Multilabel predictions with sets of probabilities: the Hamming and ranking loss cases

Sébastien Destercke

UMR CNRS 7253 Heudyasic UTC, Compiègne, France.

Abstract

In this paper, we study how multilabel predictions can be obtained when our uncertainty is described by a convex set of probabilities. Such predictions, typically consisting of a set of potentially optimal decisions, are hard to make in large decision spaces such as the one considered in multilabel problems. However, we show that when considering the Hamming or the ranking loss, outer-approximating predictions can be efficiently computed from label-wise information, as in the precise case. We also perform some first experiments showing the behaviour of the partial predictions obtained through these approximations. Such experiments also confirm that predictions become partial on those labels where the precise prediction is likely to make an error.

Keywords: multilabel, cautious predictions, weak labels, probability sets

1. Introduction

The problem of multi-label classification, which generalizes the traditional (single label) classification setting by allowing multiple labels to simultaneously belong to an instance, has recently attracted a lot of attention. Such problems indeed appear in a lot of situations: a movie can belong to multiple categories, a music can stir multiple emotions [30], proteins can possess multiple functions [33], images can have multiple elements displayed in them, [5] etc. In such problems, obtaining a complete ground truth (sets of relevant labels) for the training data and making accurate predictions is more complex than in traditional (single label) classification, in which the aim is to predict a unique label.

In such a setting the appearance of incomplete observations, i.e., instances for which we do not know whether some labels are relevant or not, is much more likely. For example, a user may be able to tag a movie as a comedy and not as a science-fiction movie, but may hesitate whether or not it should be tagged as a drama. Other examples include cases where a high number of labels are possible and where an expert cannot be expected to provide all relevant ones due to time or cost constraints. Such partial labels are commonly called weak labels [27] and are common situations in problems such as image annotation [28] or protein function prediction [33].

Email address: sebastien.destercke@hds.utc.fr (Sébastien Destercke)

Even when considering weak labels, all multilabel methods we are aware of still produce complete predictions as outputs. However, given the complexity of the prediction to make and the likely presence of missing data, it may be sensible to look for cautious yet more trustful predictions. That is it may be interesting for the learner to abstain to make a prediction about a label whose relevance is too uncertain, so that the final prediction is partial but more robust, in the sense that a prediction is made only for those labels about which we have sufficient information. Such partial predictions can help to identify which instances are hard to predict (i.e., as is the case in multi-class settings [9]), therefore pointing out where we would need to collect more information. They could therefore be used in active learning settings, or more simply to warn the user or analyst that our current information does not allow us to make a prediction for a particular instance. For example, when trying to detect protein or gene functions, it could be quite useful for the domain expert to know about which function we should collect more data.

Various approaches have been proposed in the literature to obtain such partial predictions: a classical way is to implement a reject option [3, 23], or a partial version of it [18]. More recent methods includes the use of probability sets [8] (the size of the set then reflecting our lack of knowledge) or the use of conformal prediction [26]. Finally, more recent proposals have looked at the problem of making partial predictions for the problem of label ranking [7, 6], of which multilabel problems can be seen as a special case.

In this paper, we consider the practical problem of making partial predictions in the multilabel setting when using convex sets of probabilities, or *credal sets* [20], as our model of uncertainty. Making partial predictions is one central feature of approaches using credal sets [8], and these approaches are also well-designed to cope with the problem of missing or incomplete data [34]. However, applying classical decision rules to make partial predictions with credal sets in mutlilabel setting are likely to be computationally intractable, as those decision rules involve making n^2 comparisons where n is the number of alternatives (a number that increases exponentially with the number of labels).

Our main result is to show that this tractability issue can be avoided by considering specific losses, namely the Hamming and ranking losses, and by considering approximate inferences. This is done in Section 3, first for the Hamming loss (Section 3.1), then for the ranking loss (Section 3.2). Other losses will only be discussed shortly, as nothing in their structure suggests that approximate inferences are easy to do when considering them in conjunction with credal sets. Section 4 then provides some experiments to show the behaviours of the proposed inferences and predictions. Necessary background material is given in Section 2. This paper extends a conference paper [15] that only dealt with the Hamming loss case. In addition to dealing with the ranking loss and to a discussion about the optimization of other losses, this paper also provides full details, examples and experiments.

2. Preliminaries

In this section, we introduce the multilabel setting as well as basic notions needed to deal with sets of probabilities.

X_1	X_2	X_3	X_4	y_1	y_2	y_3
107.1	25	*Blue*	60	1	0	0
-50	10	*Red*	40	1	0	1
200.6	30	*Blue*	58	$*$	1	0
107.1	5	*Green*	33	0	1	$*$
...	...	...	...	...	...	...

Table 1: Multilabel data set example

2.1. Multilabel problem setting

The usual goal of classification problems is to associate an instance $\mathbf{x}$ coming from an instance space $\mathscr{X}$ to a single (preferred) label of the space $\Lambda = \{\lambda_1, \dots, \lambda_m\}$ of possible classes. In a multilabel setting, an observation $\mathbf{x}$ is associated to an observed subset $L_{\mathbf{x}} \subset \Lambda$ of labels, often called the subset of relevant labels while its complement $\Lambda \setminus L_{\mathbf{x}}$ is considered as irrelevant. We denote by $\mathscr{Y} = \{0,1\}^m$ the set of m-dimensional binary vector, and identify a set L of relevant labels with a binary vector $y = (y_1, \dots, y_m)$ such that $y_i = 1$ if and only if $\lambda_i \in L$.

The task in a multilabel problem is the same as in usual classification: to use the training instances $(\mathbf{x}^j, y^j)$, $j = 1, \dots, n$ to estimate the theoretical conditional probability measure $P_{\mathbf{x}} : 2^{\mathscr{Y}} \to [0,1]$ associated to an instance $\mathbf{x} \in \mathscr{X}$. Ideally, observed outputs y^j should be completely specified vectors, however it may be the case that the value for some component y_i^j is unknown, which will be denoted by $y_i^j = *$. We will denote incomplete vectors by capital Y. Alternatively, an incomplete vector Y can be characterized by two sets $\underline{L} \subseteq \overline{L} \subseteq \Lambda$ of necessarily and possible relevant labels, defined as $\underline{L} := \{\lambda_i | y_i = 1\}$ and $\overline{L} := \{\lambda_i | y_i = 1 \vee y_i = *\}$ respectively. An incomplete vector Y describes a corresponding set of complete vectors, obtained by replacing each $y_i = *$ either by 1 or 0, or equivalently by considering all subsets L such that $\underline{L} \subseteq L \subseteq \overline{L}$. To simplify notations, in the sequel we will use the same notation for an incomplete vector and its associated set of complete vectors.

Example 1. Table 1 provides an example of a multilabel data set with $\Lambda = \{\lambda_1, \lambda_2, \lambda_3\}$. $Y^3 = [*\ 1\ 0]$ is an incomplete observed instance with $\underline{L}^3 = \{\lambda_2\}$ and $\overline{L}^3 = \{\lambda_1, \lambda_2\}$. Its corresponding set of complete vectors is $\{[0\ 1\ 0], [1\ 1\ 0]\}$

In multilabel problems the size of the prediction space increases exponentially with m (e.g., $|\mathscr{Y}| = 32768$ for $m = 15$), meaning that estimating directly $P_{\mathbf{x}}$ will be intractable even for limited sizes of Λ. As a means to solve this issue, different authors have proposed so-called transformation techniques [31] that reduce the initial problem into a set of simpler problems. For example

- Binary relevance (BR) consists in predicting label-wise relevance, solving independent binary problem for each label. It therefore comes down to estimate $P_{\mathbf{x}}(y_i)$ and to predicts $\hat{y}_i = 1$ if $P_{\mathbf{x}}(y_i = 1) \geq 1/2$;
- Ranking approaches such as Calibrated Ranking (CR) [17] intend to build an ordering between labels by focusing on pairwise comparisons between labels.

While such approaches reduce the inference complexity, a common critic is that they only consider partial information about $P_{\mathbf{x}}$, and are therefore likely to be sub-optimal. For instance, BR does not integrate any information about label dependencies, and many techniques have been proposed to integrate such dependencies in the predictive model [24, 25, 22].

