The Costs of Indeterminacy: How to Determine Them?

Gen Yang, Sébastien Destercke, Marie-Hélène Masson

The costs of indeterminacy: how to determine them?

Gen Yang, Sebastien Destercke, Marie-Hélène Masson

***Abstract*—Indeterminate classifiers are cautious models able to predict more than one class in case of high uncertainty. A problem that arises when using such classifiers is how to evaluate their performances. This problem has already been considered in the case where all prediction errors have equivalent costs (that we will refer as the "0/1 costs" or accuracy setting). The purpose of this work is to study the case of generic costs functions. We provide some properties that the costs of indeterminate predictions could or should follow, and review existing proposals in the light of those properties. This allows us to propose a general formula fitting our properties that can be used to produce and evaluate indeterminate predictions. Some experiments on the cost-sensitive problem of ordinal regression illustrate the behaviour of the proposed evaluation criterion.**

***Index Terms*—indeterminate classifier, evaluation, imprecise probabilities, cost-sensitive classification**

I. Introduction

The classification task consists in identifying the class of a new observation, described by a set of features, on the basis of a set of training data. However, classification errors frequently occur when multiple classes have high and similar probabilities of occurrence (uncertainty due to ambiguity), or when training data are in insufficient quantity (uncertainty due to a lack of information). One possibility to increase classifiers reliability is by allowing their outputs to better reflect these uncertain situations. *Indeterminate* classifiers, which are able to predict more than one class in case of high uncertainty, have been introduced for this purpose.

While the idea of reject option, that is of not making a prediction for some specific data, has been around for some time [1], [2], [3], the idea of producing indeterminate or partial predictions is more recent [4], [5], [6], [7], [8]. These techniques are often used for problems where being cautious or reliable is as important as being accurate. When dealing with cost-sensitive applications, two main approaches can produce indeterminate predictions. The first one, directly inspired from the reject option, is to integrate costs of indeterminacy in the decision making [9], [10]. The second approach is to consider imprecise probability estimates rather than precise models. Indeed, in the last years several extensions of classical classifiers, such as C4.5 trees [11], Naive Bayes classifiers [5], nested dichotomies [12], [13] or Bayesian Model Averaging [14] consider such a setting.

Gen Yang is with the Kpler company, situated 23 rue du Renard, 75004 Paris, France. (email: yang.gen.mail@gmail.com).

Sébastien Destercke is with Sorbonnes Universite, Universite de Technologie de Compiegne, CNRS, UMR Heudiasyc, 57 Av. de Landshut, 60203 Compiègne (email: sebastien.destercke@hds.utc.fr).

Marie-Hélène Masson is with (1) Sorbonnes Universite, Universite de Technologie de Compiegne, CNRS, UMR Heudiasyc, 57 Av. de Landshut, 60203 Compiègne and (2) Université de Picardie Jules Verne, Amiens, France. (email: massomar@hds.utc.fr).

However, a problem still largely unsolved when using such settings is to determine sensible costs of indeterminate predictions. This is also necessary when having access to an expert willing to estimate these costs, in order to provide this expert with proper guidelines and/or questions. This can also be useful in applications naturally involving costs, such as imbalanced classification problems [15] or ordinal regression [16]

The costs of indeterminate predictions are also essential to compare the classifiers or models producing such predictions, and therefore to choose an optimal model. And although some specific solutions, discussed in Section V, have been proposed in the literature [4], [17], we are unaware of any work proposing generic guidelines applicable to any cost-sensitive indeterminate predictions.

Providing and discussing the relevance of such generic guidelines is the purpose of this paper. For this, we partially take inspiration from principled approaches [18], [9] (reviewed in Section III) extending accuracy (or 0/1 costs) to indeterminate predictions. We propose our guidelines and properties for costs of indeterminacy in Section IV, going from those we perceive as the most essential to those whose adoption seems largely contextual. To make these guidelines operational, we then propose in Section VI a simple yet generic formula satisfying most of our properties and relying on a single parameter measuring the decision maker aversion to indeterminacy. Section VII presents two illustrative experiments performed on ordinal regression problems, the first one dealing with the problem of tuning an imprecise probabilistic classifier, the second one with the problem of comparing indeterminate cost-sensitive classifiers. Necessary background knowledge and notations are provided in Section II.

II. Settings: cost functions and predictions

Multi-class classification problem is about assigning a prediction $\hat{y}$ to an observation $\mathbf{x}$ issued from the input feature space $\mathbf{X} = X_1 \times \ldots \times X_m$. When the prediction $\hat{y}$ is a singleton of the output space Ω, it is what we call a *determinate prediction*.

Requiring such determinate predictions is by far the most encountered setting in classification. Yet, there are some classes of problems where making indeterminate (*i.e.*, set-valued) predictions may be useful, for instance to pre-select some possible classes with weak but efficient classifiers, or to make more reliable predictions in sensible areas (*e.g.*, medical diagnosis, risk analysis, fault detection). An indeterminate prediction then consists in predicting a set $\hat{Y} \in 2^\Omega \setminus \emptyset$ of classes. A determinate classifier then becomes a special case where all sets are reduced to singletons.

We also assume that we are in a cost-sensitive setting: each class of the output space $\hat{y} \in \Omega$ is associated to a cost function

$c_{\hat{y}} : \Omega \to \mathbb{R}^+$, such that $c_{\hat{y}}(y)$ is the cost of predicting class $\hat{y} \in \Omega$ when $y \in \Omega$ is the ground-truth, *i.e.*, the observed class. In the rest of the paper, we will refer to the specific case of 0/1 costs ($c_{\hat{y}}(y) = 1$ when $\hat{y} \neq y$, 0 otherwise) as accuracy, to differentiate it from generic costs.

Example 1. *The interest of cost functions is to model the costs of making a wrong decision. For example, consider the problem of obstacle recognition where a vehicle needs to recognize in situation* $\mathbf{x}$ *whether it faces a human (*h*), a bicycle (*b*) or nothing (*n*) (i.e.* $Y = \{h, b, n\}$*).*

As both human and bicycle are obstacles to be avoided, a confusion between h *and* b *is not very important. Predicting* h *or* b *when there is nothing becomes more costly (the vehicle makes a unnecessary manoeuvre). Finally, predicting* n *when there is an obstacle* h *or* b *is a big mistake that could cause an accident. This kind of information can easily be expressed through predictive costs. Table I provides an example of a cost matrix modelling these information.*

TABLE I
COST MATRIX DEFINED ACCORDING TO THE RISK LEVEL OF SITUATIONS

	truth		
$c_{\hat{y}}(y)$	$y = h$	$y = b$	$y = n$
$\hat{y} = h$	0	1	2
$\hat{y} = b$	1	0	2
$\hat{y} = n$	4	4	0

In learning settings, these costs can be used both to

1) determine what is the optimal prediction of a new instance $\mathbf{x}$ when a probabilistic model is given, and
2) evaluate the average cost incurred by a predictive model, in order to compare models and pick the best one.

We will first recall how such costs are used in a determinate probabilistic setting, before discussing how they can be extended to accommodate indeterminacy.

A. Making and evaluating determinate predictions

In a probabilistic model, (determinate) predictions are based on the conditional probability functions $p(\cdot/\mathbf{x}) : \Omega \to [0, 1]$ of the class given $\mathbf{x}$. The classical means to compare different predictions is the expected cost

$$\mathbb{E}[c_{\hat{y}}] = \sum_{y \in \Omega} p(y/\mathbf{x}) c_{\hat{y}}(y) \tag{1}$$

of predicting/choosing a class $\hat{y}$. The prediction

$$\hat{y}^\star = \arg\min_{\hat{y} \in \Omega} \mathbb{E}[c_{\hat{y}}] \tag{2}$$

with the lowest expected cost is then chosen. An alternative way to interpret this decision process, that will be instrumental in the rest of this paper, is that it comes down to establish a preference order $\succ$ between the classes, defined as follows:

Definition 1. *Prediction* $\hat{y}_1 \in \Omega$ *is preferred to* $\hat{y}_2 \in \Omega$*, denoted* $\hat{y}_1 \succ \hat{y}_2$*, iff*

$$\mathbb{E}[c_{\hat{y}_2} - c_{\hat{y}_1}] > 0 \Leftrightarrow \mathbb{E}[c_{\hat{y}_2}] > \mathbb{E}[c_{\hat{y}_1}]. \tag{3}$$

The $\hat{y}^\star$ of Equation (2) is then equivalent to taking the maximal element of $\succ$. Definition 1 can be interpreted in the following way: $\hat{y}_1$ is preferred to $\hat{y}_2$ when the expected cost of exchanging $\hat{y}_1$ for $\hat{y}_2$ is positive. Note that when considering simple accuracy, this procedure reduces to comparing the probabilities of each class y and choosing the one having the maximal probability.

Beyond making optimal predictions once a model is learned, the notion of cost can also be used to assess the empirical average performance of a model on a set of I.I.D. test data $\mathcal{D} = (\mathbf{x}_i, y_i)_{i \in [1;N]}$, and therefore to compare two models. Let $f_k : \mathbf{X} \to \Omega$ be the decisions taken when applying Equation (2) to conditional probabilities p_k issued from different models. Then the average loss (cost) R_k, also known as *empirical risk*, incurred by f_k on $\mathcal{D}$ is

$$R_k = \frac{1}{N} \sum_{i=1}^{N} c_{f_k(\mathbf{x}_i)}(y_i), \tag{4}$$

and we can compare any pair f_k, f_ℓ of models to choose the one incurring the minimal average loss.

