A first glance at multi-label chaining using imprecise probabilities

Yonatan-Carlos Carranza-Alarcon and Sébastien Destercke

HEUDIASYC - UMR CNRS 7253, Université de Technologie de Compiègne
57 avenue de Landshut, 60203 COMPIEGNE CEDEX - FRANCE
{yonatan-carlos.carranza-alarcon, sebastien.destercke}@hds.utc.fr
https://www.hds.utc.fr/

Abstract. In this paper, we present two different ways to extend the classical multi-label chaining approach to handle imprecise probability estimates. These estimates use convex sets of distributions (or credal sets) in order to describe our uncertainty rather than a precise one. The main reasons one could have for using such estimates are (1) to make cautious predictions (or no decision at all) when a high uncertainty is detected in the chaining and (2) to make better precise predictions by avoiding biases caused in early decisions in the chaining. We perform experiments on missing and noisy labels to investigate how accurate and how precise these predictions are in both approaches. Our experimental results indicate that while our approach produce relevant cautiousness (i.e., forget predictions likely to be erroneous), results regarding possible bias correction using a minimax approach are less encouraging, except when high adversarial noise affect the labels, in which case our approach outperform its precise counterpart.

Keywords: imprecise probabilities · multi-label · classifier chains

1 Introduction

Multi-label classification (MLC) is a generalization of traditional classification (with a single label), as well as a special case of the multi-task learning. This approach is increasingly required in different research fields, such as the classification of proteins in bioinformatics [15], text classification in information retrieval [9], object recognition in computer vision [3], and so on.

A classical issue in multi-label learning techniques is how to integrate the possible dependencies between labels while keeping the inference task tractable. Indeed, while decomposition techniques [15,9] such as Binary relevance or Calibrated ranking allow to speed up both the learning and inference tasks, they roughly ignore the label dependencies, while using a fully specified model such as probabilistic chains require, at worst, to scan all possible predictions (that grow exponentially in the number of labels). A popular technique to solve this issue, at least for the inference task, is to use a chain model [13]: this consists in using, incrementally, the predictions made on previous labels to help better predict the relevance of a current label.

To the best of our knowledge, there are only a few works of multi-label classification producing cautious predictions, such as the reject option [12], partial predictions [7,1] or abstaining labels [11]. And none of these have studied this issues in the chain model (or classifier-chains approach).

In this paper, we consider the problem of extending such an approach to the imprecise probabilistic case, and propose two different ways to extend it, based on the fact that some labels are too uncertain to be used in the chaining. The first treats the uncertain labels in a robust way, exploring all possibles path in order not to propagate early uncertain decisions, whereas the latter marginalizes the probabilistic model over the uncertain labels, in other words, the uncertain labels are not considered to infer the current label.

Section 2 introduces the notations which we will use for the multi-label setting, and give the necessary reminders about making inferences with convex sets of probabilities. In Section 3, we remind the classical classifier-chains approach and then we present our extended approaches based on imprecise probabilities.

Finally, in Section 4, we perform a set of experiments on real data sets, which are perturbed with missing and noisy labels, in order to investigate how precise (when we exchange abstained labels for precise ones) and how cautious (when we abstain on labels difficult to predict) is our approach.

2 Preliminares and basic remainders

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

2.1 Multi-label problem setting

In multi-label problem, an instance $\mathbf{x}$ of an input space $\mathscr{X} = \mathbb{R}^p$ is no longer associated with a single label m_k of an output space $\mathcal{K} = \{m_1, \ldots, m_m\}$, as in the traditional classification problem, but with a subset of labels $\Lambda_x \subseteq \mathcal{K}$ often called the set of relevant labels while its complement $\mathcal{K} \backslash \Lambda_x$ is considered as irrelevant for $\mathbf{x}$. Let $\mathscr{Y} = \{0,1\}^m$ be a m-dimensional binary space and $\boldsymbol{y} = (y_1, \ldots, y_m) \in \mathcal{Y}$ be any element of $\mathcal{Y}$ such that $y_i = 1$ if and only if $m_i \in \Lambda_x$.

