A unified view of class-selection with probabilistic classifiers

Hoel Le Capitaine

A UNIFIED VIEW OF CLASS-SELECTION WITH PROBABILISTIC CLASSIFIERS

HOEL LE CAPITAINE

Abstract. The possibility of selecting a subset of classes instead of one unique class for assignation is of great interest in many decision making systems. Selecting a subset of classes instead of singleton allows to reduce the error rate and to propose a reduced set to another classifier or an expert. This second step provides additional information, and therefore increases the quality of the result. In this paper, a unified view of the problem of class-selection with probabilistic classifiers is presented. The proposed framework, based on the evaluation of the probabilistic equivalence, allows to retrieve class-selective frameworks that have been proposed in the literature. We also describe an approach in which the decision rules are compared by the help of a normalized area under the error/selection curve. It allows to get a relative independence of the performance of a classifier without reject option, and thus a reliable class-selection decision rule evaluation. The power of this generic proposition is demonstrated by evaluating and comparing it to several state of the art methods on nine real world data sets, and four different probabilistic classifiers.

1. Introduction

The process of accurately recognizing, or discriminating, real world observations contained in a database is a fundamental task in data analysis [11]. Considered as a pattern recognition problem, there has been many approaches for the classification of the objects in known classes. Naturally, the more prior information, the more the classification algorithm can be built according to this knowledge, therefore leading to a powerful recognition

Key words and phrases. Reject options, multiple class-selection, probabilistic classification, probabilistic metric.

Laboratoire d'Informatique de Nantes Atlantique (UMR CNRS 6241), École Polytechnique de Nantes, Rue C. Pauc, 44000 Nantes, France. Email: hoel.lecapitaine@univ-nantes.fr.

system. In the special case where prior probabilities and conditional densities are known, the Bayes decision rule is known to be optimal with respect to the error rate. However, real distributions are never known in advance, so that the models do not reflect correctly the data. Moreover, in many recognition problems, the data to be classified is issued from mixed and/or noisy classes. In particular, it is not uncommon to find classes that are overlapping in the feature space, so that classification is ambiguous [21]. Due to noise or same lack of information, some samples do not even belong to any known class (known as the zero-shot learning problem [19]).

Number of investigations in the field of pattern recognition focus on problems with a large number of classes (*e.g.* face identification, image classification, character recognition and so on ...). Therefore, the possibility of selecting a small subset of classes, which can be associated to a new, larger class, containing the previous classes, shows a growing interest [8]. Another growing interest resides in multi-label classification [38], where a sample can be associated to a subset of true labels.

Selective classification must be distinguished from multi-label classification, in the sense that in the case of selective classification, the true output is still a singleton, as opposed to the several labels of samples in multi-label classification. However they show a similar behavior since they both associate samples to subset of classes, but loss function, error criterion used for the generation of the subsets are different, as we will see later. The interest of this technique is immediate in numerous applications such as text categorization [37], medical diagnosis [6], bioinformatics [13], recommender systems [34] or scene classification [2]. For instance, in character recognition, it is very useful to reduce the number of possible candidates from 26 letters to 2, and then use domain knowledge to infer the true letter with more certainty.

Faced with overlapping and or noisy classes, authors have presented a number of various propositions for multiple class-selection. These propositions differ from the usual reject

option, where a classifier may withdraw a sample instead of classifying it, see e.g. [40] [1] [16].

The two major approaches can be summarized as follows. The first approach tries to model the fact that an observation may belong to several clusters. For instance, in [20], the authors relax the constraint that hard membership vector sums to one in a mixture model, allowing 2^c possible assignments, where c is the number of classes. Note that all these approaches are built in an unsupervised setting.

The second one, that represents the vast majority of approaches, relies on decisions made on the basis of supervised classifier' outputs. In other terms, a classifier is designed without considering the possibility of selecting several classes, but posterior probabilities, or decision functions, are used to determine the number of assignments. These approaches are generally based on three families of criteria, leading to different solutions.

The first criteria is related to the cost of misclassification and the cost of selecting a large number of classes[1]. This kind of criteria has been one of the first propositions [18] aiming at selecting several classes for an unique sample. Based on the two costs, an optimal decision rule is derived, and allows to select up to c classes for incoming samples.

The second kind of criteria relies on information-theoretic considerations of the probabilistic tuple [29, 30]. For instance, one may want to maximize the dispersion around the ambiguous value of posterior probability 0.5 for a binary classification problem.

The last family of approaches is based on performance measures [15]. There are a number of different performance measures/ The most prominent works on this topic come from usual error rates [15], ROC curves [29], or precision-recall [8, 7].

In this paper, we focus our attention on the second approach, but the formulation allows to open bridges (and equivalences) to the first and the third approaches.

[1]It differs from the reject option, where the considered costs are the misclassification and the abstaining ones.

The Bayesian statistical decision theory is taken as the basis of the analysis. Therefore, we start with the following three factors: the distribution family $p(\mathbf{x}|\theta)$, prior distribution for the parameters $p(\theta)$ and a loss function $\ell(\theta, \alpha)$, where α is an action of the decision space[2] $\mathcal{A}$ [36]. An effective comparison criterion is the posterior expected loss, which can be written as

$$R(\theta, \alpha) = \int_{\Theta} \ell(\theta, \alpha) p(\theta|\mathbf{x}) \mathrm{d}\theta \tag{1}$$

where Θ is the state space (we consider in the sequel the discrete case). The posterior probability $p(\theta|\mathbf{x})$ is obtained thanks to the Bayes theorem, knowing the distribution family and priors on $\theta \in \Theta$. Under the Bayes principle, the optimal rule is obtained by choosing for $\mathbf{x}$ the action α that minimize the expected loss:

$$\alpha(\mathbf{x}) = \arg\min_{\alpha \in \mathcal{A}} R(\theta, \alpha)$$

Naturally, if the expected loss (1) is minimum for all $\mathbf{x}$, then the overall risk is also minimized. If one seeks to minimize the error probability of classification, then the zero-one loss function is used. One may also allow other actions than a binary and strict association to classes. For instance, the reject option consists in adding another action in $\mathcal{A}$ [5, 40]. The action leads to refuse, or withhold, the decision for the current sample. This is particularly useful in close cases i.e. when the largest posterior probabilities are close. Naturally, the *no decision* action must have a cost, or a loss, that needs to be modeled under the Bayes minimum risk setting.

In this paper, we are interested in an even more increased action space $\mathcal{A}$. In particular, we consider the power set of Θ. Therefore, each sample $\mathbf{x}$ can be associated to one element of the power set. The subset selection procedure is described in the next section.

[2]The decision space is composed of possible outcomes of the system.

The paper is organized as follows. The first part (Section 2) of this work follows and extends the works in [18] and [30]. The second part (Section 3) is dedicated to the probabilistic equivalence definition, extending the work proposed in [27]. The third part (Section 4) enlarges this approach by learning the probabilistic equivalence such that expected number of classes chosen for an incoming sample is close to the number of classes selected for its closest neighbor(s). In Section 5, experimental results on different data sets, using different classifier providing posterior probabilities, are given. The aim of this part is not to discuss the relative superiority of a classifier over another, but rather to observe the behavior of class-selective decision rules for a given classifier. Finally some concluding remarks are drawn in Section 6.

2. Decision rules

2.1. **Reject option, or how to abstain.** In a supervised setting, we consider n pattern $\mathbf{x}_j$ $(j = 1, \cdots, n)$ that belong to one of the c classes of a problem, *i.e.* $\exists\, i \in \{1, \cdots, c\}$ such that $\mathbf{x}_j \in \theta_i$, for every j. For writing convenience, the set $\{\theta_1, \cdots, \theta_c\}$ is denoted Θ. Let us assume that we know both the prior probabilities $p(\theta_i)$ and the conditional densities $p(\mathbf{x}|\theta_i)$. The probability that an incoming observed sample $\mathbf{x}$ belongs to the class θ_i is given by the Bayes' formula:

$$p(\theta_i|\mathbf{x}) = \frac{p(\mathbf{x}|\theta_i)p(\theta_i)}{p(\mathbf{x})}, \tag{2}$$

where $p(\mathbf{x})$ is the mixture density function, also called evidence. This function ensures that posterior probabilities over Θ sum up to one. Once all posterior probabilities have been obtained, the Bayes decision rule consists in choosing the class for which the posterior probability is the maximum (MAP rule). Under this rule, the probability of error is minimized. An elegant way to allow other actions than exclusive classification is to introduce a loss function, and to minimize the overall risk. Formally, we suppose that we have a finite set of actions α_j, the loss function $\ell(\alpha_j|\theta_i)$ denotes the loss incurred when choosing action

α_j when the true class is θ_i. Thus, the conditional risk can be written

$$R(\alpha_j|\mathbf{x}) = \sum_{i=1}^{c} \ell(\alpha_j|\theta_i)p(\theta_i|\mathbf{x}). \tag{3}$$

Denoting the decision function $\alpha(\mathbf{x})$ for each $\mathbf{x}$, the overall risk is obtained as follows

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x})d\mathbf{x}. \tag{4}$$

If $\alpha(\mathbf{x})$ is such that the conditional risk is minimum, then the overall risk is minimized. Computationally, it consists in obtaining the conditional risk for every possible action, and choose the action corresponding to the minimum. This minimum overall risk is then called the Bayes risk.

