Multi-label Chaining with Imprecise Probabilities

Yonatan Carlos Carranza-Alarcon, Sébastien Destercke

Multi-label Chaining with Imprecise Probabilities

Yonatan Carlos Carranza Alarcón[1][0000−0002−8657−6355] and Sébastien Destercke[1][0000−0003−2026−468X]

Sorbonne Universités, Université Technologique de Compiègne, CNRS, UMR 7253 - Heudiasyc, 57 Avenue de Landshut, Compiègne, France
{yonatan-carlos.carranza-alarcon, sebastien.destercke}@hds.utc.fr

Abstract. We present two different strategies 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 estimations 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 adapt both strategies to the case of the naive credal classifier, showing that this adaptations are computationally efficient. Our experimental results on missing labels, which investigate how reliable these predictions are in both approaches, indicate that our approaches produce relevant cautiousness on those hard-to-predict instances where the precise models fail.

Keywords: imprecise probabilities · multi-label · classifier chains

Multi-label classification (MLC) is a generalization of traditional classification (with a single label), as well as a special case of multi-task learning. This approach is increasingly required in different research fields, such as the classification of proteins in bioinformatics [17], text classification in information retrieval [10], 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 [17,10] 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 in probabilistic trees [6] requires, at worst, to scan all possible predictions (whose quantity grows exponentially in the number of labels). A popular technique, known as chaining [15] to solve this issue, at least for the inference task, is to use heuristic predictions: they consists in using, incrementally, the predictions made on previous labels as additional conditional features to help better predict the relevance of a current label, rather than using the prediction probabilities.

To the best of our knowledge, there are only a few works on multi-label classification producing cautious predictions, such as the reject option [14], partial

predictions [8,1] or abstaining labels [18], but none of these have studied this issue in chaining (or classifier-chains approach).

In this paper, we consider the problem of extending chaining to the imprecise probabilistic case, and propose two different extensions, as some predictive probabilities are too imprecise to use predicted labels, henceforth called abstained labels, in the chaining. The first extension treats the abstained labels in a robust way, exploring all possible conditional situation in order not to propagate early uncertain decisions, whereas the latter marginalizes the probabilistic model over those labels, ignoring them in the predictive model.

Section 1 introduces the notations that we use for the multi-label setting, and gives the necessary reminders about making inferences with convex sets of probabilities. In Section 2, we recall the classical classifier-chains approach and then we present our extended approaches based on imprecise probabilities. Section 3 then shows that in the case of the naive credal classifier (NCC) [21], those strategies can be performed in polynomial time.

Finally, in Section 4, we perform a set of experiments on real data sets, which are perturbed with missing labels, in order to investigate how cautious (when we abstain on labels difficult to predict) is our approach. In order to adhere to the page limit, all proofs and supplementary experimental results have been relegated to the appendix of an online extended version [4].

1   Problem Setting

In the multi-label problem, an instance $\boldsymbol{x}$ 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 $\boldsymbol{x}$. Let $\mathscr{Y} = \{0, 1\}^m$ be a m-dimensional binary space and $\boldsymbol{y} = (y_1, \ldots, y_m) \in \mathscr{Y}$ be any element of $\mathscr{Y}$ 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 usual classification. Given a probability distribution $\hat{\mathbb{P}}$ fitting a set of i.i.d. observations $\mathcal{D} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i) | i = 1, \ldots, N\}$ 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 $\ell(\cdot, \cdot)$:

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

where $\boldsymbol{h} : \mathscr{X} \to \mathscr{Y}$ is a m-dimensional vector function. If $\ell(\cdot, \cdot)$ is defined instance-wise, that is $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, the solution of Equation (1) is obtained by minimizing the conditional expected risk (cf. [7, eq. 3] and [9, eq. 2.21])

$$\boldsymbol{h}(\boldsymbol{x}) = \operatorname*{arg\,min}_{\boldsymbol{y} \in \mathscr{Y}} \mathbb{E}_{\hat{\mathbb{P}}_{Y|\boldsymbol{x}}} \left[\ell(Y, \boldsymbol{y}) \right] = \arg \min_{\boldsymbol{y} \in \mathscr{Y}} \sum_{\boldsymbol{y}' \in \mathscr{Y}} \hat{P}(Y = \boldsymbol{y}' | X = \boldsymbol{x}) \ell(\boldsymbol{y}', \boldsymbol{y}) \tag{2}$$

or, equivalently, by picking the undominated elements of the order relation[1] $\succeq$ over $\mathscr{Y} \times \mathscr{Y}$ for which $\boldsymbol{y}^1 \succeq \boldsymbol{y}^2$ ($\boldsymbol{y}^1$ is preferred to/dominates $\boldsymbol{y}^2$) iff

$$\mathbb{E}_{\hat{\mathbb{P}}_{Y|\boldsymbol{x}}}\left(\ell(\boldsymbol{y}^2, \cdot) - \ell(\boldsymbol{y}^1, \cdot)\right) = \mathbb{E}_{\hat{\mathbb{P}}_{Y|\boldsymbol{x}}}\left(\ell(\boldsymbol{y}^2, \cdot)\right) - \mathbb{E}_{\hat{\mathbb{P}}_{Y|\boldsymbol{x}}}\left(\ell(\boldsymbol{y}^1, \cdot)\right) \geq 0. \qquad (3)$$

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

In this paper, we are also interested in making set-valued predictions when uncertainty is too high (e.g. due to insufficient evidence to include or discard a label as relevant, see Example 1). The set-valued prediction will here be described as a partial binary vector $\boldsymbol{y}^* \in \mathfrak{Y}$ where $\mathfrak{Y} = \{0, 1, *\}^m$ is the new output space with a new element $*$ representing the abstention. For instance, a partial prediction $\boldsymbol{y}^* = (*, 1, 0)$ corresponds to two plausible binary vector solutions $\{(0, 1, 0), (1, 1, 0)\} \subseteq \mathscr{Y}$. To obtain such predictions, we will use imprecise probabilities as a well-founded framework.

1.1   Notions about Imprecise Probabilities

Imprecise probabilities consist in representing our uncertainty by a convex set of probability distributions $\mathscr{P}_X$ [19,2] (i.e. a *credal set* [12]), defined over a space $\mathscr{X}$ rather than by a precise probability measure $\mathbb{P}_X$ [16]. 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:

$$\underline{P}_X(A) = \inf_{P \in \mathscr{P}_X} P(A) \qquad \text{and} \qquad \overline{P}_X(A) = \sup_{P \in \mathscr{P}_X} P(A) \qquad (4)$$

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 [19,21] have argued that when information is lacking or imprecise, considering credal sets as our model of information better describes our actual uncertainty.

However, such an approach comes with extra challenges in the learning and inference step, especially in combinatorial domains. In this paper, we will consider making a chain of binary inferences, each inference influenced by the previous one. If we consider $\mathcal{Y} = \{0, 1\}$ as the output space and Y as a univariate random variable on $\mathcal{Y}$, a standard way to take a decision with abstention given a credal set $\mathscr{P}$ on $\mathcal{Y}$ is

$$\hat{y} = \begin{cases} 1 & \text{if } \underline{P}_{\boldsymbol{x}}(Y=1) > 0.5, \\ 0 & \text{if } \overline{P}_{\boldsymbol{x}}(Y=1) < 0.5, \\ * & \text{if } 0.5 \in \left[\underline{P}_{\boldsymbol{x}}(Y=1), \overline{P}_{\boldsymbol{x}}(Y=1)\right] \end{cases} \qquad (5)$$

The next example illustrates such notions on a multi-label example

Example 1. We consider an output space of two labels $\mathcal{K} = \{m_1, m_2\}$, a single binary feature x_1 and Table 1 with imprecise estimations of $\hat{\mathbb{P}}(Y_1, Y_2|X_1)$.