From a decision theoretic approach (DTA), the goal of the multi-label problem is the same as the usual classification problem. That means, given a probability distribution $\hat{\mathbb{P}}$ fitting a finite set of i.i.d. observations $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i) | i = 1, \ldots, N\}$ issued from a (true) theoretical probability distribution $\mathbb{P} : \mathscr{X} \times \mathscr{Y} \to [0,1]$, DTA aims to minimize the risk of getting missclassification with respect to a specified loss function $\mathcal{L}(\cdot,\cdot)$:

$$\mathcal{R}_{\mathcal{L}}(Y, h(X)) = \arg\min_{\boldsymbol{h}} \mathbb{E}_{\hat{\mathbb{P}}} \left[\mathcal{L}(Y, \boldsymbol{h}(X)) \right]. \tag{1}$$

If $\mathcal{L}(\cdot,\cdot)$ is defined instance-wise, this minimization can also be expressed as the minimization of conditional expected risk of a given unlabeled instance $\boldsymbol{x}$ (cf. [6,

eq. 3] and [8, eq. 2.21])

$$\boldsymbol{h}(\boldsymbol{x}) = \arg\min_{\mathbf{y}\in\mathscr{Y}} \mathbb{E}_{\hat{\mathbb{P}}_{Y|\boldsymbol{x}}}[\mathcal{L}(Y,\mathbf{y})] = \arg\min_{\mathbf{y}\in\mathscr{Y}} \sum_{\mathbf{y}'\in\mathscr{Y}} \hat{P}(Y=\mathbf{y}'|X=\boldsymbol{x})\mathcal{L}(\mathbf{y}',\mathbf{y}) \quad (2)$$

or, equivalently, by picking the maximal element obtained from a strict total order relation[1] $\succ$ over $\mathscr{Y}\times\mathscr{Y}$, where $\mathbf{y}^1 \succ \mathbf{y}^2$ ($\mathbf{y}^1$ is preferred to $\mathbf{y}^2$, or $\mathbf{y}^2$ is dominated by $\mathbf{y}^1$) iff

$$\mathbb{E}_{\hat{\mathbb{P}}}\left(\mathcal{L}(\mathbf{y}^2,\cdot) - \mathcal{L}(\mathbf{y}^1,\cdot)\right) = \mathbb{E}_{\hat{\mathbb{P}}}\left(\mathcal{L}(\mathbf{y}^2,\cdot)\right) - \mathbb{E}_{\hat{\mathbb{P}}}\left(\mathcal{L}(\mathbf{y}^1,\cdot)\right) \geq 0. \quad (3)$$

This amounts to saying that exchanging $\mathbf{y}^2$ for $\mathbf{y}^1$ would incur a positive expected loss (which is not desirable).

In this paper, we are interested in making set-valued predictions when uncertainty is too high (e.g. due to insufficient evidence to include or discard a relevant label, see Example 1). In our case, the set-valued prediction will be described as a partial binary vector $\mathbf{y}^* \in \mathscr{Y}^*$ where $\mathscr{Y}^* = \{0,1,*\}^m$ with $*$ standing for abstention. For instance, a partial prediction $\mathbf{y}^* = (*,1,0)$ correspond to two plausible binary vector solutions $\{(0,1,0),(1,1,0)\} \subseteq \mathscr{Y}$.

In the sequel, we will denote by $\mathscr{I}$ subset of label indices (and by $[\![j]\!] = \{1,\ldots,j\}$ set of the first j integers). Given a prediction made in the j first labels, we will denote by

1. (relevant labels) $\mathscr{I}^j_{\mathcal{R}} \subseteq [\![j]\!]$ the indices of the labels predicted as relevant among the j first, i.e. $\forall i \in \mathscr{I}^j_{\mathcal{R}}\ y_i = 1$,
2. (irrelevant labels) $\mathscr{I}^j_{\mathcal{I}} \subseteq [\![j]\!], \mathscr{I}^j_{\mathcal{I}} \cap \mathscr{I}^j_{\mathcal{R}} = \emptyset$ the indices of the labels predicted as irrelevant among the j first, i.e. $\forall i \in \mathscr{I}^j_{\mathcal{I}}\ y_i = 0$, and
3. (abstained labels) $\mathscr{I}^j_{\mathcal{A}} = [\![j]\!] \backslash (\mathscr{I}^j_{\mathcal{R}} \cup \mathscr{I}^j_{\mathcal{I}})$ the indices of the labels on which we abstained among the j first, i.e. $\forall i \in \mathscr{I}^j_{\mathcal{A}}\ y_i = \{0,1\} := *$.

Besides, for the sake of simplicity and when it is not ambiguous, we will henceforth denote probabilities conditioned on previous labels by

$$P^j_{\boldsymbol{x}}(Y_j=1) := P_{\boldsymbol{x}}(Y_j=1|Y_{\mathscr{I}^{j-1}} = \hat{\mathbf{y}}_{\mathscr{I}^{j-1}}), \quad (4)$$

where $\hat{\mathbf{y}}_{\mathscr{I}^{j-1}}$ is a $(j-1)$-dimensional vector that contains the previously inferred precise and/or abstained values of labels having indices $\mathscr{I}^{j-1}$.

