 A Probabilistic Classifier

The second classifier associates the forerunning approach of rejection by Chow [2], the work by Ha [8] and Dubuisson's one [3]. The characteristics of the resulting classifier are: probabilistic labelling, accept-first strategy (with *Step 1* consisting in thresholding $f(x)$ instead of $\mu(x)$), class-selective ambiguity rejection.

Labelling Part $L : \Re^p \to L_{fc},\ x \mapsto \mu(x)$

given prior probabilities $P(\omega_i)$ summing up to one, compute the class conditional densities $P\left(x|\omega_j\right)$, e.g. gaussian ones, and the mixture density $P(x) = \sum_{j=1}^{c} P\left(x|\omega_j\right) P\left(\omega_j\right)$

- $\mu_i(x) = P\left(\omega_i|x\right) = \frac{P(x|\omega_i)P(\omega_i)}{P(x)}$

Under gaussian assumption, the labelling parameters are the same as the previous ones, i.e. m_i and Σ_i.

Hardening Part $H : L_{fc} \to L^{0,1,a}_{hc},\ \mu(x) \mapsto l(x)$

given a density threshold s and an ambiguity reject threshold t,

- $l(x) = \left(P(x) \geq s\right)$

```
1.  if $\neg l(x)$, then                                  (*Step 1* ; $l(x) \in L^0_{hc}$)
2.          else $l(x) = \left(\mu(x) > t\right)$                (*Step 2* ; $l(x) \in L_{hc}$ or $L^{1,a}_{hc}$)
                 if $\neg l(x)$, then use the Bayes rule
                                 else
                 endif
    endif
```

It has been proven in [8] that the ambiguity threshold t should be in $\left[0, \frac{1}{2}\right]$. The density threshold s should be small enough in order to allow the probability for a pattern to belong to the distance reject area to be small, as suggested in [4]. The classifier reduces to the Bayes rule, when the hardening parameters s and t are set to 0 and $\frac{1}{2}$ respectively.

4 Experimental Results

4.1 Classification Areas

As an illustration, we present the classification areas both presented classifiers produce in both cases of overlapping and well-separated classes. Figure 4 shows the used learning sets composed of $c = 3$ classes of artificial gaussian two-dimensional samples with unit standard deviation.

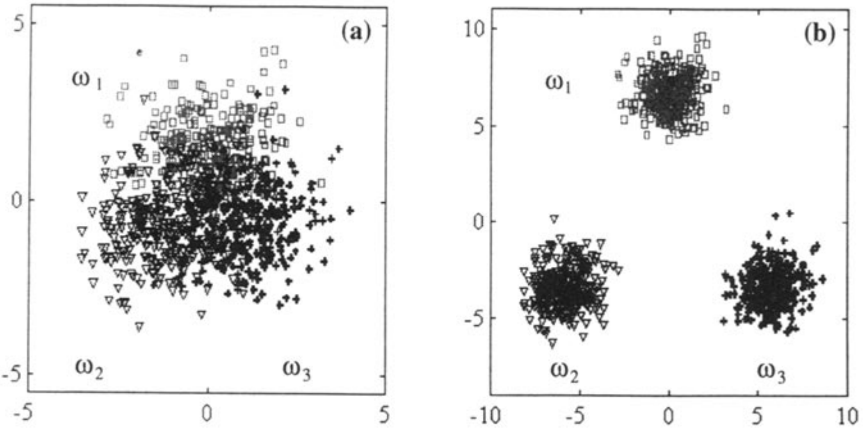

Fig. 4. Learning sets: overlapping classes (a), separated classes (b)

Figure 5 shows the classification areas obtained using the possibilistic classifier. It is worthy of note that for both cases (a) and (b), the hardening parameters were set to the same values ($\mu_{r,i} = \mu_r = 0.1$ and $\mu_{0,i} = \mu_0 = 0.2$). As expected, the proposed classifier leads to satisfactory areas, in particular in the case of well-separated classes for which no ambiguity area is performed.

The classification areas provided by the probabilistic classifier are shown in Figure 6. As well as the previous classifier, the probabilistic one gives excellent results. Again, the hardening parameters were set to the same values ($s = 0.001$ and $t = 0.1$) in both cases (a) and (b).

4.2 Classification Performance

Another interesting issue is the classification performance with respect to rejection abilities of the classifiers. We have tested both presented classifiers on the well-known *Iris* data set consisting of 150 patterns described by $p = 4$ features, divided in $c = 3$ classes of 50 patterns each. These data are such as class ω_1 is well separated from the two other ones whereas ω_2 and ω_3 slightly overlap.

