Distributionally Robust, Skeptical Binary Inferences in Multi-label Problems

Yonatan Carlos Carranza Alarcón YONATAN-CARLOS.CARRANZA-ALARCON@HDS.UTC.FR
Sébastien Destercke SEBASTIEN.DESTERCKE@HDS.UTC.FR
Sorbonne Universités, Université Technologique de Compiègne, CNRS, UMR 7253 - Heudiasyc, 57 Avenue de Landshut, Compiègne, France

Abstract

In this paper, we consider the problem of making distributionally robust, skeptical inferences for the multi-label problem, or more generally for Boolean vectors. By distributionally robust, we mean that we consider sets of probability distributions, and by skeptical we understand that we consider as valid only those inferences that are true for every distribution within this set. Such inferences will provide partial predictions whenever the considered set is sufficiently big. We study in particular the Hamming loss case, a common loss function in multi-label problems, showing how skeptical inferences can be made in this setting. We also perform some experiments demonstrating the interest of our results.

Keywords: Multi-label, Hamming loss, maximality, binary relevance.

1. Introduction

In contrast to multi-class problems where each instance is associated to one label, multi-label classification consists in associating an instance to a subset of relevant labels from a set of possible labels. Such problems can arise in different research fields, such as the classification of proteins in bioinformatics [22], text classification in information retrieval [9], object recognition in computer vision [3], etc.

Considering all possible subsets of labels as possible predictions make the estimation and decision steps of a learning problem significantly more difficult: partial observations are more likely to occur, especially when the number of labels increases, and the output space over which the probability needs to be estimated grows exponentially with the number of labels. This means that in some applications where guaranteeing the robustness and reliability of predictions is of particular importance, one may consider being cautious about such predictions, by predicting a set of possible answers rather than a single one when uncertainties are too high. In the literature, such strategies can be called partial rejection rules [19], partial abstention [18] or indeterminate classification [8, 1].

In this paper, we consider the problem of making such set-valued predictions by performing skeptic inferences when our uncertainty is described by a set of probabilities. By skeptic inference, we understand the logical procedure that consists, in the presence of multiple models, in accepting only those inferences that are true for every possible model. Such approaches are different from thresholding approaches [18, 19], and are closer in spirit to distributionally robust approaches, even if these later typically consider precise, minimax inferences, that are cautious yet not skeptic [12, 4]. We also make no assumption about the considered set of probabilities, thus departing from usual distributionally robust approaches, that typically consider precise predictions, or from existing works dealing with sets of probabilities and multi-label problems [1], that considered specific probability sets and zero/one loss function (seldom used in multi-label problems).

We first introduce in Section 2 the notations we will use for the multi-label setting, and give the necessary reminders about skeptic inferences made with sets of probabilities. Once this is done, we provide in Section 3 novel theoretical results concerning the Hamming loss and the maximality decision criterion, those results ending in an inference procedure that has an almost linear time complexity with respect to the size of the output space. We also investigate conditions under which previous heuristics using marginal probability bounds become exact.

We end the paper in Section 4 by performing some experiments whose goal is first to compare the inferences obtained by our exact procedures to previous heuristics, and second to investigate those settings where providing cautious inferences may be of interest.

2. Preliminaries

This section introduces the necessary background to understand the rest of this paper.

2.1. Multi-label Problem

In multi-label problems, given a subset $\Omega = \{\omega_1, \ldots, \omega_m\}$ of possible labels, one assumes that to each instance x of an input space $\mathscr{X} = \mathbb{R}^d$ is associated a subset $\Lambda \subseteq \Omega$ of relevant labels. In practice, we will identify such subsets with the space of Boolean vectors $\mathscr{Y} = \{0,1\}^m$, denoting a vector $\boldsymbol{y} = (y_1, \ldots, y_m)$ and having $y_i = 1$ if $\omega_i \in \Lambda$, 0 else.

We assume that observations are i.i.d. samples of a distribution $p : \mathscr{X} \times \mathscr{Y} \to [0,1]$, and denote $p_x(\boldsymbol{y}) := p(\boldsymbol{y}|x)$

the conditional probability of $\boldsymbol{y}$ given x. We denote by $\mathbf{Y} = (Y_1, \ldots, Y_m)$ the random binary vector over $\mathscr{Y}$. Given a subset $\mathscr{I} \subseteq \{1, \ldots, m\}$ of indices, we denote by $\mathscr{Y}_\mathscr{I}$ the space of binary vectors over those indices, by $Y_\mathscr{I}$ and $Y_{-\mathscr{I}}$ the marginals of $\mathbf{Y}$ over these indices and over the complementary indices $\{1, \ldots, m\} \setminus \mathscr{I}$, respectively. In particular, $Y_{\{i\}}$ will denote the marginal random variable over the ith label. Similarly, we will denote by $\boldsymbol{y}_\mathscr{I}$ the values of a vector restricted to elements indexed in $\mathscr{I}$, and by $\boldsymbol{b}_\mathscr{I}$ a particular assignment over these elements. The associated marginal probability will be

$$P_x(\boldsymbol{b}_\mathscr{I}) = \sum_{\boldsymbol{y} \in \mathscr{Y}, \boldsymbol{y}_\mathscr{I} = \boldsymbol{b}_\mathscr{I}} p_x(\boldsymbol{y}).$$

We will also consider the complement of a given vector or assignment over a subset of indices. These will be denoted by $\overline{\boldsymbol{y}}_\mathscr{I}$ and $\overline{\boldsymbol{b}}_\mathscr{I}$, respectively. Given two vectors $\boldsymbol{y}^1$ and $\boldsymbol{y}^2$, we will denote by $\mathscr{I}_{\boldsymbol{y}^1 \neq \boldsymbol{y}^2} := \{i \in \{1, \ldots, m\} : y^1_i \neq y^2_i\}$ the set of indices over which two vectors are different, and similarly by $\mathscr{I}_{\boldsymbol{y}^1 = \boldsymbol{y}^2} := \{i \in \{1, \ldots, m\} : y^1_i = y^2_i\}$ the sets of indices for which they will be equal.

Example 1 *Consider the probabilistic tree developed in Figure 1 defined over $\mathscr{Y} = \{0,1\}^2$ describing a full joint distribution over two labels. In such trees, the probability of any vector is simply the product of the probabilities along its path. We can consider the assignment $\boldsymbol{b}_2 = (1)$ and its complement $\overline{\boldsymbol{b}}_2 = (0)$ associated to the partial vectors $(\cdot, 1)$ and $(\cdot, 0)$, the first one having probability*

$$\begin{aligned} P(\boldsymbol{b}_2 = (1)) &= P((\cdot,1)) = P((0,1)) + P((1,1)) \\ &= 0.5 \cdot 0.2 + 0.5 \cdot 0.7 = 0.45. \end{aligned}$$