Example 1. We consider an output space of two labels $\mathscr{K} = \{m_1, m_2\}$, a single binary feature x_1 and the table 1 with imprecise estimations of the joint distribution $\mathbf{P}(X_1, Y_1, Y_2)$.

Based on the probabilities of Table 1, we have that $\hat{P}_0(y_1=0) = (y_1=0|x_1=0) = 1$ and that $\hat{P}_0(y_2=0) \in [0.4, 0.7]$, therefore not knowing whether $\hat{P}_0(y_2=0) > 0.5$. This leads to propose as a prediction $\hat{\mathbf{y}^*} = (0,*)$. On the contrary, the imprecision on the right hand-side is such that $\hat{P}_1(y_2=0) \in [0.6, 0.8]$, leading to the precise prediction $\hat{\mathbf{y}^*} = (1,0)$.

[1] A complete, transitive, and asymmetric binary relation

Table 1. Estimated joint probability distribution

x_1	y_1	y_2	$\hat{\mathbf{P}}$	x_1	y_1	y_2	$\hat{\mathbf{P}}$
0	0	0	[0.4,0.7]	1	0	0	0.00
0	0	1	[0.3,0.6]	1	0	1	0.00
0	1	0	0.00	1	1	0	[0.6,0.8]
0	1	1	0.00	1	1	1	[0.2,0.4]

Handling partial predictions requires a well founded strategy to do so, that can also deal with the increased complexity of the prediction space $|\mathscr{Y}^*| = 3^m$. In this paper, we will describe it by means of a set of probabilities $\mathscr{P}$ instead of a single probability distribution $\mathbb{P}$, as usually done. To this end, in what follows, we will introduce basic concepts about imprecise probabilities and decision making with it (for further details [2]).

2.2 Notions about imprecise probabilities

Imprecise probabilities consist in representing our uncertainty by a convex set of probability distributions $\mathscr{P}_X$ [16,2] (i.e. a *credal set* [10]), defined over a space $\mathscr{X}$ rather than by a precise probability measure $\mathbb{P}_X$ [14].

Given such a set of distributions $\mathscr{P}_X$ and any measurable event $A \subseteq \mathscr{X}$, we can define the notions of lower and upper probabilities as:

$$\underline{P}_X(A) = \inf_{P \in \mathscr{P}_X} P(A) \qquad \text{and} \qquad \overline{P}_X(A) = \sup_{P \in \mathscr{P}_X} P(A) \tag{5}$$

where $\underline{P}_X(A) = \overline{P}_X(A)$ only when we have sufficient information about event A. The lower probability is dual to the upper [2], in the sense that $\underline{P}_X(A) = 1 - \overline{P}_X(A^c)$ where A^c is the complement of A. Many authors [16,18] have argued that when information is lacking or imprecise, considering credal sets as our model of information better describes our actual uncertainty.

With such an approach, (1) the parametric or non-parametric estimation usually becomes more complicated computationally since we estimate a set of distributions $\mathscr{P}$, and (2) the classical decision-making framework presented in the Equation (2) needs to be extended. For the former issue, we will use a well-known classifier that extends the NBC and computes the lower and upper bound in polynomial time (see Section 4). For the latter issue, we will adapt the following binary relevance approach, described in [7, Prop. 1]:

$$\hat{y}_i = \begin{cases} 1 & \text{if } \underline{P}_{\boldsymbol{x}}(Y_j = 1) > 0.5, \\ 0 & \text{if } \overline{P}_{\boldsymbol{x}}(Y_j = 1) < 0.5, \\ * & \text{if } 0.5 \in \left[\underline{P}_{\boldsymbol{x}}(Y_j = 1), \overline{P}_{\boldsymbol{x}}(Y_j = 1)\right], \end{cases} \tag{6}$$

to the case of mutli-label chaining.

3 Multilabel chaining with imprecise probabilities

We first recall the classical precise chaining and then propose two different strategies to extend chaining to the imprecise probabilistic case.

3.1 Precise probabilistic chaining

Classifier chains is a well-known approach exploiting dependencies among labels by fitting at each step of the chain (see Figure 1) a new classifier model $h_j : \mathscr{X} \times \{0,1\}^{j-1} \to \{0,1\}$ predicting the relevance of the jth label. This classifier combines the original input space attribute and all previous predictions in the chain in order to create a new input space $\mathscr{X}_j^* = \mathscr{X} \times \{0,1\}^{j-1}, k \in \mathbb{N}^{>0}$. In brief, we consider a chain $\boldsymbol{h} = (h_1, \ldots, h_m)$ of binary classifiers resulting in the full prediction $\hat{\boldsymbol{y}}$ obtained by solving each single classifier as follows

$$\hat{y} := h_j(x) = \underset{y \in \{0,1\}}{\arg\max}\ P^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j = y). \tag{7}$$

