Multi-label Chaining using Naive Credal Classifier

16th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty

CARRANZA-ALARCON Yonatan-Carlos

Ph.D. in Computer Science

DESTERCKE Sébastien

Ph.D. in Computer Science

utc Université de Technologie Compiègne

heudiasyc

cnrs dépasser les frontières

21 to 24 Sept. 2021

Overview

- Multi-label classification problem
- Multi-label chaining with imprecise probabilities
  - Precise Probabilistic Chaining
  - Imprecise Probabilistic Chaining
    - Imprecise Probabilistic Chaining using NCC model
- Experiments
- Conclusions

Multi-label classification problem

☞ The goal of multi-label problem :

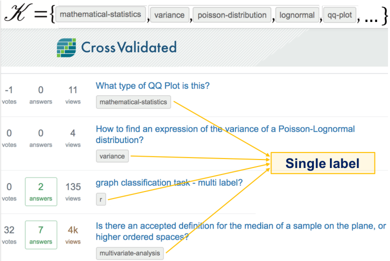
$$\text{Given a training data : } \mathscr{D} = \{\boldsymbol{x}^i, \boldsymbol{y}^i\}_{i=0}^{N} \subseteq \mathbb{R}^p \times \mathscr{Y}$$

$$\text{where : } \mathscr{Y} = \{0,1\}^m, \quad |\mathscr{Y}| = 2^m$$

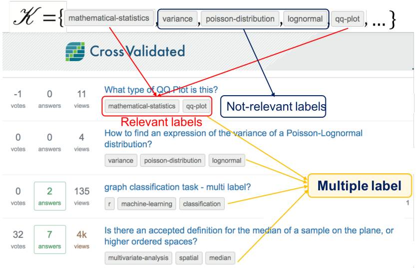
Learning a multi-label classification rule : $\varphi : \mathbb{R}^p \rightarrow \mathscr{Y}$

☞ Example :

Classical classification



Multi-label classification

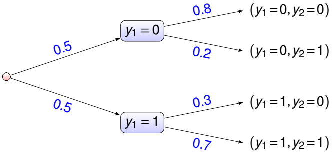


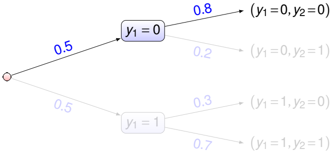
Multi-label classification problem

✌ Why imprecise multi-label chaining ?

✗ Label-wise decomposition ignores the label dependencies.

✗ Working with full probabilistic tree means exploring an exponential number of branches.





✗ Chaining heuristic introduce potential strong biases.

✗ No research on making cautious inferences in such chaining.

✌ Our contribution :

- ✓ We propose new strategies to extend the chaining multi-label problem to the imprecise probabilistic setting.
- ✓ We propose efficient procedures for NCC model.

Overview

- Multi-label classification problem
- Multi-label chaining with imprecise probabilities
  - Precise Probabilistic Chaining
  - Imprecise Probabilistic Chaining
    - Imprecise Probabilistic Chaining using NCC model
- Experiments
- Conclusions

Basic notations

Let us denote the probability of the label Y_j conditioned on previous labels

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

where $\mathscr{I}_*^{j-1}$ are indices of the j first predicted labels separated into

1. Indices of labels predicted as relevant : $\mathscr{I}_{\mathscr{R}}^{j}$
2. Indices of labels predicted as irrelevant : $\mathscr{I}_{\mathscr{I}}^{j}$
3. Indices of abstained labels : $\mathscr{I}_{\mathscr{A}}^{j}$

Example (Predict the 5ith relevant label $Y_5 = 1$)

Given the sets of indices of 4-first predicted labels: $\mathscr{I}_{\mathscr{R}}^{4} = \{2\}$, $\mathscr{I}_{\mathscr{I}}^{4} = \{1,4\}$, $\mathscr{I}_{\mathscr{A}}^{4} = \{3\}$.

$$P_{\boldsymbol{x}^*}^{[\![4]\!]}(Y_5=1) := P(Y_5=1 | Y_{\mathscr{I}_{\mathscr{R}}^{4}} = 1, Y_{\mathscr{I}_{\mathscr{I}}^{4}} = 0, Y_{\mathscr{I}_{\mathscr{A}}^{4}} = *, X=\boldsymbol{x})$$

$$= P(Y_5=1 | Y_1=0, Y_2=1, Y_3=*, Y_4=0, X=\boldsymbol{x})$$

Precise Probabilistic Chaining

Learning a multi-label chaining [Read et al. 2011]