In close cases, *i.e.* when the largest posterior probabilities are close, it may be useful to withhold the decision, depending on the cost of the action *no decision*. The first attempt to cope with such actions has been proposed by Chow [4], thereby introducing the *reject option*. This option is primarily used in particular applications where the cost of misclassification is large, with the possibility of leaving human experts classify the ambiguous cases, e.g. medical diagnosis, nuclear plant monitoring, automatic driving, and so on ... The reject option of Chow minimizes the error rate given a reject rate, and vice versa. The Chow's rule can be obtained as follows.

Proposition 1. *Consider the loss function defined by*

$$\ell(\alpha_j|\theta_i) = \begin{cases} 0 & j = i \\ C_r & j = c+1 \\ C_e & otherwise \end{cases} \tag{5}$$

where C_e is the cost of an error, and C_r is the cost of rejection. Then the corresponding minimum risk is equivalent to Chow's rule [4].

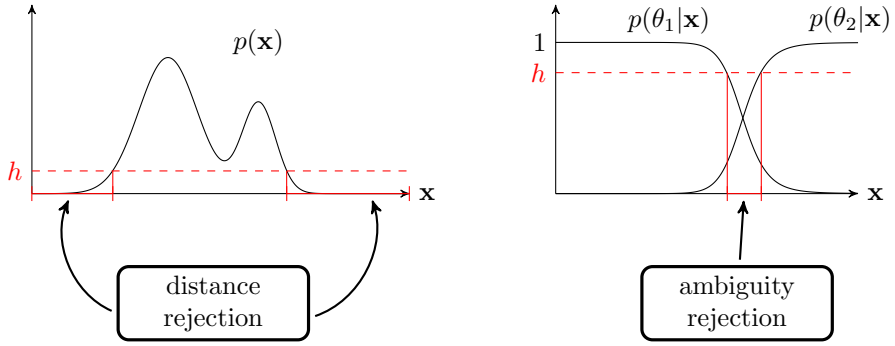
Proof. Choosing the action α_j for a sample means that the conditional risk $R(\alpha_j|\mathbf{x})$ is lower than $R(\alpha_{j+1}|\mathbf{x})$. It is straightforward to derive the following decision rule (referred to as Chow's rule):

- Decide θ_j if $p(\theta_j|\mathbf{x}) \geq p(\theta_i|\mathbf{x})$ for all i, and if $p(\theta_j|\mathbf{x}) \geq 1 - C_r/C_e = 1 - t$.
- Reject the sample $\mathbf{x}$ otherwise.

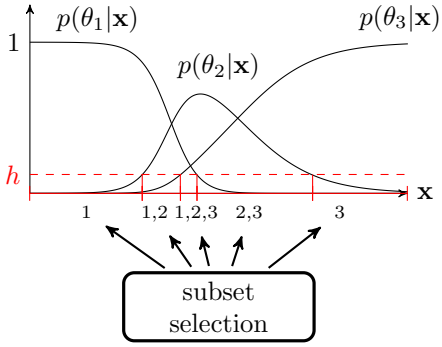
□

If $t = 0$, samples are always rejected because the cost of rejection is null. Conversely, if $t = 1$, then the cost of rejection and error are equal, and the decision rule reduces to the MAP (maximum a posteriori) rule. Naturally, t is reduced to C_r/C_e because the cost of correct classification is zero. However, the general term for the threshold is $(C_r - C_c)/(C_e - C_c)$, where C_c is the cost of correct classification. The decision rule is illustrated in Figure 1b for the case of two classes. In a second paper [5], Chow also gives a detailed analysis of the error-reject trade off. In particular, the author proves that the error probability $e(t)$ is an increasing function of t, while the reject probability $r(t)$ is a decreasing function of t. More precisely, we have $\partial e/\partial r = -t$.

In [10], Dubuisson and Masson introduced the option of distance reject. Contrary to the rule proposed by Chow, it removes the restriction that a pattern belongs to one of the c classes. In particular, a distance reject class is introduced, and identifies the samples having a small similarity to the prototypical vectors representing each classes. As the ambiguity rejection, the distance rejection allows to reduce the error probability of the classifier. According to [10], a sample $\mathbf{x}$ is rejected if its mixture density function $p(\mathbf{x})$ is lower than a threshold θ, as illustrated in Figure 1a. However, as pointed out in [32], the threshold is not related the class distributions but to the mixture density, which can be problematic with distributions having different parameters. To circumvent this problem, the authors propose to use multiple distance reject thresholds, one for each class. Their approach to rejection is related to generative models, since a pattern is rejected if its maximum conditional



(A) Distance rejection (B) Reject option



(C) Class-selective decision rules

FIGURE 1. Illustration of the distance rejection on a mixture density (a), ambiguity rejection using Chow's rule where $h = 1 - t$ (b) and selection of subset of classes using Ha's rule, with $h = t$ (c).

probability $p(\mathbf{x}|\theta_i)$ is lower than a class-dependent threshold h_i, while Chow's work adopts a discriminative approach, where posterior probabilities $p(\theta_i|\mathbf{x})$ are considered, see Figure 1b.

2.2. **Multiple class selection.** More recently, Ha introduced the class-selective schemes [18]. In contrast with the previous rule, where the two choices are classification or rejection, the class-selective schemes allow to select classes that are most likely to issue the pattern, and reject the others. The feature space is then partitioned into 2^c regions, corresponding to the power set of the c classes, see Figure 1c for an example with $c = 3$ classes. Provided posterior probabilities, there are a lot of solutions to partition the decision space. The

most commonly used is probably to set a constant number n of classes, and select the n-th larger posterior probabilities. However, the selected number of classes does not vary, and samples close to class centers may be assigned to several classes. The constant risk (CR) is another popular rule [17]. It consists in selecting the minimum number of classes such that the accumulated sum of posterior probabilities is lower than a threshold. In order to obtain a reasonable partition, Ha proposes the following loss function

$$\ell(\alpha_j|\theta_i) = \ell_e(\alpha_j|\theta_i) + \ell_n(\alpha_j) \tag{6}$$

The first part is related to usual loss functions, and helps to denote the cost of making an error:

$$\ell_e(\alpha_j|\theta_i) = \begin{cases} 0 & \text{if } \theta_i \in S_j \\ C_e & \text{otherwise} \end{cases} \tag{7}$$

The second part penalizes the selection of a large number of classes. If the cardinal of the selected subset is large, then the probability of error decreases, but many classes are remaining which does not help the final exclusive classification by the expert. The loss function is written as follows

$$\ell_n(\alpha_j) = C_n.card(S_j) \tag{8}$$

where C_n denotes the cost incurred when selecting classes. Using this loss, the conditional risk becomes

$$R(\alpha_j|\mathbf{x}) = C_e risk(\mathbf{x}) + C_n card(S_j) \tag{9}$$

where $risk(\mathbf{x})$ is the conditional probability that the true class of $\mathbf{x}$ is not in the subset S_j of selected classes. In other terms, the sum of posterior probabilities $p(\theta_i|\mathbf{x})$ such that i does not belong to S_j. Minimizing the conditional risk gives the optimal number of selected

classes with respect to the error rate (see [18] for details on derivation). Such minimization gives the optimal number of selected classes $n^\star(\mathbf{x}, t)$ for the sample $\mathbf{x}$ and a given threshold t as follows

$$n^\star(\mathbf{x}, t) = \min_{i \in [1,c]} \left\{ i \, | p(\theta_{(i+1)}|\mathbf{x}) \leq t \right\} \tag{10}$$

where $t = C_n/C_e$, and $p(\theta_{(i)}|\mathbf{x})$ is the decreasing sequence of posterior probabilities, $i = 1, \cdots, c$. The author uses the convention $p(\theta_{(c+1)}|\mathbf{x}) = 0$ so that c classes are selected if none of the posterior probabilities is strictly greater than t.