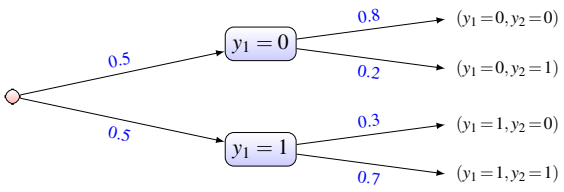


Figure 1: Probabilistic binary tree of two labels

In the sequel of this paper, we will use such trees to illustrate our results, replacing the precise probabilities on the branches by intervals[1]*. An example will be provided later. The resulting set of probabilities over $\mathscr{Y}$ will then simply be the set of all joint probabilities obtained by taking precise values within those intervals.*

As in this paper we are interested in making set-valued predictions for the multi-label problems, we will use the notation $\mathbb{Y} \subseteq \mathscr{Y}$ for generic subsets of $\mathscr{Y}$. We will use the notation $\mathfrak{Y} = \{0, 1, *\}^m$ for the specific subsets induced by partially specified binary vectors $\mathfrak{y} \in \mathfrak{Y}$, where a symbol $*$ stands for a label on which we abstain. Denoting by $\mathscr{I}^*$ the indices of such labels, we will also slightly abuse the notation $\mathfrak{y}$ and $\mathfrak{Y}$ to also denote the corresponding family of subsets over $\mathscr{Y}$, i.e.,

$$\mathfrak{y} := \{\boldsymbol{y} \in \mathscr{Y} : \forall i \notin \mathscr{I}^*, y_i = \mathfrak{y}_i\}.$$

Such subsets are indeed often used to make partial multi-label predictions, and we will refer to them on multiple occasions, calling them partial vectors. However, using only subsets within $\mathfrak{Y}$ may be insufficient if one wants to express complex partial predictions. For instance, in the case where $m = 2$, the partial prediction $\mathbb{Y} = \{(0,1), (1,0)\}$ cannot be expressed as an element of $\mathfrak{Y}$, as approximating $\mathbb{Y}$ with an element of $\mathfrak{Y}$ would result in $\mathbb{Y}$, and not the initial subset.

2.2. Skeptic Inferences with Distribution Sets

Basic representation We assume that our uncertainty is described by a convex set of probabilities $\mathscr{P}$, a.k.a. a *credal set* [16], defined over $\mathscr{Y}$. Such sets can arise in various ways: as a native result of the learning method [1, 5]; as the result of an agnostic[2] estimation in presence of imprecise data [20]; or as a neighbourhood taken over an initial estimated distribution $\hat{p}$, such as in distributionally robust approaches [4].

Skeptic inference and decision Once our uncertainty is described by a credal set $\mathscr{P}$, the next step in the learning process is to deliver an optimal prediction, given a loss function $\ell : \mathscr{Y} \times \mathscr{Y} \to \mathbb{R}$ where $\ell(\hat{\boldsymbol{y}}, \boldsymbol{y})$ is the loss incurred by predicting $\hat{\boldsymbol{y}}$ when $\boldsymbol{y}$ is the ground-truth.

When the estimate $\hat{p}$ is precise, this is classically done by picking the prediction minimizing the expected loss, i.e.

$$\hat{\boldsymbol{y}}^{\hat{p}}_\ell = \arg\min_{\boldsymbol{y}' \in \mathscr{Y}} \mathbb{E}_{\hat{p}}\left(\ell(\boldsymbol{y}', \cdot)\right) = \arg\min_{\boldsymbol{y}' \in \mathscr{Y}} \sum_{\boldsymbol{y} \in \mathscr{Y}} \hat{p}(\boldsymbol{y}) \ell(\boldsymbol{y}', \boldsymbol{y}) \quad (1)$$

or, equivalently, by picking the maximal elements of the linear ordering $\succeq^{\hat{p}}_\ell$ where $\boldsymbol{y}'' \succeq^{\hat{p}}_\ell \boldsymbol{y}'$ if

$$\begin{aligned} \mathbb{E}_{\hat{p}}\left(\ell(\boldsymbol{y}', \cdot) - \ell(\boldsymbol{y}'', \cdot)\right) &= \sum_{\boldsymbol{y} \in \mathscr{Y}} \hat{p}(\boldsymbol{y}) \left(\ell(\boldsymbol{y}', \boldsymbol{y}) - \ell(\boldsymbol{y}'', \boldsymbol{y})\right) \\ &= \mathbb{E}_{\hat{p}}\left(\ell(\boldsymbol{y}', \cdot)\right) - \mathbb{E}_{\hat{p}}\left(\ell(\boldsymbol{y}'', \cdot)\right) \geq 0, \end{aligned} \quad (2)$$

Since $\succeq^{\hat{p}}_\ell$ is a complete pre-order, picking any of the possibly indifferent maximal elements will be equivalent with respect to expected loss minimization.

When considering a set $\mathscr{P}$ as our uncertainty representation, there are many ways [21] to extend Equation (2). In

1. While we will use IP trees for illustrative purpose, our results hold for any credal set, not only those given by IP trees, that cannot represent all possible credal sets over $\mathscr{Y}$, in contrast with the precise case.

2. With respect to the missingness process.

this paper, we will consider the main decision rule that may return more than one decision in case of insufficient information: maximality. This rule follows a skeptical strategy, in the sense that the returned set of predictions is guaranteed to contain the optimal prediction, whatever the true distribution within $\mathscr{P}$.

Definition 1 *Maximality consists in returning the maximal, non-dominated elements of the partial order $\succ_\ell^{\mathscr{P}}$ such that $\boldsymbol{y} \succ_\ell^{\mathscr{P}} \boldsymbol{y}'$ if*

$$\underline{\mathbb{E}}\left(\ell(\boldsymbol{y}',\cdot)-\ell(\boldsymbol{y},\cdot)\right) := \inf_{P\in\mathscr{P}} \mathbb{E}_P\left(\ell(\boldsymbol{y}',\cdot)-\ell(\boldsymbol{y},\cdot)\right) > 0, \quad (3)$$

that is if exchanging $\boldsymbol{y}'$ for $\boldsymbol{y}$ is guaranteed to give a positive expected loss. The maximality rule returns the prediction set

$$\hat{\mathbb{Y}}^M_{\ell,\mathscr{P}} = \left\{ \boldsymbol{y} \in \mathscr{Y} \,\middle|\, \nexists \boldsymbol{y}' \in \mathscr{Y} \;\; s.t. \;\; \boldsymbol{y}' \succ_\ell^{\mathscr{P}} \boldsymbol{y} \right\}. \quad (4)$$

Since $\succ_{\ell,\mathscr{P}}$ is in general a partial order with incomparabilities, $\hat{\mathbb{Y}}^M_{\ell,\mathscr{P}}$ may result in a set of multiple, incomparable elements. Clearly, the more imprecise is $\mathscr{P}$, the larger is the set $\hat{\mathbb{Y}}^M_{\ell,\mathscr{P}}$. Computing $\hat{\mathbb{Y}}^M_{\ell,\mathscr{P}}$ can be a computationally demanding task, thus making the prediction step critical when considering combinatorial spaces, such as the one considered in this paper. Obtaining $\hat{\mathbb{Y}}^M_{\ell,\mathscr{P}}$ may indeed require at worst to perform $(|\mathscr{Y}|)(|\mathscr{Y}|-1)/2$ comparisons, where $|\mathscr{Y}| = 2^m$, ending up with a complexity of $\mathscr{O}(2^{2m})$ that quickly becomes untractable even for small values of m.