However, it has been proven that these methods are actually theoretically optimal under some specific loss functions. The main goal of this paper is to show that similar results hold when the model $P_{\mathbf{x}}$ becomes a set of probabilities.

Indeed, making a precise and accurate estimation of $P_{\mathbf{x}}$ is an extremely difficult problem given the number 2^m of alternatives and the possible presence of missing data. This problem is even more severe if little data are available, and this is why making cautious inferences (i.e., partial predictions) using as model a (convex) set $\mathscr{P}_{\mathbf{x}}$ of probability distributions may be interesting in the multilabel setting.

2.2. *Notions about probability sets*

We assume that our uncertainty is described by a convex set of probabilities $\mathscr{P}_{\mathbf{x}}$, a *credal set* [20], defined over $\mathscr{Y}$ rather than by a precise probability measure $P_{\mathbf{x}}$. Such a set is usually defined either by a collection of linear constraints on the probability masses or by a set of extreme probabilities. Many authors [32, 4, 13] have argued that when information is lacking or imprecise, considering credal sets as our model of information better describes our actual uncertainty. Credal sets are also convenient models in frequentist settings, when one do not want to stick to point-valued parameters [2, 21].

Given such a set, we can define for any event $A \subseteq \mathscr{Y}$ the notions of lower and upper probabilities $\underline{P}_{\mathbf{x}}(A)$ and $\overline{P}_{\mathbf{x}}(A)$, respectively as

$$\underline{P}_{\mathbf{x}}(A) = \inf_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}} P_{\mathbf{x}}(A) \text{ and } \overline{P}_{\mathbf{x}}(A) = \sup_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}} P_{\mathbf{x}}(A).$$

Lower and upper probabilities are dual, in the sense that $\underline{P}(A) = 1 - \overline{P}(A^c)$. Similarly, if we consider a real-valued bounded function $f : \mathscr{Y} \to \mathbb{R}$, the lower and upper expectations $\underline{\mathbb{E}}_{\mathbf{x}}(f)$ and $\overline{\mathbb{E}}_{\mathbf{x}}(f)$ are defined as

$$\underline{\mathbb{E}}_{\mathbf{x}}(f) = \inf_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}} \mathbb{E}_{\mathbf{x}}(f) \text{ and } \overline{\mathbb{E}}_{\mathbf{x}}(f) = \sup_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}} \mathbb{E}_{\mathbf{x}}(f),$$

where $\mathbb{E}_{\mathbf{x}}(f)$ is the expectation of f w.r.t. $P_{\mathbf{x}}$. Lower and upper expectations are also dual, in the sense that $\underline{\mathbb{E}}(f) = -\overline{\mathbb{E}}(-f)$. They are also scale and translation invariant in the sense that given two numbers $\alpha \in \mathbb{R}^+, \beta \in \mathbb{R}$, we have $\underline{\mathbb{E}}(\alpha f + \beta) = \alpha \underline{\mathbb{E}}(f) + \beta$.

2.3. *Decision and predictions with probability sets*

Once a space $\mathscr{Y}$ of possible observations is defined, selecting a prediction, or equivalently making a decision, requires to define:

- a space $\mathscr{A} = \{a_1, \ldots, a_d\}$ of possible alternatives;
- a loss function $\ell : \mathscr{A} \times \mathscr{Y} \to \mathbb{R}$ such that $\ell(a, y)$ defines the loss incurred by predicting a when y is the ground-truth.

In this paper, we differentiate the space of predictions or alternatives $\mathscr{A}$ from the space of output observations $\mathscr{Y}$, since they may not always coincide. Actually, we will have $\mathscr{A} = \mathscr{Y}$ in Section 3.1, but not in Section 3.2, since the ranking loss compares a vector $y \in \mathscr{Y}$ to a ranking over the labels Λ (in which case $\mathscr{A}$ is the set of all permutations over Λ).

In a precise probability setting, given an instance $\mathbf{x}$ and a probability $\hat{P}_{\mathbf{x}}$, a decision a will be preferred to a decision a' under loss function ℓ, denote $a \succ_\ell a'$, if

$$\begin{aligned}\mathbb{E}_{\hat{P}_{\mathbf{x}}}\left(\ell(a',\cdot)-\ell(a,\cdot)\right) &= \sum_{y\in\mathscr{Y}} \hat{P}_{\mathbf{x}}(y)\left(\ell(a',y)-\ell(a,y)\right) \\ &= \sum_{y\in\mathscr{Y}} \hat{P}_{\mathbf{x}}(y)\ell(a',y) - \sum_{y\in\mathscr{Y}} \hat{P}_{\mathbf{x}}(y)\ell(a,y) \\ &= \mathbb{E}_{\hat{P}_{\mathbf{x}}}\left(\ell(a',\cdot)\right) - \mathbb{E}_{\hat{P}_{\mathbf{x}}}\left(\ell(a,\cdot)\right) > 0,\end{aligned} \tag{1}$$

where $\mathbb{E}_{\hat{P}_{\mathbf{x}}}$ is the expectation w.r.t. $\hat{P}_{\mathbf{x}}$, and $\ell(a,\cdot) : \mathscr{Y} \to \mathbb{R}$ the cost function of choosing a' as our prediction. This equation means that exchanging a' for a would incur a positive expected loss, or equivalently that the expected loss of choosing a' is higher than the one of choosing a, therefore a should be preferred to a'. In the case of a precise estimate $\hat{P}_{\mathbf{x}}$, $\succ_\ell$ is a complete pre-order and the optimal prediction comes down to take the maximal element of this pre-order, i.e.,

$$\hat{a}_\ell = \arg\min_{a\in\mathscr{A}} \mathbb{E}_{\hat{P}_{\mathbf{x}}}\left(\ell(a,\cdot)\right) = \arg\min_{a\in\mathscr{A}} \sum_{y\in\mathscr{Y}} \hat{P}_{\mathbf{x}}(y)\ell(a,y) \tag{2}$$

that is to minimize the expected loss (ties can be broken arbitrarily, as they will lead to the same expected loss). This means that finding the best alternative (or prediction) will require d computations of expectations.

When considering a set $\mathscr{P}_{\mathbf{x}}$ as cautious estimate, there are many ways [29] to extend Equation (1). The concept of maximality [32, Sec. 3.9.] is the one we will consider here. Under this criterion, we have that $a \succ_\ell a'$ if

$$\underline{\mathbb{E}}_{\mathscr{P}_{\mathbf{x}}}\left(\ell(a',\cdot)-\ell(a,\cdot)\right) = \inf_{P_{\mathbf{x}}\in\mathscr{P}_{\mathbf{x}}} \mathbb{E}_{P_{\mathbf{x}}}\left(\ell(a',\cdot)-\ell(a,\cdot)\right) > 0, \tag{3}$$

that is if exchanging a' for a is guaranteed to give a positive expected loss. Our lack of information is then reflected by the fact that the relation $a \succ_\ell a'$ will be a partial order, hence the maximal set

$$\hat{A}_\ell = \{a \in \mathscr{A} \mid \nexists a' \in \mathscr{A} \text{ s.t. } a' \succ_\ell a\}. \tag{4}$$

of alternatives can be a set of values that will form our prediction. Clearly, the more imprecise is $\mathscr{P}_{\mathbf{x}}$, the larger is the set $\hat{A}_\ell$. It should be noted that having $\underline{\mathbb{E}}_{\mathscr{P}_{\mathbf{x}}}\left(\ell(a',\cdot)-\ell(a,\cdot)\right) > 0$ only implies that $\underline{\mathbb{E}}_{\mathscr{P}_{\mathbf{x}}}(\ell(a',\cdot) > \overline{\mathbb{E}}_{\mathscr{P}_{\mathbf{x}}}\ell(a,\cdot)$, due to the fact that we only have

$$\underline{\mathbb{E}}(f+g) \geq \underline{\mathbb{E}}(f) + \underline{\mathbb{E}}(g), \tag{5}$$

i.e., $\underline{\mathbb{E}}$ is sub-additive. Saying that $a \succ_\ell a'$ if $\underline{\mathbb{E}}_{\mathscr{P}_{\mathbf{x}}}(\ell(a',\cdot) > \overline{\mathbb{E}}_{\mathscr{P}_{\mathbf{x}}}\ell(a,\cdot)$ is then a quite weaker criterion to build the partial order $\succ_\ell$, known under the name of interval dominance. The goal of such partial prediction is not to do "better" then precise predictions, but to point out those cases where information is not sufficient to make a reliable precise prediction. They are therefore useful in applications where it is interesting to know that we do not know. The next example illustrates this

Example 2. Consider a multi-class setting where $\mathscr{Y} = \{y_1, y_2, y_3\}$, and two different input instances x^1, x^2 for which we must make a prediction. Assume that, for x^1, we have observed in the training set $30, 20$ and 10 times y_1, y_2 and y_3, respectively. For x^2, we have observed in the training set $3, 2$ and 1 times y_1, y_2 and y_3, respectively. Assume furthermore that we are considering a classical 0/1 loss, whose associated optimal prediction is the mode of $\hat{P}_{\mathbf{x}}$.