Remark 1. *In practice, Equation 4 can be derived from any decision function* f*, not necessarily obtained from conditional probabilities. In this case, the misclassification costs have to be integrated in the learning process to obtain optimal models, in contrast with probabilistic approaches. In the rest of the paper, to make the discussion easier to follow, we keep assuming a probabilistic model. Yet, our results extend easily to non-probabilistic decision functions allowing for indeterminate predictions.*

B. Making and evaluating indeterminate predictions with precise probabilities

In principle, adapting the cost-sensitive setting to indeterminate predictions is rather straightforward: we just have to define, for each set-valued prediction $\hat{Y} \in 2^\Omega \backslash \emptyset$, a cost function $c_{\hat{Y}} : \Omega \to \mathbb{R}^+$ where $c_{\hat{Y}}(y)$ is the cost of predicting the set $\hat{Y}$ when y is the observed class. This means that the equivalent cost matrix is no longer a square matrix, but has a number of rows equal to $2^{|\Omega|} - 1$.

Example 2. *We consider Example 1, but now allow for the possibility to predict sets of classes. That is, we may decide to predict that the potential obstacle may be "human or bicycle" (*$\{h, b\}$*), "bicycle or nothing" (*$\{b, n\}$*), "human or nothing" (*$\{h, n\}$*) or even "uncertain" (*$\{h, b, n\}$*). Table II summarizes the extended cost matrix.*

TABLE II
COST MATRIX WITH INDETERMINATE PREDICTIONS

	truth		
$c_{\hat{Y}}(y)$	$y = h$	$y = b$	$y = n$
$\hat{Y} = \{h\}$	0	1	2
$\hat{Y} = \{b\}$	1	0	2
$\hat{Y} = \{n\}$	4	4	0
$\hat{Y} = \{h, b\}$	$c_{\{h,b\}}(h)$	$c_{\{h,b\}}(b)$	$c_{\{h,b\}}(n)$
$\hat{Y} = \{b, n\}$	$c_{\{b,n\}}(h)$	$c_{\{b,n\}}(b)$	$c_{\{b,n\}}(n)$
$\hat{Y} = \{h, n\}$	$c_{\{h,n\}}(h)$	$c_{\{h,n\}}(b)$	$c_{\{h,n\}}(n)$
$\hat{Y} = \{h, b, n\}$	$c_{\{h,b,n\}}(h)$	$c_{\{h,b,n\}}(b)$	$c_{\{h,b,n\}}(n)$

Once this cost matrix is defined, producing optimal indeterminate predictions or evaluating an indeterminate classifier can be done by simply extending the domain of the output space from Ω to $2^\Omega \backslash \emptyset$ in Equations (2) and (4).

It should be noted that while such a matrix is useful to visualise how costs of indeterminacy can be defined, it is not computationally practical if there are many classes, as an exponentially increasing number of alternatives is involved in Equation (2). This is why most existing probabilistic proposals [9], [19], [4] (reviewed in next sections) focus on providing formulas for which only a limited number of indeterminate predictions have to be considered to obtain the optimal one.

C. Making and evaluating indeterminate predictions with imprecise probabilities

Indeterminate predictions, by being more cautious, are meant to be more reliable than determinate ones. In practice, this reliability need can also be addressed when representing the uncertainty, prior to the decision-making stage. This is one of the core idea of the imprecise probability framework [20], where uncertainty is modelled by a (convex) set $\mathcal{P}$ of possible probability distributions instead of a single one. This set, also called *credal set* [21], represents the uncertainty about the true distribution that cannot be perfectly identified (*e.g.*, due to noises, biases, lack or imprecision of data, ...). Example 3 shows how our obstacle recognition illustrative problem can be represented in the imprecise probability framework.

Example 3. *We consider again Example 1. A standard probabilistic classifier could yield precise estimates such as:*

$$p(h/\mathbf{x}) = 0.1, \quad p(b/\mathbf{x}) = 0.3, \quad p(n/\mathbf{x}) = 0.6.$$

In the imprecise probability framework, a classifier could return interval-valued estimates such as:

$$p(h/\mathbf{x}) \in [0; 0.2], \quad p(b/\mathbf{x}) \in [0.3; 0.4], \quad p(n/\mathbf{x}) \in [0.4; 0.6].$$

The probabilities are now interval-valued, the width of which represent the uncertainty about the estimations. A large interval means that we have poor or inconsistent information about the class. On the contrary, the classifier may output "precise" estimations (or narrow interval) when it has enough information (data).

The credal set $\mathcal{P}$ is the set of all possible precise probabilities $(p(h/\mathbf{x}), p(b/\mathbf{x}), p(n/\mathbf{x}))$ within these interval bounds. Here $\mathcal{P}$ is a polytope defined by the convex hull of its four vertices in a three dimensional space:

$$\begin{aligned}\mathcal{P} = CH\{&(0, 0.4, 0.6); (0.2, 0.3, 0.5)\\ &; (0.2, 0.4; 0.4); (0.1, 0.3, 0.6)\}\end{aligned}$$

where CH stands for convex hull.

As $\mathcal{P}$ is a polytope and the expectation $\mathbb{E}$ is a linear operator, expectations on $\mathcal{P}$ can be represented by lower and upper bounds $[\underline{\mathbb{E}}; \overline{\mathbb{E}}]$, on which we can base our decision process. Given a function $c : \Omega \to \mathbb{R}$, the lower expectation reads

$$\underline{\mathbb{E}}[c] = \min_{p \in \mathcal{P}} \mathbb{E}[c] = \min_{p \in \mathcal{P}} \sum_{y \in \Omega} p(y/\mathbf{x}) c(y), \tag{5}$$

The upper expected cost $\overline{\mathbb{E}}$ is obtained by replacing $\min$ with $\max$ and both are dual, in the sense that $\underline{\mathbb{E}}(c) = -\overline{\mathbb{E}}(-c)$.

There are then several ways to extend the notion of expected cost and classical decision making to credal sets [22], some of which still producing determinate predictions, other producing indeterminate ones, on which we will focus. More precisely, we will use the notion of maximality, that builds a partial order $\succ_\mathcal{M}$ over the classes using the following definition:

Definition 2 (Maximality)**.**

$$\hat{y}_i \succ_\mathcal{M} \hat{y}_j \Leftrightarrow \underline{\mathbb{E}}[c_{\hat{y}_j} - c_{\hat{y}_i}] > 0 \Leftrightarrow \mathbb{E}[c_{\hat{y}_j} - c_{\hat{y}_i}] > 0 \ \ \forall p \in \mathcal{P}. \tag{6}$$

This is a formal extension of Definition 1, as we retrieve it when $\mathcal{P}$ is reduced to a singleton. Equation (6) can be interpreted as follows: $\hat{y}_i$ is preferred to $\hat{y}_j$ if exchanging $\hat{y}_i$ for $\hat{y}_j$ always has a positive cost whatever the given probability distribution. Note that obtaining the order $\succ$ requires to perform at worst $K(K-1)$ computations (where $K = |\Omega|$ is the cardinal of Ω), one for each pair of classes. As in the classical case, we can take as prediction $\hat{Y}$ the maximal elements of the induced order:

$$\hat{Y} = \Big\{ \hat{y}_i \in \Omega \,|\, \nexists \hat{y}_j : \hat{y}_j \succ_\mathcal{M} \hat{y}_i \Big\}. \tag{7}$$

There are other extensions of Definition 1 using credal sets and producing indeterminate predictions (i.e., interval dominance, E-admissibility), yet for the purpose of this paper, it is enough to focus on the most used and well-founded one. We refer to Troffaes [22] for a detailed discussions of the different rules.

Example 4. *Let us use the maximality criterion on the interval-valued probabilities given in Example 3 and the costs given in Example 1 to infer $\hat{Y}_\mathcal{M}$.*

Let us first consider the pair $\{b, n\}$ and the difference $c_n - c_b$. We have

$$\begin{aligned}\underline{\mathbb{E}}[c_{\{n\}} - c_{\{b\}}] &= \min \big(3p(h/\mathbf{x}) + 4p(b/\mathbf{x}) - 2p(n/\mathbf{x})\big)\\ &= 3 * 0.1 + 4 * 0.3 - 2 * 0.6 = 0.3\end{aligned}$$

which is obtained for the extreme point $(0.1, 0.3, 0.6)$ of Example 3. As this is positive, we can infer $b \succ_\mathcal{M} n$. Furthermore, for the pair $\{h, b\}$ we have

$$\begin{aligned}\underline{\mathbb{E}}[c_{\{h\}} - c_{\{b\}}] &= \min \big(-1p(h/\mathbf{x}) + 1p(b/\mathbf{x}) + 0p(n/\mathbf{x})\big)\\ &= -0.2 + 0.3 = 0.1\end{aligned}$$

obtained for the extreme point $(0.2, 0.3, 0.5)$. This is again positive, so $b \succ_\mathcal{M} h$, and b ends up being the only non-dominated class, hence Equation (7) gives us $\hat{Y} = \{b\}$.