Proposition 2. *If one keeps the loss function ℓ_n as defined by (8), and defines the more general loss function ℓ_e as*

$$\ell_e(\alpha_j|\theta_i) = \begin{cases} C_c & \text{if } \theta_i \in S_j \\ C_e & \text{otherwise} \end{cases} \tag{11}$$

where C_c is the cost of correct classification, then the optimum number of classes is given by (10), where t is given by

$$t = \frac{C_n}{C_e - C_c} \tag{12}$$

Proof. Using both losses, the conditional risk can be written as follows

$$R(\alpha_j|\mathbf{x}) = C_e \sum_{\theta_i \notin S_j} p(\theta_i|\mathbf{x}) + C_c \sum_{\theta_i \in S_j} p(\theta_i|\mathbf{x}) + C_n n \tag{13}$$

where $n = card(S_j)$. By convention, we have $C_e > C_c$, so that the conditional risk is minimum when the n selected classes correspond to n maximum posterior probabilities. This leads to the following conditional risk

$$R(\alpha_j|\mathbf{x}) = C_e \sum_{i=n+1}^{c} p(\theta_{(i)}|\mathbf{x}) + C_c \sum_{i=1}^{n} p(\theta_{(i)}|\mathbf{x}) + C_n n, \tag{14}$$

which can be simplified into

$$\begin{aligned} R(\alpha_j|\mathbf{x}) &= -(C_e - C_c)\sum_{i=1}^{n} p(\theta_{(i)}|\mathbf{x}) + C_e + C_n n, \\ &= -\sum_{i=1}^{n} p(\theta_{(i)}|\mathbf{x}) + \frac{C_e}{C_e - C_c} + n\frac{C_n}{C_e - C_c} \end{aligned} \tag{15}$$

Let us denote $F_n = \sum_{i=1}^{n} p(\theta_{(i)}|\mathbf{x})$, $s = \frac{C_e}{C_e - C_c}$ and $t = \frac{C_n}{C_e - C_c}$. The conditional risk can now be casted as a function of the number of selected classes and the sample $\mathbf{x}$, *i.e.* $R(n, \mathbf{x})$.

$$R(n, \mathbf{x}) = -F_k + s + tn \tag{16}$$

Letting $n^\star$ denoting the optimal number of selected classes, we have

$$R(n^\star + 1, \mathbf{x}) \geq R(n^\star, \mathbf{x}) \tag{17}$$

$$R(n^\star - 1, \mathbf{x}) \geq R(n^\star, \mathbf{x}) \tag{18}$$

Thanks to the convexity of (15), and using (17–18), we obtain

$$p(\theta_{(n^\star+1)}|\mathbf{x}) \leq t \tag{19}$$

$$p(\theta_{(n^\star)}|\mathbf{x}) \geq t \tag{20}$$

The two inequalities can be converted into (10), which concludes the proof. □

Naturally, setting the cost of correct classification to zero enables to retrieve the decision rule proposed in [18]. The cost of correct classification is generally negative, since it is a gain. If C_c is negative, then the threshold t decreases for fixed C_e and C_n, implying that the accuracy increases.

In [22], the author proposes a new scheme for class selection. Starting from the observation that some outputs of Ha's decision rule produce unwanted results, a new one is proposed. The author does not write a loss function but directly defines the number of

selected classes as a function of the difference of successive posterior probabilities

$$n^{\star}(\mathbf{x},t) = \min_{i\in[1,c]} \left\{ i \,|p(\theta_{(i)}|\mathbf{x}) - p(\theta_{(i+1)}|\mathbf{x}) \geq s \right\} \tag{21}$$

One should note that (21) can be rewritten as

$$n^{\star}(\mathbf{x},t) = \min_{i\in[1,c]} \left\{ i \,|1 - (p(\theta_{(i)}|\mathbf{x}) - p(\theta_{(i+1)}|\mathbf{x})) \leq t \right\} \tag{22}$$

where we use $s = 1 - t = (C_e - C_n)/C_e$, in order to obtain a decision rule consistent with the formulation adopted by Ha. In the sequel, (22) is used when referring to Horiuchi's rule. Naturally, the rule is not optimal with respect to the error rate given an average number of selected classes. However, the author proves that it is optimum with respect to the maximum distance between selected classes, given the average number of selected classes. Another heuristic, proposed in [28], defined by

$$n^{\star}(\mathbf{x},t) = \min_{i\in[1,c]} \left\{ i | \frac{p(\theta_{(i+1)}|\mathbf{x})}{p(\theta_{(i)}|\mathbf{x})} \leq t \right\}, \tag{23}$$

also uses a notion of similarity between consecutive posterior probabilities. The ratio comes from the interpretation of consequent and antecedent in a logical implication. This rule is denoted as LC (for Logical Confidence) in the sequel.

We restrict in this paper to standard approaches of class-selection (i.e. probability based selection), but it should be noted that other strategies based on blockwise similarities [30] or support vector machines [15] have been proposed.

2.3. **A generic formulation.** We propose to formulate adaptive class-selection decision rules in a generic manner as follows

Proposition 3. *An adaptive class-selective decision rule can be written as follows*

$$n^{\star}(\mathbf{x},t) = \min_{i\in[1,c]} \{ i \,|\phi_i(\mathbf{p}(\theta|\mathbf{x})) \leq t \} \tag{24}$$

where $\mathbf{p}(\theta|\mathbf{x})$ *is the vector* $[p(\theta_{(1)}|\mathbf{x}), \cdots, p(\theta_{(c)}|\mathbf{x})]^T$*, and* ϕ_i *a function of individual posterior probabilities that evaluates the ambiguity lying in* $\mathbf{p}(\theta|\mathbf{x})$*.*

Clearly, when ϕ_i is large, which mean that there is a high doubt concerning the decision, then the corresponding decision, *i.e.* select i classes, is not convenient. This rule can be interpreted as follows: take the minimum number of classes such that the corresponding decision ambiguity is below a given threshold. Note that the threshold t can also be a function of the observation $\mathbf{x}$, the probabilistic vector $\mathbf{p}(\theta|\mathbf{x})$, or both. Now, we consider the different class-selective schemes that have been proposed, and how they can be written with a convenient ϕ_i function. It is interesting to see that even the decision rule of Chow, despite it is not a class-selective decision rule, can be written using (24). By setting $\phi_i = 1 - p(\theta_{(1)}|\mathbf{x})$, we obtain the Chow's decision rule, keeping the convention that if there is no i such that the inequality holds implies that $n^\star(\mathbf{x}, t)$ is set to c. The constant risk rule can also be written under this form, taking

$$\phi_i = 1 - \sum_{j=1}^{i} p(\theta_{(j)}|\mathbf{x}) \tag{25}$$

Naturally, it is straightforward to obtain the rules of Ha and Horiuchi by taking $\phi_i = p(\theta_{(i+1)}|\mathbf{x})$ and $\phi_i = 1 - p(\theta_{(i)}|\mathbf{x}) + p(\theta_{(i+1)}|\mathbf{x})$, respectively. There exists many other ambiguity measures, *e.g.* entropy, measure of fuzziness, and so on. Using these information-theoretic based measures for class selection is quite straightforward by using Proposition 3. For instance, in this paper, and for comparison purpose, we propose to use the normalized entropy as the measure ϕ_i:

$$\phi_i = 1 - \frac{1}{\log c} \sum_{j=1}^{i} p(\theta_{(j)}|\mathbf{x}) \log(p(\theta_{(j)}|\mathbf{x})) \tag{26}$$

The normalization factor $\frac{1}{\log c}$ is here to ensure that ϕ_i lies into the unit interval, but does not change the dynamic of the measure. In the next section, a new decision rule that includes all traditional decision rules is presented.

3. Probabilistic equivalence

3.1. **Probabilistic metric spaces.** In this section, we propose to design a new decision rule based on the equivalence of posterior probabilities. More precisely, the equivalence is obtained by considering a probabilistic metric (PM) space where a convenient metric is chosen between two values.