Example 2 *Figure 2(a) illustrates the computation of an expected loss in the case of a probabilistic tree and the zero/one loss function when comparing $\boldsymbol{y}' = (0,1)$ and $\boldsymbol{y}'' = (1,0)$. Global expectation is reached by making local, backward computations. In this case, we have that $(1,0) \succ^p_{\ell_{0/1}} (0,1)$, the expectation being positive.*

Figure 2(b) pictures an imprecise probabilistic tree for the same situation, with interval probabilities. The computation of the corresponding lower expectation is done in the same way as in the precise case, starting from the leaves and iterating local computations. In the example $(1,0) \succ^{\mathscr{P}}_{\ell_{0/1}} (0,1)$ as the final lower expectation is positive.

Thus, simply enumerating elements of $\mathscr{Y}$ is not practically possible, and other strategies need to be adopted. We next show that in the case of Hamming loss, one of the most common loss used in multi-label and binary problems, we can use an efficient algorithmic procedure to perform skeptic inferences. This is done both for general sets $\mathscr{P}$ and for specific sets induced from binary relevance models.

3. Skeptic Inference for the Hamming Loss

The Hamming loss, that we will denote ℓ_H, is a commonly used loss in multi-label problems. It simply amounts to compute the Hamming distance between the ground truth $\boldsymbol{y}$ and a prediction $\hat{\boldsymbol{y}}$, that is

$$\ell_H(\hat{\boldsymbol{y}},\boldsymbol{y}) = \sum_{i=1}^m \mathbb{1}_{(\hat{y}_i \neq y_i)} = |\mathscr{I}_{\hat{\boldsymbol{y}}\neq\boldsymbol{y}}| \quad (5)$$

where $\mathbb{1}_{(A)}$ denotes the indicator function of the event A. Note that in contrast with the subset loss $\ell_{0/1}(\hat{\boldsymbol{y}},\boldsymbol{y}) = \mathbb{1}_{(\hat{\boldsymbol{y}}\neq\boldsymbol{y})}$, the Hamming loss differentiates the situations where only some mistakes are made from the ones where a lot of them are made (being maximum when $\hat{\boldsymbol{y}} = \overline{\boldsymbol{y}}$ is the complement of $\boldsymbol{y}$).

In the case of precise probabilities, it is also useful to recall that the optimal prediction for the Hamming loss [7], i.e. the vector $\hat{\boldsymbol{y}}_{\ell_H,p}$ satisfying Equation (1) is

$$\hat{y}_{i,\ell_H,p} = \begin{cases} 1 & \text{if } p(Y_{\{i\}}=1) \geq \frac{1}{2} \\ 0 & \text{else.} \end{cases} \quad (6)$$

When considering a set $\mathscr{P}$ of distribution, one is immediately tempted to adopt as partial prediction the partial vector $\hat{\mathfrak{y}}_{\ell_H,\mathscr{P}} \in \mathfrak{Y}$ such that

$$\hat{\mathfrak{y}}_{i,\ell_H,\mathscr{P}} = \begin{cases} 1 & \text{if } \underline{P}(Y_{\{i\}}=1) > \frac{1}{2} \\ 0 & \text{if } \underline{P}(Y_{\{i\}}=0) > \frac{1}{2} \\ * & \text{if } \frac{1}{2} \in [\underline{P}(Y_{\{i\}}=1), \overline{P}(Y_{\{i\}}=1)]. \end{cases} \quad (7)$$

It has however been proven that $\hat{\mathfrak{y}}_{\ell_H,\mathscr{P}}$ is an outer-approximation of $\hat{\mathbb{Y}}^M_{\ell,\mathscr{P}}$ (i.e., $\hat{\mathbb{Y}}^M_{\ell,\mathscr{P}} \subseteq \hat{\mathfrak{y}}_{\ell_H,\mathscr{P}}$), thus providing a quick heuristic to get an approximate answer [8].

The next sections study the problem of providing exact skeptic inferences, first for any possible probability set $\mathscr{P}$, then for the specific case where $\mathscr{P}$ is built from marginal models on each label, that corresponds to binary relevance approaches in multi-label learning.

3.1. General Case

In this section, we demonstrate that for the Hamming loss, we can use inference procedures that are much more efficient than an exhaustive, naive enumeration. Let us first simplify the expression of the expected value.

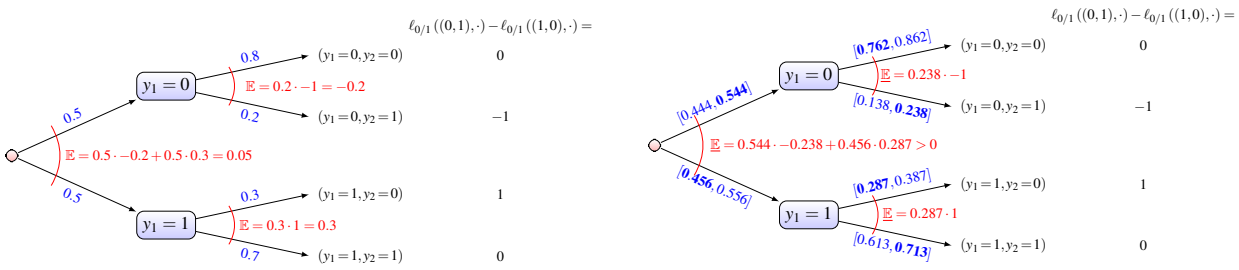
Lemma 2 *In the case of Hamming loss and given $\boldsymbol{y}^1, \boldsymbol{y}^2$, we have*

$$\mathbb{E}\left[\ell_H(\boldsymbol{y}^2,\cdot) - \ell_H(\boldsymbol{y}^1,\cdot)\right] = \sum_{i=1}^m P(Y_i = y_i^1) - P(Y_i = y_i^2) \quad (8)$$

If we consider a set of indices $\mathscr{I}_{\boldsymbol{y}^1=\boldsymbol{y}^2}$ for which Equation (8) is cancelled, it can be rewritten

$$\sum_{i\in\mathscr{I}_{\boldsymbol{y}^1\neq\boldsymbol{y}^2}} P(Y_i = y_i^1) - P(Y_i = y_i^2). \quad (9)$$

The next proposition shows that this expression can be leveraged to perform the maximality check of Equation (3) on a limited number of vectors.