---

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

y_1	y_2	x_1	$\hat{\mathscr{P}}_{Y_1,Y_2|X_1=0}$	y_1	y_2	x_1	$\hat{\mathscr{P}}_{Y_1,Y_2|X_1=1}$
0	0	0	[0.4,0.7]	0	0	1	0.00
0	1	0	[0.3,0.6]	0	1	1	0.00
1	0	0	0.00	1	0	1	[0.6,0.8]
1	1	0	0.00	1	1	1	[0.2,0.4]

Table 1. Estimated conditional probability distributions $\hat{\mathbb{P}}(Y_1,Y_2|X_1) \in \hat{\mathscr{P}}_{Y_1,Y_2|X_1}$.

Based on the probabilities of Table 1, we have that $\hat{P}_0(y_1=0) := \hat{P}(y_1 = 0|x_1=0) = 1$ and $\hat{P}_0(y_2=0) \in [0.4, 0.7]$, therefore not knowing whether $\hat{P}_0(y_2 = 0) > 0.5$. This could lead us to propose as a prediction $\hat{\boldsymbol{y}}^* = (0, *)$. 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 $\hat{\boldsymbol{y}}^* = (1, 0)$.

2   Multilabel Chaining with Imprecise Probabilities

Solving (2) can already be computationally prohibitive in the precise case [6], which is why heuristic to approximate inferences done on the full joint models such as the chain model have been proposed. This section recalls its basics and presents our proposed extension. To do so, we will need a couple notations: we will denote by $\mathscr{I}$ subsets of label indices and by $[\![j]\!] = \{1, \ldots, j\}$ the set of the first j integers. Given a prediction made in the j first labels, we will denote by

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

and of course $\mathscr{I}^j = \mathscr{I}_{\mathcal{A}}^j \cup \mathscr{I}_{\mathcal{R}}^j \cup \mathscr{I}_{\mathcal{I}}^j = [\![j]\!]$. Besides, for the sake of simplicity, we will use the notation

$$P_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j=1) := P(Y_j=1|Y_{\mathscr{I}^{j-1}}= \hat{\boldsymbol{y}}_{\mathscr{I}^{j-1}}, X = \boldsymbol{x}), \qquad (6)$$

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

2.1   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 $\mathscr{X}_{j-1}^* = \mathscr{X} \times \{0, 1\}^{j-1}, j \in \mathbb{N}^{>0}$. 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

$$\hat{y}_j := h_j(x) = \arg\max_{y \in \{0,1\}} P_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j=y). \qquad (7)$$

The classical multi-label chaining then works as follows:

1. RANDOM LABEL ORDERING. We randomly pick an order between labels and assume that the indices are relabelled in an increasing order.
2. PREDICTION j^{th} LABEL. For a label y_j and the previous predictions on labels $y_1, \ldots, y_{j-1}$ 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 with $\mathscr{I}_{\mathcal{R}}^{j-1} \cap \mathscr{I}_{\mathcal{I}}^{j-1} = \emptyset$. Then, the prediction of $\hat{y}_j$ (or $h_j(\boldsymbol{x})$) for a new instance $\boldsymbol{x}$ is

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

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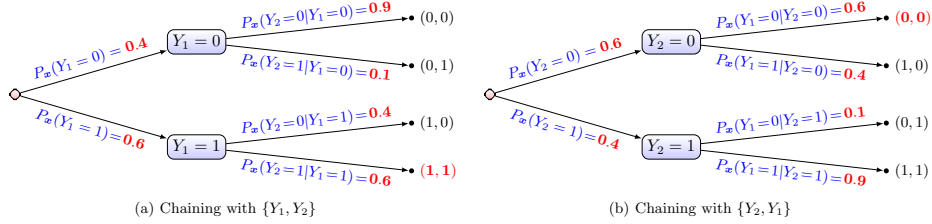

(a) Chaining with $\{Y_1, Y_2\}$ (b) Chaining with $\{Y_2, Y_1\}$

Fig. 1. Precise chaining

Figure 1 shows that using this heuristic can lead to strong biases, as two different orderings of the same joint model can lead to shift from one prediction to its opposite. Intuitively, adding some robustness and cautiousness in the process could help to avoid unwarranted biases.

In what follows, we propose two different extensions of precise chaining based on imprecise probability estimates, in which the final prediction belongs to the output space $\mathfrak{Y}$ of partial bianry vectors.

2.2 Imprecise Probabilistic Chaining

We now consider that the estimates $P_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = 1)$ can become imprecise, that is, we now have $[P_{\boldsymbol{x}}^{[\![j-1]\!]}](Y_j = y_j) := [\underline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = y_j), \overline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = y_j)]$. The basic idea of using such estimates is that in the chaining, we should be cautious when the classifier is unsure about the most probable prediction. In this section, we describe two different strategies (or extensions) in a general way, and we will propose an adaptation of those strategies to the NCC in the next section.

Let us first formulate the generic procedure to calculate the probability bound of the jth label,

1. RANDOM LABEL ORDERING. As in the precise case.

2. PREDICTION j^{th} LABEL. For a given label y_j, we assume we have (possibly imprecise) predictions for $y_1, \ldots, y_{j-1}$ such that $\mathscr{I}_{\mathcal{A}}^{j-1}$ are the indices of labels on which we abstained $\{*\}$ so far, and $\mathscr{I}_{\mathcal{R}}^{j-1}$ and $\mathscr{I}_{\mathcal{I}}^{j-1}$ remain the 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 $[P_{\boldsymbol{x}}^{\llbracket j-1 \rrbracket}](Y_j = 1)$ in order to predict the label $\hat{y}_j$ as

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

where this last equation is a slight variation of Equation (5) by using the new input space $\mathscr{X}_{j-1}^*$.

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

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

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

So we consider all 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 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 2(b) where we have abstained on labels (Y_2, Y_4) and we want to compute lower and upper probability bounds of the label $Y_5 = 1$.
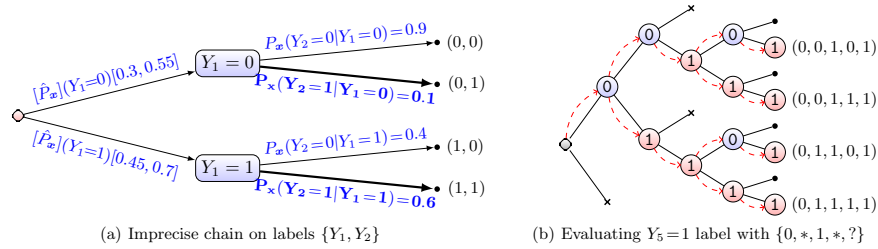

(a) Imprecise chain on labels $\{Y_1, Y_2\}$ (b) Evaluating $Y_5 = 1$ label with $\{0, *, 1, *, ?\}$

Fig. 2. Imprecise branching strategy

In Figure $2(a)$, we will consider the previous example (see Figure 1) in order to study in detail how we should calculate probability bounds $[\underline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = 1), \overline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = 1)]$. For the sake of simplicity, we assume that probabilities about Y_2 are precise and that the probability bounds for $Y_1 = 1$ is $\hat{P}_{\boldsymbol{x}}^{[\![0]\!]}$ $(Y_1 = 1) \in [0.45, 0.70]$. Taking all possible paths (bold on Figure $2(a)$) for which $Y_2 = 1$, we get

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

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

which means that in this case we would abstain on both labels, i.e. $(\hat{y}_1, \hat{y}_2) = (*, *)$.