Formally, a metric space consists of a set X and a metric d allowing to compute distances between two points u, v lying in X. A PM space replaces the distance $d(u, v)$ between the two points by considering a distribution function F_{uv}, whose value $F_{uv}(x)$, for any x in X, corresponds to the probability that $d(u, v) \leq x$. However, one of the most important property of distances is that they hold the triangle inequality $d(u, w) \leq d(u, v)+d(v, w)$, for $(u, v, w) \in X^3$. The corresponding problem with distribution function relies on the comparison and relationships of F_{uw}, F_{uv} and F_{vw}. This is the rationale under the proposition of Menger, introducing the following inequality:

$$F_{uw}(x + y) \geq T(F_{uv}(x), F_{vw}(y)),$$

where T is a triangular norm (t-norm), *i.e.* a commutative, associative and monotone binary function, having 1 as identity, see [25] for details. In this paper, the equivalence, or the probabilistic metric is obtained with the Schweizer and Sklar triangular norm. More precisely, the t-norm defined by

$$T(x, y) = \Big(\max\Big(x^\lambda + y^\lambda - 1, 0\Big)\Big)^{1/\lambda} \tag{27}$$

where $(x, y) \in [0, 1]^2$ and $\lambda \in [-\infty, \infty]$, leads to the following residual implication between x and y

$$I(x, y) = \begin{cases} \left(1 + y^\lambda - x^\lambda\right)^{1/\lambda} & \text{if } x \geq y \\ 1 & \text{otherwise} \end{cases} \tag{28}$$

Based on this implication, a T-equivalence is obtained by

$$E(x, y) = \min\left(I(x, y), I(y, x)\right) \tag{29}$$

The Figure 2 illustrates the T-equivalence with various values of λ. A reference point is chosen, and the other is varying in the unit interval. As can be seen, the metric induced is not always symmetric. In particular, it only defines a pseudo-metric on the unit interval when $\lambda = 1$. Otherwise, the operator defines a pre-metric. Moreover, depending on the value of λ, the usual properties of a metric may not hold. In particular, the triangular inequality holds if and and only if λ is set to 1. The absence of triangular inequality may be interesting when evaluating similarity in specific spaces, as described in [24].

Another important point is the tendency to favor high values. In other terms, large probabilities give rise to large similarities. This possibility is highly wanted in the context of pattern recognition, because large probabilities reflect the confidence of the decision step.

3.2. **Probabilistic equivalence.** Let us denote F_i the probability of being correct when selecting the best i classes, so that

$$F_i = \sum_{j=1}^{i} p(\theta_{(j)}|\mathbf{x}) \tag{30}$$

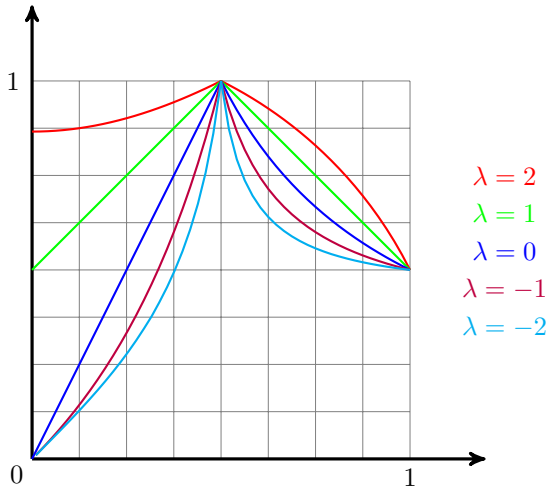


FIGURE 2. T-equalities for various values of λ with respect to typical reference value $x = 0.5$ and $y \in [0, 1]$. Symmetry over 0.5 is only obtained for $\lambda = 1$.

and $F_0 = \sum_{j=1}^{0} p(\theta_{(j)}|\mathbf{x}) = 0$. We define the power difference d_λ between two values $x \geq y$, $(x, y) \in [0, 1]^2$ as

$$d_\lambda(x, y) = (x - y)^\lambda \tag{31}$$

Now, we are interested in inspecting the behavior, in terms of risk (or in terms of probability of being correct), when the number of selected classes varies. We consider three cases, corresponding to different selections:

- changes when the number of considered classes is increased by one, $d_\lambda(F_{i+1}, F_i)$
- changes when the number of considered classes is decreased by one, $d_\lambda(F_i, F_{i-1})$
- changes when zero (distance rejection) and c (no decision) classes are selected, $d_\lambda(F_c, F_0)$.

Keeping in mind that i is increasing when searching its optimal value (i.e. the number of selected classes increases), we associate to $d_\lambda(F_c, F_0)$ and $d_\lambda(F_{i+1}, F_i)$ a positive influence (weight), while $d_\lambda(F_i, F_{i-1})$, already considered in a previous step, has a negative influence.

Consequently, we define the generalized decision function as

$$\phi_i = \sqrt[\lambda]{d_\lambda(F_c, F_0) + d_\lambda(F_{i+1}, F_i) - d_\lambda(F_i, F_{i-1})}, \tag{32}$$

where the λth root ensure values in the unit interval. Therefore, the generalized class-selective decision is obtained by using the decision rule (24) where ϕ_i is defined by (32). The decision function (32) can be written using T-equivalences. In particular, it has been shown in [26] that T-equivalences can be interpreted as similarity measures. In this context, let us consider the equivalence, or similarity, between $p(\theta_{(i)}|\mathbf{x})$ and $p(\theta_{(i+1)}|\mathbf{x})$. It is straightforward to obtain

$$E\left(p(\theta_{(i)}|\mathbf{x}), p(\theta_{(i+1)}|\mathbf{x})\right) = \left(1 + p(\theta_{(i+1)}|\mathbf{x})^\lambda - p(\theta_{(i)}|\mathbf{x})^\lambda\right)^{1/\lambda} \tag{33}$$

by ordering property of equivalences, which is equal to ϕ_i as defined by (32). Applied to posterior probabilities, we propose to define the new decision rule as

$$n^\star(\mathbf{x}) = \min_{i\in[1,c]} \left\{ i | E\big(p(\theta_{(i)}|\mathbf{x}), p(\theta_{(i+1)}|\mathbf{x})\big) \leq t \right\}, \tag{34}$$

called here after PE for Probabilistic Equivalence.

Now, we consider some particular cases of Equation (34) that uses the equivalence defined by Equation (29), when using the implication (28).

Proposition 4. *Using the decision function ϕ_i as defined by (32) with $\lambda = -\infty$ within (24), or equivalently by (34), leads to the decision rule*

$$n^\star(\mathbf{x}, t) = \min_{i\in[1,c]} \left\{ i \,|p(\theta_{(i+1)}|\mathbf{x}) \leq t \right\}, \tag{35}$$

which is the decision rule of Ha [18].

Proof. By definition, $F_c = 1$ and $F_0 = 0$. We give the proof by studying the function

$$f_\lambda(x, y) = \left(1 + y^\lambda - x^\lambda\right)^{1/\lambda} \tag{36}$$

when $\lambda \to -\infty$, giving $\lim_{\lambda\to-\infty} f_\lambda(x,y) = y$. Consequently, ϕ_i reduces to $F_{i+1} - F_i$, which is equal to $p(\theta_{(i+1)}|\mathbf{x})$), concluding the proof. □

Proposition 5. *Using the decision function ϕ_i as defined by (32) with $\lambda = 1$ within (24) leads to the decision rule*

$$n^\star(\mathbf{x}, t) = \min_{i\in[1,c]} \left\{ i \,|1 - (p(\theta_{(i)}|\mathbf{x}) - p(\theta_{(i+1)}|\mathbf{x})) \leq t \right\}, \tag{37}$$

which is the decision rule of Horiuchi [22].

Proof. When $\lambda = 1$, we have

$$\phi_i = 1 + F_{i+1} - F_i - F_i + F_{i-1} \tag{38}$$

Therefore, ϕ_i reduces to $1 + p(\theta_{(i+1)}|\mathbf{x}) - p(\theta_{(i)}|\mathbf{x}) = 1 - (p(\theta_{(i)}|\mathbf{x}) - p(\theta_{(i+1)}|\mathbf{x}))$. □

Proposition 6. *Using the decision function ϕ_i as defined by (32) with $\lambda = 0$ within (24) leads to the decision rule*

$$n^\star(\mathbf{x}, t) = \min_{i\in[1,c]} \left\{ i \,|(p(\theta_{(i+1)}|\mathbf{x})/p(\theta_{(i)}|\mathbf{x})) \leq t \right\}, \tag{39}$$

which is the decision rule LC [28].