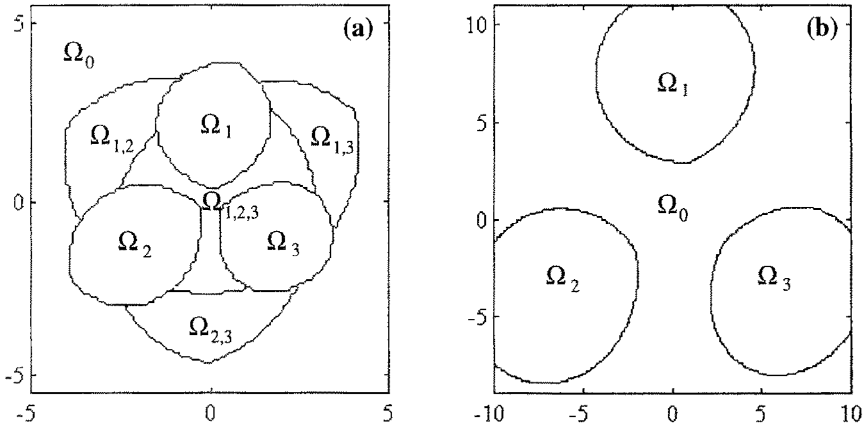

Fig. 5. Possibilistic classification areas: overlapping classes (a), separated classes (b)

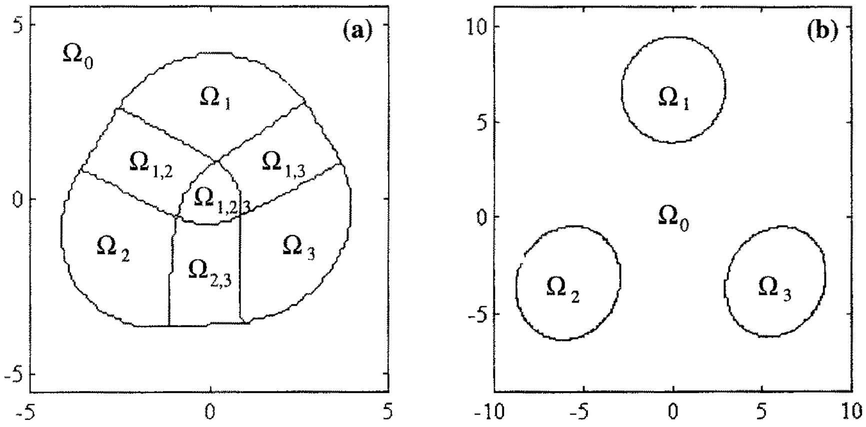

Fig. 6. Probabilistic classification areas: overlapping classes (a), separated classes (b)

A 1-fold cross-validation (*leave-one-out*) procedure has been used in order to estimate the different probabilities. $\hat{P}_c$ (correct classification), $\hat{P}_a$ (ambiguity rejection), $\hat{P}_e$ (error or misclassification) and $\hat{P}_0$ (distance rejection) by the empirical rates. Table 3 shows the results obtained. The hardening parameters have been set to $\mu_{r,i} = \mu_r = 0.1$ and $\mu_{0,i} = \mu_0 = 0.2$ for the possibilistic classifier and to $s = 0$ and $t = 0.1$ for the probabilistic one.

For comparison, the probabilities estimates obtained using the Bayes rule and the 1-NN one are shown in Table 4. As expected, $\hat{P}_e$ is lower for the rejection-based classifiers. Not surprisingly, the hard labels of all ambiguity rejected patterns have always been $(0, 1, 1)$, i.e. the ambiguity reject area reduced to $\Omega_a = \Omega_{2,3}$ whatever the classifier is.

Table 3. Rejection-based classifiers

%	$\hat{P}_c$	$\hat{P}_e$	$\hat{P}_a$	$\hat{P}_0$
Possibilistic	96	2	2	0
Probabilistic	94.67	2	2	1.33

Table 4. Classifiers

%	$\hat{P}_c$	$\hat{P}_e$
Bayes	96.67	3.33
1-NN	94.67	5.33

5 Conclusion

In this paper, we have proposed a general formulation of two-fold rejection based classifiers as a couple (L, H) of a labelling and a hardening function, each involving independently a set of parameters. Within this framework, it is possible to unify the presentation of classifiers. With respect to the label space, possibilistic classifiers are shown to encompass probabilistic / fuzzy ones and crisp classifiers. Of course, for particular values of the hardening parameters, so-defined classifiers reduce to classifiers with no reject option.

A possibilistic and a probabilistic classifiers have been presented within this framework. Their ability to deal with both reject options (distance, ambiguity) has been shown on simple examples of overlapping and well-separated classes.

References

1. Bezdek, J.C., Reichherzer, T.R., Lim, G., Attikiouzel, Y.: Classification with multiple prototypes. Proc. 5th IEEE International Conference on Fuzzy Systems (1996) 626-632
2. Chow, C.K., An optimum character recognition system using decision functions. IRE Transactions on Electronic Computers **6** (1957) 247-254
3. Dubuisson, B., Masson, M.H.: A statistical decision rule with incomplete knowledge about classes. Pattern Recognition **26** (1993) 155-165
4. Dubuisson, B., Masson, M.H., Frélicot, C.: Some topics in using pattern recognition for system diagnosis. Engineering Simulation **13** (1996) 863-888
5. Frélicot, C.: Learning rejection thresholds for a class of fuzzy classifiers from possibilistic clustered noisy data. Proceedings of 7th International Fuzzy Systems Association World-congress **3** (1997) 111-116
6. Frélicot, C.: A rejection-based possibilistic classifier and its parameters learning. Proc. 7th IEEE International Conference on Fuzzy Systems (1998)
7. Fukunaga, K.: Introduction to statistical pattern recognition. Academic Press (1990) Second edition
8. Ha, T.M.: The optimum class-selective rejection rule. IEEE Transactions on Pattern Analysis and Machine Intelligence **19** (1997) 608-615