1. Learning a binary classifier at each step of the chaining :

$$\varphi_j : \mathbb{R}^p \times \{0,1\}^{j-1} \rightarrow \{0,1\}$$

2. Decision step under a binary classifier $\ell(y_j, \widehat{y}_j) \rightarrow$

$$\text{“Optimal” decision : } \varphi_j := \widehat{y}_j = \begin{cases} 1 & P_{\boldsymbol{x}^*}^{[\![j-1]\!]}(Y_j = 1) \geq 0.5 \\ 0 & P_{\boldsymbol{x}^*}^{[\![j-1]\!]}(Y_j = 1) < 0.5 \end{cases}$$

An example of multi-label chaining

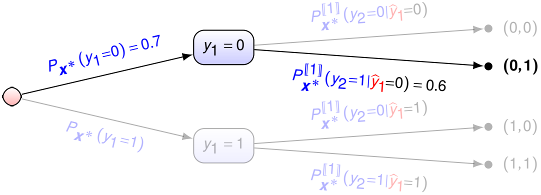


FIGURE – *Precise multi-label chaining with two labels.*

Imprecise Probabilistic Chaining

☞ Learning a multi-label chaining using imprecise probabilities (IP)

❶ Learning an imprecise classifier model at each step of the chaining :

$$[P_{\boldsymbol{x}^*}^{[\![j-1]\!]}] : \mathbb{R}^p \times \{0,1\}^{j \leq m} \rightarrow [\underline{P}_{\boldsymbol{x}^*}^{[\![j-1]\!]}, \overline{P}_{\boldsymbol{x}^*}^{[\![j-1]\!]}]$$

❷ Making a cautious decision

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

☞ An example of imprecise chaining

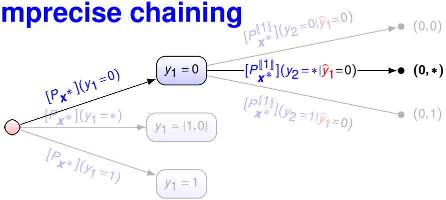


FIGURE – *An example of multi-label chaining using IP.*

How to get $[\underline{P}_{\mathbf{x}^*}^{[\![j-1]\!]}, \overline{P}_{\mathbf{x}^*}^{[\![j-1]\!]}]$? Strategy ❶ : Imprecise branching

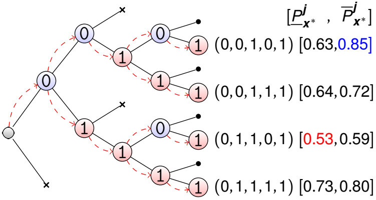
Considering all possible branching in the chaining as soon as there is an abstained label.

$$\underline{P}_{\mathbf{x}^*}^{[\![j-1]\!]}(Y_j = 1) = \min_{\mathbf{y}\in\{0,1\}^{|\mathscr{I}_{\mathscr{A}}^{j-1}|}} \underline{P}_{\mathbf{x}^*}(Y_j = 1 | Y_{\mathscr{I}_{\mathscr{R}}^{j-1}} = 1, Y_{\mathscr{I}_{\mathscr{I}}^{j-1}} = 0, Y_{\mathscr{I}_{\mathscr{A}}^{j-1}} = \mathbf{y}),$$
$$\overline{P}_{\mathbf{x}^*}^{[\![j-1]\!]}(Y_j = 1) = \max_{\mathbf{y}\in\{0,1\}^{|\mathscr{I}_{\mathscr{A}}^{j-1}|}} \overline{P}_{\mathbf{x}^*}(Y_j = 1 | Y_{\mathscr{I}_{\mathscr{R}}^{j-1}} = 1, Y_{\mathscr{I}_{\mathscr{I}}^{j-1}} = 0, Y_{\mathscr{I}_{\mathscr{A}}^{j-1}} = \mathbf{y}). \quad \text{(IB)}$$

Example :

Computing the probability of the label $Y_5 = 1$ conditioned on previous labels

$$\{\widehat{Y}_1 = 0, \widehat{Y}_2 = *, \widehat{Y}_3 = 1, \widehat{Y}_4 = *\}$$



Strategy ❷ : Marginalization

Ignore unsure predictions chaining in the interests of not propagating imprecision in the tree.

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

where $\mathscr{P}^*$ is the set of full joint probability distributions described by the imprecise probabilistic tree [De Cooman et al. 2008].

✗ The optimization problem can be tricky, since the probability space of $\mathscr{P}^*$ is not the same as $P'_{\boldsymbol{x}^*}$.

Imprecise Chaining with Naive Credal Classifier

The class-conditional probability bounds evaluated for $Y_j = 1$ ($Y_j = 0$ can be directly calculated using duality) can be calculated as follows

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

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

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

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

where $a \in \{0, 1\}$. → use factorization properties at our advantage !