Proof. Here again, considering (36), we obtain $\lim_{\lambda\to0} f_\lambda(x,y) = y/x$. Therefore, ϕ_i can be written as $(F_{i+1} - F_i)/(F_i - F_{i-1})$, which reduces to $p(\theta_{(i+1)}|\mathbf{x})/p(\theta_{(i)}|\mathbf{x})$, concluding the proof. □

For illustration purpose, let us consider a toy problem. The dataset contains 1500 samples (see Figure 3) that can belong to three different classes θ_1, θ_2 and θ_3 of 500 samples each. The first class (red) is a mixture of two equi-weighted dimensional normal distributions $\mathcal{N}(\mu, \Sigma)$ with parameters

$$\mu_{1,1} = [0.4, 0.9]^T \;\; \mu_{1,2} = [4.5, 1.95]^T,$$

and

$$\Sigma_{1,1} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \quad \Sigma_{1,2} = \begin{pmatrix} 1.2 & 0.7 \\ 0 & 1.8 \end{pmatrix}$$

The second (green) and the third (blue) class are two-dimensional normal distributions with the following parameters:

$$\mu_2 = [-1.5, 3]^T \quad \mu_3 = [-2.25, 0]^T,$$

and

$$\Sigma_2 = \begin{pmatrix} 3 & 0 \\ 0 & 0.5 \end{pmatrix} \quad \Sigma_3 = \begin{pmatrix} 1 & 0 \\ -1.2 & 1.5 \end{pmatrix}, \text{ respectively.}$$

The corresponding data is plotted in Figure 3. Knowing the distribution and prior probabilities, posterior probabilities are computed using (2). Given a threshold, one can define

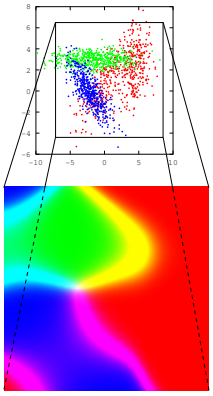
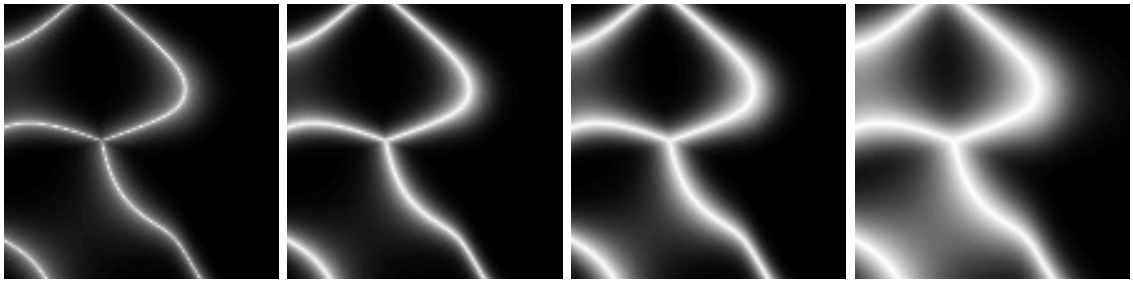


FIGURE 3. Posterior probabilities are encoded as RGB values, corresponding to their membership to the three classes: red, green and blue. For instance, the pink area is the mixture of blue and red classes.

a partition of the feature space that corresponds to the different colors. In this case, the white color corresponds to the entire set of classes $\{\theta_1, \theta_2, \theta_3\}$. Geometrically, it defines the center of the three classes. For illustration purpose, we give in Figures 4 and 5 the value of the probabilistic equivalence for the first order (i.e. the probabilistic equivalence of



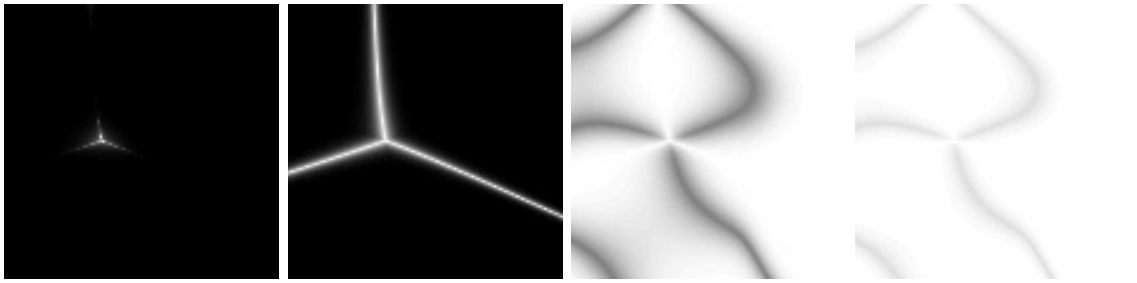
(A) $\lambda = -2$ (B) $\lambda = 0$ (C) $\lambda = 1$ (D) $\lambda = 2$

FIGURE 4. First order probabilistic equivalence values for different values of λ.

$p(\theta_{(1)}|\mathbf{x})$ and $p(\theta_{(2)}|\mathbf{x})$) and the second order, respectively. The values range from 0 (black) to 1 (white). In Figure 4, depending on the threshold, one or two classes are selected. Since one class is selected if the equivalence is below the threshold and two or more are selected if the equivalence is larger than the threshold, it is easy to see that the more the white, the more we select two classes. As can be seen, when the value of λ increases the area of possible multiple class selection increases, but still follows the boundaries observed in Figure 3. Inversely, when the value of λ is rather low, the bandwidth of multiple selection decreases, due to the property of creating sharp boundaries for low values of λ, as it has been observed in Figure 2. One can see that in Figure 5, large values of equivalences are located where it is interesting to select three classes, when the value of λ is rather low (e.g. $\lambda = -2, 0$). When λ is increased, one can observe that the area where three classes can be selected is augmented, except areas where selecting two classes is more appropriate, where the probabilistic equivalence remains rather low.

4. Learning the equivalence function

The framework provided in the previous section is quite generic, and allows to retrieve the majority of class-selective decision rules that have been proposed so far. However, now comes the question of selecting the convenient free parameter λ in practice. A first



(A) $\lambda = -2$ (B) $\lambda = 0$ (C) $\lambda = 1$ (D) $\lambda = 2$

FIGURE 5. Second order probabilistic equivalence values for different values of λ.

straightforward solution, adopted in [27], consists in making a grid search through a cross-validation procedure in order to find the optimal value of λ. This is computationally expensive and does not meet the requirements of a fast, accurate, and flexible classification system. Therefore, in order to find the optimal value, we propose to learn it from the available data. More precisely, we want a value λ such that the correct class belongs to the selected subset, while keeping the number of selected classes as small as possible. The procedure is separated into two phases. The first one, that corresponds to the learning of λ values for each training sample, is as follows

- From the learning set X_{train}, one can determine the optimal and necessary number of selected classes (nnc) based on computed posterior probabilities. In other terms, one must select at least nnc classes in order to have the correct target into the predicted set of classes. This number is set as the number of selected classes such that the sum of posterior probabilities of selected classes, given that the sample belongs to a selected class, is minimum.
- Once the nnc have been computed, one must select the value λ such that the decision rule gives the correct number of classes. To do so, it must respect the two inequalities

$$\left(1 + p(\theta_{(i+1)}|\mathbf{x}_m)^{\lambda_m} - p(\theta_{(i)}|\mathbf{x}_m)^{\lambda_m}\right) \leq t^{\lambda_m} \tag{40}$$

and

$$\Big(1 + p(\theta_{(i)}|\mathbf{x}_m)^{\lambda_m} - p(\theta_{(i-1)}|\mathbf{x}_m)^{\lambda_m}\Big) > t^{\lambda_m} \tag{41}$$

The corresponding algorithm is given in **Algorithm 1**

The second phase corresponds to the test step, where each test sample is associated to a subset of classes. It is built as follows

- In the test set X_{test}, we look for the closest sample of the learning set. The term *closest* depends on a notion of proximity defined by the user. The distance can be computed in the feature space or in the probabilistic space. Moreover, the distance chosen can be a simple Euclidean one, or a more sophisticated Mahalanobis distance where feature vectors are projected into new representation spaces. In this paper, an Euclidean distance is used for simplicity.
- The λ value λ_ℓ of the closest sample computed in **Algorithm 1** is used for the test sample, and the number of selected classes for a sample $\mathbf{x}_j$ is given by

$$n^\star(\mathbf{x}_j) = \min_{i\in[1,c]} \Big\{ i | \left(1 + p(\theta_{(k+1)}|\mathbf{x}_j)^{\lambda_\ell} - p(\theta_{(k)}|\mathbf{x}_j)^{\lambda_\ell}\right)^{1/\lambda_\ell} \leq t \Big\} \tag{42}$$

The corresponding algorithm is given in **Algorithm 2**.