In contrast with the precise case of Section II-B, it is not necessary to expand the original determinate costs $c_{\hat{y}}$ to generic sets $c_{\hat{Y}}$ in order to produce the prediction $\hat{Y}$ from a credal set. This means, in particular, that with generic costs the computational complexity of producing indeterminate predictions increases quadratically in the number of classes, rather than exponentially. However, we still needs to define costs $c_{\hat{Y}}$ to be able to compare the predictions of two classifiers based on credal sets.

If defining cost of determinate predictions can already be difficult, defining the cost functions of indeterminate predictions is even more difficult. For instance, costs of determinate prediction can be extracted from expert information, from the actual costs (in money) of making mistakes, or from the structure of Ω (examples include ordinal classification, multilabel problems, hierarchical classification, ...). Such sources cannot usually be used when adding imprecision to the predictions, making more difficult the task to assess the cost of indeterminacy.

This assessment is the question we address in this paper, which can take the following form: considering Example 2, how to fill up the missing numbers in matrix in Table II so that they make sense? Although some specific proposals exist [4], [17], there are to our knowledge no general axiomatic guidelines applicable to any cost functions that would help to define principled costs of indeterminate classifications. We provide such guidelines in Section IV, while the next section reviews the case of standard accuracy, for which convincing, principled answers already exist.

III. Standard accuracy

Accuracy for determinate classification corresponds to the cost function $c_{\hat{y}}(y) = 0$ if $y = \hat{y}$, and $c_{\hat{y}}(y) = 1$ otherwise. A common proposal to adapt this to indeterminate classification is the so-called discounted accuracy, such that $c_{\hat{Y}}(y) = 1 - 1/|\hat{Y}|$ (where $|\hat{Y}|$ stands for the cardinal of $\hat{Y}$) if $y \in \hat{Y}$, $c_{\hat{Y}}(y) = 1$ else. Yet such a choice has been strongly criticized by Zaffalon *et al.* [18], on the basis that it considers the value of predicting $\hat{Y}$ to be the same as choosing randomly within $\hat{Y}$. In other words, it gives no value to cautiousness, which is confused with randomness. We will see that a similar argument holds for costs of indeterminate predictions in Section IV.

A. Utility discounted accuracy

Rather than adopting the discounted accuracy, Zaffalon *et al.* [18] propose to keep $c_{\hat{Y}}(y) = 1$ if $y \notin \hat{Y}$, but to take $c_{\hat{Y}}(y) = 1 - g(1/|\hat{Y}|)$ if $y \in \hat{Y}$, where g is a utility function on $[0, 1]$ such that $g(1/|\hat{Y}|) \geq 1/|\hat{Y}|$, $g(1) = 1$ and $g(0) = 0$. They interpret g as a concave function modelling the risk-aversion (i.e., the utility), or the cautiousness-seeking attitude of the decision maker. In particular, they propose specific quadratic forms of g fitted by specifying one additional point, among which the following one:

$$g(x) = -0.6x^2 + 1.6x \tag{8}$$

That corresponds to fixing $g(1/2) = 0.65$, a small increase of the initial discounted accuracy $1/2$. The function is pictured in Figure 1. This specific utility keeps some appealing properties of the discounted accuracy: when the correct class is not in $\hat{Y}$ the cost should stay at 1, when the prediction is both precise and accurate ($|\hat{Y}| = 1$, $y \in \hat{Y}$) then the cost should be 0. Moreover, this utility function should express our risk-aversion by giving a smaller cost (or higher accuracy) to imprecise but correct predictions compared to the discounted accuracy.

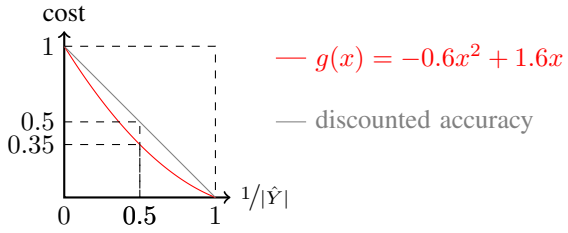


Fig. 1. Risk-averse discounted accuracy

Table III provides the matrix obtained when Equation (8) is applied to Example 1. The very basic properties we will propose in the next sections will rely on a similar observation concerning the costs of indeterminate predictions. However, there will be some differences with accuracy, notably because the costs of making mistakes in the prediction will not always be the same.

TABLE III
Cost matrix defined using the utility discounted accuracy

$c_{\hat{Y}}(y)$	truth $y = h$	$y = b$	$y = n$
$\hat{Y} = \{h\}$	0	1	1
$\hat{Y} = \{b\}$	1	0	1
$\hat{Y} = \{n\}$	1	1	0
$\hat{Y} = \{h, b\}$	0.35	0.35	1
$\hat{Y} = \{b, n\}$	1	0.35	0.35
$\hat{Y} = \{h, n\}$	0.35	1	0.35
$\hat{Y} = \{h, b, n\}$	0.54	0.54	0.54

B. F_β measure

Del Coz and Bahamonde [9] make a similar proposal to evaluate indeterminate predictions based on the well-known F_β measure, which computes the harmonic mean (weighted by the coefficient β) between *precision* P and *recall* R:

$$F_\beta = \frac{(1 + \beta^2)P \cdot R}{\beta^2 P + R}. \tag{9}$$

This approach has been shown in [18] to be equivalent to choosing a specific instance of utility function g. The *precision* P measures how many predicted classes in $\hat{Y}$ are relevant, and the *recall* R measures how many relevant classes are predicted. Therefore, for an indeterminate prediction $\hat{Y}$ and the truth $y \in \Omega$, we can compute the following contingency table:

	$y = z$	$y \neq z$	$\sum$
$z \in \hat{Y}$	TP	FP	$\lvert\hat{Y}\rvert$
$z \notin \hat{Y}$	FN	TN	$\lvert\Omega\rvert - \lvert\hat{Y}\rvert$
$\sum$	1	$\lvert\Omega\rvert - 1$	

The table expresses four situations: the sets of "true positive" $\{z \in \hat{Y} : y = z\}$ and "false positive" $\{z \in \hat{Y} : y \neq z\}$ on one hand, with cardinality TP and FP, respectively; and the sets of "false negative" and "true negative" on the other hand, with cardinality FN and TN, respectively. Therefore, $TP+FP$ gives the cardinal of the predicted set $\hat{Y}$ and, as there is only one true class, $TP + FN$ is equal to one. Therefore

precision and recalls simplify into

$$P(\hat{Y}, y) = \frac{TP}{TP + FP} = \frac{\mathbb{1}_{\hat{Y}}(y)}{|\hat{Y}|},$$
$$R(\hat{Y}, y) = \frac{TP}{TP + FN} = TP = \mathbb{1}_{\hat{Y}}(y),$$

where $\mathbb{1}_{\hat{Y}}$ is the indicator function of $\hat{Y}$. As F_β measures the reward of a prediction, and not its cost, the costs of indeterminate prediction are given by $1-F_\beta$. Table IV provides the matrix that would be obtained in Example 1 when β is set to 1.

TABLE IV
COST MATRIX DEFINED USING F_1 MEASURE

$c_{\hat{Y}}(y)$	truth $y=h$	$y=b$	$y=n$
$\hat{Y}=\{h\}$	0	1	1
$\hat{Y}=\{b\}$	1	0	1
$\hat{Y}=\{n\}$	1	1	0
$\hat{Y}=\{h,b\}$	1/3	1/3	1
$\hat{Y}=\{b,n\}$	1	1/3	1/3
$\hat{Y}=\{h,n\}$	1/3	1	1/3
$\hat{Y}=\{h,b,n\}$	0.5	0.5	0.5

In the case of accuracy, both approaches correspond to reward cautiousness when making indeterminate predictions containing the true class. Because mistake costs are constant, this reward is always the same for a given cardinality (e.g., $c_{\{h,n\}}(h) = c_{\{h,n\}}(n)$ in our examples), and similarly the cost of making a mistake is also constant (e.g., $c_{\{h,b\}}(n) = c_{\{b,n\}}(h) = c_{\{h,n\}}(b)$). In the next section we shall propose guidelines for general costs that, as we shall discuss in Section V, are consistent with 0/1 case.

IV. THE COST OF INDETERMINACY: GUIDELINES

This section proposes guidelines to build the function $c_{\hat{Y}}$, by defining properties that it should or could follow. We will start with properties that it should follow, those that make indeterminate predictions possible and that belong to common sense. We will then discuss some properties that may be desirable in some settings, and undesirable in others. This will allow us to revisit some existing proposals, before proposing some generic formulation (similar to the g function in Equation (8) for accuracy).

A. Making indeterminacy possible

The logic of discounted accuracy (Section III) gives a first way to define costs for indeterminate predictions, that we will call *discounted costs*.

Definition 3. *Given costs $c_{\hat{y}}$ for determinate predictions $\hat{y} \in \Omega$, a discounted cost is defined as the average cost (noted $\bar{c}$) over its precise components $\hat{y} \in \hat{Y}$, such that*

$$\bar{c}_{\hat{Y}}(y) = \frac{\sum_{\hat{y}\in\hat{Y}} c_{\hat{y}}(y)}{|\hat{Y}|}. \tag{10}$$

Discounted costs reduce to discounted accuracy when considering 0/1 costs. Table V gives the completed matrix of Example 2 with the discounted costs of Definition 3.