(*a*) Probabilistic tree and expected loss

(*b*) Imprecise probabilistic tree and lower expected loss

Figure 2: Precise and imprecise probabilistic trees

Proposition 3 *For a given set $\mathscr{I}$ of indices, let us consider an assignment $\boldsymbol{a}_{\mathscr{I}}$ and its complement $\overline{\boldsymbol{a}}_{\mathscr{I}}$. Then, for any two vectors $\mathbf{y}^1,\mathbf{y}^2$ such that $\mathbf{y}^1_{\mathscr{I}}=\boldsymbol{a}_{\mathscr{I}}$, $\mathbf{y}^2_{\mathscr{I}}=\overline{\boldsymbol{a}}_{\mathscr{I}}$ and $\mathbf{y}^1_{-\mathscr{I}}=\mathbf{y}^2_{-\mathscr{I}}$, we have*

$$\mathbf{y}^1 \succ_M \mathbf{y}^2 \iff \inf_{P\in\mathscr{P}} \sum_{i\in\mathscr{I}} P(Y_i=a_i) > \frac{|\mathscr{I}|}{2} \tag{10}$$

In the remaining of the paper, given a partial assignment $\boldsymbol{b}_{\mathscr{I}}$ over a subset of indices $\mathscr{I}$, we will define the partial Hamming loss between $\boldsymbol{b}_{\mathscr{I}}$ and an observation $\mathbf{y}$ as

$$\ell^*_H(\boldsymbol{b}_{\mathscr{I}},\mathbf{y}) = \sum_{i\in\mathscr{I}} \mathbb{1}_{(b_i\neq y_i)}\,. \tag{11}$$

When $\mathscr{I}=\{1,\ldots,m\}$, we simply retrieve the usual Hamming loss. The next proposition shows that the condition of Proposition 3 actually comes down to minimize the expected partial Hamming loss.

Proposition 4 *For a given set $\mathscr{I}$ of indices, let us consider an assignment $\boldsymbol{a}_{\mathscr{I}}$ and its complement $\overline{\boldsymbol{a}}_{\mathscr{I}}$. We have*

$$\inf_{P\in\mathscr{P}} \sum_{i\in\mathscr{I}} P(Y_i=a_i) = \underline{\mathbb{E}}[\ell^*_H(\overline{\boldsymbol{a}}_{\mathscr{I}},\cdot)] \tag{12}$$

This allows us to use Algorithm 1 to find $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}$. The following result provides the time complexity of the algorithm.

Proposition 5 *Algorithm 1 has to perform 3^m-1 computations, and its complexity is in $\mathscr{O}(3^m)$*

Proposition 5 tells us that, in the case of Hamming loss, finding $\hat{\mathbb{Y}}^M_{\ell,\mathscr{P}}$ can be done almost linearly with respect to the size of $\mathscr{Y}$. This is to be compared to a naive enumeration, that requires $(2^m)(2^m-1)$ computations. Figure 3 plots the two curves as a function of the number m of labels, demonstrating that our result allows a significant gain in computations. In later experiments, we shall study the differences between $\hat{\mathbb{Y}}^M_{\ell,\mathscr{P}}$ and the crude approximation of Equation (7). Also, such a strategy can be optimized by using well-known techniques [2, algo. 16.4]. As said before, the set $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}$ will in general not be exactly described by a partial vector within $\mathfrak{Y}$, as shows the next example.

Algorithm 1: Maximal solutions under Hamming loss and general set

Data: $\mathscr{P}$ (convex set of distributions)
Result: $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}$ (set of undominated solutions)
$S=\mathscr{Y}$;
for *i in 1:m* **do**
  $\mathscr{Z}_i=\{\mathscr{I}:\mathscr{I}\subseteq\{1,\ldots,m\},|\mathscr{I}|=i\}$; // Index sets of size i
  forall $z\in\mathscr{Z}_i$ **do**
    forall $\boldsymbol{a}_z\in\mathscr{Y}_z$; // Binary vectors over indices in z **do**
      if $\inf_{P\in\mathscr{P}}\sum_{j\in z}P(Y_j=a_j)>\frac{i}{2}$ **then**
        $S=S\setminus\{\mathbf{y}\in\mathscr{Y}:\mathbf{y}_z=\overline{\boldsymbol{a}}_z\}$;
    end
  end
end

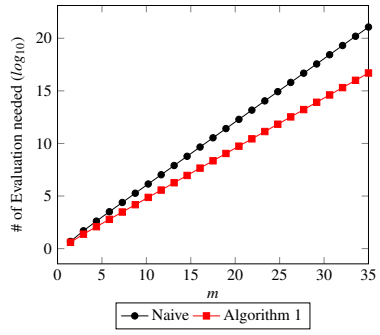


Figure 3: Comparison of Algorithm 1 with naive enumeration (log-ordinate scale).

Example 3 *Consider again the tree provided in Figure 2(b). The result of applying Algorithm 1 provides the following results:*

$$\underline{\mathbb{E}}[\ell_H((1,*),\cdot)]=0.444>0.5 \implies (0,*)\not\succ_M(1,*),$$
$$\underline{\mathbb{E}}[\ell_H((0,*),\cdot)]=0.456>0.5 \implies (1,*)\not\succ_M(0,*),$$

$$\underline{\mathbb{E}}[\ell_H((*,1),\cdot)] = 0.498 > 0.5 \implies (*,0) \not\succ_M (*,1),$$
$$\underline{\mathbb{E}}[\ell_H((*,0),\cdot)] = 0.354 > 0.5 \implies (*,1) \not\succ_M (*,0),$$
$$\underline{\mathbb{E}}[\ell_H((1,1),\cdot)] = 0.942 > 1.0 \implies (0,0) \not\succ_M (1,1),$$
$$\underline{\mathbb{E}}[\ell_H((1,0),\cdot)] = 0.846 > 1.0 \implies (0,1) \not\succ_M (1,0),$$
$$\underline{\mathbb{E}}[\ell_H((0,1),\cdot)] = 1.001 > 1.0 \implies (\mathbf{1},\mathbf{0}) \succ_M (\mathbf{0},\mathbf{1}),$$
$$\underline{\mathbb{E}}[\ell_H((0,0),\cdot)] = 0.810 > 1.0 \implies (1,1) \not\succ_M (0,0),$$

where for two partial vectors $\boldsymbol{y}^1, \boldsymbol{y}^2$ such that $\mathscr{I}^*_{\boldsymbol{y}^1} = \mathscr{I}^*_{\boldsymbol{y}^2}$, we use the short-hand notation $\boldsymbol{y}^1 \succ_M \boldsymbol{y}^2$ to say that the dominance relation given by Definition 1 holds for any fixed replacement of the abstained labels.

About this example, we can first note that $3^2 - 1 = 8$ comparisons are performed (in accord with Proposition 5). Secondly, also note that the final solution which is the set

$$\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}} = \{(1,0),(0,0),(1,1)\}$$

does not belong to $\mathcal{Y}$.

Remark 6 *A key finding of the results of this section, illustrated by Example 3, is that when considering sets of distributions and skeptic inferences, it is not sufficient to consider marginal probabilities in order to get optimal, exact predictions. This contrasts heavily with the case of precise distributions, in which having only the marginal information allows to get optimal predictions for a number of loss functions, including the Hamming loss, but also precision@k, micro- and macro-F measure, as well as others [14, 15].*

Remark 7 *Despite our best efforts and except for some few tweaks, we were not really able to significantly lower the complexity of finding $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}$, i.e., to go from the still exponential complexity of Proposition 5 to a polynomial one in the number of labels. This contrasts with the precise case, where one can use the marignal information to obtain the result in a polynomial time in the number of labels. The fact that we cannot rely on the marginal bounds of $P(Y_i)$ suggest us that reaching such a polynomial complexity for exact inferences over generic credal sets $\mathscr{P}$ may not be doable.*

3.2. Binary Relevance and Partial Vectors

The previous section looked at the very general case where the set $\mathscr{P}$ is completely arbitrary and proposed some rather efficient inference methods (almost linear in the size of $\mathscr{Y}$) for this case. In this section, we are interested in conditions imposed upon $\mathscr{P}$ that guarantee the sets $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}$ to be partial vectors, that is to belong to $\mathcal{Y}$. In particular, we show that this is the case when considering models that generalize binary relevance notions by using imprecise marginals with an assumption of independence. The interest in studying such models is that they constitute baseline models when it comes to multi-label problems.