The value λ can take particular values of the t-norm. In this case the measure used is the one corresponding to the associated t-norm. For instance, if $\lambda = 0$, then the equivalence obtained by the product is used, thanks to the continuity around 0 of (27). In practice, we set the range of values for λ to $[-10, 10]$, because the difference of $E_{-\infty}$ and E_{-10}, E_∞ and E_{10} is small, and thanks again to the continuity of the equivalence function.

5. Comparative study

In theory, the rule defined by Ha is the optimum decision rule in the sense that there are no other rules yielding a lower error rate for a given average number of selected classes.

Algorithm 1 Learning

Input: X_{train}, $p(\Theta|X_{train})$ and t
for each training sample $\mathbf{x}_m$ in X_{train}, $m = 1, \cdots, N$ **do**
    compute the necessary number of classes of $\mathbf{x}_m$ nnc_m as follows

$$nnc_m = \text{argmin}_i \left\{ \sum_i p(\theta_{(i)}|\mathbf{x}_m) | \mathbf{x}_m \in \theta_i \right\}$$

    $i \leftarrow nnc_m$
    select λ_m such that (40) and (41) holds
end for
return the vector Λ of N lambda values

Algorithm 2 Predict

Input: X_{test}, $p(\Theta|X_{test})$, t and Λ
for each test sample $\mathbf{x}_j$ in X_{test} **do**
    set the closest training sample $\mathbf{x}_m$ of $\mathbf{x}_j$ according to a pre-defined distance d:
    $\ell \leftarrow \text{argmin}_{m \in \{1,\cdots,N\}} d(\mathbf{x}_j, \mathbf{x}_m)$
    $\lambda_\ell = \Lambda(\ell)$
    set the number of selected classes, by using (42)
end for

This optimum is reached when the distribution of the data is known and true which is rarely the case in practice: posterior probabilities are often inaccurate, they are not known in advance and are estimated by algorithms that try to optimize the accuracy on the test sets. Moreover, this is an optimum rule with respect to the average number of selected classes, which may not be the only criteria that must be considered.

In this section, we conduct a detailed study on the behavior of each of the class-selective decision rules presented above, namely the constant risk, Ha, Horiuchi, LC, and the two variants of the new generalized decision rule.

5.1. **Experimental setup.** In order to compare the various decision rules that have been presented in the previous sections, we consider several datasets available online. Experiments are carried out on nine real datasets briefly described in Table 1. The datasets are publicly available from the UCI repository [14]. As can be seen in the Table, the datasets

present a larger variety in terms of number of classes, features and samples. Moreover, they are coming from various application domain such handwriting recognition, image segmentation, bioinformatics

For each data set, some samples are randomly selected as the learning set (see Table 1 for details). For comparison purpose, four different classifiers are used:

- a Linear Bayes classifier (LB),
- a Quadratic Bayes classifier assuming normal densities (QB),
- a multi layer perceptron, with one hidden layer having 20 hidden units (MLP),
- and a multi-class support vector classifier (SVM). According to comparative studies provided in [35], the one-against-all strategy is employed for the construction of the models.

The last two classifiers do not provide posterior probabilities, therefore a transformation (or a calibration) is needed to obtain an approximation of posterior probabilities. It is important to note that what is evaluated here is the selection rule, and not the classifiers. Several classifiers are used in order to assess the consistency of a possible superiority of a decision rule for a given classifier. Individual components such as the SVM classifier comes from the implementation given in [3], and calibration is operated with a gradient descent based method described in [39].

The source code used in this paper for the selection part is available online[3].

5.2. **Evaluation.** Reject options, and more generally class-selective decision rules cannot be evaluated by considering only their corresponding accuracy. This is due to two major reasons. The first is that they generally use a specified threshold, therefore giving different classification and rejection rates. The second reason comes from the tradeoff they imply. For the reject option, the tradeoff to find is between the error rate and the rejection rate. One wants to minimize the error rate, and to keep the reject rate low. For class-selective

[3] `http://www.polytech.univ-nantes.fr/lecapitaine/RPC`.

TABLE 1. Datasets used in the experiments.

Dataset	#training	#testing	#classes	#features
Vehicle	846	0	4	18
Segment	2310	0	7	19
Vowel	528	462	11	10
Letter	15000	5000	26	16
Shuttle	43500	14500	7	9
USPS	7291	2007	10	256
MNIST	60000	10000	10	780
Satimage	4435	2000	6	36
Dna	2000	1186	3	180

decision, the tradeoff is between a low error rate, and a low average number of selected classes.

Therefore, a common quality measure is to evaluate the area under the curve (Error(Rate) for the reject options, Error(Average class selection) for class-selective rules [18]). This curve is composed of operating points that can be attained when setting a specific threshold t. Naturally, if t is a function of the input, this kind of evaluation is not adapted, and one should find another way to evaluate a class-selective scheme. More recently, in [33], the authors present three different relationships between two error/rejection curves. Although this method can be adapted for class-selective evaluation, it relies on a graphical interpretation, which may be difficult when curves are somewhat similar, in particular when they are crossing each other.

A quantitative measure has been proposed in [29], however this evaluation measure is not adapted for the comparison of decision rules used on different classifiers, because each classifier provides different accuracies without rejecting samples. In other terms, depending on the classifier, the analysis of decision rules is not on the same scale, so that results are not interpretable. Instead, we propose to use the normalized area under the curve (see Figure 6) presented in [27] in order to overcome this problem. The normalized area is obtained by considering the curve and its associated area, but starting on the baseline,

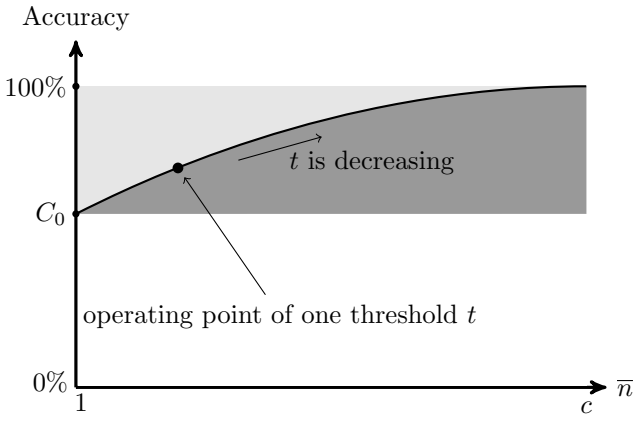


FIGURE 6. Area under the curve: the performance score is given by the ratio of the dark gray area over light gray area and itself (i.e. the rectangle defined by $(C_0, 1)$ and $(1, \overline{n})$).

i.e. the classification accuracy without rejection of the considered classifier. Therefore, the normalized area under the curve is defined by

$$nAUC = \frac{\int_0^1 (C(t) - C_0)\, dt}{\int_0^1 (1 - C_0)\, dt} \tag{43}$$

where C_0 is the baseline accuracy of the classifier (i.e. the classification rate without reject option), and t is the threshold used in the decision rule (34). The term $C(t)$ is the classification rate obtained by selecting subset of classes using (34). The convention is to say that the classification is correct if one label of the subset is the true label of the sample.

It can be proved that $C_0 \leq C(t) \leq 1$ for any t in the unit interval (we have in particular $C(0) = 1$ and $C(1) = C_0$), so that $0 \leq nAUC \leq 1$. A $nAUC$ equal to 1 means that for any t, the error rate with subset selection is equal to zero, while $nAUC$ value equal to zero means that adding subset of classes does not increase the classification rate at all (i.e. exclusive classification is the best choice). Therefore, the higher the better for $nAUC$.

The performance of a selection rule is a function of the threshold t that can be manually selected by the user or the experts as a function error and selection costs (see section 2.2).