TABLE V
COST MATRIX WITH DISCOUNTED COSTS

$c_{\hat{Y}}(y)$	truth $y=h$	$y=b$	$y=n$
$\hat{Y}=\{h\}$	0	1	2
$\hat{Y}=\{b\}$	1	0	2
$\hat{Y}=\{n\}$	4	4	0
$\hat{Y}=\{h,b\}$	1/2	1/2	2
$\hat{Y}=\{b,n\}$	5/2	2	1
$\hat{Y}=\{h,n\}$	2	5/2	1
$\hat{Y}=\{h,b,n\}$	5/3	5/3	4/3

We will say that an indeterminate prediction is *possible* if it satisfies the following definition:

Definition 4 (Possibility of an indeterminate prediction $\hat{Y}$)**.** *An imprecise prediction $\hat{Y}$ is said to be possible if there exists a probability distribution p such that:*

$$\mathbb{E}[c_{\hat{Y}}] < \min_{\hat{y}\in\hat{Y}} \mathbb{E}[c_{\hat{y}}].$$

Now, if we want to make indeterminate predictions, a first obvious requirement is that at least one of them should be possible. This can be translated by the following property, that we call *Possibility of indeterminate predictions*.

Property 1 (Possibility of indeterminate predictions)**.** *Costs are said to make indeterminate predictions possible if there is at least one indeterminate prediction $\hat{Y} \in 2^\Omega \setminus \emptyset$ that is possible, according to Definition 4.*

This is an essential property, since if we do not have it, then speaking of indeterminate predictions makes no sense at all, as they would be impossible to obtain. In light of this, we can easily show that discounted costs, or any cost function c such that $c_{\hat{Y}} > \bar{c}_{\hat{Y}}$, do not make indeterminate predictions possible.

Proposition 1. *The discounted costs $\bar{c}_{\hat{Y}}(y)$ are such that for any $\hat{Y} \in 2^\Omega \setminus \emptyset$*

$$\mathbb{E}[\bar{c}_{\hat{Y}}(y)] \geq \min_{\hat{y}\in\hat{Y}} \mathbb{E}[c_{\hat{y}}].$$

Proof: We have that

$$\mathbb{E}[\bar{c}_{\hat{Y}}] = \sum_{y\in\Omega} p(y/\mathbf{x})\bar{c}_{\hat{Y}}(y) = \frac{1}{|\hat{Y}|}\sum_{y\in\Omega} p(y/\mathbf{x}) \sum_{\hat{y}\in\hat{Y}} c_{\hat{y}}(y)$$
$$= \frac{1}{|\hat{Y}|}\sum_{\hat{y}\in\hat{Y}}\sum_{y\in\Omega} p(y/\mathbf{x})c_{\hat{y}}(y) = \frac{1}{|\hat{Y}|}\sum_{\hat{y}\in\hat{Y}} \mathbb{E}[c_{\hat{y}}].$$

As $\mathbb{E}[\bar{c}_{\hat{Y}}]$ is the average of values $\mathbb{E}[c_{\hat{y}}]$ for $\hat{y} \in \hat{Y}$, it cannot by definition be lower than $\min_{\hat{y}\in\hat{Y}} \mathbb{E}[c_{\hat{y}}]$. ∎

Hence, $c_{\hat{Y}} \not\geq \bar{c}_{\hat{Y}}$ is a necessary constraint for indeterminate prediction $\hat{Y}$ to be possible, showing that discounted costs, similarly to the discounted accuracy in the 0/1 case, are not ideal choices to make indeterminate predictions.

Example 5. *Consider Table II with the following values for $\hat{Y} = \{h, b\}$:*

$$c_{\{h,b\}}(h) = c_{\{h,b\}}(b) = 1 \text{ and } c_{\{h,b\}}(n) = 2$$

which are higher than the discounted costs of Table V. Since $c_{\{h,b\}} \geq c_{\{h\}}$ and $c_{\{h,b\}} \geq c_{\{b\}}$, then $\mathbb{E}[c_{\{h,b\}}] \geq \mathbb{E}[c_{\{h\}}]$ and

$\mathbb{E}[c_{\{h,b\}}] \geq \mathbb{E}[c_{\{b\}}]$ *for any probability. On the converse, if we now take the value*

$$c_{\{h,b\}}(h) = c_{\{h,b\}}(b) = 1/4 \text{ and } c_{\{h,b\}}(n) = 2$$

that are lower than the discounted cost, then $\{h, b\}$ *becomes a possible prediction. Indeed, the uniform probability gives* $\mathbb{E}[c_{\{h,b\}}] = 5/6$*, which is lower than* $\mathbb{E}[c_{\{h\}}] = \mathbb{E}[c_{\{b\}}] = 1$.

A stronger property is to require every possible indeterminate prediction to satisfy the necessary condition

Property 2 (indeterminacy permissiveness)**.** *Costs are said to be indeterminacy permissive if, for all* $\hat{Y} \in 2^\Omega \setminus \emptyset$*, we have*

$$c_{\hat{Y}} \not\geq \bar{c}_{\hat{Y}}$$

where $c_{\hat{y}_1} \leq c_{\hat{y}_2}$ *denote element-wise inequality, that is :*

$$c_{\hat{y}_1} \leq c_{\hat{y}_2} \Leftrightarrow \forall y \in \Omega : c_{\hat{y}_1}(y) \leq c_{\hat{y}_2}(y). \tag{11}$$

Unless we want some particular indeterminate prediction to be impossible, or strongly penalized during evaluation, this property looks sensible. Let us now enumerate some properties that are not related to allowing for indeterminate predictions, but can be seen as common-sense, and should therefore be satisfied in our opinion.

B. Common-sense properties

A first common sense property, similar to those addressed in the case of accuracy (Section III), is that cautiousness should be rewarded to some extent. Said in other words, an indeterminate prediction $\hat{Y}$ should be rewarded if it contains the true value y, *i.e.*, if $y \in \hat{Y}$.

Property 3 (Reward rightful cautiousness)**.** *Costs are said to reward rightful cautiousness if, for any* $\hat{Y}$*, we have*

$$y \in \hat{Y} \Rightarrow c_{\hat{Y}}(y) < \bar{c}_{\hat{Y}}(y).$$

This property complements the one of indeterminacy permissiveness (Property 2), as it specifies at least a subset of values for which the inequality $c_{\hat{Y}} \not\geq \bar{c}_{\hat{Y}}$ should be satisfied. For example, Property 3 imposes $c_{\{h,b\}}(h) \leq 1/2$ in Table II, but does not constraint $c_{\{h,b\}}(n)$, as $n \notin \{h, b\}$.

Up to now, we have explored properties making indeterminate predictions possible, yet another natural requirement is that determinate predictions should also remain possible. In particular, this means that none of the indeterminate prediction should have a cost always lower than the minimum of the determinate costs of its components.

Property 4 (Non dominance of indeterminate predictions)**.** *Indeterminate predictions are said to be non-dominant if, for all* $\hat{Y} \in 2^\Omega \setminus \emptyset$*, we have*

$$c_{\hat{Y}}(y) \geq \min_{\hat{y} \in \hat{Y}} c_{\hat{y}}(y).$$

For example, we should have $c_{\{b,n\}}(h) \geq 1$ or $c_{\{h,b\}}(n) \geq 2$ in Table II. Put together, Properties 2 and 4 already provide some constraints that $c_{\hat{Y}}$ should follow. Another one is that, if the cost vectors formed by the elements of a indeterminate predictions are the same, up to a permutation, for two observed classes y and y', then $c_{\hat{Y}}(y)$ and $c_{\hat{Y}}(y')$ should be identical, due to symmetry reasons. Before giving the related property, let us introduce some notation. If $\hat{Y} = \{\hat{y}_1, \ldots, \hat{y}_n\}$ is an indeterminate prediction and y a class, we will denote by $C_{(\hat{Y})}(y) = (c_{\hat{y}_{(1)}}(y), \ldots, c_{\hat{y}_{(n)}}(y))$ the ordered vector such that $c_{\hat{y}_{(i)}}(y) \leq c_{\hat{y}_{(i+1)}}(y)$.

Property 5 (permutation invariance)**.** *Costs of indeterminate prediction* $\hat{Y}$ *are said to be permutation invariant if, for two different classes* y *and* y'*, we have*

$$C_{(\hat{Y})}(y) = C_{(\hat{Y})}(y') \Rightarrow c_{\hat{Y}}(y) = c_{\hat{Y}}(y').$$

For instance, in Example 2, we do have that $C_{\{h,b,n\}}(h) = C_{\{h,b,n\}}(b) = (0, 1, 4)$, hence if we require permutation invariance, we should have $c_{\{h,b,n\}}(h) = c_{\{h,b,n\}}(b)$ in Table II. We do not see any reason to not require permutation invariance, hence we also consider it as a desirable property.

C. Context-dependent properties

We will now study properties whose desirability may depend on the specific use of indeterminate predictions. In particular, we will differentiate two different settings:

- The **filtering** setting, where the goal is to filter some classes (possibly with computationally cheap methods) before applying a costlier procedure. In this case, it seems essential that the indeterminate prediction contains the true class, otherwise the costly procedure is applied for nothing;
- The **choice** setting, in which case the indeterminate prediction aims at giving cautious predictions to the decision maker, that can then make a choice among them if she/he desires or takes other actions such as gathering more information. In this case, an indeterminate prediction is a valuable information by itself, as it may indicate to the decision maker an ambiguous or poorly informed situation, therefore it may be desirable even if the true class is not within it.