The classical multi-label chaining then works as follows:

1. Random order of labels.-. We randomly pick an order between labels $\mathscr{I}^*$ (possibly different from the original indices $\mathscr{I} = [\![m]\!]$) and assume that the index are relabelled in an increasing order.
2. Prediction j^{th} label.- For a given label y_j, let us assume that we have previously predicted labels of lower index $y_1, \ldots, y_{j-1}$ and let $\mathscr{I}_{\mathcal{R}}^{j-1}, \mathscr{I}_{\mathcal{I}}^{j-1} \subseteq [\![j-1]\!]$ be set of indices of relevant and irrelevant labels, such that $\mathscr{I}_{\mathcal{R}}^{j-1} \cap \mathscr{I}_{\mathcal{I}}^{j-1} = \emptyset$. Then, the prediction of $\hat{y}_j$ (or $h_j(\boldsymbol{x})$) for a new instance $\boldsymbol{x}$ is

$$\hat{y}_j = \begin{cases} 1 & \text{if } P_{\boldsymbol{x}}(Y_j = 1 | Y_{\mathscr{I}_{\mathcal{R}}^{j-1}} = 1, Y_{\mathscr{I}_{\mathcal{I}}^{j-1}} = 0) \geq 0.5 \\ 0 & \text{if } P_{\boldsymbol{x}}(Y_j = 0 | Y_{\mathscr{I}_{\mathcal{R}}^{j-1}} = 1, Y_{\mathscr{I}_{\mathcal{I}}^{j-1}} = 0) < 0.5 \end{cases} \tag{8}$$

Figure 1 summarizes the procedure presented above, as well as the obtained predictions for a specific case (in bold red predicted labels and probabilities).

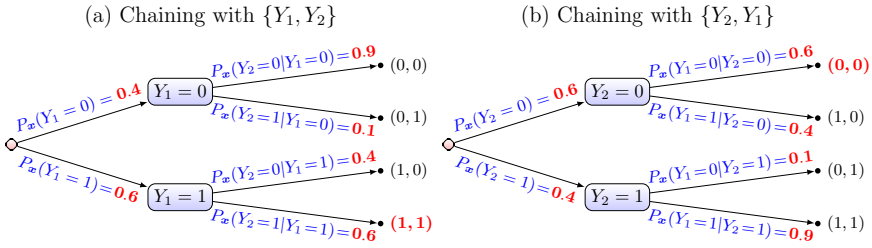


Fig. 1. Precise chaining

From the figure, it is clear that the ordering and the fact of choosing a single branch at each step can have a significant impact on the final predictions, as in our example it shifts from one prediction to its opposite. Intuitively, adding some robustness and cautiousness in the process could halpe to avoid unwarranted biases.

In what follows, we propose two different extensions of precise chaining based on imprecise probability estimates. By this, we mean that it is be based on binary cautious classifiers, which consider a new output space $\mathscr{Y}=\{0,1,*\}^m$ from which to pick the predictions.

3.2 Imprecise probabilistic chaining

When considering imprecise probabilities, the estimates $P_{\boldsymbol{x}}^{\boldsymbol{j}}(Y_j=1)$ become imprecise, that is, we now have $[\hat{P}_{\boldsymbol{x}}^{\boldsymbol{j}}](Y_j=y_j) := [\underline{P}_{\boldsymbol{x}}^{\boldsymbol{j}}(Y_j=y_j),\ \overline{P}_{\boldsymbol{x}}^{\boldsymbol{j}}(Y_j=y_j)]$. The basic idea of using such estimates is that in the chaining, we should be cautious when the classifier is unsure about which is the most probable prediction. In this section, we describe two different strategies (or extensions) in a general way, and we will apply them to the naive credal classifier (an extension of the Naive Bayes classifier) in the next section.

Let us first formulate the generic procedure to calculate the probability bound of j^{th} label,

1. Random order of labels.- As before (in precise version), randomly pick an order between labels, assuming again that index are relabelled in increasing order.
2. Prediction j^{th} label.- For a given label y_j, let us assume we have made possibly imprecise predictions for $y_1,\ldots,y_{j-1}$ such that $\mathscr{I}_{\mathcal{A}}^{j-1}$ contains the set of indices of labels on which we abstained $\{*\}$, and hence, $\mathscr{I}_{\mathcal{R}}^{j-1}$ and $\mathscr{I}_{\mathcal{I}}^{j-1}$ are the set of indices of relevant and irrelevant labels, such that $\mathscr{I}_{\mathcal{A}}^{j-1}\cup\mathscr{I}_{\mathcal{R}}^{j-1}\cup\mathscr{I}_{\mathcal{I}}^{j-1}=\mathscr{I}^{j-1}$. Then, we calculate $[P_{\boldsymbol{x}}^{\boldsymbol{j}}](Y_j=1)$ (we will show after the possible ways to obtain this interval) in order to predict the label $\hat{y}_j$ as