In this section, we consider that the joint probability p over $\mathscr{Y}$ and its imprecise extension are built as follows: we have information on the marginal probability $p_i \in [0,1]$ of y_i being positive, and define the probability of a vector $\boldsymbol{y}$ as

$$p(\boldsymbol{y}) = \prod_{\{i|y_i=1\}} p_i \prod_{\{i|y_i=0\}} (1-p_i). \tag{13}$$

Without loss of generality, the imprecise version then amounts to consider that the information we have is an interval $[\underline{p}_i, \overline{p}_i]$, as every convex set of probabilities on a binary space (here, $\{0,1\}$) is an interval. We then consider that a probability set $\mathscr{P}_{BR}$ over $\mathscr{Y}$ amounts to consider the robust version of Equation (13), that is

$$p(\boldsymbol{y}) \in \left\{ \prod_{\{i|y_i=1\}} p_i \prod_{\{i|y_i=0\}} (1-p_i) | p_i \in [\underline{p}_i, \overline{p}_i] \right\}. \tag{14}$$

In this specific case, we can show that $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}$ can be exactly described by a partial vector.

Proposition 8 *Given a probability set $\mathscr{P}_{BR}$ and the Hamming loss, the set $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}_{BR}} \in \mathcal{Y}$*

Remark 9 *As the optimal prediction for the 0/1 or subset loss $\ell_{0/1}$ in the precise case is the same as Equation (6) when $p(\boldsymbol{y})$ is of the kind (13), it follows that Proposition 8 is also true for this loss.*

4. Experiments

In this section, we perform some empirical experiments investigating the interest of using skeptical inferences rather than precisely-valued inferences when uncertainties are too high. More precisely, after formalizing inferences in trees (such as the one used in Example 2), we first evaluate, through simulation, the difference between exact inferences and the approximation of Equation (7). We then investigate, under an assumption of binary relevance (i.e. independent binary models), the interest of using IP to produce partial, skeptic inferences. We investigate in particular how such a setting cope with missing labels.

4.1. Inference in Binary Trees

As we saw in Proposition 3 and Algorithm 1, estimating $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}$ implies the calculation of the infimum expectation $\underline{\mathbb{E}}_{\mathbf{Y}}[\ell_H(\cdot,\overline{\boldsymbol{a}}_{\mathscr{I}})]$ given an assignment $\boldsymbol{a}_{\mathscr{I}}$. One possibility to compute it is to write it as an iterated conditional expectation over the chain of labels, i.e.,

$$\mathbb{E}_{\mathbf{Y}}[\ell_H(\cdot,\overline{\boldsymbol{a}}_{\mathscr{I}})] = \inf_{P\in\mathscr{P}} \mathbb{E}_{Y_1}\left[\mathbb{E}_{Y_2}\left[\ldots \mathbb{E}_{Y_m}\left[\ell_H(\cdot,\overline{\boldsymbol{a}}_{\mathscr{I}}) \middle| Y_{\mathscr{I}_{[|m-1|]}}\right]\ldots\right]\middle|\right], \tag{15}$$

where $[|j|] = \{1,2,\ldots,j-1,j\}$ is a set of previous indices and $Y_{\mathscr{I}_{[|m-1|]}} = \{Y_1,\ldots,Y_{m-1}\}$ is a random binary

vector. While in general such an expectation has to be computed globally, it has been shown by Hermans and De Cooman [11] that in the specific case of tree structures, it can be computed recursively. using the law of iterated lower expectations[3]

$$\underline{\mathbb{E}}_{\mathbf{Y}}[\ell_H(\cdot,\bar{a}_{\mathscr{I}})]=\underline{\mathbb{E}}_{Y_1}\left[\underline{\mathbb{E}}_{Y_2}\left[\ldots\underline{\mathbb{E}}_{Y_m}\left[\ell_H(\cdot,\bar{a}_{\mathscr{I}})\middle|Y_{\mathscr{I}_{[|m-1|]}}\right]\ldots\right]\right]. \tag{16}$$

Equation (16) allows one to compute global infimum expectation using local models and backward recursion, i.e., we first compute the local lower expectations starting from the leaves of the tree and proceed iteratively (for further details see [24]). Figure 2(*b*) is an illustration of this procedure.

Finally, let us note that computing marginals $\underline{P}(Y_{\{i\}}=0)$ and $\underline{P}(Y_{\{i\}}=1)$ used in Equation (7) is equally easy, as it amounts to compute the lower expectation of the indicator functions $\mathbb{1}_{(y_i=0)}$ and $\mathbb{1}_{(y_i=1)}$, respectively.

4.2. Exact vs Approximate Skeptic Inference

In this section, we want to assess how good is the outer-approximation given by Equation (7), by comparing it to an exact estimation of the set $\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}$. Such an estimate is essential to know in which situation Equation (7) is likely to give a too conservative outer-approximation, and in which cases it can safely be used.

To perform this study, we simulate credal sets $\mathscr{P}$ over $\mathscr{Y}$ by generating binary trees in the following way: we choose an $\varepsilon\in[0,0.5]$, and for a label Y_i and a path $y_1,\ldots,y_{i-1}$, we generate a random $\theta\sim\mathscr{U}([0,1])$ to obtain the interval

$$\begin{aligned}\underline{P}_{\boldsymbol{x}}(Y_{\{i\}}=1|y_1,\ldots,y_{i-1})&=\max(0,\theta-\varepsilon)\\ \overline{P}_{\boldsymbol{x}}(Y_{\{i\}}=1|y_1,\ldots,y_{i-1})&=\min(\theta+\varepsilon,1)\end{aligned}$$

where $\mathscr{U}([0,1])$ is a uniform distribution and ε is a parameter representing the imprecision level of our interval. The value of parameter ε impacts directly the width of the interval and therefore the precision of the obtained prediction.