The proposed evaluation measure does not allow to proceed to a point wise comparison of decision rules, but gives an average performance, by computing the mean performances of a rule for all possible thresholds. If one already have specified costs, then the simplest way to select a decision rule is a point evaluation, with a constant t value. If the mean number of classes to be selected is known, then one have simply to select the corresponding threshold. Finally, the threshold can be selected using methods of operating point selection of ROC curves [31].

5.3. **Results.** The results for each dataset and each decision rule are given in Table 2, where best scores for each dataset are reported in bold font. The last column is the average rank of each selection rule over all datasets for a given classifier. The considered decision rules are the following : CR (for *constant risk*, defined by (25)), Ha (defined by (10)), Horiuchi (defined by (22)), LC (for *logical confidence*, defined by (23)), Entropy (defined by (26)), PE and PE* (defined by (34)) which are the proposed decision rules without or with equivalence learning, respectively.

As can be seen in the Table 2, the proposed methods are superior to the other rules : the average ranks of both rules are the lowest for all classifiers.

Looking more in depth the results, one can say that one can observe a larger difference between rules for LB than for QB, which is explained by the quality of estimation of posterior probabilities. The $nAUC$ score is generally better for QB than for NB, due to the performances rates of individual methods. Moreover, the nAUC score follows the quality of classification of each classifier, as one can observe the general tendency of ranking LB $\prec$ QB $\prec$ MLP $\prec$ SVM.

It is interesting to note that one can distinguish two groups of datasets, one set (composed of Vehicle, Segment, Letter and Satimage) for which the results of all methods are quite similar, and another set (composed of Vowel, Shuttle, USPS, MNIST and Dna) for which the results are clearly in advantage for PE and its variant PE*.

TABLE 2. Normalized AUC (nAUC) for all datasets and all decision rules. Right column indicates the average rank of decision rules over all datasets for a given classifier.

		Datasets									
Classifier	Rule	Vehicle	Segment	Vowel	Letter	Shuttle	USPS	MNIST	Satimage	Dna	Avg. Rank
LB	CR	66.74	63.60	85.11	83.05	88.87	73.18	77.41	57.01	88.08	4.33
	Ha	66.94	63.61	84.77	81.61	88.63	72.79	76.54	57.02	87.29	4.77
	Horiuchi	53.36	64.80	65.04	70.59	82.01	77.85	75.35	56.82	85.04	6.11
	LC	65.52	64.67	82.17	79.89	89.01	78.13	79.44	57.22	86.78	4.22
	Entropy	52.10	62.72	82.67	85.10	83.52	77.89	77.61	56.56	85.27	5.33
	PE	67.75	77.96	85.09	84.94	89.87	84.77	80.08	70.19	88.41	2.22
	PE⋆	**73.64**	**90.00**	**89.69**	**87.08**	**97.77**	**94.36**	**91.00**	**89.82**	**97.72**	**1.00**
QB	CR	88.09	92.49	59.22	84.24	95.36	74.80	81.06	88.07	85.33	5.00
	Ha	88.05	92.50	59.25	83.44	95.42	73.54	80.65	89.61	85.39	4.66
	Horiuchi	84.09	89.04	53.27	85.09	92.23	80.17	78.82	80.08	83.15	6.22
	LC	88.63	92.88	60.33	90.79	95.52	79.50	82.74	89.42	84.73	3.44
	Entropy	82.90	90.39	54.84	90.62	92.77	80.18	81.50	81.80	82.56	5.44
	PE	89.56	93.19	60.21	91.07	95.97	85.73	84.22	89.58	85.79	2.22
	PE⋆	**95.80**	**96.14**	**73.59**	**94.36**	**97.85**	**94.61**	**91.35**	**90.68**	**97.89**	**1.00**
MLP	CR	96.05	96.70	61.48	93.49	85.73	78.84	72.70	92.11	59.03	5.22
	Ha	96.23	96.72	61.26	92.23	85.88	77.04	72.26	92.12	59.83	5.11
	Horiuchi	90.86	95.88	58.72	84.58	87.77	81.62	76.46	83.99	60.42	5.55
	LC	96.01	97.33	62.46	90.84	86.60	82.01	76.45	91.93	60.62	4.22
	Entropy	89.82	96.50	59.34	94.19	86.27	82.51	75.06	85.22	60.68	4.77
	PE	96.58	97.50	62.79	94.87	88.05	82.19	84.79	92.18	83.96	2.11
	PE⋆	**96.74**	**98.72**	**78.24**	**95.16**	**99.55**	**96.18**	**94.80**	**94.95**	**97.42**	**1.00**
SVM	CR	96.83	99.89	83.01	94.27	89.94	66.47	82.68	91.31	82.54	5.22
	Ha	96.99	99.91	85.80	92.31	89.95	82.81	80.59	91.79	91.41	4.22
	Horiuchi	92.75	99.56	66.11	94.65	96.41	93.82	69.40	86.03	87.26	5.77
	LC	96.58	99.90	73.60	95.35	93.21	91.37	79.07	88.68	90.74	4.77
	Entropy	91.79	99.72	79.07	96.46	96.52	94.22	77.41	86.04	86.52	4.77
	PE	97.01	99.94	85.87	95.44	96.59	93.88	84.18	91.81	91.44	2.22
	PE⋆	**97.99**	**99.97**	**87.09**	**98.91**	**99.94**	**97.73**	**95.07**	**95.82**	**98.06**	**1.00**

For this second group, one can remark that they are constituted by strongly overlapping classes, and do not necessarily follows normal distributions. This demonstrates the utility of the local adaptation of our proposition, based both on probabilities and local distances.

TABLE 3. Ranking statistics

Classifier	χ^2_F	F_F
LB	36.95	17.34
QB	40.38	23.72
MLP	35.42	15.26
SVM	34.85	14.56
ALL	142.43	67.76

As a side note, one should remark that the best rank of Ha's measure is obtained for the SVM classifier. Since this measure is theoretically the best one, it may seems that the SVM classifier provides the best approximation of the true distribution of the data (or at least the best approximation of posterior probabilities).

In order to test the significance of differences between decision rules, we propose to use the non-parametric Friedman test, as suggested in [9]. Let R^i_j be the rank of the j-th selection rule on the i-th dataset. The Friedman test compares the average ranks R_j over all datasets (last column of Table 2). Under the null-hypothesis, stating that two selection rules are equivalent, their ranks should be equal (here R_j should be equal to 4 for all j). The Friedman statistic is given by

$$\chi^2_F = \frac{12N}{ns(ns+1)} \left(\sum_j R_j^2 - \frac{ns(ns+1)^2}{4} \right) \tag{44}$$

where N, the number of datasets, and ns the number of selection rules are big enough, typically $N > 10$ and $ns > 5$. A derived and better statistic proposed in [23] is given by

$$F_F = \frac{(N-1)\chi^2_F}{N(ns-1) - \chi^2_F} \tag{45}$$

The Table 3 gives χ^2_F and F_F rank statistics for each classifier.

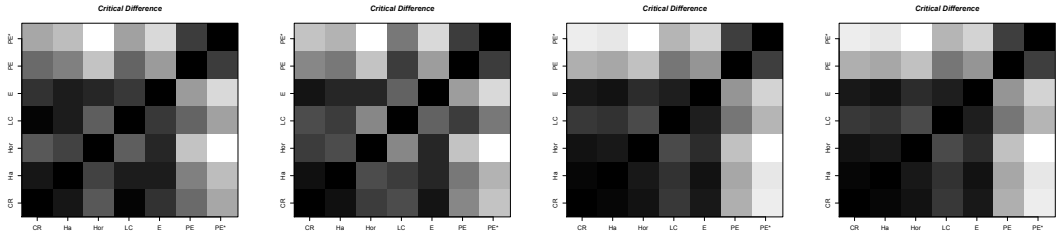
F_F is distributed according to the F distribution with $7-1=6$ and $(7-1)\times(9-1)=48$ degrees of freedom. The critical value at a probability level of 5 % is 2.29, so that the null-hypothesis of the equivalence between selection rules is rejected (compared to the F_F values of Table 3).

If the null hypothesis is rejected, two post-hoc tests can be considered, the Nemenyi or the Bonferroni-Dunn tests. However, as stated in [9], one should prefer the later because it does not make pairwise comparisons for the proposed method. The performance of two selection rules is significantly different if their corresponding average ranks differ by at least the critical difference, defined by

$$CD = q_\alpha \sqrt{\frac{ns(ns+1)}{6N}}, \tag{46}$$

where q_α values are based on the Studentized range statistic divided by $\sqrt{2}$, (see [9] for details). For a confidence level of 95%, we have $q_{0.05} = 2.638$, so that the critical difference for this experiment is equal to 2.68. Therefore, one can say that the selection rule PE* is statistically better than all the other rules (except its non learned variant PE), for all classifiers. Although one can see that PE* performs better than PE, one cannot conclude on the significant difference the two rules at $\alpha = 0.05$.