Again, we will illustrate these two settings with Example 7.

1) Mistake behaviour: When making mistakes with indeterminate predictions, we can have two behaviours: either penalizing such mistakes, or seeking cautiousness even when making mistakes. This can be translated in the following two properties:

Property 6 (Mistake averse)**.** *Costs are said to be mistake averse if, for any* $\hat{Y}$*, we have*

$$y \notin \hat{Y} \Rightarrow c_{\hat{Y}}(y) \geq \bar{c}_{\hat{Y}}(y).$$

Property 7 (Cautiousness seeking)**.** *Costs are said to be cautiousness seeking if, for any* $\hat{Y}$*, we have*

$$y \notin \hat{Y} \Rightarrow c_{\hat{Y}}(y) \leq \bar{c}_{\hat{Y}}(y).$$

Clearly, these two properties complement Property 3 in opposite ways, by specifying the desirable behaviour of indeterminate predictions in case of mistake. In the filtering setting, making a mistake when being indeterminate is clearly penalizing, as we will incur the cost of the additional procedure

without any benefit, therefore Property 6 is more adapted than Property 7 to this setting.

In the choice setting, we think that the choice is not so obvious. Of course the decision maker could desire to punish the fact of being indeterminate and wrong by considering that this is worse than being determinate and wrong. Yet, it could also be the case that the decision maker prefers to be cautious when missing some information, even if the prediction is wrong. Indeed, by explicitly stating (through indeterminacy) that the situation is ambiguous, further introspection could lead to the right results, while expressing certainty would not lead to such introspection. A similar remark could be made about automatic systems where indeterminacy triggers a warning towards a human operator, while determinacy let the automatic system handle the situation. Therefore, we think that in the choice setting, both Properties 6 and 7 can be conceived.

Example 6. *Consider Table II with the following values for* $\hat{Y} = \{h, b\}$*:*

$$c_{\{h,b\}}(h) = c_{\{h,b\}}(b) = 1/4 \text{ and } c_{\{h,b\}}(n) = 3$$

which satisfy all previous properties as well as Property 6. In this, $\{h, b\}$ *is still a possible prediction (for example by taking* $p(h/\mathbf{x}) = p(b/\mathbf{x}) = 1/2$*), but the uniform probability gives* $\mathbb{E}[c_{\{h,b\}}] = 7/6$*, which is now higher than* $\mathbb{E}[c_{\{h\}}] = \mathbb{E}[c_{\{b\}}] = 1$*, in contrast with what happens in Example 5.*

2) Cost-sensitivity to correctness: Property 5 tells us that identical cost vectors should result into identical costs for the related indeterminate predictions. Yet, it does not specify any behaviour when the vectors are different. The two following properties address this situation.

Property 8 (Correctness cost-insensitivity)**.** *Costs are said to be insensitive in case of correctness if, for any* $\hat{Y}$ *and any two classes* $y, y' \in \hat{Y}$*, we have* $c_{\hat{Y}}(y) = c_{\hat{Y}}(y')$*.*

Property 9 (Correctness cost-sensitivity)**.** *Costs are said to be correctness sensitive if we can have a* $\hat{Y}$ *and two different classes* y *and* y'*, such that*

$$C_{(\hat{Y})}(y) \neq C_{(\hat{Y})}(y') \text{ and } c_{\hat{Y}}(y) \neq c_{\hat{Y}}(y').$$

In a fully automatic procedure like in the filtering setting where the goal is to filter the relevant classes and to leave the final decision to a more specialized classifier, we think that the first property fits better, as the cost then corresponds to the cost of using the specialized classifier, which is independent of the true class. However, Property 9 could be useful in some situations, namely when an expert is able to assign different costs according to the truth. Example 7 gives an illustration of these two settings.

Example 7. *To simplify things, assume that in the problem of obstacle recognition, only the two classes* h *and* n *are present. The cost matrix given in Table VI satisfies properties 2, 3 and 8. Using simple computations, we can express the associated decision rule as a function of* $p(n/\mathbf{x})$*. This decision rule is represented in the upper part of Figure 2.*

We consider now that the opinion of an expert is that deciding $\{h, n\}$ *when the truth is* $\{n\}$ *is a kind of false alarm and thus should be more costly than deciding* $\{h, n\}$ *when the truth is* $\{h\}$*. Assume that the expert roughly states that it should be* 3 *times more costly, and that to fix cost values we impose the expected cost of* $\{h, n\}$ *to be* 0.5 *when* $p(h) = p(n)$ *(as in Table VI). This gives the cost matrix expressed in Table VII, which now follows properties 2, 3 and 9.*

TABLE VI
COST MATRIX WITH INDETERMINATE PREDICTIONS

$c_{\hat{Y}}(y)$	truth $y = h$	truth $y = n$
$\hat{Y} = \{h\}$	0	2
$\hat{Y} = \{n\}$	4	0
$\hat{Y} = \{h, n\}$	0.5	0.5

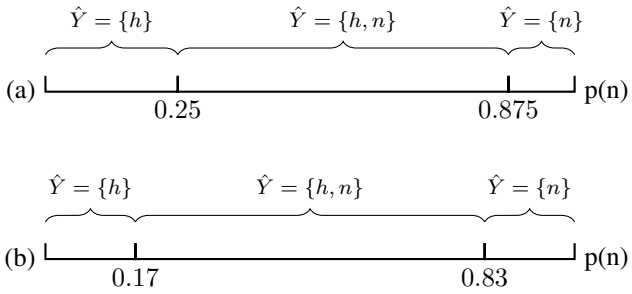


Fig. 2. Graphical representation of the decision as a function of $p(n/\mathbf{x})$; (a) : decision with a cost matrix in Table VI ; (b) : decision with a cost matrix in Table VII.

TABLE VII
COST MATRIX WITH INDETERMINATE PREDICTIONS

$c_{\hat{Y}}(y)$	truth $y = h$	truth $y = n$
$\hat{Y} = \{h\}$	0	2
$\hat{Y} = \{n\}$	4	0
$\hat{Y} = \{h, n\}$	0.25	0.75

In this case, the decision is slightly modified and is represented in the lower part of Figure 2. As could be expected, the boundary between $\{h, n\}$ *and* $\{h\}$ *has been shifted to the left due to the decrease of* $c_{\{h,n\}}(h)$*: the decision* $\{h, n\}$ *is more favoured for small values of* $p(n/\mathbf{x})$ *than with the cost matrix in Table VI. On the other hand, as the cost* $c_{\{h,n\}}(n)$ *has been raised, the boundary between* $\{h, n\}$ *and* $\{n\}$ *has also been shifted to the left, what penalizes the decision* $\{h, n\}$ *for high values of* $p(n/\mathbf{x})$*.*

From Example 7, we can also notice that fulfilling Property 8 is more constraining in terms of cost definition, especially in a binary setting where it means that the expected cost of the (only) indeterminate prediction does not depend at all on the values of $p(n)$ and $p(h)$. While this certainly makes easier the determination of those costs, it may not always be a desirable value, as illustrated in Example 7.

3) Upper boundedness: Similarly to Property 4, it may be desirable to bound the cost of indeterminate predictions by above:

Property 10 (Upper bounded)**.** *An indeterminate prediction*

is upper-bounded if for all $\hat{Y} \in 2^\Omega \setminus \emptyset$*, we have*

$$c_{\hat{Y}}(y) \leq \max_{\hat{y} \in \hat{Y}} c_{\hat{y}}(y).$$

That this property is as desirable as Property 4 is not always obvious. Indeed, in the filtering setting, the cost of the additional procedure may go beyond the maximal cost, as in case of mistake we may want to give a very strong penalty to avoid using the procedure for nothing. Requiring this property, however, makes perfect sense in the choice setting, as whatever the final choice of the decision maker, the incurred cost cannot be bigger than the maximal one, whatever the truth.

It should be noted that if one requires both Properties 4 and 10, then it implies that when $c_{\hat{y}}(y) = c_{\hat{y}'}(y)$ for any $\hat{y}, \hat{y}'$ in some indeterminate prediction $\hat{Y}$, we must have $c_{\hat{Y}}(y) = c_{\hat{y}}(y)$. For instance, in the matrix of Table II, this enforces $c_{\{h,b\}}(n) = 2$.

Table VIII provides a synthetic view about the different properties discussed in this paper, as well as whether we perceive their satisfaction as necessary, optional or even unwarranted in some setting. Of course, this table should be considered as indicative of a general case, and not in any way mandatory, as some peculiar situations may require peculiar cost definitions.

V. Review of related works

In this section, we will re-examine some existing proposals in the light of our properties.

a) Utility-discounted accuracy : The proposal of Zaffalon *et al.* [18] is quite interesting, as it satisfies absolutely all our proposed properties (apart from Property 9 which does not happen in case of 0/1 costs), yet this is mostly due to the inherent symmetry and constant values of the considered costs.