In a precise setting, our prediction for x^1 and x^2 would both be y_1, a natural choice given the observations and the fact that a frequentist evaluation of $\hat{P}_{\mathbf{x}}$ would give $\hat{P}_{\mathbf{x}}(y_1) = 0.5$, $\hat{P}_{\mathbf{x}}(y_1) = 1/3$, $\hat{P}_{\mathbf{x}}(y_1) = 1/6$ for x^1 and x^2, since $\hat{P}_{\mathbf{x}}(y_i) = n_i/n$, where n_i is the number of times y_i was observed, and n the total number of observations.

Instead of a precise evaluation, we could use the Imprecise Dirichlet Model, that provides the following bounds

$$\underline{\hat{P}}_{\mathbf{x}}(y_i) = \frac{n_i}{n+s} \quad \overline{\hat{P}}_{\mathbf{x}}(y_i) = \frac{n_i+s}{n+s}$$

over $\hat{P}_{\mathbf{x}}(y_i)$, with s an hyper-parameter, often set to $s = 2$. This kind of model is used, e.g., in imprecise probabilistic decision trees [1]. With this approach, we get the following bounds for x^1 and x^2

		y_1	y_2	y_3
x^1	$\overline{\hat{P}}$	$32/62$	$22/62$	$12/62$
	$\underline{\hat{P}}$	$30/62$	$20/62$	$10/62$
x^2	$\overline{\hat{P}}$	$5/8$	$4/8$	$3/8$
	$\underline{\hat{P}}$	$3/8$	$2/8$	$1/8$

In this case, we still predict $\hat{A} = \{y_1\}$ for x^1, but $\hat{A} = \{y_1, y_2\}$ for x^2 (as there are distributions within the bounds $\underline{\hat{P}}_{x^2}, \overline{\hat{P}}_{x^2}$ where y_1 or y_2 are modal values), due to the fact that we have little information in the second case.

Making predictions with probability sets is usually harder than with precise ones, as they require solving multiple linear optimization problems. For instance, computing $\hat{A}_\ell$ requires at worst $d(d-1)$ computations, a quadratic number of comparisons with respect to the number of alternatives. This makes the prediction step in a multilabel setting even more difficult. In the next sections, we explore how the multilabel problem can be solved with such credal sets. We discuss the problem, usually computationally intensive, of making partial decision and show that it can be simplified when considering the Hamming loss or the ranking loss as our loss functions. Using these results, we then perform some experiment based on label-wise or pairwise decomposition and k-nn algorithm to assess the interest of making partial predictions based on credal sets.

3. Making multilabel predictions with probability sets

Computing $\hat{A}_\ell$ in a multilabel setting will be intractable in most cases, as the worse number of computations to achieve will then be 2^{2m} if $\mathscr{A} = \mathscr{Y}$ ($m = 15$ labels means at worst $\sim 10^9$ comparisons). It is therefore important to search how this computational burden can be diminished. One way to do so, that we explore here, is to provide efficient inference methods taking advantage of the formulations of particular loss functions.

In the next subsection, we show that for the Hamming loss ℓ_H and the ranking loss ℓ_R, we can get an outer approximation of $\hat{A}_\ell$ at a very affordable computational cost. Offering such efficient way to make cautious predictions based on $\mathscr{P}_\mathbf{x}$ is essential to be able to use such kind of models in complex problems.

3.1. The Hamming loss

Hamming loss ℓ_H considers as set of alternatives $\mathscr{A} = \mathscr{Y}$ the set of possible outputs of the multilabel problem. Given an observation y and a prediction $\hat{y}$, it reads

$$\ell_H(\hat{y}, y) = \frac{1}{k} \sum_{i=1,\ldots,k} \mathbf{1}_{(\hat{y}_i \neq y_i)} \,. \tag{6}$$

It counts the number of labels for which our prediction is wrong, and normalize it. When the estimate $P_\mathbf{x}$ is precise, it is known [12] that the optimal decision is the vector $\hat{y}$ such that $\hat{y}_j = 1$ if $P_\mathbf{x}(y_j = 1) \geq 1/2$ and $\hat{y}_j = 0$ else. In particular, this means that optimal decision can be derived from the sole knowledge of the marginals $P_\mathbf{x}(y_j = 1)$, $j = 1, \ldots, n$. This means that the theoretical number of estimates to obtain to have an optimal prediction is m, which compared to the number 2^m in the general case is a drastic reduction.

The question is then to know if such a reduction is still possible when working with a credal set $\mathscr{P}_\mathbf{x}$. Let us denote $\hat{Y}_{\ell_H}$ the maximal set of vectors that would be obtained using Equation (4). The next proposition shows that in contrast with the precise case, $\hat{Y}_{\ell_H}$ can only be outer-approximated using the marginals of the cautious estimate $\mathscr{P}_\mathbf{x}$.

Proposition 1. *Let $\mathscr{P}_\mathbf{x}$ be our estimate, then the imprecise vector $\hat{Y}^*$ such that*

$$\hat{Y}^*_j = \begin{cases} 1 & \text{if } \underline{P}(y_j = 1) > 1/2 \\ 0 & \text{if } \underline{P}(y_j = 0) > 1/2 \\ * & \text{if } \underline{P}(y_j = 1) \leq 1/2 \leq \overline{P}(y_j = 1) \end{cases} \qquad \text{for } j = 1, \ldots, m$$

is an outer approximation of $\hat{Y}_{\ell_H}$, in the sense that $\hat{Y}_{\ell_H} \subseteq \hat{Y}^*$.

Proof. To prove Proposition 1, we will simply show that if $\underline{P}(y_j = 1) > 1/2$, then any alternative y where $y_j = 1$ dominates (in the sense of Equation 3) the prediction y' where only the jth component is changed (from 1 to 0). Consider a given $j \in \{1, \ldots, m\}$ and two alternatives $\hat{y}$ and $\hat{y}'$ such that $\hat{y}_j = 1 \neq \hat{y}'_j$ and $\hat{y}_i = \hat{y}'_i$ for any $i \neq j$. Let us now look at the value of $\ell_H(\hat{y}', \cdot) - \ell_H(\hat{y}, \cdot)$: for any y such that $y_j = 1$ we have

$$\begin{aligned} \ell_H(\hat{y}', y) - \ell_H(\hat{y}, y) &= \left(\sum_{k \neq j} \mathbf{1}_{(\hat{y}'_k \neq y_k)} + \mathbf{1}_{(\hat{y}'_j \neq y_j)} \right) - \left(\sum_{k \neq j} \mathbf{1}_{(\hat{y}_k \neq y_k)} + \mathbf{1}_{(\hat{y}_j \neq y_j)} \right) \\ &= \mathbf{1}_{(\hat{y}'_j = 0)} - \mathbf{1}_{(\hat{y}_j = 0)} = 1, \end{aligned}$$

and for any y such that $y_j = 0$ we have

$$\ell_H(\hat{y}', y) - \ell_H(\hat{y}, y) = \left(\sum_{k \neq j} \mathbf{1}_{(\hat{y}'_k \neq y_k)} + \mathbf{1}_{(\hat{y}'_j \neq y_j)} \right) - \left(\sum_{k \neq j} \mathbf{1}_{(\hat{y}_k \neq y_k)} + \mathbf{1}_{(\hat{y}_j \neq y_j)} \right)$$
$$= \mathbf{1}_{(\hat{y}'_j = 1)} - \mathbf{1}_{(\hat{y}_j = 1)} = -1.$$