$$\hat{y}_j=\begin{cases}1 & \text{if } \underline{P}_{\boldsymbol{x}}^{\boldsymbol{j}}(Y_j=1)>0.5,\\ 0 & \text{if } \overline{P}_{\boldsymbol{x}}^{\boldsymbol{j}}(Y_j=1)<0.5,\\ * & \text{if } 0.5\in\left[\underline{P}_{\boldsymbol{x}}^{\boldsymbol{j}}(Y_j=1),\overline{P}_{\boldsymbol{x}}^{\boldsymbol{j}}(Y_j=1)\right],\end{cases}, \tag{9}$$

where this last equation is a slight variation of Equation (6).

We then propose two different extensions of how to calculate $[P_{\boldsymbol{x}}^{\boldsymbol{j}}](Y_j=1)$ at each inference step of the imprecise chaining.

Imprecise branching The first strategy treats unsure predictions in a robust way, considering all possible branching in the chaining as soon as there is an abstained label. Thus, the estimation of $[\underline{P}_{\boldsymbol{x}}^{\boldsymbol{j}}(Y_j=1),\overline{P}_{\boldsymbol{x}}^{\boldsymbol{j}}(Y_j=1)]$ (for $Y_j=0$, it directly obtains as $\underline{P}_{\boldsymbol{x}}^{\boldsymbol{j}}(Y_j=1)=1-\overline{P}_{\boldsymbol{x}}^{\boldsymbol{j}}(Y_j=0)$, and similarly for the upper bound) comes down to compute

$$\begin{aligned}\underline{P}_{\boldsymbol{x}}^{\boldsymbol{j}}(Y_j=1) &= \min_{\mathbf{y}\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}} \underline{P}_{\boldsymbol{x}}(Y_j=1|Y_{\mathscr{I}_{\mathcal{R}}^{j-1}}=1,Y_{\mathscr{I}_{\mathcal{I}}^{j-1}}=0,Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=\mathbf{y}),\\ \overline{P}_{\boldsymbol{x}}^{\boldsymbol{j}}(Y_j=1) &= \max_{\mathbf{y}\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}|}} \overline{P}_{\boldsymbol{x}}(Y_j=1|Y_{\mathscr{I}_{\mathcal{R}}^{j-1}}=1,Y_{\mathscr{I}_{\mathcal{I}}^{j-1}}=0,Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=\mathbf{y}).\end{aligned} \tag{IB}$$

That is to consider every possible replacements of variables for which we have abstained so far. This corresponds to a very robust version of the chaining, where every possible path is explored. It will therefore propagate imprecision along the tree, and may produce quite imprecise evaluations, especially if we abstain on the first labels.

Illustrations providing some intuition about this strategy can be seen in Figure 2b where we have abstained on labels (Y_2, Y_4) and we want to compute lower and upper probability bounds of the label $Y_5 = 1$.

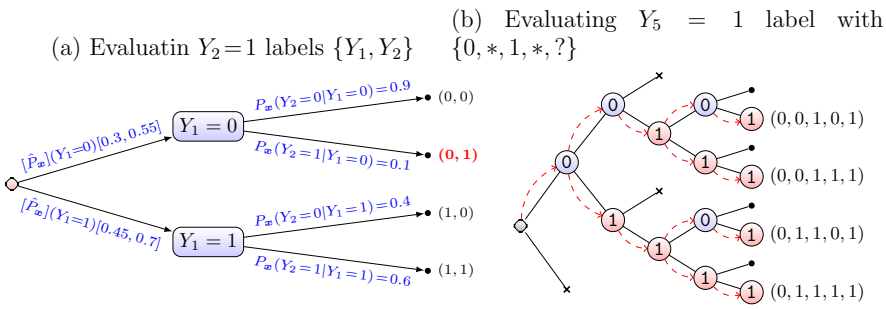


Fig. 2. Imprecise branching strategy

In the Figure 2a, we will consider the previous example (see Figure 1) in order to study in details how we should calculate probability bounds $[\underline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j = 1), \overline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j = 1)]$. For the sake of simplicity, we assume that probabilities about Y_2 are precise and that probability bounds of $Y_1 = 1$ is $[\hat{P}^{\boldsymbol{j}}_{\boldsymbol{x}}](Y_1 = 1) \in [0.45, 0.70]$. This last result would correspond to the following tree where we would consider the first two branches as possibles paths hence

$$\underline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_2 = 1) = \min_{y_1 \in \{0,1\}} P_{\boldsymbol{x}}(Y_2 = 1 | Y_1 = y_1) = \min(0.1, 0.6) = 0.1, \tag{10}$$

$$\overline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_2 = 1) = \max_{y_1 \in \{0,1\}} P_{\boldsymbol{x}}(Y_2 = 1 | Y_1 = y_1) = \max(0.1, 0.6) = 0.6, \tag{11}$$

which means that in this case we would abstain on both labels.