We evaluate skeptic inferences on five different samples of 2000 binary trees, each sample having a fixed ε (i.e. 10^3 instances). For each instance, we evaluate the quality of the outer-approximation by computing the number of added elements in the corresponding set of binary vectors, i.e.,

$$d^{\varepsilon}_{(\hat{\mathfrak{y}},\hat{\mathbb{Y}})}=|\hat{\mathfrak{y}}_{\ell_H,\mathscr{P}}|-|\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}|. \tag{17}$$

As we have that $\hat{\mathfrak{y}}_{\ell_H,\mathscr{P}}\supseteq\hat{\mathbb{Y}}^M_{\ell_H,\mathscr{P}}$, Equation (17) will never be negative. Also, since different number of labels will induce different upper bounds for Equation (17), we uniformize the results across different numbers by partitioning the results in four bins:

$$q_0=\#\left\{(\hat{\mathfrak{y}},\hat{\mathbb{Y}})_i^{(2000)}\;\middle|\;d^{\varepsilon}_{(\hat{\mathfrak{y}},\hat{\mathbb{Y}})_i}=0\right\},$$

$$q_{\leq 0.25}=\#\left\{(\hat{\mathfrak{y}},\hat{\mathbb{Y}})_i^{(2000)}\;\middle|\;0<d^{\varepsilon}_{(\hat{\mathfrak{y}},\hat{\mathbb{Y}})_i}\leq 2^{|\Omega|}/4\right\},$$

$$q_{\leq 0.5}=\#\left\{(\hat{\mathfrak{y}},\hat{\mathbb{Y}})_i^{(2000)}\;\middle|\;2^{|\Omega|}/4<d^{\varepsilon}_{(\hat{\mathfrak{y}},\hat{\mathbb{Y}})_i}\leq 2^{|\Omega|}/2\right\},$$

$$q_{\leq 1}=\#\left\{(\hat{\mathfrak{y}},\hat{\mathbb{Y}})_i^{(2000)}\;\middle|\;2^{|\Omega|}/2<d^{\varepsilon}_{(\hat{\mathfrak{y}},\hat{\mathbb{Y}})_i}\leq 2^{\Omega}\right\}.$$

Finally, we perform the computer simulations on a discretization of the parameter $\varepsilon\in\{0.05,0.15,\ldots,0.45\}$. Thus, the results obtained, in percentage and with confidence interval (of the five repetitions), for each ε value and partitions q_* are shown in the Table 1. We omitted the results of $\varepsilon=0.45$ since it always yields $q_0=100\%$ for all labels.

The main findings of those simulations are as follows:

- globally, $\hat{\mathfrak{y}}_{\ell_H,\mathscr{P}}$ provides a quite accurate approximation of the true set, as it is exact (i.e., in q_0) most of the time;
- the quality of $\hat{\mathfrak{y}}_{\ell_H,\mathscr{P}}$ decreases as the number of labels increases, making it unfit for applications having a high number of labels [13];
- the quality of $\hat{\mathfrak{y}}_{\ell_H,\mathscr{P}}$ seems to be the worst for moderate imprecision, probably because a high imprecision will tend to provide more empty vectors as predictions;
- there are a few cases where $\hat{\mathfrak{y}}_{\ell_H,\mathscr{P}}$ provides bad (i.e., are in $q_{\leq 0.5}$) to really bad approximation (i.e., are in $q_{\leq 1}$). This indicates that having exact inference methods may be helpful to identify those cases.

We now perform other experimental studies on real data sets in order to check how skeptic inferences for multi-label problems behave in presence of noisy or missing labels.

4.3. Skeptic Inference with Binary Relevance

In this subsection, we perform a set of experiments to investigate the usefulness of using skeptic inferences in multi-label problems. In particular, we investigate what happens when some labels are noisy or missing. To that end, we use a set of standard real-word data sets from the MULAN repository[4] (c.f. Table 2), following a 10×10 cross-validation procedure to fit the model.

Evaluation As we perform set-valued predictions, usual measures used in multi-label problems cannot be adopted here. We thus consider it appropriate to use an incorrectness measure (IC), coupled with a completeness (CP) measure [8, §4.1], defined as follows

$$IC(\hat{\mathbb{Y}},\mathfrak{y})=\frac{1}{|Q|}\sum_{\hat{\mathfrak{y}}_i\in Q}\mathbb{1}_{(\hat{\mathfrak{y}}_i\neq y_i)}, \tag{18}$$

$$CP(\hat{\mathbb{Y}},\mathfrak{y})=\frac{|Q|}{m}, \tag{19}$$

3. In general, there is only an inequality between Equations (15) and (16)

4. http://mulan.sourceforge.net/datasets.html

#label	ε	$d^{\varepsilon}_{\hat{\mathfrak{y}},\hat{\mathfrak{Y}}}$			
		q_0	$q_{\leq 0.25}$	$q_{\leq 0.5}$	$q_{\leq 1}$
2	0.05	**100.0 ± 0.00%**	0.00 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%
	0.15	**98.93 ± 0.11%**	0.00 ± 0.00%	**1.07 ± 0.11%**	0.00 ± 0.00%
	0.25	**98.98 ± 0.18%**	0.00 ± 0.00%	**1.02 ± 0.18%**	0.00 ± 0.00%
	0.35	**100.0 ± 0.00%**	0.00 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%
4	0.05	**97.05 ± 0.25%**	**2.95 ± 0.25%**	0.00 ± 0.00%	0.00 ± 0.00%
	0.15	**95.85 ± 0.38%**	**2.97 ± 0.24%**	**1.17 ± 0.17%**	**0.01 ± 0.02%**
	0.25	**99.02 ± 0.17%**	**0.08 ± 0.05%**	**0.90 ± 0.18%**	0.00 ± 0.00%
	0.35	**100.0 ± 0.00%**	0.00 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%
6	0.05	**90.26 ± 0.44%**	**9.74 ± 0.44%**	0.00 ± 0.00%	0.00 ± 0.00%
	0.15	**91.44 ± 0.63%**	**4.75 ± 0.35%**	**2.79 ± 0.19%**	**1.02 ± 0.23%**
	0.25	**97.98 ± 0.18%**	**1.28 ± 0.06%**	**0.71 ± 0.12%**	**0.03 ± 0.02%**
	0.35	**100.0 ± 0.00%**	0.00 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%

(*a*)

#label	ε	$d^{\varepsilon}_{\hat{\mathfrak{y}},\hat{\mathfrak{Y}}}$			
		q_0	$q_{\leq 0.25}$	$q_{\leq 0.5}$	$q_{\leq 1}$
8	0.05	**78.61 ± 1.35%**	**19.9 ± 1.31%**	**1.49 ± 0.25%**	0.00 ± 0.00%
	0.15	**91.66 ± 0.33%**	**5.97 ± 0.31%**	**1.78 ± 0.15%**	**0.59 ± 0.14%**
	0.25	**97.70 ± 0.21%**	**1.66 ± 0.21%**	**0.64 ± 0.20%**	0.00 ± 0.00%
	0.35	**99.67 ± 0.04%**	0.00 ± 0.00%	0.33 ± 0.04%	0.00 ± 0.00%
10	0.05	**74.28 ± 0.92%**	**25.03 ± 0.96%**	**0.69 ± 0.07%**	0.00 ± 0.00%
	0.15	**93.43 ± 0.32%**	**4.44 ± 0.34%**	**1.38 ± 0.33%**	**0.75 ± 0.25%**
	0.25	**98.50 ± 0.15%**	0.00 ± 0.00%	**1.50 ± 0.15%**	0.00 ± 0.00%
	0.35	**100.0 ± 0.00%**	0.00 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%
11	0.05	**73.63 ± 0.60%**	**24.99 ± 0.66%**	**1.38 ± 0.13%**	0.00 ± 0.00%
	0.15	**93.72 ± 0.64%**	**4.20 ± 0.55%**	**2.08 ± 0.56%**	0.00 ± 0.00%
	0.25	**97.20 ± 0.20%**	**2.80 ± 0.20%**	0.00 ± 0.00%	0.00 ± 0.00%
	0.35	**100.0 ± 0.00%**	0.00 ± 0.00%	0.00 ± 0.00%	0.00 ± 0.00%

(*b*)

Table 1: Average partitions amounts q_* (%) with confidence interval.