We therefore have $(\ell_H(\hat{y}', \cdot) - \ell_H(\hat{y}, \cdot) + 1)/2 = \mathbf{1}_{(y_j = 1)}$, hence

$$\underline{P}(y_j = 1) = \underline{\mathbb{E}}\left(\frac{\ell_H(\hat{y}', \cdot) - \ell_H(\hat{y}, \cdot) + 1}{2} \right)$$
$$= \frac{1}{2} \underline{\mathbb{E}}\left(\ell_H(\hat{y}', \cdot) - \ell_H(\hat{y}, \cdot) \right) + \frac{1}{2}$$

the last equality coming from scale and translation invariance of lower expectations. Hence $\underline{\mathbb{E}}(\ell_H(\hat{y}', \cdot) - \ell_H(\hat{y}, \cdot)) > 0$ if and only if $\underline{P}(y_j = 1) > 1/2$. This means that, if $\underline{P}(y_j = 1) > 1/2$, any vector $\hat{y}'$ with $\hat{y}'_j = 0$ is dominated (in the sense of Equation (3)) by the vector $\hat{y}$ where only the j-th element is modified, hence no vector with $\hat{y}'_j = 0$ is in the maximal set $\hat{Y}_{\ell_H}$. The proof showing that if $\underline{P}(y_j = 0) > 1/2$, then no vector with $\hat{y}'_j = 1$ is in the maximal set is similar. ■ □

We now provide an example showing that the inclusion of Proposition 1 is usually strict.

Example 3. Consider the 2 label case $\Lambda = \{\lambda_1, \lambda_2\}$ with the following constraints:

$$0.4 \leq P(y_1 = 1) = P(\{[1\ 0]\}) + P(\{[1\ 1]\}) \leq 0.6$$
$$0.9\,(P(\{[1\ 0]\}) + P(\{[1\ 1]\})) = P(\{[1\ 0]\})$$
$$0.84\,(P(\{[0\ 1]\}) + P(\{[0\ 0]\})) = P(\{[0\ 1]\})$$

These constraints describe a convex set $\mathscr{P}$, whose extreme points (obtained by saturating the first inequality one way or another) are summarized in Table 2. The first constraint induces that $\underline{P}(y_1 = 1) = 0.4$ and $\overline{P}(y_1 = 0) = 0.6$, while the bounds $\underline{P}(y_2 = 1) = 0.396, \overline{P}(y_2 = 1) = 0.544$, are reached by the extreme distributions $P([1\ 1]) = 0.06$, $P([0\ 1]) = 0.336$ and $P([1\ 1]) = 0.04$, $P([0\ 1]) = 0.504$, respectively. Given these bounds, we have that $\hat{Y}^* = [*\ *]$ corresponds to the whole space $\mathscr{Y}$ (i.e., the empty prediction). Yet we have that

$$\underline{\mathbb{E}}\left(\ell_H([1\ 1], \cdot) - \ell_H([0\ 0], \cdot)\right) = 0.0008 \geq 0$$

also obtained with the distribution $P([1\ 1]) = 0.06$, $P([0\ 0]) = 0.064$. This means that the vector $[0\ 0]$ is not in the maximal set $\hat{Y}_{\ell_H}$, while it is included in $\hat{Y}^*$.

Proposition 1 shows that we can rely on marginal information to provide an outer-approximation of $\hat{Y}_{\ell_H}$ that is efficient to compute, as it requires to compute $2m$ values, which are to be compared to the 2^{2m} usually required to assess $\hat{Y}_{\ell_H}$. It also indicates that extensions of the binary relevance approach are well adapted to provide partial predictions from credal sets when considering the Hamming loss, and that in this case global models integrating label dependencies are not necessary, thus saving a lot of heavy computations.

$P(\{[0\,0]\})$	$P(\{[1\,0]\})$	$P(\{[0\,1]\})$	$P(\{[1\,1]\})$
0.096	0.36	0.504	0.04
0.064	0.54	0.336	0.06

Table 2: Extreme points of $\mathscr{P}$ of Example 3

3.2. The ranking loss

In the multilabel setting, it is also quite common to consider predicting a ranking over the labels (ideally from the most relevant to the least relevant) instead of a vector of relevant labels. This is quite natural if the classifier returns for each label λ_i a score $s_{\mathbf{x}}(i) \in \mathbb{R}$ [19] (in which case the labels are ranked according to their scores), or if one uses methods especially tailored to return rankings [17, 16]. In such a case, the set of alternatives $\mathscr{A}$ is the set $\mathscr{R}_\Lambda$ of rankings over Λ, or equivalently the set $\mathscr{S}_m$ of all permutations over $[m] := \{1,\ldots,m\}$, meaning that the number of alternative $|\mathscr{A}| = m!$ is in general much higher than 2^m. For easiness of notation, we will denote by $r : [m] \to [m]$ the permutation of indices corresponding to a ranking r, and by $\succ_r$ the corresponding order between elements of Λ.

In such cases, the ranking loss can be used. Given a predicted ranking r and an observed output y, it reads as

$$\ell_R(r,y) = \sum_{\substack{(i,j)\in[m]\times[m] \\ y_i > y_j}} \mathbf{1}_{(\lambda_i \prec_r \lambda_j)} \ . \tag{7}$$

The ranking loss counts the number of pairs of labels that disagree between r and the partial order induced by y (assuming that all relevant labels are preferred to non-relevant ones). When the estimate $P_{\mathbf{x}}$ is precise, the optimal decision [12] is the ranking $\hat{r}$ that ranks items in Λ according to the values $P_{\mathbf{x}}(y_j = 1)$, that is $\lambda_i \succeq_{\hat{r}} \lambda_j$ if $\hat{P}_{\mathbf{x}}(y_i = 1) \geq \hat{P}_{\mathbf{x}}(y_j = 1)$. Again, the optimal decision can be derived from the sole knowledge of the marginals $P_{\mathbf{x}}(y_j = 1)$, $j = 1,\ldots,n$.

Example 4. Consider the space $\Lambda = \{\lambda_1,\lambda_2,\lambda_3\}$ and the predicted ranking $\lambda_2 \succ_{\hat{r}} \lambda_3 \succ_{\hat{r}} \lambda_1$, then $\hat{r}(2) = 1$, $\hat{r}(3) = 2$ and $\hat{r}(1) = 3$. If the observed multilabel output is $y = [110]$, then $\ell_R(\hat{r},y) = 1$ because $\lambda_3 \succ_{\hat{r}} \lambda_1$ while $y_1 > y_3$.

Before showing that we can extend the precise case prediction technique to compute an outer approximating prediction of the optimal set $\hat{R}_{\ell_R}$ that would be obtained by using Equation (4), we need to first establish a small result.