Strategy ❶ : Imprecise branching (IB) with NCC

In a nutshell :

1. Finding bounds usually requires searching $2^{|\text{Abstained}|}$ values
2. Using the fact that

$$P_a(\mathbf{Y}_{\mathscr{I}_{j-1}}=\mathbf{y}_{\mathscr{I}_{j-1}}) := \prod_{k=1}^{j-1} P(Y_k=\hat{y}_k | Y_j=a),$$

   we can drastically reduce this search by optimizing the terms separately.

Proposition 1

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

Strategy ❶ : Marginalization (IB) with NCC

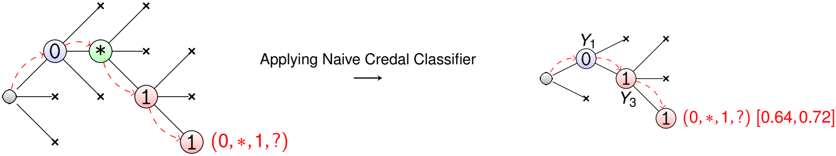
We recall that the conditional upper probability on the (j-1)th first labels is

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

Thanks to NCC, the abstained labels can be removed of the conditioning

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

Graphically, if we use the NCC model to compute $P_{\boldsymbol{x}^*}$, the probabilistic chaining comes down to :



✓ The global time complexity of the MARGINALIZATION strategy is $\mathscr{O}(m)$.

Overview

- Multi-label classification problem
- Multi-label chaining with imprecise probabilities
  - Precise Probabilistic Chaining
  - Imprecise Probabilistic Chaining
    - Imprecise Probabilistic Chaining using NCC model
- Experiments
- Conclusions

Dataset and experimental setting

Material/Imprecise Classifier/Metrics

- The data set issued from MULAN repository.

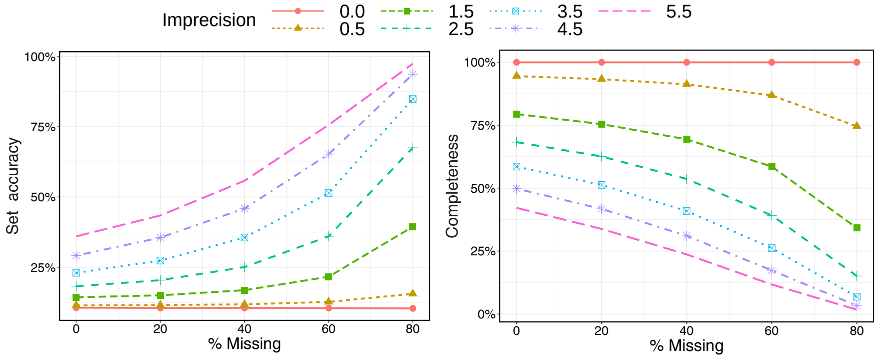
Data set	#Features	#Labels	#Instances	#Cardinality	#Density
emotions	72	6	593	1.90	0.31
⋮	⋮	⋮	⋮	⋮	⋮
yeast	**103**	**14**	**2417**	**4.23**	**0.30**

- Naive credal classifier (NCC) [ZAFFALON 2002]
- Metric evaluations : (Q denotes the set of predicted label s.t. $\widehat{y}_i = 1$ or $\widehat{y}_i = 0$)

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

Missing labels

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	0	*
200.6	30	Blue	58	1	*	0	0
...	...	...	...	...	...	...	...



Imprecise Branching. Evolution of the set-accuracy (left) and the completeness (right) in average (%) for each level of imprecision (a curve for each one), with respect to the % of missingness.

- ✗ The precise model (with imprecision = 0.0) is not really affected by randomly missing labels.
- ✌ However, our proposal provide some level of protection as the number of missing labels increases, although it requires sometime a high amount of imprecision to get the ground-truth solution within the set-valued prediction.

Overview

- Multi-label classification problem
- Multi-label chaining with imprecise probabilities
  - Precise Probabilistic Chaining
  - Imprecise Probabilistic Chaining
    - Imprecise Probabilistic Chaining using NCC model
- Experiments
- Conclusions

Conclusions and Perspective

❶ Works done in this paper :

- We propose two new strategies (IB and MAR) to adapt the chaining multi-label problem to the case of handling imprecise probability estimates.
- We propose efficient procedures to solve such strategies by using the NCC model.

❷ What remains to do

- ✘ How to come up with general but efficient optimisation methods to solve the strategies IB and MAR
- ✘ Investigating the performance of our proposed strategies on other imprecise classifier.


Thank
You!

References

ZAFFALON, Marco (2002). “The naive credal classifier”. In : *Journal of statistical planning and inference* 105.1, p. 5-21.

DE COOMAN, Gert et Filip HERMANS (2008). “Imprecise probability trees : Bridging two theories of imprecise probability”. In : *Artificial Intelligence* 172.11, p. 1400-1427.

READ, Jesse et al. (2011). “Classifier chains for multi-label classification”. In : *Machine learning* 85.3, p. 333.