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 $[\underline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = 1), \overline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = 1)]$ comes down to computing

$$\underline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j\!=\!1)\!=\!\underline{P}_{\boldsymbol{x}}(Y_j = 1 | Y_{\mathscr{I}_{\mathcal{R}}^{j-1}}\!=\!1, Y_{\mathscr{I}_{\mathcal{I}}^{j-1}}\!=\!0, Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}\!=\!\{0,1\}^{|\mathscr{I}_{\mathcal{A}}^{j-1}|}),$$

$$\overline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j\!=\!1)\!=\!\overline{P}_{\boldsymbol{x}}(Y_j = 1 | Y_{\mathscr{I}_{\mathcal{R}}^{j-1}}\!=\!1, Y_{\mathscr{I}_{\mathcal{I}}^{j-1}}\!=\!0, Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}\!=\!\{0,1\}^{|\mathscr{I}_{\mathcal{A}}^{j-1}|}), \quad \text{(MAR)}$$

where $\mathscr{I}_{\mathcal{A}}^{j-1} = \{i_1, \ldots, i_k\}$ denotes the set of indices of abstained labels and the last conditional term of probability bounds can be defined as

$$\left(Y_{\mathscr{I}_{\mathcal{A}}^{j-1}} = \{0,1\}^{|\mathscr{I}_{\mathcal{A}}^{j-1}|}\right) := (Y_{i_1} = 0 \cup Y_{i_1} = 1) \cap \cdots \cap (Y_{i_k} = 0 \cup Y_{i_k} = 1). \ (10)$$

The MAR formulation can be reduced by using Bayes's theorem in conjunction with the law of total probability. That is, for instance, given abstained labels $(Y_1 = *, Y_3 = *)$ and the precise prediction $(Y_2 = 1)$, inferring $Y_4 = 1$ comes down to computing $P_{\boldsymbol{x}}(Y_4\!=\!1 | (Y_1\!=\!0 \cup Y_1\!=\!1), Y_2\!=\!1, (Y_3\!=\!0 \cup Y_3\!=\!1))$ as follows

$$\frac{\sum\limits_{y_3, y_1 \in \{0,1\}^2} P_{\boldsymbol{x}}(Y_4\!=\!1, Y_1\!=\!y_1, Y_2\!=\!1, Y_3\!=\!y_3)}{\sum\limits_{y_3, y_1 \in \{0,1\}^2} P_{\boldsymbol{x}}(Y_1\!=\!y_1, Y_2\!=\!1, Y_3\!=\!y_3)} = \frac{P_{\boldsymbol{x}}(Y_4\!=\!1, Y_2\!=\!1)}{P_{\boldsymbol{x}}(Y_2\!=\!1)} = P_{\boldsymbol{x}}(Y_4\!=\!1 | Y_2\!=\!1),$$

An illustration providing some intuition about this last example can be seen in Fig. $3(a)$, in which we draw the possible path to infer the label Y_4 (considering a third branch in the chain to represent abstained labels).

The results of the last example can easily be generalized, and hence, MAR comes down to calculating the new formulation called (MAR*)

$$\underline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = 1) = \min_{P \in \mathscr{P}^*} P_{\boldsymbol{x}}(Y_j = 1 | Y_{\mathscr{I}_{\mathcal{R}}^{j-1}} = 1, Y_{\mathscr{I}_{\mathcal{I}}^{j-1}} = 0), \qquad (11)$$

$$\overline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = 1) = \max_{P \in \mathscr{P}^*} P_{\boldsymbol{x}}(Y_j = 1 | Y_{\mathscr{I}_{\mathcal{R}}^{j-1}} = 1, Y_{\mathscr{I}_{\mathcal{I}}^{j-1}} = 0). \qquad (12)$$

where $\mathscr{P}^*$ is simply the set of joint probability distributions described by the imprecise probabilistic tree (we refer to [5] for a detailed analysis). (MAR) comes down to restrict the conditional chain model to consider only those labels on which we have not abstained (for which we have sufficient reliable information). In general, both (IB) and (MAR) can lead to burdensome computations. In the next section, we propose to adapt their principle to the NCC, showing that this can be done efficiently.

Remark 1. Note that another strategy that would be computationally efficient would be to simply considers all precise chaining paths consistent with local intervals, pruning dominated branches. However, this would also mean that each explored branch would have all previous predicted labels as conditioning features, thus still being impacted by the bias of picking those predictions.

3   Imprecise Chaining with NCC

NCC extends the classical naive Bayes classifier (NBC) on a set of distributions. NCC preserves the assumption of feature independence made by NBC, and relies on the Imprecise Dirichlet model (IDM) [20] to estimate class-conditional imprecise probabilities, whose imprecision level is controlled through a hyper-parameter $s \in \mathbb{R}$. Therefore, the class-conditional probability bounds evaluated for $Y_j = 1$ ($Y_j = 0$ can be directly calculated using duality) can be calculated as follows[2]

$$\underline{P}(Y_j=1|\mathbf{X}=\boldsymbol{x},Y_{\mathscr{G}j-1}=\hat{\boldsymbol{y}}_{\mathscr{G}j-1})=\left(1+\frac{P(Y_j=0)\overline{P}_0(\mathbf{X}=\boldsymbol{x})\overline{P}_0(Y_{\mathscr{G}j-1}=\hat{\boldsymbol{y}}_{\mathscr{G}j-1})}{P(Y_j=1)\underline{P}_1(\mathbf{X}=\boldsymbol{x})\underline{P}_1(Y_{\mathscr{G}j-1}=\hat{\boldsymbol{y}}_{\mathscr{G}j-1})}\right)^{-1} \tag{13}$$

$$\overline{P}(Y_j=1|\mathbf{X}=\boldsymbol{x},Y_{\mathscr{G}j-1}=\hat{\boldsymbol{y}}_{\mathscr{G}j-1})=\left(1+\frac{P(Y_j=0)\underline{P}_0(\mathbf{X}=\boldsymbol{x})\underline{P}_0(Y_{\mathscr{G}j-1}=\hat{\boldsymbol{y}}_{\mathscr{G}j-1})}{P(Y_j=1)\overline{P}_1(\mathbf{X}=\boldsymbol{x})\overline{P}_1(Y_{\mathscr{G}j-1}=\hat{\boldsymbol{y}}_{\mathscr{G}j-1})}\right)^{-1} \tag{14}$$

where conditional upper probabilities of $[\underline{P}_1, \overline{P}_1]$ and $[\underline{P}_0, \overline{P}_0]$ are defined as

$$\overline{P}_a(\mathbf{X}=\boldsymbol{x}):=\prod_{i=1}^{p}\overline{P}(X_i=x_i|Y_j=a) \text{ and } \overline{P}_a(\mathbf{Y}_{\mathscr{G}j-1}=\boldsymbol{y}_{\mathscr{G}j-1}):=\prod_{k=1}^{j-1}\overline{P}(Y_k=\hat{y}_k|Y_j=a), \tag{15}$$

where $a \in \{0, 1\}$. Conditional lower probabilities are obtained similarly. Using Equations (13) and (14), we now propose efficient procedures to solve the afore-mentioned strategies.