Even if that was not needed (due to the strong theoretical foundations of the initial proposal), this shows that this proposal is very sensible in the case of 0/1 costs.

b) The optimum class-selective rejection rule: Ha [4] elaborated an indeterminate classifier based on partial rejection rules. Its goal is to find the optimal error-reject trade-off using a specific loss structure (cost matrix) defined as follows :

$$c_{\hat{Y}}(y) = L_{\hat{Y}}(y) + L_{ip}(\hat{Y}),$$

where $L_{\hat{Y}}(y) = 0$ if the true class y is included in the prediction $\hat{Y}$, and otherwise $L_{\hat{Y}}(y) = \eta(y)$, with $\eta(y)$ modelling the loss of missing the true class y. $L_{ip}(\hat{Y}) = \delta(|\hat{Y}| - 1)$ where δ is a constant parameter representing the cost of being imprecise, with the condition that $\delta < 1/2\eta(y)$ for all y. The obtained cost matrix when $\Omega = \{a, b, c\}$ is given by Table IX. The condition $\delta < 1/2\eta(y)$ ensures that Properties 2 and 3 are satisfied. Furthermore, the proposal also satisfies basic properties 4 and 5.

Regarding context-dependent properties, this proposal satisfies Properties 8 and 6, and does not satisfy Property 10. In the light of our discussion of those properties, it is clear that this proposal is better fitted to the filtering than to the choice setting, which is precisely the setting in which Ha [4] sets his work.

It should however be noted that the proposal only considers specific cost matrices where the misclassification cost only depends on the true class, and not on the mistaken prediction. As recalled in Section II-B, the main reason for this is that such a constraint allows one to use efficient algorithms to find the prediction having minimal expected cost.

c) Cost matrix defined using different assumptions: Abellan and Masegosa [17] propose another measurement for imprecise classifiers. Their main goal is to be able to compare indeterminate classifiers, since the indeterminacy of predictions follows in their case from using imprecise probabilities. This means that they do not face the problem of an exponentially increasing complexity when making the predictions. The cost matrix they propose, once a linear change is applied to have $c_y(y) = 0$ (this is done to compare with previous approaches), is as follows:

- if $y \in \hat{Y}$, then
$$c_{\hat{Y}}(y) = \log |\hat{Y}|;$$
- if $y \notin \hat{Y}$, then
$$c_{\hat{Y}}(y) = \log |\Omega| \left(\frac{\max_{\hat{y} \in \hat{Y}} c_{\hat{y}}(y)}{|\Omega| - 1} + 1 \right).$$

This satisfies basic Properties 4, 10 and 5, yet there are no guarantee that Properties 2 and 3 will be satisfied, a clear potential drawback. The proposal also does modify the initial costs (values $c_{\hat{y}}(y)$ are modified), another potential drawback. It satisfies Properties 6 and 8, which seems to indicate that it is more fitted to a filtering setting, yet it is not entirely clear that Abellan and Masegosa [17] considered such a purpose, as the paper only uses the costs for comparing cost-sensitive indeterminate classifiers.

VI. Generic formulation for cost of indeterminacy

In this section, we propose a general way of instantiating costs for indeterminate classifiers from the initial determinate costs $c_{\hat{y}}$, based on the notion of utility discounted cost introduced in Section III, that we will call "$p-$discounted costs". Of course, as for usual costs, costs of indeterminacy should be defined on a case-by-case basis in applications, with the help of experts. The goal of the presented formula is to provide an easy way to perform systematic investigation and comparison of indeterminate classifiers. For instance, this can concern problems where the space Ω possess a structure and where costs are derived from this structure (e.g., multi-label, label ranking, ordinal regression).

To make our proposal, we start from the simple observation that discounted costs of the form

$$\bar{c}_{\hat{Y}}(y) = \frac{\sum_{\hat{y} \in \hat{Y}} c_{\hat{y}}(y)}{|\hat{Y}|}$$

provide a kind of "baseline" according to Property 1. As they correspond to a simple arithmetic averaging, a natural way to extend them is to use generalized mean, that is

$$\bar{c}^p_{\hat{Y}}(y) = \left(\frac{1}{|\hat{Y}|} \sum_{\hat{y} \in \hat{Y}} c_{\hat{y}}(y)^p \right)^{\frac{1}{p}}, \tag{12}$$

TABLE VIII
SUMMARY OF PROPERTIES DESIRABILITY, ACCORDING TO THE SETTING. N=NECESSARY, D=DESIRABLE, A=ANALYST CHOICE/APPLICATION DEPENDENT, U=UNDESIRABLE.

	properties									
	1	2	3	4	5	6	7	8	9	10
Filtering	N	D	N	N	D	N	U	D	U	A
Choice	N	D	N	N	D	A	A	U	D	N

TABLE IX
COST MATRIX WHERE COSTS ARE BUILT ACCORDING TO HA'S DEFINITION

	truth		
$c_{\hat{Y}}(y)$	$y=a$	$y=b$	$y=c$
$\hat{Y}=a$	0	$\eta(b)$	$\eta(c)$
$\hat{Y}=b$	$\eta(a)$	0	$\eta(c)$
$\hat{Y}=c$	$\eta(a)$	$\eta(b)$	0
$\hat{Y}=\{a,b\}$	δ	δ	$\eta(c)+\delta$
$\hat{Y}=\{a,c\}$	δ	$\eta(b)+\delta$	δ
$\hat{Y}=\{b,c\}$	$\eta(a)+\delta$	δ	δ
$\hat{Y}=\{a,b,c\}$	2δ	2δ	2δ

with $p \in]-\infty,\infty[$. We retrieve $\bar{c}_{\hat{Y}}(y)$ when $p=1$. In our case, an interesting feature of generalized means is that if $q<p$, then $\bar{c}^q_{\hat{Y}}(y) \leq \bar{c}^p_{\hat{Y}}(y)$, with the two being equal if and only if $c_{\hat{y}}(y)=c_{\hat{y}'}(y)$ for any $\hat{y},\hat{y}' \in \hat{Y}$. Hence, if we want something lower than $\bar{c}_{\hat{Y}}$, we just have to choose $p<1$, and $p>1$ for something higher.

It should however be noted that we can not define $\bar{c}^p_{\hat{Y}}(y)$ for $p<0$ if we have $c_{\hat{y}}(y)=0$ for some values of $\hat{y}$ and y, because of division by zero. In practice, there are ways to solve this issue. As the formula is used mainly for comparing classifiers, we can just add a translation ϵ to the cost matrix, so that there is no more 0 in the matrix. As the translation affects uniformly all classifiers, it will not impact on their comparison. Moreover, it is not uncommon to have a cost matrix that has no null element: in a problem involving money, choosing the right class also has a cost (the basic cost of fabrication, the minimum buying price,...).

For $p=0$, we adopt the convention that $\bar{c}^0_{\hat{Y}}(y)$ is the geometric mean

$$\bar{c}^0_{\hat{Y}}(y) = \big(\prod_{\hat{y}\in\hat{Y}} c_{\hat{y}}(y)\big)^{\frac{1}{|\hat{Y}|}}. \tag{13}$$

Therefore, choosing a positive real value $r \in [0;1]$, we then propose the two following instantiations to introduce the $p-$discounted costs

- Cautiousness seeking, where for all y, we define
$$c_{\hat{Y}}(y) = \bar{c}^{1-r}_{\hat{Y}}(y); \tag{14}$$
- Mistake averse, where for all $y \in \hat{Y}$, we define
$$c_{\hat{Y}}(y) = \bar{c}^{1-r}_{\hat{Y}}(y), \tag{15}$$
and for all $y \notin \hat{Y}$, we define
$$c_{\hat{Y}}(y) = \bar{c}^{1+r}_{\hat{Y}}(y). \tag{16}$$

This naturally satisfies basic Properties 2, 3 and 5. Since we have[1] that $\bar{c}^{-\infty}_{\hat{Y}} = \min$ and $\bar{c}^{+\infty}_{\hat{Y}} = \max$, Properties 10 and 4 are also naturally satisfied for any choice of r. With respect to context-dependent properties, we satisfy Property 9 (in contrast with other solutions), and each instantiation either satisfy Property 7 or 6. Given this, it seems that this proposal is more adapted to a choice setting. Table X provides the costs obtained for Table II when $r=0.5$ and for the cautiousness seeking version.

[1]At least when there is no $c_{\hat{y}}(y)=0$ for some values of $\hat{y}$ and y.

TABLE X
COST MATRIX WITH $r=0.5$ AND CAUTIOUSNESS SEEKING APPROACH

	truth		
$c_{\hat{Y}}(y)$	$y=h$	$y=b$	$y=n$
$\hat{Y}=\{h\}$	0	1	2
$\hat{Y}=\{b\}$	1	0	2
$\hat{Y}=\{n\}$	4	4	0
$\hat{Y}=\{h,b\}$	0.25	0.25	2
$\hat{Y}=\{b,n\}$	2.25	1	0.5
$\hat{Y}=\{h,n\}$	1	2.25	0.5
$\hat{Y}=\{h,b,n\}$	1	1	0.89

The above formulation also has the advantage that the value of r (a single parameter) should be able to calibrate how "cautiousness-friendly" we are. Indeed, the higher r, the more cautiousness is rewarded (and mistakes penalized, if we pick the mistake-averse version), with no reward at all for $r=0$.