Lemma 1. *Given a model $\mathscr{P}_{\mathbf{x}}$ on $\mathscr{Y}$, we have that*

$$\underline{\mathbb{E}}\left(\mathbf{1}_{(y_i=1,y_j=0)} - \mathbf{1}_{(y_i=0,y_j=1)}\right) = \underline{\mathbb{E}}\left(\mathbf{1}_{(y_i=1)} - \mathbf{1}_{(y_j=1)}\right)$$

Proof. Simply consider that

$$\begin{aligned}
\inf_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}} \mathbb{E}\left(\mathbf{1}_{(y_i=1)} - \mathbf{1}_{(y_j=1)}\right) &= \inf_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}} P(y_i = 1) - P(y_j = 1) \\
&= \inf_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}} \left(P(y_i = 1, y_j = 0) + P(y_i = 1, y_j = 1)\right) - \\
&\quad \left(P(y_j = 1, y_i = 0) + P(y_j = 1, y_i = 1)\right) \\
&= \inf_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}} P(y_i = 1, y_j = 0) - P(y_j = 1, y_i = 0) \\
&= \inf_{P_{\mathbf{x}} \in \mathscr{P}_{\mathbf{x}}} \mathbb{E}\left(\mathbf{1}_{(y_i=1, y_j=0)} - \mathbf{1}_{(y_i=0, y_j=1)}\right)
\end{aligned}$$

□

We are now ready to prove the main proposition of this section

Proposition 2. *Let $\mathscr{P}_{\mathbf{x}}$ be our estimate, then the set $\hat{R}^*$ of linear extensions of the partial order R such that*

$$\lambda_i \succ_R \lambda_j \text{ if } \underline{P}(y_i = 1) > \overline{P}(y_j = 1)$$

is an outer approximation of $\hat{R}_{\ell_R}$, in the sense that $\hat{R}_{\ell_R} \subseteq \hat{R}^*$.

Proof. Similarly to the proof of Proposition 1, we will show that any alternative where $\lambda_j \succ \lambda_i$ is dominated by another one where $\lambda_i \succ \lambda_j$ if $\underline{P}(y_i = 1) > \overline{P}(y_j = 1)$. To do this, consider a ranking r where $\lambda_i \succ_r \lambda_j$ and the ranking $r^{i|j}$ where only the ranks of labels λ_i, λ_j are swapped, the others being left untouched, i.e., $r(k) = r^{i|j}(k)$ for any $k \neq i, j$, $r(i) = r^{i|j}(j)$ and $r(j) = r^{i|j}(i)$. Let us now look at the form of $\ell_R(r, \cdot) - \ell_R(r^{i|j}, \cdot)$. Depending on y, three cases can occur:

I if $y_i = 1$ and $y_j = 0$, we have

$$\begin{aligned}
\ell_R(r, y) - \ell_R(r^{i|j}, y) &= \sum_{\substack{(p,q) \in [m] \times [m] \\ y_p > y_q}} \mathbf{1}_{(\lambda_p \prec_r \lambda_q)} - \sum_{\substack{(p,q) \in [m] \times [m] \\ y_p > y_q}} \mathbf{1}_{(\lambda_p \prec_{r^{i|j}} \lambda_q)} \\
&= r(i) - r(j),
\end{aligned}$$

that is, the number of elements in the ranking r (or $r^{i|j}$) between λ_i and λ_j. To see this, consider that the only pair that could agree in r and disagree in $r^{i|j}$, outside of the pair (i, j) itself, are those involving either λ_i or λ_j and one element ranked between λ_i and λ_j. Now, for any element λ_k ranked between λ_i and λ_j, if $y_k = 0$, then the pair (i, k) is disagreeing in $r^{i|j}$ and not in r, and if $y_k = 1$, then the pair (j, k) is disagreeing in $r^{i|j}$ and not in r. This means that the difference is indeed the number of elements between λ_i and λ_j ($r(i) - r(j) - 1$), plus the pair (i, j) itself;

II if $y_i = 0$ and $y_j = 1$, we have

$$\begin{aligned}
\ell_R(r, y) - \ell_R(r^{i|j}, y) &= \sum_{\substack{(p,q) \in [m] \times [m] \\ y_p > y_q}} \mathbf{1}_{(\lambda_p \prec_r \lambda_q)} - \sum_{\substack{(p,q) \in [m] \times [m] \\ y_p > y_q}} \mathbf{1}_{(\lambda_p \prec_{r^{i|j}} \lambda_q)} \\
&= r(j) - r(i),
\end{aligned}$$

for reasons similar to the previous case.

III if $y_i = y_j$, then

$$\ell_R(r,y) - \ell_R(r^{i|j},y) = \sum_{\substack{(p,q)\in[m]\times[m] \\ y_p > y_q}} \mathbf{1}_{(\lambda_p \prec_r \lambda_q)} - \sum_{\substack{(p,q)\in[m]\times[m] \\ y_p > y_q}} \mathbf{1}_{(\lambda_p \prec_{r^{i|j}} \lambda_q)}$$
$$= 0$$

since the swap between λ_i and λ_j in r does not make new disagreeing pairs in this case.

Putting these three cases together, we have that

$$\begin{aligned} \ell_R(r,\cdot) - \ell_R(r^{i|j},\cdot) &= (r(i)-r(j))\,\mathbf{1}_{(y_i=1,y_j=0)} + (r(j)-r(i))\,\mathbf{1}_{(y_i=0,y_j=1)} \\ &= (r(i)-r(j))\left(\mathbf{1}_{(y_i=1,y_j=0)} - \mathbf{1}_{(y_i=0,y_j=1)}\right) \end{aligned}$$

Now, r dominates $r^{i|j}$ in the sense of Equation 3 if $\underline{\mathbb{E}}\left(\ell_R(r,\cdot) - \ell_R(r^{i|j},\cdot)\right) > 0$. Since we have that

$$\begin{aligned} \underline{\mathbb{E}}\left(\ell_R(r,\cdot) - \ell_R(r^{i|j},\cdot)\right) &= (r(i)-r(j))\,\underline{\mathbb{E}}\left(\mathbf{1}_{(y_i=1,y_j=0)} - \mathbf{1}_{(y_i=0,y_j=1)}\right) \\ &= (r(i)-r(j))\left(\underline{\mathbb{E}}\left(\mathbf{1}_{(y_i=1)} - \mathbf{1}_{(y_j=1)}\right)\right) \\ &\geq (r(i)-r(j))\left(\underline{\mathbb{E}}\left(\mathbf{1}_{(y_i=1)}\right) + \underline{\mathbb{E}}\left(-\mathbf{1}_{(y_j=1)}\right)\right) \\ &= (r(i)-r(j))\left(\underline{P}(y_i=1) - \overline{P}(y_j=1)\right) \end{aligned}$$

is positive if $\underline{P}(y_i = 1) > \overline{P}(y_j = 1)$, which is sufficient to show the proposition (the second equality is obtained by applying Lemma 1). $\square$

Proposition 2 shows that the partial order $\hat{R}^*$ obtained by considering the interval order induced by $[\underline{P}(y_j = 1), \overline{P}(y_j = 1)]$, that is to state that $\lambda_i \succ_R \lambda_j$ if $\underline{P}(y_i = 1) > \overline{P}(y_j = 1)$, approximates the optimal prediction. This is a straightforward extension of the optimal ranking obtained in the precise case. That the inclusion of Proposition 2 can be strict already follows from the fact that the ordering is based on an upper bound of the value $\underline{\mathbb{E}}\left(\mathbf{1}_{(y_i=1)} - \mathbf{1}_{(y_j=1)}\right)$. Proposition 2 tells us that one can also rely on marginal information on each label to optimize the rank loss, meaning that the same learning techniques can be used to approximate both $\hat{R}_{\ell_R}$ and $\hat{Y}_{\ell_H}$, which from a computational view point is a good news.

There is a more general version of the ranking loss than Equation 7, namely

$$\ell_R(r,y) = w(y) \sum_{\substack{(i,j)\in[m]\times[m] \\ y_i > y_j}} \mathbf{1}_{(\lambda_i \prec_r \lambda_j)} \,. \tag{8}$$

where each possible ground-truth y is weighted by $w(y)$. In the precise case, it has been shown [10] that such a generalization is not too problematic, and that producing an optimal ranking remains relatively easy. Unfortunately, Proposition 2 do not extend as easily when considering sets of probabilities, mainly because $\underline{\mathbb{E}}$ is a sub-additive function.

3.3. Discussion about other losses

There exist many other losses that have been defined within the multilabel setting [12, 31], some of them being the average for each label over an entire data set (micro measures), others being computed for each instance and then averaged (macro measures). Equation (3) only applies to these latter losses, among which one can find the 0/1 loss

$$\ell_{0/1}(\hat{y}, y) = \mathbf{1}_{(y \neq y)} \tag{9}$$

which is a straightforward extension of the 0/1 loss used in classification. Also commonly used is the F_1-measure

$$\ell_{F_1}(\hat{y}, y) = 1 - \frac{2\sum_{i=1}^{m} \mathbf{1}_{(y_i = \hat{y}_i = 1)}}{\sum_{i=1}^{m} \mathbf{1}_{(y_i=1)} + \sum_{i=1}^{m} \mathbf{1}_{(\hat{y}_i=1)}} \tag{10}$$

and other ones such as accuracy, recall, precision, ...