3.1   Imprecise Branching

In the specific case where we use the NCC, we can efficiently reduce the optimization problems of Equations (IB), as expressed in the proposition below.

---

[2] For reviewer convenience, details are given in the supplementary material.

Proposition 1. *Optimisation problems of the imprecise branching* (IB) *can be reduced by using probability bounds obtained from the NCC, namely Equations* (13) *and* (14)*, as follows*

$$\underline{P}^{[\![j-1]\!]}_{\boldsymbol{x}}(Y_j=1)\propto\max_{\boldsymbol{y}\in\{0,1\}^{|\mathscr{I}^{j-1}_{\mathcal{A}}|}}\frac{\overline{P}_0(Y_{\mathscr{I}^{j-1}_{\mathcal{A}}}=\boldsymbol{y})}{\underline{P}_1(Y_{\mathscr{I}^{j-1}_{\mathcal{A}}}=\boldsymbol{y})}\ \text{ and }\ \overline{P}^{[\![j-1]\!]}_{\boldsymbol{x}}(Y_j=1)\propto\min_{\boldsymbol{y}\in\{0,1\}^{|\mathscr{I}^{j-1}_{\mathcal{A}}|}}\frac{\underline{P}_0(Y_{\mathscr{I}^{j-1}_{\mathcal{A}}}=\boldsymbol{y})}{\overline{P}_1(Y_{\mathscr{I}^{j-1}_{\mathcal{A}}}=\boldsymbol{y})}.$$

Besides, applying the equations derived from the imprecise Dirichlet model, we have that the values of abstained labels for which the previous optimisation problems are solved are, respectively

$$\hat{\underline{\boldsymbol{y}}}_{\mathscr{I}^{j-1}_{\mathcal{A}}}:=\arg\max_{\boldsymbol{y}\in\{0,1\}^{|\mathscr{I}^{j-1}_{\mathcal{A}}|}}\prod_{y_i\in\boldsymbol{y}}\frac{n(y_i|y_j=0)+s}{n(y_i|y_j=1)}\ \text{ and }\ \hat{\overline{\boldsymbol{y}}}_{\mathscr{I}^{j-1}_{\mathcal{A}}}:=\arg\min_{\boldsymbol{y}\in\{0,1\}^{|\mathscr{I}^{j-1}_{\mathcal{A}}|}}\prod_{y_i\in\boldsymbol{y}}\frac{n(y_i|y_j=0)}{n(y_i|y_j=1)+s}\ (16)$$

where $\mathscr{I}^{j-1}_{\mathcal{A}}$ is the set of indices of the $(j-1)th$ first predicted abstained labels, $n(\cdot)$ is a count function that counts the number of occurrences of the event $y_i|y_j$ and $n(y_i|y_j=1)$ is always strictly positive.

Proposition 1 says that it is not necessary to know the original input features $\mathscr{X}$ and neither the $(j-1)th$ first precise predicted labels, in order to get the lower and upper probability bound of Equations (IB). However, it is necessary to keep track of the estimates made on all abstained labels, which is consistent with the fact that we want to capture the optimal lower and upper bounds of the conditional probability over all possible paths on which we have abstained.

Proposition 1 allows us to propose an algorithm below that can calculate Equations (16) linearly in the number of abstained labels.

Proposition 2. *The chain of labels $\hat{\underline{\boldsymbol{y}}}_{\mathscr{I}^{j-1}_{\mathcal{A}}}$ and $\hat{\overline{\boldsymbol{y}}}_{\mathscr{I}^{j-1}_{\mathcal{A}}}$ can be obtained in a time complexity of $\mathcal{O}(|\mathscr{I}^{j-1}_{\mathcal{A}}|)$.*

The following proposition provides the time complexity of the inference step of the imprecise branching strategy, jointly with the NCC and previous results.

Proposition 3. *The global time complexity of the* IMPRECISE BRANCHING *strategy in the worst-case is $\mathcal{O}(m^2)$ and in the best-case is $\mathcal{O}(m)$.*

3.2 Marginalization

When the NCC is considered, nothing needs to be optimized in the marginalization strategy, thanks to the assumption of independence applied between the binary conditional models of the chain.

We recall that the marginalization strategy needs to compute the conditional models described in Equations (MAR). These latter can be solved by simply ignoring the abstained labels in Equations (13) and (14) of the NCC. We thus focus on adapting Equation (14) (Equation (13) can be treated similarly), in order to show that the abstained labels can be removed of the conditioning and to get the expression presented in Equation (12). Based on Equation (14), we can

only focus on the conditional upper probability on labels, namely Equation (15), and rewrite it as follows:

$$\overline{P}_0(\mathbf{Y}_{\mathscr{I}^{j-1}} = \hat{\boldsymbol{y}}_{\mathscr{I}^{j-1}}) := \overline{P}_0(\mathbf{Y}_{\mathscr{I}_*^{j-1}} = \hat{\boldsymbol{y}}_{\mathscr{I}_*^{j-1}}, \mathbf{Y}_{\mathscr{I}_{\mathcal{A}}^{j-1}} = \{0,1\}^{|\mathscr{I}_{\mathcal{A}}^{j-1}|}) \qquad (17)$$

where $\mathscr{I}_*^{j-1} = \mathscr{I}_{\mathcal{R}}^{j-1} \cup \mathscr{I}_{\mathcal{I}}^{j-1}$ is the set of indices of relevant and irrelevant inferred labels, and the right side of last equation can be stated as

$$\max_{\substack{P \in \left\{ \mathscr{P}_{Y_k|Y_j}, \mathscr{P}_{Y_a|Y_j} \right\} \\ k \in \mathscr{I}_*^{j-1}, a \in \mathscr{I}_{\mathcal{A}}^{j-1}}} \prod_{k \in \mathscr{I}_*^{j-1}} P(Y_k = \hat{y}_k | Y_j = 0) \prod_{a \in \mathscr{I}_{\mathcal{A}}^{j-1}} P(Y_a = 0 \cup Y_a = 1 | Y_j = 0).$$

Thanks to the product assumption in the NCC, each term can be treated separately, making it possible to decouple the multiplication in two parts;

$$\max_{\substack{P \in \mathscr{P}_{Y_k|Y_j} \\ k \in \mathscr{I}_*^{j-1}}} \prod_{k \in \mathscr{I}_*^{j-1}} P(Y_k = \hat{y}_k | Y_j = 0) \times \max_{\substack{P \in \mathscr{P}_{Y_a|Y_j} \\ a \in \mathscr{I}_{\mathcal{A}}^{j-1}}} \prod_{a \in \mathscr{I}_{\mathcal{A}}^{j-1}} P(Y_a = 0 \cup Y_a = 1 | Y_j = 0),$$

where $P(Y_a=0 \cup Y_a=1|Y_j=0)=1$, and hence the second part becomes 1. Replacing this result in Equation (17), and then this latter in Equation (14), we get

$$\overline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = 1) = \max_{P \in \mathscr{P}_{Y_j|Y_{\mathscr{I}_{\mathcal{R}}^{j-1}}, Y_{\mathscr{I}_{\mathcal{I}}^{j-1}}}} P_{\boldsymbol{x}}(Y_j = 1 | Y_{\mathscr{I}_{\mathcal{R}}^{j-1}} = 1, Y_{\mathscr{I}_{\mathcal{I}}^{j-1}} = 0).$$

Therefore, at each inference step, we can apply Equations (13) and (14) on the reduced new formulation of the marginalization strategy (MAR*). An illustration providing some intuition about this reduction can be seen in Figure 3.
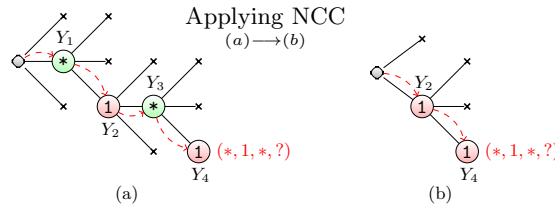

Fig. 3. Marginalization strategy applied to NCC for four labels $\{Y_1, Y_2, Y_3, Y_4\}$

4   Experiments

In this section, we perform experiments[3] on 6 data sets issued from the MULAN repository[4] (c.f. Table 2), following a 10×10 cross-validation procedure.