Now we consider all the classifiers at once for the comparison, and rank statistics are given in the last line (ALL) of Table 3. Here again the null hypothesis of equivalence between selection rules is highly rejected, since the critical value at a probability level of 5 % is 2.14, compared to the value 67.76 given in the Table. Since there are more experiments involved, the critical difference of the Bonferroni-Dunn test is 1.345 for a confidence level of 95%. Therefore, the difference between PE and $PE^\star$ is almost equal to the theoretical critical difference, showing a statistical superiority of $PE^\star$ over PE, and naturally over all the other decision rules.



(A) Linear Bayes classifier (B) Quadratic Bayes classifier (C) Multi Layer Perceptron (D) Support Vector Machine

FIGURE 7. Critical differences of the selection rules for the four classifiers used in this study, the whitest the largest.

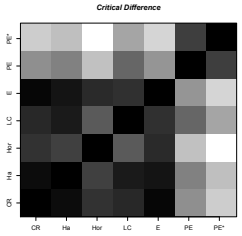


FIGURE 8. Critical differences of the selection rules for all classifiers

Finally, we propose a visual interpretation of the critical differences between the performances of the selection rules for a specified classifier (Figure 7) and for all classifiers (Figure 8). These plots should be read as follows. Each cell has grey value corresponding the normalized critical difference of all selection rules. In other terms, the larger critical difference leads to a value of 1 (white), and the lowest to a value of 0 (black). The order retained is the same as the one used in the tables, i.e. CR, Ha, Horiuchi, LC, Entropy, PE and PE*. As can be seen, two categories (or two blocks in the Figures) of selection rules can be distinguished, PE and PE* against the others.

6. Conclusion

In pattern recognition or related problems, the possibility of selecting a subset of classes instead of singletons for assignation is of great interest. It exists many ways of selecting

possible subsets among the power set of classes, and this implies that one cannot explore the entire solution space. To this aim, several heuristics have been proposed, but they cannot handle the diversity and variety of specific datasets.

In this paper, a generalized approach to class-selection is presented. Given a classifier providing posterior probabilities outputs, the proposed rule allows to retrieve the three class-selective decision rules proposed in the literature. The new decision rule has a logical justification, since the class-selection is made upon the evaluation of the equivalence between posterior probabilities. We also describe an approach in which the decision rules are compared by the help of a normalized area under the error/selection curve. It allows to get a relative independence of the performance of a classifier without reject option, and thus a reliable class-selection decision rule evaluation.

As a potential future work, we could cite another method of evaluation for set-valued classifiers proposed in [8], which is based on usual information retrieval quality measures. Moreover, the basis for a theoretical analysis of characterization of risk-coverage trade-off given in [12] can help to provide new elements with respect to the optimality of decision rules. Let us also mention the multi-label multi-class classification, where each sample may actually belong to several classes. In this case, there is no more tradeoff of error/selection, and the set selection must be made using other loss functions, e.g. from information retrieval applications.

References

[1] P. Bartlett and M. Wegkamp. Classification with a reject option using a hinge loss. *Journal of Machine Learning Research*, 9, 2008.

[2] M.R. Boutell, J. Luo, X. Shen, and C.M. Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[3] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27:1–27:27, 2011.

[4] C.K. Chow. An optimum character recognition system using decision functions. *IRE Transactions on Electronic Computers*, 6(4):247–254, 1957.

[5] C.K. Chow. On optimum error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

[6] A. Clare and R. King. Knowledge discovery in multi-label phenotype data. *Principles of Data Mining and Knowledge Discovery*, pages 42–53, 2001.

[7] G. Corani and M. Zaffalon. Learning reliable classifiers from small or incomplete data sets: The naive credal classifier 2. *Journal of Machine Learning Research*, 9:581–621, 2008.

[8] J. del Coz, J. Diez, and A. Bahamonde. Learning nondeterministic classifiers. *Journal of Machine Learning Research*, 10:2273–2293, 2009.

[9] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

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

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

[12] R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010.

[13] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *Advances in Neural Information Processing Systems 14*, 2002.

[14] A. Frank and A. Asuncion. UCI machine learning repository, 2010. University of California, Irvine, School of Information and Computer Sciences.

[15] E. Grall-Maes and P. Beauseroy. Optimal decision rule with class-selective rejection and performance constraints. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):2073–2082, 2009.

[16] Yves Grandvalet, Alain Rakotomamonjy, Joseph Keshet, and Stephane Canu. Support vector machines with a reject option. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 537–544. 2009.

[17] S. Gupta. On some multiple decision (selection and ranking) rules. *Technometrics*, 7:225–245, 1965.

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

[19] B. Hariharan, S. V. N. Vishwanathan, and M. Varma. Efficient max-margin multi-label classification with applications to zero-shot learning. *Machine Learning Journal*, 88(1):127–155, June 2012.

[20] K. A. Heller and Z. Ghahramani. A nonparametric Bayesian approach to modeling overlapping clusters. In *AISTATS*, 2007.

[21] Tin Kam Ho and M. Basu. Complexity measures of supervised classification problems. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3):289 –300, mar 2002.

[22] T. Horiuchi. Class-selective rejection rule to minimize the maximum distance between selected classes. *Pattern Recognition*, 31(10):1579–1588, 1998.

[23] R. L. Iman and J. M. Davenport. Approximations of the critical region of the friedman statistic. *Communications in Statistics*, 9(6):571–595, 1980.

[24] F. Jakel, B. Scholkopf, and F.A. Wichmann. Similarity, kernels, and the triangle inequality. *Journal of Mathematical Psychology*, 52(5):297–303, 2008.

[25] E. Klement, R. Mesiar, and E. Pap. *Triangular Norms*. Kluwer Academic, 2000.

[26] H. Le Capitaine. A relevance-based learning model of fuzzy similarity measures. *IEEE Transactions on Fuzzy Systems*, 20(1):57–68, 2012.

[27] H. Le Capitaine. Set-valued Bayesian inference with probabilistic equivalence. In *21st International Conference on Pattern Recognition*, Tsukuba, Japan, 2012.

[28] H. Le Capitaine and C. Frelicot. Classification with reject options in a logical framework: a fuzzy residual implication approach. In *Int. Fuzzy Systems Association World Congress and European Society for Fuzzy Logic and Technology IFSA/EUSFLAT*, pages 855–860, 2009.

[29] H. Le Capitaine and C. Frélicot. An optimum class-rejective decision rule and its evaluation. In *20th International Conference on Pattern Recognition, 2010*, pages 3312–3315, Istanbul, Turkey, 2010.

[30] H. Le Capitaine and C. Frélicot. A family of measures for best top-n class-selective decision rules. *Pattern Recognition*, 45(1):552–562, 2012.

[31] Charles E. Metz. Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8:283–298, 1978.

[32] R. Muzzolini, Y-H. Yang, and R. Pierson. Classifier design with incomplete knowledge. *Pattern Recognition*, 31(4):345–369, 1998.

[33] Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. Accuracy-rejection curves (arcs) for comparing classification methods with a reject option. *Journal of Machine Learning Research - Proceedings Track*, 8:65–81, 2010.

[34] F. Pachet and P. Roy. Improving multilabel analysis of music titles: A large-scalevalidation of the correction approach. *IEEE Transactions on Audio, Speech and Language Processing*, 17(2):335–343, 2009.

[35] R. Rifkin and A. Klautau. In defense of one-vs-all classification. *Journal of Machine Learning Research*, 5:101–141, 2004.

[36] C.P. Robert. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Verlag, 2007.

[37] Robert Schapire and Yoram Singer. Boostexter: A boosting-based system for text categorization. *Machine Learning*, 39:135–168, 2000.

[38] G. Tsoumakas and I. Katakis. Multi-label classification. *International Journal of Data Warehousing & Mining*, 3(3):1–13, 2007.

[39] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.

[40] M. Yuan and B. Wegkamp. Classification methods with reject option based on convex risk minimization. *Journal of Machine Learning Research*, 11:111–130, 2010.