All these loss functions cannot be easily decomposed in pairwise or labelwise components, and computing the optimal prediction for such losses is already far from being trivial in the precise case (see [11] for the case of the F-measure). There is therefore very little hope to obtain easy ways to approximate their prediction sets when considering a credal set $\mathscr{P}_{\mathbf{x}}$ as our uncertainty model, and further efforts should focus on developing adequate heuristic algorithms.

4. Experiments

In this section, we provide first experimentations illustrating the effect of making partial predictions with a decreasing amount of information. These experiment show the typical behaviour we can expect from imprecise probabilistic methods in a multilabel setting: that the labels over which we abstain to make predictions are those for which we make the most mistakes. This means that the percentage of correct answer over the labels for which we still make a prediction increases.

While those predictions are not ”better” than precise ones, they are more cautious, as they do inform us when information is insufficient to make a reliable prediction. Those experiments mainly serve to illustrate how the theoretical results of Section 3 can be applied and evaluated in practice.

4.1. Evaluation

Usual loss functions such as Equations (6) or (7) are based on complete predictions. When making partial predictions, the quality of a classifier cannot be computed by simply making an average of such measures, and new measures must therefore be proposed. This can be done, for instance, by decomposing the quality into two components [7], one measuring the accuracy or correctness of the made prediction, the other measuring its completeness.

If the partial prediction is an incomplete vector such as the prediction $\hat{Y}^*$ obtained by Proposition 1, then Hamming loss can be easily split into these two components. Given the prediction $\hat{Y}^*$ characterized by subsets $\underline{L}, \overline{L}$, let us denote $Q = \Lambda \setminus (\underline{L} \cap \overline{L})$ the set of predicted labels (i.e., labels such that $\hat{Y}_j^* = 1$ or $\hat{Y}_j^* = 0$). Then, if the observed set is y, we define incorrectness (IC_H) and completeness (CP_H) as

$$IC_H(\hat{Y}^*, y) = \frac{1}{|Q|} \sum_{\lambda_i \in Q} \mathbf{1}_{(\hat{y}_i \neq y_i)} \; ; \tag{11}$$

$$CP_H(\hat{Y}^*, y) = \frac{|Q|}{m}. \tag{12}$$

when predicting complete vectors, then $CP_H = 1$ and IC_H equals the Hamming loss (6). When predicting the empty vector, then $CP_H = 0$ and by convention $IC_H = 0$.

Example 5. Consider the space $\Lambda = \{\lambda_1, \lambda_2, \lambda_3\}$, the predicted vector $\hat{Y}^* = [10*]$ and the observed vector $y = [110]$. Then we have $IC_H(\hat{Y}^*, y) = 1/2$ as only one predicted vector is correct, and $CP_H(\hat{Y}^*, y) = 2/3$ as one label remains unpredicted.

If the partial prediction is the ranking $\hat{R}^*$ obtained by Proposition 2, then we can just adopt the setting proposed by Cheng *et al.* [7]. y being the set of observed relevant label, let

$$C = |\{(i,j) | (i,j) \in [m] \times [m], y_i > y_j \wedge \lambda_i \succ_{\hat{R}^*} \lambda_j\}|$$

and

$$D = |\{(i,j) | (i,j) \in [m] \times [m], y_i > y_j \wedge \lambda_i \prec_{\hat{R}^*} \lambda_j\}|$$

be the number of label pairs compared in $\hat{R}^*$ that are respectively concordant (C) and discording (D) with the partial order induced by y. We then define incorrecteness (IC_R) and completeness (CP_R) as

$$IC_R(\hat{R}^*, y) = \frac{D}{C+D}; \tag{13}$$

$$CP_R(\hat{R}^*, y) = \frac{C+D}{(\sum_{i=1}^m y_i)(m - \sum_{i=1}^m y_i)}. \tag{14}$$

IC_R is simply the fraction of discordant pairs over the number of predicted pairs, as proposed by [7], while CP_R measure the number of predicted pairs over the total number of pairwise comparison induced by y. When predicting the empty vector, then $CP_R = 0$ and by convention $IC_R = 0$.

Example 6. Consider again the space $\Lambda = \{\lambda_1, \lambda_2, \lambda_3\}$ and the observed vector $y = [110]$ corresponding to the partial ranking $\lambda_1 \succ \lambda_3, \lambda_2 \succ \lambda_3$. If the predicted partial ranking $\hat{R}^*$ is $\lambda_2 \succ_{\hat{R}^*} \lambda_3$, then we have $C = 1$, $D = 0$ from which follows that $IC_R(\hat{R}^*, y) = 0$ (no predicted relations contradict the observed vector) and $CP_R(\hat{R}^*, y) = 1/2$ (only one of the two observed preference is predicted).

4.2. Method

Note that, in practice, the main results of this paper can be applied to any method estimating a credal set over the space $\mathscr{Y}$, as long as we can retrieve marginal bounds over each label. Whatever the complexity of $\mathscr{P}$, Propositions 1 and 2 tells us that we can use the marginal information to derive efficient predictions.

The method we used was to apply, label-wise, the k-nn method using lower probabilities introduced in [14]. This means that from an initial training data set $\mathscr{D}$, m data sets $\mathscr{D}_j$ corresponding to binary classification problems are built, this decomposition being illustrated in Figure 1. Given an instance $\mathbf{x}$, the result of the k-nn method on data set $\mathscr{D}_j$ provides an estimate of $[\underline{P}(y_j = 1), \overline{P}(y_j = 1)]$ and by duality an estimate of $\underline{P}(y_j = 0) = 1 - \overline{P}(y_j = 1)$ and $\overline{P}(y_j = 0) = 1 - \underline{P}(y_j = 1)$. Algorithm 1 provides the details about how the bounds are derived from the instances $\mathbf{x}^1, \ldots, \mathbf{x}^n$ in this particular case.

Algorithm 1: Computation of probability bounds by k-nn method

---

Input: Data set $\mathscr{D}_j$, $\beta > 0$, ε_0, k, instance $\mathbf{x}$
Output: Interval $[\underline{P}(y_j = 1), \overline{P}(y_j = 1)]$
$\underline{P}(y_j = 1) \leftarrow 0$;
$\overline{P}(y_j = 1) \leftarrow 0$;
Compute average distance $\overline{d}_{y_j=1}$ between instances $\{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ where $y_j = 1$;
Compute average distance $\overline{d}_{y_j=1}$ between instances $\{\mathbf{x}^1, \ldots, \mathbf{x}^n\}$ where $y_j = 0$;
Order instances $x^{(1)}, \ldots, x^{(n)}$ such that $d(\mathbf{x}^(i), x) \leq d(\mathbf{x}^{(i+1)}, \mathbf{x})$;
for $i = 1, \ldots, k$ **do**
    $Disc = \varepsilon_0 \cdot \exp^{\frac{-d(\mathbf{x}^{(i)}, \mathbf{x})^\beta}{\overline{d}_{y_j=1}}}$;
    if $y_j^{(i)} = 1$ **then**
        $\overline{P}(y_j = 1) \leftarrow \overline{P}(y_j = 1) + 1/k$;
        $\underline{P}(y_j = 1) \leftarrow \underline{P}(y_j = 1) + Disc/k$;
    if $y_j^{(i)} = 0$ **then**
        $\overline{P}(y_j = 1) \leftarrow \overline{P}(y_j = 1) + 1/k - Disc/k$;
    if $y_j^{(i)} = *$ **then**
        $\overline{P}(y_j = 1) \leftarrow \overline{P}(y_j = 1) + 1/k$;

As we have that $\overline{P}(y_j = 1) = 1 - \underline{P}(y_j = 0)$ and $\underline{P}(y_j = 1) = 1 - \overline{P}(y_j = 0)$, the algorithm process as follows:

- if 1 is observed, then $\underline{P}(y_j = 1)$ is increased slightly less than $\overline{P}(y_j = 1)$, thus lowering the plausibility ($\overline{P}(y_j = 0)$) of having 0;
- if 0 is observed, $\overline{P}(y_j = 1)$ is increased only slightly to reflect the fact that the neighbours only brings an imperfect information about $\mathbf{x}$, leaving $\underline{P}(y_j = 1)$ (and $\overline{P}(y_j = 0)$) untouched;
- if a missing variable is observed, then $\overline{P}(y_j = 1)$ is increased with maximal value to maximally increase the gap $\overline{P}(y_j = 1) - \underline{P}(y_j = 1)$ for this observation.