VII. ILLUSTRATIVE APPLICATION: CALIBRATION AND SELECTION OF AN IMPRECISE PROBABILISTIC CLASSIFIER

To visualize the behaviour of the formula proposed in Section VI, we show in this section two experiments, respectively dealing with the calibration and selection of an imprecise probabilistic classifier. The first experiment has two main goals: to experimentally confirm that the calibration of the r parameter of $p-$discounted costs can truly reflect our aversion for indeterminate predictions, and to show how $p-$discounted costs can be used to adjust an imprecision parameter involved when using an imprecise probabilistic classifier. The second experiment illustrates a potential use of our framework.

As it is difficult to find data sets which comes naturally with predetermined error costs, we use ordinal data sets for our experiments. For these data, the finite set of possible labels is naturally ordered. For instance, the rating of movies can be done using the following labels: *Very-Bad*, *Bad*, *Average*, *Good*, *Very-Good* that are ordered from the worst situation to the best. This will give us an easy way to establish an ordering, and therefore a metric, over the classes.

Experiments will be conducted on three ordinal data sets of the UCI Machine Learning Repository, whose details are given in Table XI. As our aim is not to validate numerically any model, but to illustrate the behaviours and use of our proposed formula, using three data sets seems sufficient.

TABLE XI
DATA SETS DETAILS

Name	#instances	#features	#classes
ERA	1000	5	9
ESL	488	5	9
LEV	1000	5	5

A. ℓ_1 norm as cost measurement and specificities of ordinal data

As the classes in the data sets are ordered, we can use the ℓ_1 distance between the rank of classes as an initial cost function. Let y_i and y_j be two classes in the space of possible ordinal classes $(y_1, \ldots, y_K)$ (indexed according to their order). The cost $c_{y_i}(y_j)$ of predicting y_i when y_j is the true class and the cost $c_{y_j}(y_i)$ of predicting y_j when y_i is true are both defined as :

$$c_{y_i}(y_j) = c_{y_j}(y_i) = |i - j|.$$

Of course, any ℓ_k norm ($k \neq 1$) can also be used if there is a specific reason to do so. In our experiments, we will use ℓ_1.

Due to this specific ordering of classes, we may also propose a way to restrict the space of possible predictions to simplify the decision-making process specified in Section II-B. For instance, given the labels {*Bad*, *Average*, *Good*}, grouping *Bad* and *Good* together and leaving *Average* aside is contradictory to the given order, and it may be argued that only contiguous classes should be predicted as indeterminate decisions. This is for instance what is done by Alonso *at al.* [19]. While this hypothesis seems intuitive and sensible, it may not always be true. Indeed, for a controversial film, it is very likely that there will be two major tendencies that contradict each other (for example *Very Bad* and *Very Good*). Therefore, restricting to contiguous classes may cause some loss of information, which may be valuable especially when we try to build a reliable classifier.

B. Naive Credal Classifier

We have chosen for our experiments to use the Naive Credal Classifier (NCC) [5], which is an extension of the Naive Bayesian Classifier (NBC) to the imprecise probability framework. We summarize in this section the main features of this credal classifier, and refer to Zaffalon [5] for further details. The NCC preserves the main properties of the NBC, i.e. the assumption of attribute independence conditionally on the class, which can be written as:

$$p(x_1, \ldots, x_m/y) = \prod_{i=1}^{m} p(x_i/y),$$

where $(x_1, \ldots, x_m) \in (X_1, \ldots, X_m)$ are the input features and $y \in \Omega$. Using Bayes law, we can easily compute the posterior probabilities:

$$p(y/x_1, \ldots, x_m) = \frac{p(y, x_1, \ldots, x_m)}{p(x_1, \ldots, x_m)} \tag{17}$$

$$= \frac{p(y)\prod_{i=1}^{m} p(x_i/y)}{\sum_{y' \in \Omega} p(y') \prod_{i=1}^{m} p(x_i/y')}. \tag{18}$$

In the NCC, each $p(y)$ is supposed to belong to a set $\mathcal{P}_y$, and each conditional probability $p(x_i/y)$ to a set $\mathcal{P}^y_{X_i}$ (these sets are referred to as local credal sets) so that the model is characterized by the set $\mathcal{P}$ of joint distributions obtained from all possible combinations of the local credal sets.

Using the Imprecise Dirichlet Model (IDM) [23], one can then define the local credal sets by probability intervals [24]. These intervals can be computed using the training data by simply counting occurrences:

$$\underline{p}(x_i/y) = \frac{occ_{i,y}}{occ_y + s}, \qquad \overline{p}(x_i/y) = \frac{occ_{i,y} + s}{occ_y + s}, \tag{19}$$

$$\underline{p}(y) = \frac{occ_y}{occ_\Omega + s}, \qquad \overline{p}(y) = \frac{occ_y + s}{occ_\Omega + s}, \tag{20}$$

where $occ_{i,y}$ is the number of instances in the training set where the attribute X_i is equal to x_i and the class value is y. occ_y is the number of instances in the training set where the class value is in y. occ_Ω refers to the size of the training data set. The hyper-parameter s sets the imprecision level of the IDM, which means that the greater is s, the more indeterminate the predictions of NCC will be, and *vice versa*. Note that to use the IDM, a discretization of the attribute values is needed. The data set LEV is natively discrete (each attribute is an integer between 1 and 5). For the two other data sets ERA and ESL, a discretization into five levels of equal frequencies has been performed.

Rather than computing Equations (5) by spanning all possible marginals and conditional within $\mathcal{P}_y$ and $\mathcal{P}^\Omega_{X_i}$, we will simplify the problem[2] by first computing bounds over each singletons of the posterior probability $p(y|x_1, \ldots, x_m)$. For this, we need to solve the following minimization/maximization problem over the local credal sets:

$$\underline{p}(y/x_1, \ldots, x_m) = \min_{p(y)\in\mathcal{P}_y} \min_{\substack{p(x_i/\omega)\in\mathcal{P}^\Omega_{X_i},\\ i=1,\ldots,m}} \frac{p(y)\prod_{i=1}^{m} p(x_i/y)}{\sum_{\omega\in\Omega}\prod_{i=1}^{m} p(x_i/\omega)p(\omega)}, \tag{21}$$

$$\overline{p}(y/x_1, \ldots, x_m) = \max_{p(y)\in\mathcal{P}_y} \max_{\substack{p(x_i/\omega)\in\mathcal{P}^\Omega_{X_i},\\ i=1,\ldots,m}} \frac{p(y)\prod_{i=1}^{m} p(x_i/y)}{\sum_{\omega\in\Omega}\prod_{i=1}^{m} p(x_i/\omega)p(\omega)}. \tag{22}$$

It can be shown that the former problems are equivalent to:

$$\underline{p}(y/x_1, \ldots, x_m) = \min_{p(y)\in\mathcal{P}_y} \left(1 + \frac{\sum_{y'\neq y} p(y') \prod_{i=1}^{m} \overline{p}(x_i/y')}{p(y)\prod_{i=1}^{m} \underline{p}(x_i/y)} \right)^{-1}, \tag{23}$$

[2]As our purpose is illustrative, using such an approximation is reasonable, even if it may give more conservative predictions than solving the exact problem.

$$\overline{p}(y/x_1,\ldots,x_m) = \max_{p(y)\in\mathcal{P}_y} \left(1 + \frac{\sum\limits_{y'\neq y} p(y') \prod\limits_{i=1}^{m} \underline{p}(x_i/y'}{p(y) \prod\limits_{i=1}^{m} \overline{p}(x_i/y)} \right)^{-1}, \quad (24)$$

which in turn can be solved by enumerating the extreme points of $\mathcal{P}_y$. Computational details can be found in [5]. Finally, as shown in [24], since the probability intervals are Choquet capacities of order 2, the computation of the lower and upper expected costs can be simply obtained using Choquet integrals.

C. Classifier calibration

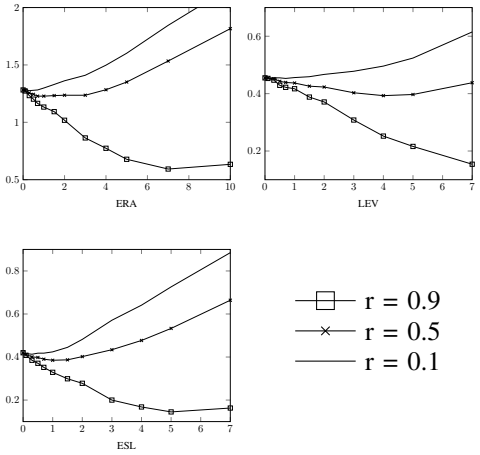


Fig. 3. $p-$discounted costs (cautiousness-seeking approach, y axis) of ERA, LEV and ESL with different combinations of r and s (x axis)

In Figure 3, we show, for $r \in \{0.1, 0.5, 0.9\}$, the $p-$discounted costs with s ranging from 0.001 to 10. A first trivial remark is the respective position of the curves which confirms that r can truly calibrate our aversion for indeterminate predictions: the higher r, the lower is the cost for indeterminate predictions, for a given level of imprecision.

More interestingly, we can notice that the evolution of the $p-$discounted costs follows the same general behaviour. That is, the curves are approximately concave: firstly the cost decreases when we increase s, because the added imprecision allows to obtain more correct predictions for "hard to classify" instances, then, after a given value of s, the cost starts to increase, as too many (unnecessary) imprecision is added.