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 appro-

---

[3] Implemented in Python, see https://github.com/sdestercke/classifip
[4] http://mulan.sourceforge.net/datasets.html

Data set	#Domain	#Features	#Labels	#Instances	#Cardinality	#Density
emotions	music	72	6	593	1.90	0.31
scene	image	294	6	2407	1.07	0.18
yeast	biology	103	14	2417	4.23	0.30
cal500	music	68	174	502	26.04	0.15
medical	text	1449	45	978	1.25	0.03
enron	text	1001	53	1702	3.38	0.06

Table 2. Multi-label data sets summary

priate to use the set-accuracy (SA) and completeness (CP) [8, §4.1], as follows

$$SA(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \mathbb{1}_{(\boldsymbol{y} \in \hat{\boldsymbol{y}})} \quad \text{and} \quad CP(\hat{\boldsymbol{y}}, \boldsymbol{y}) = \frac{|Q|}{m},$$

where $\hat{\boldsymbol{y}}$ is the partial binary prediction (i.e. the set of all possibles binary vectors) and Q denote the set of non-abstained labels. When predicting complete vectors, then $CP = 1$ and SA equals the $0/1$ loss function and when predicting the empty vector, i.e. all labels $\hat{y}_i = *$, then $CP = 0$ and by convention $SA = 1$. The reason for using SA is that chaining is used as an approximation of the optimal prediction for a $0/1$ loss function.

Imprecise Classifier As mentioned in Section 3, we will use the *naive credal classifier* (NCC). Note that NCC needs discretized input spaces, so we discretize data sets to $z = 6$ intervals (except for Medical and Enron data sets). Besides, we restrict the values of the hyper-parameter of the imprecision to $s \in \{0.0, 0.5, \dots, 4.5, 5.5\}$ (when $s = 0.0$, NCC becomes the precise classifier NBC). At higher values of s, the NCC model will make mostly vacuous predictions (i.e. abstain in all labels $\forall i, Y_i = *$) for the data sets we consider here.

Missing Labels To simulate missingness during the training step, we uniformly pick at random a percentage of labels $Y_{j,i}$ (the jth label of the ith instance), which are then removed from the training data used to fit the conditional models in the chain. In this paper, we set up five different percentages of missingness: $\{0, 20, 40, 60, 80\}\%$.

4.1 Experimental Results

Figure 4, we provide the results for 3 data sets, showing set-accuracy and completeness measures in average[5](%) obtained by fitting the NCC model for different percentages of missing labels, respectively, applied to the data sets of Table 2 and using the imprecise branching strategy (trends were similar for the marginalization strategies, not displayed due to lack of space)[6].

The results show that when the percentage of missing labels increases, the set-accuracy (SA) increases (regardless of the amount of imprecision s we inject) as we abstain more and more (as completeness decreases). This means that the more

---

[5] The confidence intervals obtained on the experimental results are very small and we therefore prefer not to display them in the figures in order not to overcharge them.

[6] The supplementary results can be found in the online extend version [4]
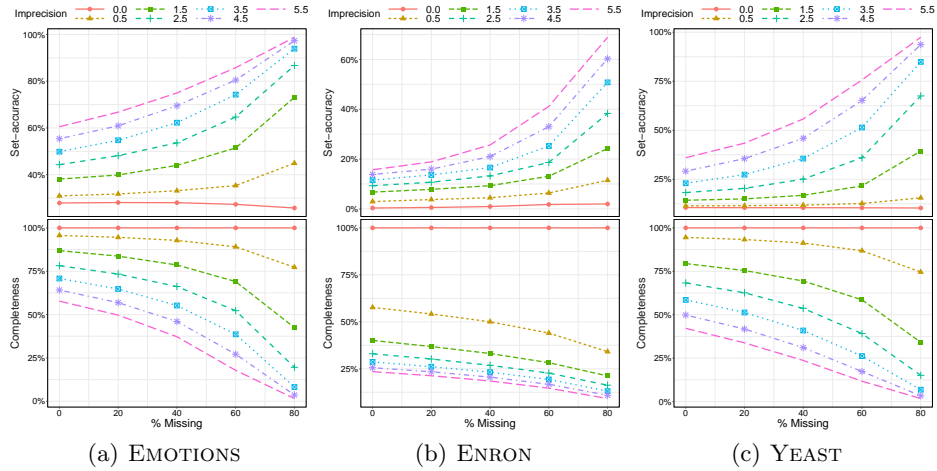
(a) EMOTIONS   (b) ENRON   (c) YEAST

Fig. 4. Missing labels - Imprecise Branching Evolution of the average (%) set-accuracy (top) and completeness (down) for each level of imprecision (a curve for each one) and a discretization $z = 6$, with respect to the percentage of missing labels.

imprecise we get, the more accurate are those predictions we retain. However, a high amount of imprecision is sometimes required to include the ground-truth solution within the set-valued prediction (this may be due to the very restrictive $0/1$ loss metric). For instance, with $s = 5.5$ and 40% of missingness, we get a $> 65\%$ of set-accuracy versus a $< 50\%$ of completeness, in the Emotions data set.

Overall, the results are those that we expect and are also sufficient to show that cautious inferences with probability sets may provide additional benefits when dealing with missing labels.

5   Conclusions

In this paper, we have proposed two new strategies to adapt the classical chaining multi-label problem to the case of handling imprecise probability estimates. Such strategies come with daunting challenges to obtain cautious and reliable predictions, and have been successfully resolved using the NCC model.

While the NCC makes easy to solve the strategies thanks to its assumptions, the same restrictive assumptions may also be the reason why the initial accuracy is rather low (especially for Yeast). Indeed, it seems reasonable to think that the independence assumptions somehow limit the benefits of including label dependencies information through the chaining. It seems therefore essential, in future works, to investigate other classifiers as well as to solve optimisation issues in a general or approximative way.

Another open issue is how we can use or extend the existing heuristics of probabilistic classifier approaches on our proposal strategies, such as epsilon-approximate inference, A^* and beam search methods [13,11].