The value $Disc$ is an increasing function of the distance $d(\mathbf{x}^{(i)}, \mathbf{x})$, meaning that the more distant is a neighbour, the bigger is the value $1/k - 1/Disc$. As this corresponds to the imprecision added to $\overline{P}(y_j = 1) - \underline{P}(y_j = 1)$ when observing either a 1 or a 0 in a neighbour, this means that the further away is a neighbour, the more imprecise is its

X_1	X_2	X_3	X_4	y_1	y_2	y_3
107.1	25	*Blue*	60	1	0	0
-50	10	*Red*	40	1	0	1
200.6	30	*Blue*	58	$*$	1	0
...	...	...	...	...	...	...

data set $\mathscr{D}_1$

X_1	...	X_4	y_1
107.1		60	1
-50		40	1
200.6		58	$*$
...		...	...

data set $\mathscr{D}_2$

X_1	...	X_4	y_2
107.1		60	0
-50		40	0
200.6		58	1
...		...	...

data set $\mathscr{D}_3$

X_1	...	X_4	y_3
107.1		60	0
-50		40	1
200.6		58	0
...		...	...

Figure 1: Label-wise decomposition of data set $\mathscr{D}$

contribution to our knowledge of the current label. For example, if $\mathbf{x}^{(i)}$ is very close to $\mathbf{x}$, almost no change will be done on $[\underline{P}(y_j = 1), \overline{P}(y_j = 1)]$ if $y_j^{(i)} = 0$, while values close to $^1/_k$ will be added to both if $y_j^{(i)} = 1$.

The width of $[\underline{P}(y_j = 1), \overline{P}(y_j = 1)]$ is also increasing with the number k, hence a higher number of neighbours will provide more cautious answers. The method also automatically takes account of missing label information, as show the case $y_j^{(i)} = *$, and treat such missing data in a conservative way, considering them as completely vacuous information (that is, we treat them as non-MAR variables [37]).

4.3. Results

In the experiments, the parameters of the k-nn algorithm were set to $\beta = 0.5$ and $\varepsilon_0 = 0.99$, so that results obtained when fixing the number k of neighbors to 1 display a sufficient completeness. ε_0 settles the initial imprecision, while β determines how much imprecision increases with distance (details about the role of these parameters can be found in [14]).

We ran experiments on well-known multilabel data sets having real-valued features. Their characteristics are summarized in Table 3. Recall that the cardinality is the average number of relevant labels per instance, while the density is the average percentage of relevant labels among all labels per instance. For data set, we ran a 10-fold cross validation with the number k of neighbors varying from 1 to 3, and with various percentages of missing labels in the training data set (0%, 20% and 40%). Varying k in the algorithm allows us to control the completeness of the prediction: the higher k is, the more imprecise become the estimations.

Results on those data sets for the Hamming and ranking losses are shown in Tables 4 and 5, respectively. Note that the results display the expected behaviour, since when k increases, the incorrectness (IC) and the completeness both decrease on all data sets. This means that those labels for which we abstain to make a prediction are more

Name	# Features	# Labels	# Instances	Cardinality	Density
emotion	72	6	593	1.90	0.31
scene	294	6	2407	1.07	0.18
yeast	103	14	2417	4.23	0.30
CAL500	68	174	502	26.04	0.15
mediamill	120	101	43907	4.38	0.04
NUS-WIDE	128	81	269648	1.86	0.02

Table 3: Multilabel data sets summary

likely to be labels on which a mistake was made (otherwise, incorrectness would not decrease). In addition to that, we can make a couple of interesting remarks:

- missing data mainly affects the completeness of the predictions, but have almost no effect on incorrectness (particularly for the Hamming loss). This can be explained by the conservative approach we adopt in Algorithm 1, that is equivalent to consider every possible replacement of missing data;

- the decrease in completeness for the Hamming loss when one neighbour is added can be strong (emotions) or quite limited (scene or mediamill), and can be explained by the data set density. Yet, it is clear that when data are missing, considering multiple neighbours will quickly give quite incomplete results, which is again due to the conservativeness of our approach;

- the completeness values are much lower when considering the rank loss, and the completeness decrease is usually more important than for the Hamming loss when considering missing data or additional neighbours. This can go to the point where results are almost completely incomplete (e.g., $k = 3$ and 40% missing data for the sparser data sets CAL500, yeast, emotions and NUS-WIDE). This can be explained by the following considerations: consider a precise and fully correct model where n labels y_j are such that $P(y_j = 1) = 1$ and $m - n$ labels y_j are such that $P(y_j = 1) = 0$, hence give fully precise predictions. Now, consider that our knowledge on one of the label y_j such that $P(y_j = 1) = 1$ becomes vacuous, i.e., $[\underline{P}(y_j = 1), \underline{P}(y_j = 1)] = [0, 1]$. In this case, the value of CP_H will decrease by $1/m$, while the value of CP_R will decrease by $(m-1)/m$, a much higher number. So, the decrease of completeness for the ranking loss will usually be much quicker than for Hamming loss. Also, the produced predictions $\hat{R}^*$ for the ranking loss can be expected to be more conservative than for the Hamming loss, as we explicitly use a conservative approximation of the value $\underline{\mathbb{E}}\left(\mathbf{1}_{(y_i=1)} - \mathbf{1}_{(y_j=1)}\right)$.

These results are sufficient to show the good behaviour of the proposed predictions, as well as what can be expected from the completeness and incorrectness evaluation measures. In future works, it would be interesting to study the behaviour of additional imprecise probabilistic classifiers such as the naive credal classifier [35] or its recent extensions [8].

		Hamming loss					
		k=1		k=2		k=3	
Data set	% missing	IC_H	CP_H	IC_H	CP_H	IC_H	CP_H
CAL500	0	0.23	0.79	0.12	0.61	0.09	0.52
	20	0.23	0.63	0.12	0.40	0.08	0.28
	40	0.22	0.48	0.12	0.21	0.08	0.11
emotions	0	0.23	0.97	0.14	0.72	0.10	0.60
	20	0.22	0.78	0.15	0.49	0.09	0.31
	40	0.22	0.60	0.10	0.29	0.08	0.11
scene	0	0.11	1.00	0.06	0.90	0.04	0.81
	20	0.11	0.80	0.06	0.57	0.04	0.41
	40	0.11	0.60	0.06	0.31	0.03	0.19
yeast	0	0.25	1.00	0.16	0.76	0.13	0.61
	20	0.24	0.80	0.16	0.50	0.13	0.31
	40	0.25	0.60	0.16	0.30	0.11	0.12
mediamill	0	0.05	0.87	0.03	0.81	0.02	0.79
	20	0.05	0.72	0.03	0.51	0.02	0.40
	40	0.05	0.50	0.03	0.29	0.02	0.19
NUS-WIDE	0	0.03	0.91	0.01	0.81	0.005	0.75
	20	0.03	0.78	0.01	0.61	0.005	0.49
	40	0.03	0.67	0.01	0.49	0.005	0.36

Table 4: Experiment results: Hamming loss

5. Conclusions

Producing sets of optimal predictions in the multilabel setting when uncertainty is modeled by convex probability sets is computationally hard. The main contribution of this paper is to show that some results coming from the precise case can be partially transferred to the imprecise probabilistic case. Precisely, sets of optimal predictions under the Hamming and the ranking loss can be easily outer-approximated by considering the marginal probabilities of each label being relevant. This makes both computation and learning issues easier, as one can focus on estimating such marginals (instead of the whole joint model). We can consider that as an important result, as it shows that imprecise probabilistic approaches can be computationally affordable (at least under some conditions).