This is particularly visible for $r = 0.5$ and $r = 0.9$, where we can see clearly the inflection point (except for LEV where the inflection point for $r = 0.9$ is beyond $s = 10$). For other cases, the inflection point is still present but less remarkable in the graphics. The same behaviour can be observed when we use the *mistake-aversed* approach, the only difference is that the value of s which minimize the score is smaller than with the *cautiousness-seeking* approach.

We can also note that the range which makes indeterminate predictions interesting is when $r \in]0; 1[$. When $r = 0$, which means that we give no reward for cautiousness, then the curves of $p-$discounted costs are strictly increasing. Similarly for $r = 1$, which means that the cost is null each time the truth is included in the prediction, the curve is strictly decreasing.

However, for a same level of r, the optimal s is not always the same from one data set to another. So it is not possible to state a direct relationship between the two. Once r is set, we suggest techniques such as cross-validation to automatically determine the optimal value of s from the data set.

TABLE XII
CLASSIFIER AVERAGE INDETERMINATE COSTS. RANKS ARE BETWEEN PARENTHESIS.

	NCC	2Tr	Fo
ERA	3.79 (3)	2 (2)	1.93 (1)
ESL	2.2 (3)	0.87 (2)	0.81 (1)
LEV	3.43 (3)	1.18 (2)	0.88 (1)

D. Classifier comparison

Our next illustrative experiment concern the choice of a particular classifier or method producing indeterminate predictions. Still using the ordinal data of Table XI, we fix the value $r = 0.5$, and consider Equation (14) to complete the cost matrix. However, we now consider the NCC of Section VII-B with a fixed parameter $s = 2$, but in different settings:

- used as in the previous experiments (NCC)
- used as a base classifier in a binary-tree decomposition approach (2Tr), detailed in [13]
- used as a base classifier in a forest of such binary trees (Fo), also detailed in [13]

Results in terms of average indeterminacy costs are reported in Table XII. They are obtained through a ten-fold cross validation procedure. The forest of binary tree (Fo) is obtained through a uniform sampling of 50 trees among all possible decompositions, while the single binary tree (2Tr) is the one obtaining the best result on the learning set among the sampled ones. At least on the three considered data sets, it seems that the binary tree methods outperform the NCC, suggesting that using such trees is quite interesting for ordinal regression problems. This would have to be confirmed by complementary experiments, yet we remind that comparing methods for imprecise ordinal regression is not our purpose here, but an illustration of a possible use of Equation (12).

VIII. CONCLUSION

In this paper, we have addressed an important issue when using indeterminate classifiers for deriving reliable and cautious predictive models, which is how to deal with cost-sensitive problems. We have proposed some generic properties and guidelines that we thought are essential and/or relevant for either producing or evaluating indeterminate predictions. We have elaborated a sensible formula ($p-$discounted costs) to compare determinate and/or indeterminate classifiers when generic cost functions are involved. Experiments have shown that this formula can be calibrated according to our risk aversion to allow fair comparisons.

In practice, both our formula as well as a part of our properties (e.g., Property 2) are inspired from the initial work

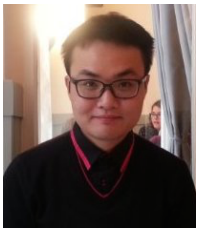
of Zaffalon *et al.* [18] done in the simpler setting of 0/1 costs or accuracy. While our work is consistent with theirs, it took a different way to discuss the relevance of defining costs of indeterminacy, as we rely on properties to justify our choices, not on a betting framework. Extending the results of Zaffalon *et al.* [18] to cost-sensitive settings is clearly a desirable complement to our and their work.

ACKNOWLEDGEMENTS

This work was carried out in the framework of the Labex MS2T, which was funded by the French Government, through the program "Investments for the future" managed by the National Agency for Research (Reference ANR-11-IDEX-0004-02).

REFERENCES

[1] C. Chow, "On optimum recognition error and reject tradeoff," *Information Theory, IEEE Transactions on*, vol. 16, no. 1, pp. 41–46, Jan 1970.
[2] B. Dubuisson and M. Masson, "A statistical decision rule with incomplete knowledge about classes," *Pattern Recognition*, vol. 26, no. 1, pp. 155 – 165, 1993.
[3] R. Herbei and M. Wegkamp, "Classification with reject option," *Canadian Journal of Statistics*, vol. 34, no. 4, pp. 709–721, 2006.
[4] T. Ha, "The optimum class-selective rejection rule," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, no. 6, pp. 608–615, Jun. 1997.
[5] M. Zaffalon, "The naive credal classifier," *Journal of Statistical Planning and Inference*, vol. 105, no. 1, pp. 5–21, 2002.
[6] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *The Journal of Machine Learning Research*, vol. 9, pp. 371–421, 2008.
[7] Z.-G. Liu, Q. Pan, G. Mercier, and J. Dezert, "A new incomplete pattern classification method based on evidential reasoning," *IEEE transactions on cybernetics*, vol. 45, no. 4, pp. 635–646, 2015.
[8] T. Denœux and M.-H. Masson, "Evclus: evidential clustering of proximity data," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 34, no. 1, pp. 95–109, 2004.
[9] J. del Coz and A. Bahamonde, "Learning nondeterministic classifiers," *Journal of Machine Learning Researchs*, vol. 10, pp. 2273–2293, 2009.
[10] P. L. Bartlett and M. H. Wegkamp, "Classification with a reject option using a hinge loss," *The Journal of Machine Learning Research*, vol. 9, pp. 1823–1840, 2008.
[11] C. Mantas and J. Abellán, "Credal-c4.5: Decision tree based on imprecise probabilities to classify noisy data," *Expert Syst. Appl.*, vol. 41, no. 10, pp. 4625–4637, 2014.
[12] S. Destercke and G. Yang, "Cautious ordinal classification by binary decomposition," in *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part I*, 2014, pp. 323–337.
[13] G. Yang, S. Destercke, and M. Masson, "Nested dichotomies with probability sets for multi-class classification," in *ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014)*, 2014, pp. 363–368.
[14] G. Corani and M. Zaffalon, "Credal model averaging: An extension of bayesian model averaging to imprecise probabilities." in *ECML/PKDD (1)*, ser. Lecture Notes in Computer Science, W. Daelemans, B. Goethals, and K. Morik, Eds., vol. 5211. Springer, 2008, pp. 257–271.
[15] S. Wang and X. Yao, "Multiclass imbalance problems: Analysis and potential solutions," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 1119–1130, Aug 2012.
[16] A. Riccardi, F. Fernández-Navarro, and S. Carloni, "Cost-sensitive adaboost algorithm for ordinal regression based on extreme learning machine," *IEEE Transactions on Cybernetics*, vol. 44, no. 10, pp. 1898–1909, Oct 2014.
[17] J. Abellán and A. R. Masegosa, "Imprecise Classification With Credal Decision Trees," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 20, no. 05, pp. 763–787, Oct. 2012.
[18] M. Zaffalon, G. Corani, and D. Mauá, "Evaluating credal classifiers by utility-discounted predictive accuracy," *International Journal of Approximate Reasoning*, vol. 53, no. 8, pp. 1282 – 1301, 2012.
[19] J. Alonso, J. J. Del Coz, J. Díez, O. Luaces, and A. Bahamonde, "Learning to predict one or more ranks in ordinal regression tasks," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2008, pp. 39–54.
[20] P. Walley, *Statistical reasoning with imprecise probabilities*. Chapman and Hall, 1991.
[21] I. Levi, *The enterprise of knowledge: An essay on knowledge, credal probability, and chance*. MIT press, 1983.
[22] M. Troffaes, "Decision making under uncertainty using imprecise probabilities," *International Journal of Approximate Reasoning*, vol. 45, no. 1, pp. 17–29, May 2007.
[23] J.-M. Bernard, "An introduction to the imprecise Dirichlet model for multinomial data," *International Journal of Approximate Reasoning*, vol. 39, no. 2-3, pp. 123–150, 2005.
[24] L. de Campos, J. Huete, and S. Moral, "Probability intervals: a tool for uncertain reasoning," *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 2, pp. 167–196, 1994.



Gen Yang graduated as an engineer in computer science from Ecole Nationale Supérieure d'Informatique pour l'Industrie et l'Entreprise in Evry (France). He received his P.h.D. degree in computer science in the Heuristic and Diagnostic of Complex Systems (Heudiasyc) Laboratory in 2016, on the topic of efficiently learning imprecise models. He is now working as a data scientist in the Kpler company.

Sebastien Destercke received his Ph.D. degree in computer science from *Université Paul Sabatier*, in Toulouse (France) in 2008. From 2008 to 2011, he was a research engineer at *Centre de coopération internationale en recherche agronomique pour le développement*. He is currently a researcher with the *French National Centre for Scientific Research*, in the joint unit *Heuristic and Diagnosis for Complex System*. His research focuses on uncertainty reasoning in presence of imprecision, including as particular interests information fusion, machine learning, uncertainty representation.

Marie-Hélène Masson graduated as an engineer from the University of Technology of Compiègne and earned a Ph.D in computer science in 1992 and a "Habilitation à diriger des Recherches" in 2005 from the same institution. She is now assistant Professor at the Université de Picardie Jules Verne, France and is a member of Heudiasyc Laboratory of the Université de Technologie de Compiègne, France since 1993. Her research interests include statistical pattern recognition, data analysis, uncertainty modelling and information fusion.