Marginalization The second strategy simply ignores unsure predictions in the chaining. Its interest is that it will not propagate imprecision in the tree. Thus, we begin by presenting the general formulation (which will after lead to the formulation without unsureness) which takes into account unsure predicted labels conditionally, so the estimation of probability bounds $[\underline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j = 1), \overline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j = 1)]$

comes down to compute

$$\begin{aligned}\underline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j=1)&=\underline{P}_{\boldsymbol{x}}(Y_j=1|Y_{\mathscr{I}^{j-1}_{\mathcal{R}}}=1,Y_{\mathscr{I}^{j-1}_{\mathcal{I}}}=0,Y_{\mathscr{I}^{j-1}_{\mathcal{A}}}=\{0,1\}^{|\mathscr{I}^{j-1}_{\mathcal{A}}|}),\\ \overline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j=1)&=\overline{P}_{\boldsymbol{x}}(Y_j=1|Y_{\mathscr{I}^{j-1}_{\mathcal{R}}}=1,Y_{\mathscr{I}^{j-1}_{\mathcal{I}}}=0,Y_{\mathscr{I}^{j-1}_{\mathcal{A}}}=\{0,1\}^{|\mathscr{I}^{j-1}_{\mathcal{A}}|}),\end{aligned} \tag{MAR}$$

Which comes down to calculate

$$\begin{aligned}\underline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j=1)&=\min_{P\in\mathscr{P}^*}P_{\boldsymbol{x}}(Y_j=1|Y_{\mathscr{I}^{j-1}_{\mathcal{R}}}=1,Y_{\mathscr{I}^{j-1}_{\mathcal{I}}}=0),\\ \overline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j=1)&=\max_{P\in\mathscr{P}^*}P_{\boldsymbol{x}}(Y_j=1|Y_{\mathscr{I}^{j-1}_{\mathcal{R}}}=1,Y_{\mathscr{I}^{j-1}_{\mathcal{I}}}=0).\end{aligned} \tag{MAR*}$$

where $\mathscr{P}^*$ is simply the set of joint probability distributions described by the imprecise probabilistic tree (we refer to de Cooman and Herman [5] for a detailed analysis of those). In general, such an optimisation can be computationally quite intensive, but remains easy in the case of the Naive credal classifier, thanks to its independence assumption (a full detail of those computation is not given here, due to space constraints).

Note that, once any of the two strategies has been applied, we can either keep the prediction as it is, producing an incomplete vector where label $Y_{\mathscr{I}_{\mathcal{A}}}$ become imprecise, or we can consider precise estimations of labels $j \in \mathscr{I}_{\mathcal{A}}$ by considering a minimax robust strategy, i.e., picking $\hat{y}_j = \arg\max_{y\in\{0,1\}} \underline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j = y)$ to replace the label Y_j by the corresponding prediction.

4 Experiments

In this section, we perform experiments on 3 data sets issued from the MULAN repository[2] (c.f. Table 2), following a 10×10 cross-validation procedure (at every j^{th}-fold, we proceed randomly to shuffle the set of labels).

Table 2. Multi-label data sets summary

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

4.1 Evaluation and setting

The usual metrics used in multi-label problems are not adapted at all when we infer set-valued predictions. Thus, we consider appropriate to use the incorrectness (IC) and completeness (CP) metrics proposed in [7, §4.1], as follows

$$IC(\hat{\boldsymbol{y}},\boldsymbol{y}) = \frac{1}{|Q|}\sum_{\hat{y}_i\in Q}\mathbb{1}_{(\hat{y}_i\neq y_i)} \quad \text{and} \quad CP(\hat{\boldsymbol{y}},\boldsymbol{y}) = \frac{|Q|}{m},$$

[2] http://mulan.sourceforge.net/datasets.html

where $\hat{\boldsymbol{y}}$ is the partial binary prediction and Q denote the set of non-abstained labels. When predicting complete vectors, then CP$=1$ and IC equals the Hamming loss and when predicting the empty vector, i.e. all labels $\hat{y}_i = *$, then CP$=0$ and by convention IC$=0$.