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. Carranza Alarcón, Y.C., Destercke, S.: Multi-label chaining with imprecise probabilities (2021), https://arxiv.org/abs/2107.07443
5. De Cooman, G., Hermans, F.: Imprecise probability trees: Bridging two theories of imprecise probability. Artificial Intelligence **172**(11), 1400–1427 (2008)
6. Dembczynski, K., Cheng, W., Hüllermeier, E.: Bayes optimal multilabel classification via probabilistic classifier chains. In: ICML (2010)
7. 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)
8. 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)
9. Friedman, J., Hastie, T., Tibshirani, R.: The elements of statistical learning. Springer New York Inc. (2001)
10. Fürnkranz, J., Hüllermeier, E., Mencía, E.L., Brinker, K.: Multilabel classification via calibrated label ranking. Machine Learning **73**(2), 133–153 (2008)
11. Kumar, A., Vembu, S., Menon, A.K., Elkan, C.: Beam search algorithms for multilabel learning. Machine learning **92**(1), 65–89 (2013)
12. Levi, I.: The enterprise of knowledge: An essay on knowledge, credal probability, and chance. MIT press (1983)
13. Mena, D., Montañés, E., Quevedo, J.R., del Coz, J.J.: An overview of inference methods in probabilistic classifier chains for multilabel classification. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **6**(6), 215–230 (2016)
14. Pillai, I., Fumera, G., Roli, F.: Multi-label classification with a reject option. Pattern Recognition **46**(8), 2256–2266 (2013)
15. Read, J., Pfahringer, B., Holmes, G., Frank, E.: Classifier chains for multi-label classification. Machine learning **85**(3), 333–359 (2011)
16. Taylor, S.J.: Introduction to measure and integration. CUP Archive (1973)
17. Tsoumakas, G., Katakis, I.: Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM) **3**(3), 1–13 (2007)
18. Vu-Linh Nguyen, E.H.: Reliable multilabel classification: Prediction with partial abstention. In: Thirty-Fourth AAAI Conference on Artificial Intelligence (2019)
19. Walley, P.: Statistical reasoning with imprecise Probabilities. Chapman and Hall (1991)
20. Walley, P.: Inferences from multinomial data: learning about a bag of marbles. Journal of the Royal Statistical Society: Series B (Methodological) **58**(1), 3–34 (1996)
21. Zaffalon, M.: The naive credal classifier. Journal of statistical planning and inference **105**(1), 5–21 (2002)

Appendix

A   NCC derivation for multi-label chaining

As the purpose of the imprecise chaining is to compute binary conditional dependence models, we need only get conditional probability bounds of the probability $P(Y_j = y_j | \mathbf{X} = \boldsymbol{x}, Y_{\mathscr{I}j-1} = \hat{\boldsymbol{y}}_{\mathscr{I}j-1})$, so by using Bayes' theorem and naive Bayes' attribute independence assumption, it can be written as follows

$$\frac{P(Y_j = y_j) \prod_{i=1}^{p} P(X_i = x_i | Y_j = y_j) \prod_{k=1}^{j-1} P(Y_k = \hat{y}_k | Y_j = y_j)}{\sum_{y_l \in \{0,1\}} P(Y_j = y_l) \prod_{i=1}^{p} P(X_i = x_i | Y_j = y_l) \prod_{k=1}^{j-1} P(Y_k = \hat{y}_k | Y_j = y_l)}. \quad (18)$$

Computing lower and upper probability bounds $[\underline{P}, \overline{P}]$ over all possible marginals $\mathscr{P}_{Y_j}$ and conditional distributions $\mathscr{P}_{X_i|Y_j}, \mathscr{P}_{Y_k|Y_j}$ can be performed by solving the following minimization/maximization problem of Equation (18) as follows

$$\underline{P}(Y_j = y_j | X = \boldsymbol{x}, Y_{\mathscr{I}j-1} = \hat{\boldsymbol{y}}_{\mathscr{I}j-1}) =$$
$$\min_{P \in \mathscr{P}_{Y_j}} \min_{\substack{P \in \left\{ \mathscr{P}_{X_i|Y_j}, \mathscr{P}_{Y_k|Y_j} \right\} \\ i=1,\dots,d \\ k=1,\dots,j-1}} P(Y_j = y_j | X = \boldsymbol{x}, Y_{\mathscr{I}j-1} = \hat{\boldsymbol{y}}_{\mathscr{I}j-1}), \quad (19)$$

$$\overline{P}(Y_j = y_j | X = \boldsymbol{x}, Y_{\mathscr{I}j-1} = \hat{\boldsymbol{y}}_{\mathscr{I}j-1}) =$$
$$\max_{P \in \mathscr{P}_{Y_j}} \max_{\substack{P \in \left\{ \mathscr{P}_{X_i|Y_j}, \mathscr{P}_{Y_k|Y_j} \right\} \\ i=1,\dots,d \\ k=1,\dots,j-1}} P(Y_j = y_j | X = \boldsymbol{x}, Y_{\mathscr{I}j-1} = \hat{\boldsymbol{y}}_{\mathscr{I}j-1}). \quad (20)$$

In practice, we assume a precise estimation of the marginal distribution $\mathbb{P}_{Y_j}$ in lieu of a credal set $\mathscr{P}_{Y_j}$, so optimization problems over the credal set of marginal distributions $\mathscr{P}_{Y_j}$ can be ignored. Therefore, one can easily show that last equations evaluated to $Y_j = 1$ ($Y_j = 0$ can be directly calculated using duality) are equivalent to

$$\underline{P}(Y_j = 1 | \mathbf{X} = \boldsymbol{x}, Y_{\mathscr{I}j-1} = \hat{\boldsymbol{y}}_{\mathscr{I}j-1}) = \left(1 + \frac{P(Y_j = 0) \overline{P}_0(\mathbf{X} = \boldsymbol{x}) \overline{P}_0(Y_{\mathscr{I}j-1} = \hat{\boldsymbol{y}}_{\mathscr{I}j-1})}{P(Y_j = 1) \underline{P}_1(\mathbf{X} = \boldsymbol{x}) \underline{P}_1(Y_{\mathscr{I}j-1} = \hat{\boldsymbol{y}}_{\mathscr{I}j-1})} \right)^{-1} \tag{21}$$

$$\overline{P}(Y_j = 1 | \mathbf{X} = \boldsymbol{x}, Y_{\mathscr{I}j-1} = \hat{\boldsymbol{y}}_{\mathscr{I}j-1}) = \left(1 + \frac{P(Y_j = 0) \underline{P}_0(\mathbf{X} = \boldsymbol{x}) \underline{P}_0(Y_{\mathscr{I}j-1} = \hat{\boldsymbol{y}}_{\mathscr{I}j-1})}{P(Y_j = 1) \overline{P}_1(\mathbf{X} = \boldsymbol{x}) \overline{P}_1(Y_{\mathscr{I}j-1} = \hat{\boldsymbol{y}}_{\mathscr{I}j-1})} \right)^{-1} \tag{22}$$

where probability bounds $[\underline{P}_1, \overline{P}_1]$ and $[\underline{P}_0, \overline{P}_0]$ of each different conditional event are defined as follows

$$\overline{P}_0(\mathbf{X} = \boldsymbol{x}) := \prod_{i=1}^{p} \overline{P}(X_i = x_i | Y_j = 0) \;\; \text{and} \;\; \overline{P}_0(\mathbf{Y}_{\mathscr{I}^{j-1}} = \boldsymbol{y}_{\mathscr{I}^{j-1}}) := \prod_{k=1}^{j-1} \overline{P}(Y_k = \hat{y}_k | Y_j = 0),$$
(23)

$$\underline{P}_1(\mathbf{X} = \boldsymbol{x}) := \prod_{i=1}^{p} \underline{P}(X_i = x_i | Y_j = 1) \;\; \text{and} \;\; \underline{P}_1(\mathbf{Y}_{\mathscr{I}^{j-1}} = \boldsymbol{y}_{\mathscr{I}^{j-1}}) := \prod_{k=1}^{j-1} \underline{P}(Y_k = \hat{y}_k | Y_j = 1).$$
(24)

The last conditional probability bounds are derived using the Imprecise Dirichlet model (IDM) [20]

$$\underline{P}(X_i = x_i | Y_j = y_j) = \frac{n(x_i | y_j)}{n(y_j) + s} \;\; \text{and} \;\; \overline{P}(X_i = x_i | Y_j = y_j) = \frac{n(x_i | y_j) + s}{n(y_j) + s}$$
(25)

where $n(\cdot)$ is a count function that counts the number of occurrences of events $x_i | y_j$ and y_j in the observed data set. For instance; $n(x_i | y_j)$ is the number of instances in the training set where $X_i = x_i$ and the label value is $Y_i = y_i$.