Data set	#Features	#Labels	#Instances	#Cardinality	#Density
emotions	72	6	593	1.90	0.31
scene	294	6	2407	1.07	0.18
yeast	103	14	2417	4.23	0.30

Table 2: Multi-label data sets summary

where Q denotes the set of predicted label such that $\hat{\mathfrak{y}}_i = 1$ or $\hat{\mathfrak{y}}_i = 0$ (in other words any abstained predicted label $\hat{\mathfrak{y}}_i = *$ is not in Q). When predicting complete vectors, then $CP = 1$ and IC equals the Hamming loss (i.e. Equation (5)), and when predicting the empty vector, i.e. all labels equals to $\hat{\mathfrak{y}}_i = *$, then $CP = 0$ and by convention $IC = 0$. Since those measures are adapted to partial vectors, we will use a simple binary relevance strategy in the experiments.

Naive Credal classifier To obtain probability intervals over each label, we use an imprecise classifier called the naïve credal classifier (NCC)[5] [25], which extends the classical naive Bayes classifier (NBC). We refer to Zaffalon [25] for details, and will only recall here that the imprecision of this classifier is regulated by a value $s \in \mathbb{R}$, with the imprecision being higher as s increases (for $s = 0$, we retrieve basic empirical frequencies estimate).

In this paper, we restrict the values of the hyper-parameter of the imprecision to $s \in \{0, 0.5, 1.5, 2.5, 3.5, 4.5\}$. Our purpose here is not to find the "optimal" value of s, but to show the effectiveness of injecting imprecision (i.e. to provide robust and skeptical inferences). As the NCC requires discrete features, when those were continuous we simply discretized in z equal-width intervals, with two levels of discretization $z=5$ and $z=6$.

Missing labels To simulate missingness, we uniformly pick at random a percentage of missing labels, with five different levels of missingness: $\{0, 20, 40, 60, 80\}$. Missing values are removed from the training data. Table 3 illustrates a data set data with missing values.

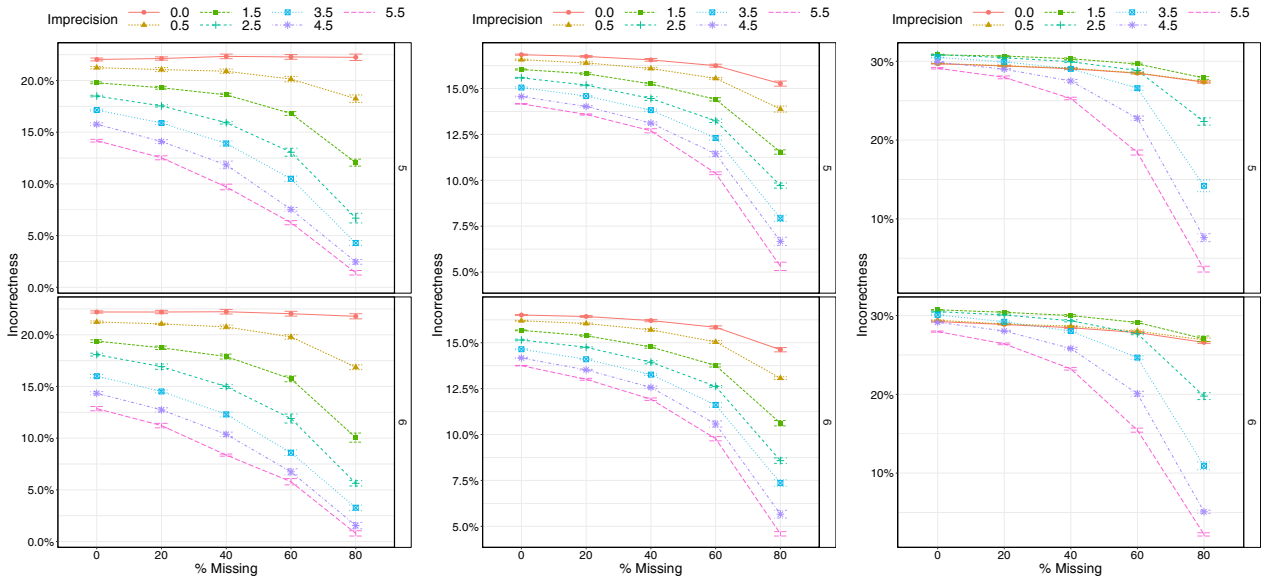
Features					**Missing**		
X_1	X_2	X_3	X_4	X_5	Y_1	Y_2	Y_3
107.1	25	Blue	60	1	1	0	*
-50	10	Red	40	0	1	*	1
200.6	30	Blue	58	1	*	1	0
107.1	5	Green	33	0	*	1	0
...	...	...	...	...	...	...	...

Table 3: Missing labels illustration

In Figures 4 and 5, we provide the results of the incorrectness and incompleteness measures obtained by fitting the NCC model on different percentages of missing labels and data sets of Table 2. While it may be surprising to see that the precise model is not really affected by randomly missing labels, the figures show that IP models behave as expected: as more labels are missing, our model becomes more cautious but also more accurate on those prediction is still makes. Moreover, for moderate values of missingness (20 or 40%) and moderate imprecision ($s = 2.5$ or below), completeness remain reasonable and above 50%, with important variations accross data sets that we will investigate. Of those, one quite noticeable result is that for the Emotions data set, even with 80% of missing label, a light imprecision ($s = 0.5$) allows us to reach a reasonable completeness of about 80% with a gain of 5% in terms of correct predictions.

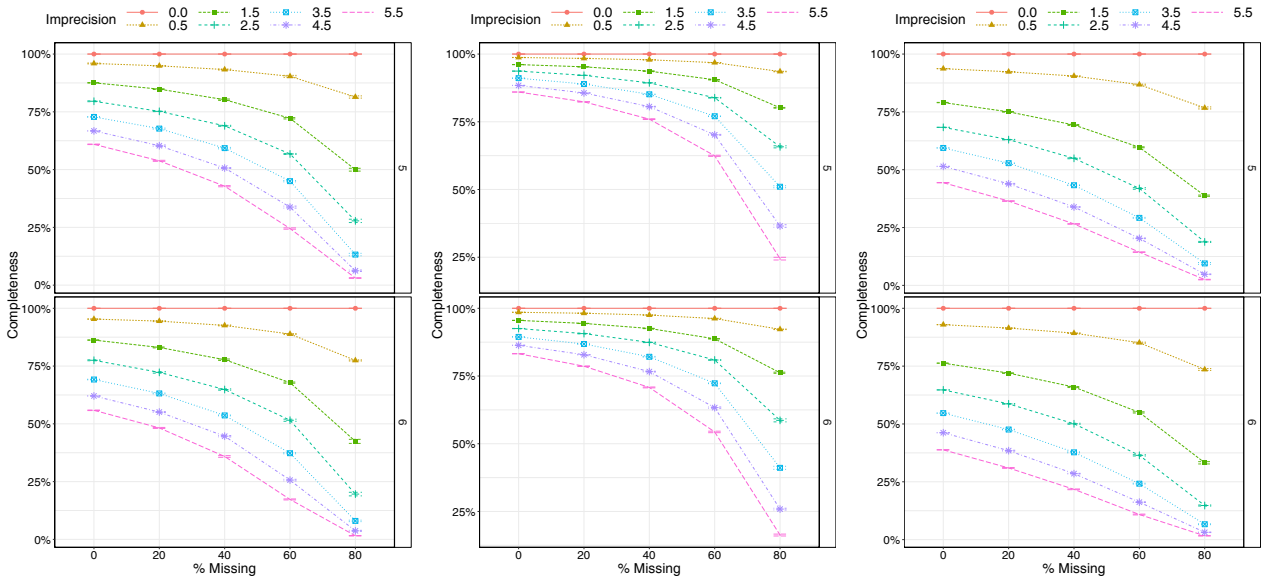
Results obtained are sufficient to show that skeptic inferences with probability sets may provide additional benefits when dealing with missing labels. Those results could, of course, be improved by picking other classifiers, such as the NCC2 [6], an extension of the NCC tailored for missing values.