Imprecise classifier As was mentioned earlier, and for practical purposes, we chose to use the so-called *naïve credal classifier* (NCC)[18][3] in order to compute the probability bounds. NCC is an extension of the classical naive Bayes classifier (NBC) on a set of probability distributions. That means, NCC preserves the assumption of feature independence given the class of NBC, and relies on the Imprecise Dirichlet model (IDM) [17] to estimate class-conditional probabilities, whose imprecision level is controlled through a hyper-parameter $s \in \mathbb{R}$. The higher s, the wider the intervals $[\underline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j = 1), \overline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j = 1)]$ will be. For $s = 0$, we retrieve the classical NBC with precise predictions, and for high enough values of $s >>> 0$, the NCC model will make vacuous predictions. Thus, we restrict the values of the hyper-parameter of the imprecision to $s \in \{2.0, 2.5, 3.0, 3.5, 4.0, 4.5\}$ (starting from $s = 2$ as advised in [17]).

Missing and Noise labels In this paper, we consider three different settings in order to compare the quality of performance of both approaches. In all settings, we apply to each label $Y_{j,i}$ (the jth label of the ith instance) the following changes with a chance of either 40% or 80%:

1. **Missing (miss)** in this case, the label becomes $Y_{j,i} = *$, and is not included in the training samples of the conditional mmodels.
2. **Noise** in this case, we consider two different type of changes:
   (a) **Reversing (rev)** we reverse the current value of the label. In other words, if $Y_{j,i} = 1$ it becomes $Y_{j,i} = 0$ (and similarly $Y_{j,i} = 0 \rightarrow Y_{j,i} = 1$). This setting can be seen as an adversarial one, where the adversary can swtich a number of labels,
   (b) **Flipping (flip)** in contrast to previous case, for each chosen label $Y_{j,i}$, we proceed to throw a Bernoulli trial with probability $\beta := P(Y_{j,i} = 1)$, i.e. $Y_{j,i} \sim \mathcal{B}er(\beta)$, with $\beta \in \{0.2, 0.8\}$.

4.2 Results

The average performances of the minimax approach for (IB) and (MAR) strategies obtained in terms of the IC measure are shown in Tables 3.a and 3.b respectively, with an imprecise level[4] $s = 2$, applied to our imprecise approach (ICC) (resp. precise approach (CC)).

[3] For further details of NCC, we refer to Zaffalon's work [18] and also [4].
[4] We could have optimised on s, but it seemed unfair compared to the precise approach that does not benefit from this hyper-parameter.

Table 3. Average (%) of the IC on missing and noise settings for $s=2$ and $\beta=0.8$

(a) Imprecise Branching

Data set	%	Missing		Reversing		Flipping	
		CC	*ICC*	*CC*	*ICC*	*CC*	*ICC*
Emotion	0.0	**21.87**	22.70	—	—	—	—
	0.4	**21.82**	23.02	**32.02**	32.75	**27.71**	27.74
	0.8	**21.61**	23.17	74.64	**73.58**	39.51	**34.80**
Scene	0.0	**16.03**	16.94	—	—	—	—
	0.4	**15.74**	17.21	**30.38**	31.54	**28.22**	28.70
	0.8	**14.07**	18.38	74.92	**73.68**	38.33	**34.91**
Yeast	0.0	**29.59**	33.00	—	—	—	—
	0.4	**28.96**	34.54	**40.50**	41.85	**36.45**	38.34
	0.8	**26.17**	40.10	67.49	**64.58**	53.15	**50.55**

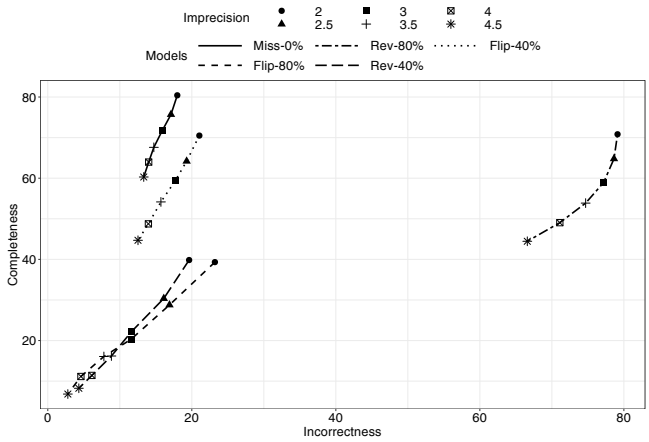
(b) Marginalization

Data set	%	Missing		Reversing		Flipping	
		CC	*ICC*	*CC*	*ICC*	*CC*	*ICC*
Emotion	0.0	**21.76**	22.83	—	—	—	—
	0.4	**21.84**	23.24	**31.71**	32.59	**27.75**	27.79
	0.8	**21.64**	24.35	74.74	**73.72**	40.04	**35.29**
Scene	0.0	**16.03**	16.98	—	—	—	—
	0.4	**15.73**	17.31	**30.62**	31.73	**28.20**	28.74
	0.8	**14.14**	18.77	74.92	**73.67**	38.37	**34.85**
Yeast	0.0	**29.67**	33.69	—	—	—	—
	0.4	**28.86**	34.80	**40.50**	41.84	**36.45**	38.19
	0.8	**26.17**	42.29	67.54	**64.73**	53.17	**50.59**