In the same way as precedent equations, we can obtain probability bounds $[\underline{P}(Y_k = \hat{y}_k | Y_j = y_j), \overline{P}(Y_k = \hat{y}_k | Y_j = y_j)]$.

Remark 2. The denominator of Equations (21) and (22) may in some cases (depending on the training data) become zero, and therefore, the division would not be defined. In these cases, we adopt the "Laplace smoothing" technique for the precise probabilities $P(Y_j = 0)$ and $P(Y_j = 1)$. Furthermore, if we obtain an undefined division, i.e. $0/0$, we manually put 0.

Remark 3. Note that the choice of the marginal distribution $\mathbb{P}_{Y_j}$ in lieu of a credal set $\mathscr{P}_{Y_j}$ does not change the theoretical results obtained in Proposition 1. However, if the credal $\mathscr{P}_Y$ is considered, we can obtain a very small, but not significant, improvement in the experimental results[7].

B Proofs for our results

*Proof (**of Proposition 1**).* Let us begin to prove the optimization problem of the lower probability of (IB) evaluated to $Y_j = 1$

$$\underline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = 1) = \min_{\boldsymbol{y} \in \{0,1\}^{\left|\mathscr{I}_{\mathcal{A}}^{j-1}\right|}} \underline{P}_{\boldsymbol{x}}(Y_j = 1 | Y_{\mathscr{I}_{\mathcal{R}}^{j-1}} = 1, Y_{\mathscr{I}_{\mathcal{I}}^{j-1}} = 0, Y_{\mathscr{I}_{\mathcal{A}}^{j-1}} = \boldsymbol{y}).$$
(26)

Let us to define $\mathscr{I}_*^{j-1} = \mathscr{I}_{\mathcal{R}}^{j-1} \cup \mathscr{I}_{\mathcal{I}}^{j-1}$ as the set of indices of relevant and irrelevant predicted labels down to the $(j-1)th$ index. By applying Equation (21)

---

[7] Anyone can reproduce these results by using our online implementation.

to the right side of last equation, we get

$$\min_{\boldsymbol{y}\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}^{j-1}|}} \left(1 + \frac{P(Y_j=0)\overline{P}_0(\mathbf{X}=\boldsymbol{x})\overline{P}_0(Y_{\mathscr{I}_*^{j-1}}=\hat{\boldsymbol{y}}_{\mathscr{I}_*^{j-1}}, Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=\boldsymbol{y})}{P(Y_j=1)\underline{P}_1(\mathbf{X}=\boldsymbol{x})\underline{P}_1(Y_{\mathscr{I}_*^{j-1}}=\hat{\boldsymbol{y}}_{\mathscr{I}_*^{j-1}}, Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=\boldsymbol{y})}\right)^{-1},$$

where $\hat{\boldsymbol{y}}_{\mathscr{I}_*^{j-1}}$ is the binary vector with predicted relevant and irrelevant values. So, using the fact that minimizing $\frac{1}{1+x}$ is equal to maximizing x, we therefore get

$$\max_{\boldsymbol{y}\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}^{j-1}|}} \frac{P(Y_j=0)\overline{P}_0(\mathbf{X}=\boldsymbol{x})\overline{P}_0(Y_{\mathscr{I}_*^{j-1}}=\hat{\boldsymbol{y}}_{\mathscr{I}_*^{j-1}})\overline{P}_0(Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=\boldsymbol{y})}{P(Y_j=1)\underline{P}_1(\mathbf{X}=\boldsymbol{x})\underline{P}_1(Y_{\mathscr{I}_*^{j-1}}=\hat{\boldsymbol{y}}_{\mathscr{I}_*^{j-1}})\underline{P}_1(Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=\boldsymbol{y})}.$$

The first three terms of the numerator (and of the denominator) of the last equation can be omitted from the optimization problem since they do not change the optimal solution, and by applying Equations (23), (24) and (25), to the last term, we get what we sought

$$\max_{\boldsymbol{y}\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}^{j-1}|}} \frac{\overline{P}_0(Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=\boldsymbol{y})}{\underline{P}_1(Y_{\mathscr{I}_{\mathcal{A}}^{j-1}}=\boldsymbol{y})}$$

$$\Longleftrightarrow \max_{\boldsymbol{y}\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}^{j-1}|}} \frac{\prod_{k\in\mathscr{I}_{\mathcal{A}}^{j-1}}\overline{P}(Y_k=y_k|Y_j=0)}{\prod_{k\in\mathscr{I}_{\mathcal{A}}^{j-1}}\underline{P}(Y_k=y_k|Y_j=1)}$$

$$\Longleftrightarrow \max_{\boldsymbol{y}\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}^{j-1}|}} \frac{\prod_{k\in\mathscr{I}_{\mathcal{A}}^{j-1}}\frac{n(y_k|y_j=0)+s}{n(y_j=0)+s}}{\prod_{k\in\mathscr{I}_{\mathcal{A}}^{j-1}}\frac{n(y_k|y_j=1)}{n(y_j=1)+s}}$$

$$\Longleftrightarrow \max_{\boldsymbol{y}\in\{0,1\}^{|\mathscr{I}_{\mathcal{A}}^{j-1}|}} \left[\frac{n(y_j=1)+s}{n(y_j=0)+s}\right]^{|\mathscr{I}_{\mathcal{A}}^{j-1}|} \prod_{y_i\in\boldsymbol{y}}^{i\in\mathscr{I}_{\mathcal{A}}^{j-1}} \frac{n(y_i|y_j=0)+s}{n(y_i|y_j=1)},$$

in which it is easy see that: (1) the term $[\cdots]^{|\mathscr{I}_{\mathcal{A}}^{j-1}|}$ can be omitted, and hence, we can get $\underline{\hat{\boldsymbol{y}}}_{\mathscr{I}_{\mathcal{A}}^{j-1}}$, and (2) using similar arguments as above we can easily get the labels $\overline{\hat{\boldsymbol{y}}}_{\mathscr{I}_{\mathcal{A}}^{j-1}}$ which maximise the upper probability $\overline{P}_{\boldsymbol{x}}(Y_j=1)$ of (IB).