5. Bearing in mind that it can be replaced by any other (credal) imprecise classifiers, see [2, §10].



(*c*) EMOTIONS (*d*) SCENE (*e*) YEAST

Figure 4: **Incorrectness** evolution for each level of imprecision (one curve each) and discretization $z=5$ .pdf(top) and $z=6$ (down), with respect the percentage of missing labels.



(*a*) EMOTIONS (*b*) SCENE (*c*) YEAST

Figure 5: **Incompleteness** evolution for each level of imprecision (one curve each) and discretization $z=5$ (top) and $z=6$ (down), with respect the percentage of missing labels.

5. Conclusion and Discussion

In this paper, we investigated the problem of providing cautious, skeptical multi-label inferences when considering the well-known Hamming loss and generic probabiltiy sets. We provided efficient algorithmic procedure to do so in the general case, and showed that in the Binary relevance scheme, those same predictions were reduced to partial vectors computable from marginal probability bounds over the labels.

Experiments on simulated data sets show that this last solution, when used as an outer-approximation in the general

case, degrades in quality as the number of labels increases and the level of imprecision is mild. On the other hand, experiments on various real data sets show that making skeptical inferences generally provide quite satisfactory results when considering missing labels.

In future works, it would be interesting to compare our skeptical inference approach against those rejecting and abstaining approaches, for instance the recently proposed abstention approach in [23]. Such comparisons would nevertheless require a deep analysis of the models, decision rules as well as instances on which each approach abstains, and is out of the scope of the present paper, whose main focus was how to derive cautious predictions over binary vectors when considering probability sets as our uncertainty model.

Another natural next step will be to solve the maximality criterion using other loss functions commonly used in multi-label problems, e.g. ranking loss, Jaccard loss, F-measure, and so on. As noticed in Remark 6, such problems are likely to be much more intricate when considering sets of probabilities. Finally, let us notice that while this paper focused on the issue of multi-label learning problems, our results readily apply to any Boolean vectors of m items. As Boolean vectors and structures as well as probability bounds naturally appear in a number of other applications, including occupancy grids [17] or data bases [10], a future work would be to investigate how our present findings can help in such problems.

References

[1] Alessandro Antonucci and Giorgio Corani. The multilabel naive credal classifier. *International Journal of Approximate Reasoning*, 83:320–336, 2017.

[2] Thomas Augustin, Frank PA Coolen, Gert de Cooman, and Matthias CM Troffaes. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.

[3] Matthew R Boutell, Jiebo Luo, Xipeng Shen, and Christopher M Brown. Learning multi-label scene classification. *Pattern recognition*, 37(9):1757–1771, 2004.

[4] Ruidi Chen and Ioannis Ch Paschalidis. A robust learning approach for regression models based on distributionally robust optimization. *The Journal of Machine Learning Research*, 19(1):517–564, 2018.

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

[6] Giorgio Corani and Marco Zaffalon. Learning reliable classifiers from small or incomplete data sets: the naive credal classifier 2. *Journal of Machine Learning Research*, 9(Apr):581–621, 2008.

[7] Krzysztof Dembczyński, Willem Waegeman, Weiwei Cheng, and Eyke Hüllermeier. On label dependence and loss minimization in multi-label classification. *Machine Learning*, 88(1-2):5–45, 2012.

[8] Sebastien 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.

[9] Johannes Fürnkranz, Eyke Hüllermeier, Eneldo Loza Mencía, and Klaus Brinker. Multilabel classification via calibrated label ranking. *Machine Learning*, 73 (2):133–153, 2008.

[10] Wolfgang Gatterbauer and Dan Suciu. Oblivious bounds on the probability of boolean functions. *ACM Transactions on Database Systems (TODS)*, 39(1): 1–34, 2014.

[11] Filip Hermans, Erik Quaeghebeur, et al. Imprecise markov chains and their limit behavior. *Probability in the Engineering and Informational Sciences*, 23(4): 597–635, 2009.

[12] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037, 2018.

[13] Himanshu Jain, Yashoteja Prabhu, and Manik Varma. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 935–944, 2016.

[14] Wojciech Kotlowski and Krzysztof Dembczyński. Surrogate regret bounds for generalized classification performance metrics. In *Asian Conference on Machine Learning*, pages 301–316, 2016.

[15] Oluwasanmi O Koyejo, Nagarajan Natarajan, Pradeep K Ravikumar, and Inderjit S Dhillon. Consistent multilabel classification. In *Advances in Neural Information Processing Systems*, pages 3321–3329, 2015.

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

[17] Hafida Mouhagir, Véronique Cherfaoui, Reine Talj, François Aioun, and Franck Guillemard. Using evidential occupancy grid for vehicle trajectory planning

under uncertainty with tentacles. In *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*, pages 1–7. IEEE, 2017.

[18] Vu-Linh Nguyen and Eyke Hüllermeier. Reliable multi-label classification: Prediction with partial abstention. *arXiv preprint arXiv:1904.09235*, 2019.

[19] Ignazio Pillai, Giorgio Fumera, and Fabio Roli. Multi-label classification with a reject option. *Pattern Recognition*, 46(8):2256–2266, 2013.

[20] Julia Plass, Marco EGV Cattaneo, Thomas Augustin, Georg Schollmeyer, and Christian Heumann. Reliable inference in categorical regression analysis for non-randomly coarsened observations. *International Statistical Review*, 87(3):580–603, 2019.

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

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

[23] Eyke Hüllermeier Vu-Linh Nguyen. Reliable multil-abel classification: Prediction with partial abstention. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*, 2019.

[24] Gen Yang, Sébastien Destercke, and Marie-Hélène Masson. Nested dichotomies with probability sets for multi-class classification. In *ECAI*, pages 363–368, 2014.

[25] Marco Zaffalon. The naive credal classifier. *Journal of statistical planning and inference*, 105(1):5–21, 2002.