The obtained results with unchanged labels are similar. In the case of the missing labels, it seems that the minimax strategy and the addition of imprecision actually impairs the results, which is surprising and worthy of further investigation. Interestingly though, our strategy seems to be more robust to the presence of high noise in the data, as we systematically outperform the precise cahining when 80% of the labels are affected.

In Figure 4.2, we provide the evolution of IC and CP in average (%), with a set-up of $\beta = 0.8$ for the flipping setting. The results displayed are those that we expect, since when s increases, the incorrectness (IC) decreases as we forget more and more (as completeness or CP decreases). We can note that as the data set becomes worse (80% noise), completeness decrease in a quicker way, and could maybe used as an indicator of the data quality.

Fig. 3. Evolution of the incorrectness and completeness for the imprecise branching strategy and Emotion dataset.



All those results however only provide a proof of concept for our methodology, and are also obtained with a classifier which, through its independence

assumption, makes imprecise chaining computationally efficient but limits the benefits of using a chaining approach.

5 Conclusions

In this paper, we have introduced initial ideas to adapt the classical chaining algorithms of multi-label problems to the case of imprecise or set-valued probabilities. Such an idea is indeed promising to tamper the usual biases of picking a particular branch in the chain.

However, much remains to be done, as how to come up with general but efficient optimisation methods to solve Equations (IB) and (MAR). Indeed, while the Naive credal classifier makes them easy to solve thanks to its assumptions, the same assumptions may be the reason of our mitigated results. It seems therefore essential, in future works, to investigate other classifiers as well as to solve optimisation issues.

References

1. Antonucci, A., Corani, G.: The multilabel naive credal classifier. International Journal of Approximate Reasoning **83**, 320–336 (2017)
2. Augustin, T., Coolen, F.P., de Cooman, G., Troffaes, M.C.: Introduction to imprecise probabilities. John Wiley & Sons (2014)
3. Boutell, M.R., Luo, J., Shen, X., Brown, C.M.: Learning multi-label scene classification. Pattern recognition **37**(9), 1757–1771 (2004)
4. Corani, G., Benavoli, A.: Restricting the idm for classification. In: Hüllermeier, E., Kruse, R., Hoffmann, F. (eds.) Information Processing and Management of Uncertainty in Knowledge-Based Systems. Theory and Methods. pp. 328–337. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
5. De Cooman, G., Hermans, F.: Imprecise probability trees: Bridging two theories of imprecise probability. Artificial Intelligence **172**(11), 1400–1427 (2008)
6. Dembczyński, K., Waegeman, W., Cheng, W., Hüllermeier, E.: On label dependence and loss minimization in multi-label classification. Machine Learning **88**(1-2), 5–45 (2012)
7. Destercke, S.: Multilabel prediction with probability sets: the hamming loss case. In: Information Processing and Management of Uncertainty in Knowledge-Based Systems. pp. 496–505. Springer (2014)
8. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning. Springer New York Inc. (2001)
9. Fürnkranz, J., Hüllermeier, E., Mencía, E.L., Brinker, K.: Multilabel classification via calibrated label ranking. Machine learning **73**(2), 133–153 (2008)
10. Levi, I.: The enterprise of knowledge: An essay on knowledge, credal probability, and chance. MIT press (1983)
11. Nguyen, V.L., Hüllermeier, E.: Reliable multi-label classification: Prediction with partial abstention. arXiv preprint arXiv:1904.09235 (2019)
12. Pillai, I., Fumera, G., Roli, F.: Multi-label classification with a reject option. Pattern Recognition **46**(8), 2256–2266 (2013)

13. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine learning **85**(3), 333 (2011)
14. Taylor, S.J.: Introduction to measure and integration. CUP Archive (1973)
15. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) **3**(3), 1–13 (2007)
16. Walley, P.: Statistical reasoning with imprecise Probabilities. Chapman and Hall (1991)
17. Walley, P.: Inferences from multinomial data: learning about a bag of marbles. Journal of the Royal Statistical Society. Series B (Methodological) pp. 3–57 (1996)
18. Zaffalon, M.: The naive credal classifier. Journal of statistical planning and inference **105**(1), 5–21 (2002)