Some preliminary experiments with a nearest neighbour approach also indicate the interest of producing such partial predictions, showing that making more cautious predictions lead to more correct predictions. Nevertheless, those same results show that it can be hard to finely control the trade-off between completeness ad incorrectness. Experimental studies with other well known imprecise probabilistic classifiers should thus be performed.

Several other questions of interest remain. We can for example mention the case of the generalized form (8), for which it is not clear if efficient approximations can be obtained. This contrasts with the precise probabilistic case, where results easily extends to this form. Similarly, it would be desirable to develop methods using imprecise prob-

		Ranking loss					
		k=1		k=2		k=3	
Data set	% missing	IC_R	CP_R	IC_R	CP_R	IC_R	CP_R
CAL500	0	0.25	0.3	0.14	0.12	0.11	0.06
	20	0.26	0.20	0.15	0.05	0.15	0.02
	40	0.24	0.11	0.20	0.02	0.13	0.01
emotions	0	0.24	0.62	0.07	0.40	0.05	0.30
	20	0.20	0.40	0.05	0.29	0.03	0.10
	40	0.14	0.21	0.03	0.05	0.01	0.01
scene	0	0.29	0.73	0.15	0.72	0.11	0.68
	20	0.17	0.50	0.08	0.49	0.04	0.38
	40	0.11	0.25	0.04	0.24	0.02	0.12
yeast	0	0.25	0.61	0.15	0.39	0.09	0.28
	20	0.25	0.4	0.12	0.18	0.05	0.10
	40	0.25	0.21	0.07	0.04	0.01	0.01
mediamill	0	0.11	0.51	0.05	0.42	0.05	0.39
	20	0.11	0.32	0.05	0.23	0.04	0.20
	40	0.13	0.19	0.04	0.10	0.03	0.08
NUS-WIDE	0	0.11	0.49	0.02	0.31	0.01	0.27
	20	0.10	0.25	0.02	0.13	0.01	0.02
	40	0.07	0.13	0.01	0.02	0.01	0.01

Table 5: Experiment results: Ranking loss

abilistic ideas and dedicated to producing partial predictions corresponding to a given loss function. Another issue is that we currently use two dimensions (completeness and incorrectness) to analyze our results, making it difficult to compare them with classical multilabel methods producing determinate predictions. Note that how to meaningfully do such a comparison is already non-trivial when considering usual classification and 0/1 loss [36], and that no principled solution currently exists when considering complex outputs (here, vectors of relevant labels) or losses different from the 0/1.

Acknowledgements

Work carried out in the framework of the Labex MS2T, funded by the French Government, through the National Agency for Research (Reference ANR-11-IDEX-0004-02)

[1] J. Abellan and A. R. Masegosa. Imprecise classification with credal decision trees. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 20(05):763–787, 2012.

[2] M. S. Balch. Mathematical foundations for a theory of confidence structures. *International Journal of Approximate Reasoning*, 53(7):1003–1019, 2012.

[3] P. Bartlett and M. Wegkamp. Classification with a reject option using a hinge loss. *The Journal of Machine Learning Research*, 9:1823–1840, 2008.

[4] J. O. Berger. An overview of robust Bayesian analysis. *Test*, 3:5–124, 1994. With discussion.

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

[6] W. Cheng, E. Hüllermeier, W. Waegeman, and V. Welker. Label ranking with partial abstention based on thresholded probabilistic models. In *Advances in Neural Information Processing Systems 25 (NIPS-12)*, pages 2510–2518, 2012.

[7] W. Cheng, M. Rademaker, B. De Baets, and E. Hüllermeier. Predicting partial orders: ranking with abstention. *Machine Learning and Knowledge Discovery in Databases*, pages 215–230, 2010.

[8] G. Corani, A. Antonucci, and M. Zaffalon. Bayesian networks with imprecise probabilities: Theory and application to classification. *Data Mining: Foundations and Intelligent Paradigms*, pages 49–93, 2012.

[9] G. Corani and A. Mignatti. Credal model averaging for classification: representing prior ignorance and expert opinions. *International Journal of Approximate Reasoning*, 56:264–277, 2015.

[10] K. Dembczynski, W. Kotlowski, and E. Hüllermeier. Consistent multilabel ranking through univariate losses. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1319–1326, 2012.

[11] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hüllermeier. An exact algorithm for f-measure maximization. In *Advances in neural information processing systems*, pages 1404–1412, 2011.

[12] K. Dembczynski, W. Waegeman, W. Cheng, and E. Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1-2):5–45, 2012.

[13] A. Dempster. Upper and lower probabilities induced by a multivalued mapping. *Annals of Mathematical Statistics*, 38:325–339, 1967.

[14] S. Destercke. A k-nearest neighbours method based on imprecise probabilities. *Soft Comput.*, 16(5):833–844, 2012.

[15] S. Destercke. Multilabel prediction with probability sets: the hamming loss case. In *Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 496–505. Springer, 2014.

[16] A. Elisseeff and J. Weston. A kernel method for multi-labelled classification. In *Advances in neural information processing systems*, pages 681–687, 2001.

[17] J. Fürnkranz, E. Hüllermeier, E. Loza Mencía, and K. Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73(2):133–153, 2008.

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

[19] T. Joachims. Text categorization with suport vector machines: Learning with many relevant features. In *ECML*, pages 137–142, 1998.

[20] I. Levi. *The Enterprise of Knowledge*. MIT Press, London, 1980.

[21] C. F. Manski. *Partial identification of probability distributions*. Springer, 2003.

[22] E. Montañes, R. Senge, J. Barranquero, J. Ramón Quevedo, J. José del Coz, and E. Hüllermeier. Dependent binary relevance models for multi-label classification. *Pattern Recognition*, 47(3):1494–1508, 2014.

[23] I. Pillai, G. Fumera, and F. Roli. Multi-label classification with a reject option. *Pattern Recognition*, 46(8):2256–2266, 2013.

[24] J. Ramón Quevedo, O. Luaces, and A. Bahamonde. Multilabel classifiers with a probabilistic thresholding strategy. *Pattern Recognition*, 45(2):876–883, 2012.

[25] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine learning*, 85(3):333–359, 2011.

[26] G. Shafer and V. Vovk. A tutorial on conformal prediction. *The Journal of Machine Learning Research*, 9:371–421, 2008.

[27] Y.-Y. Sun, Y. Zhang, and Z.-H. Zhou. Multi-label learning with weak label. In *Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

[28] F. Tian and X. Shen. Image annotation with weak labels. In *Web-Age Information Management*, pages 375–380. Springer, 2013.

[29] M. Troffaes. Decision making under uncertainty using imprecise probabilities. *Int. J. of Approximate Reasoning*, 45:17–29, 2007.

[30] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. P. Vlahavas. Multi-label classification of music into emotions. In *ISMIR*, volume 8, pages 325–330, 2008.

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

[32] P. Walley. *Statistical reasoning with imprecise Probabilities*. Chapman and Hall, New York, 1991.

[33] G. Yu, C. Domeniconi, H. Rangwala, and G. Zhang. Protein function prediction using dependence maximization. In *Machine Learning and Knowledge Discovery in Databases*, pages 574–589. Springer, 2013.

[34] M. Zaffalon. Exact credal treatment of missing data. *Journal of Statistical Planning and Inference*, 105(1):105–122, 2002.

[35] M. Zaffalon. The naive credal classifier. *J. Probabilistic Planning and Inference*, 105:105–122, 2002.

[36] M. Zaffalon, G. Corani, and D. Mauá. Evaluating credal classifiers by utility-discounted predictive accuracy. *International Journal of Approximate Reasoning*, 53(8):1282–1301, 2012.

[37] M. Zaffalon and E. Miranda. Conservative inference rule for uncertain reasoning under incompleteness. *Journal of Artificial Intelligence Research*, 34(2):757, 2009.