*Proof (**of Proposition** 2).* This proof can be performed using a dichotomy algorithm (equivalent to a binary search tree), starting with y_k last abstained label (i.e. $k=|\mathscr{I}_{\mathcal{A}}|-1$) and calculating the values $\frac{n(y_k=1|\cdot)+s}{n(y_k=1|\cdot)}$ and $\frac{n(y_k=0|\cdot)+s}{n(y_k=0|\cdot)}$, then we retain the maximal value of these last two terms (or the minimal value, whichever applies) and we go forward with second-to-last label y_{k-1}, but this time multiplied by the last term retained, and so on. After having obtained the lower binary path $\underline{\hat{\boldsymbol{y}}}_{\mathscr{I}_{\mathcal{A}}^{j-1}}$(or the upper binary path $\overline{\hat{\boldsymbol{y}}}_{\mathscr{I}_{\mathcal{A}}^{j-1}}$), we can directly

calculate the values $\underline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = 1)$ and $\overline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = 1)$ (and by duality of the lower and upper probability bounds $[\underline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = 0), \overline{P}_{\boldsymbol{x}}^{[\![j-1]\!]}(Y_j = 0)]$).

*Proof (**of Proposition** 3).* The proof for the best-case is straighforward, because if there is not any abstained labels, the time complexity is the same than precise chaining $\mathcal{O}(m)$. The worst-case complexity, in which all inferred labels are abstained, is also easy to calculate: the first label performs a single operation, i.e. $\mathcal{O}(1)$, then the second label is also inferred in a single operation due to the number of previous abstained labels being equal to 1 (c.f. Proposition 2), then the third label takes into account two previous abstained labels and performs two operations (c.f. Proposition 2), and the fourth label performs three operations, and so on. We therefore obtain $\mathcal{O}(m(m-1)/2 + 1)$ operations which is equal to $\mathcal{O}(m^2)$ asymptotically.

C   Supplementary results

(a) EMOTIONS   (b) ENRON   (c) YEAST
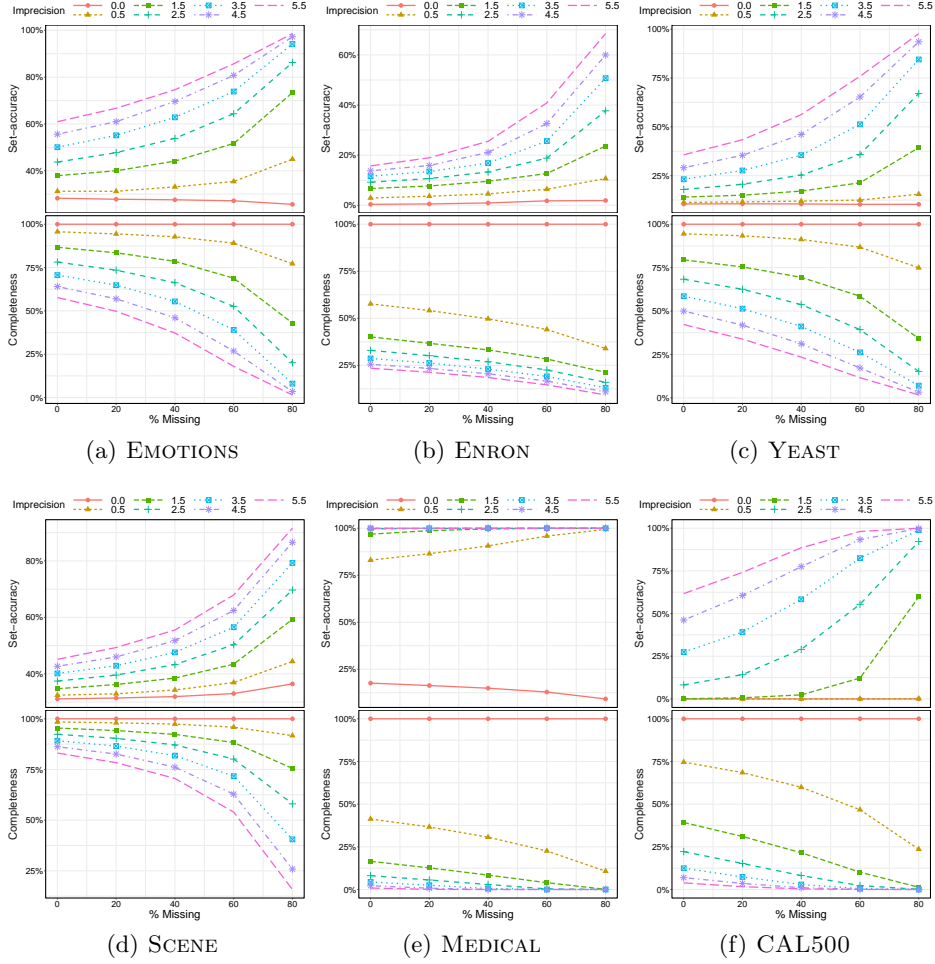
(d) SCENE   (e) MEDICAL   (f) CAL500

Fig. 5. Missing labels - Marginalization Evolution of the average (%) set-accuracy (top) and completeness (down) for each level of imprecision (a curve for each one) and discretization $z = 6$, with respect the percentage of missing labels.
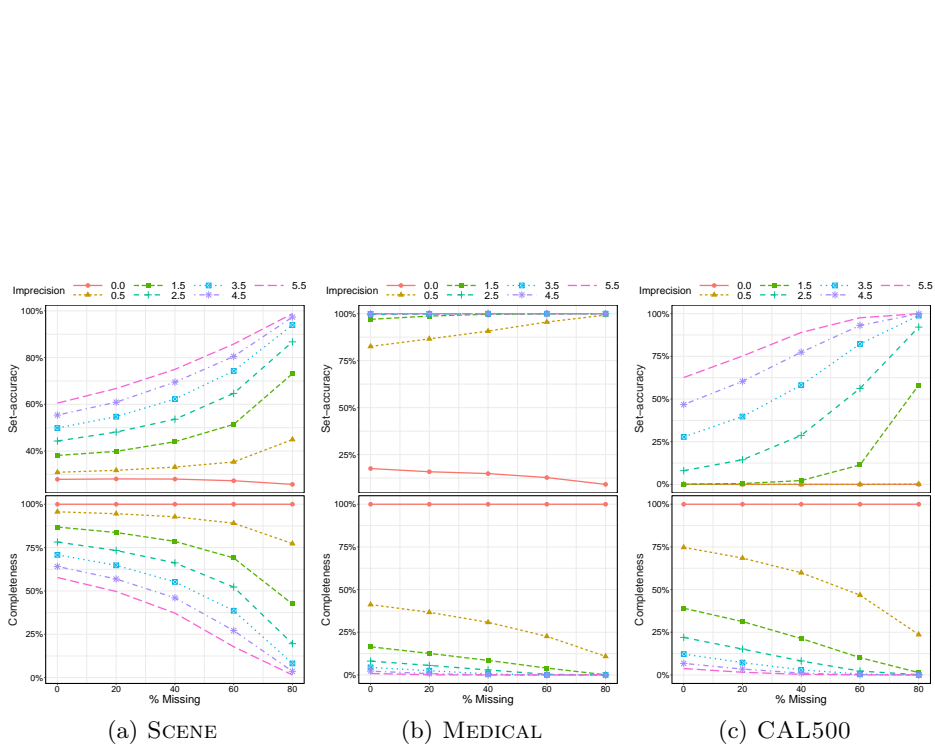
(a) SCENE   (b) MEDICAL   (c) CAL500

Fig. 6. Missing labels - Imprecise Branching Evolution of the average (%) set-accuracy (top) and completeness (down) for each level of imprecision (a curve for each one) and a discretization $z = 6$, with respect to the percentage of missing labels.