A first glance at multi-label chaining using imprecise probabilities

ECML/PKDD 2020 Tutorial and Workshop on Uncertainty in Machine Learning

<u>**CARRANZA-ALARCON Yonatan-Carlos**</u>
Ph.D. Candidate in Computer Science

DESTERCKE Sébastien
Ph.D Director

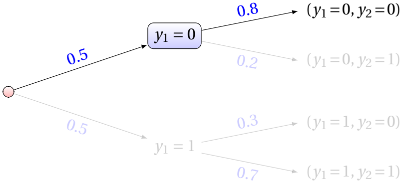
utc Université de Technologie Compiègne
heudiasyc
cnrs dépasser les frontières

18 September 2020

Our approach in a nutshell

What ?

Multi-label chaining using a set of probability models [5] instead of a single probability model.

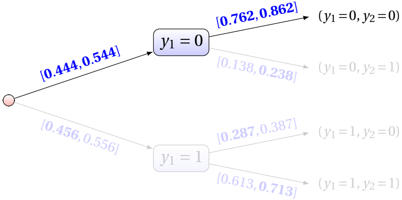


Chaining with precise probabilistic models

$\mathbb{P}$

$P(Y_j | Y_1, \ldots, Y_{j-1}, X)$

widely studied !



Chaining with imprecise probabilistic models

$\mathscr{P}$

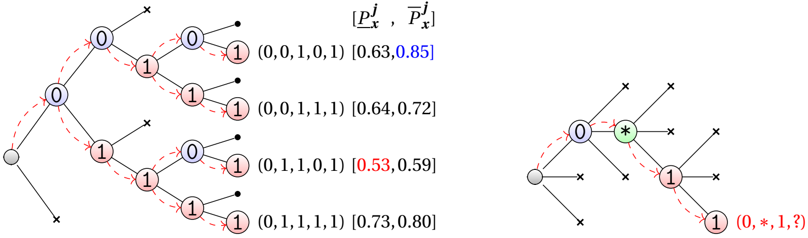
$[\underline{P}(Y_j | Y_1, \ldots, Y_{j-1}, X), \overline{P}(Y_j | Y_1, \ldots, Y_{j-1}, X)]$

how can we do it ?

Our approach in a nutshell

How can doing it ?

- We propose two strategies to get probability bounds $[\underline{P}, \overline{P}]$
  - Imprecise Branching
  - Marginalization



(a) Imprecise Branching

(b) Marginalization ($* = \{0, 1\}$)

Our approach in a nutshell

Why ?

☛ Recognizing hard instances to predict in order to avoid making mistakes → **Making a cautious decision.**

☛ Trying to avoid to propagate unsure predictions in the chaining.

Results ?

☛ Our proposal overcomes precise ones in noisy setting.

☛ Good balance between abstained labels and performance.

Overview

- Introduction to multi-label classification
- Multi-label chaining with imprecise probabilities
- Evaluation
  - Settings and Datasets
  - Experimental results
- Conclusions and Perspectives

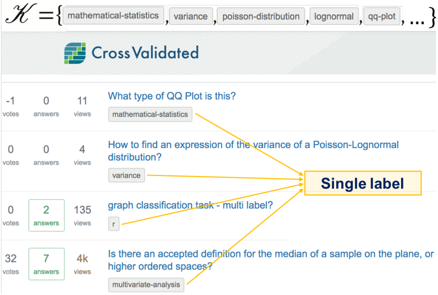
Multi-label classification problem

☞ **Problem statement :**

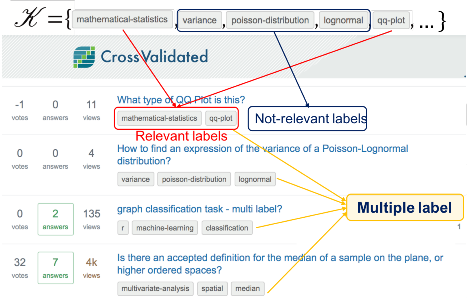
Let $\mathscr{K}=\{m_1,\ldots,m_K\}$ be a set of labels and let $\boldsymbol{x}\in\mathbb{R}^p$ be an unlabelled instance, attribute it a set of relevant labels $\mathscr{S}(\boldsymbol{x})\subseteq\mathscr{K}$.

☞ **Example :**

Classical classification



Multi-label classification



Multi-label classification problem

☞ **The goal of multi-label problem :**

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

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

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

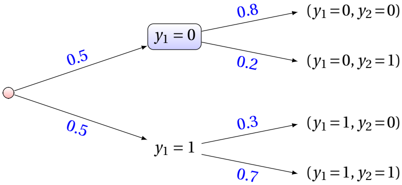
☞ Example of training data :

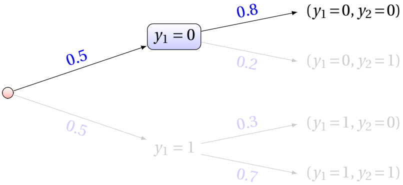
X_1	X_2	X_3	X_4	y_1	y_2	y_3
107.1	25	Blue	60	1	0	0
-50	10	Red	40	1	0	1
200.6	30	Blue	58	0	1	0
...	...	...	...	...	...	...

Multi-label classification problem

✌ **Why we want to use the multi-label chaining.**

✗ Decomposition techniques ignore the label dependencies.
✗ Probabilistic tree chains require to scan all possible predictions.





Overview

- Introduction to multi-label classification
- Multi-label chaining with imprecise probabilities
- Evaluation
  - Settings and Datasets
  - Experimental results
- Conclusions and Perspectives

Basic notations

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

$$P^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j=1) := P_{\boldsymbol{x}}(Y_j=1|Y_{\mathscr{I}^{j-1}_{\mathscr{R}}}=1, Y_{\mathscr{I}^{j-1}_{\mathscr{I}}}=0) \quad (1)$$

where $\mathscr{I}^j_*$ is the set of indices of the labels among the j first predicted as

1. (relevant labels) $\mathscr{I}^j_{\mathscr{R}} \subseteq [\![j]\!]$ [1], $\forall i \in \mathscr{I}^j_{\mathscr{R}},\ y_i = 1$,
2. (irrelevant labels) $\mathscr{I}^j_{\mathscr{I}} \subseteq [\![j]\!], \mathscr{I}^j_{\mathscr{I}} \cap \mathscr{I}^j_{\mathscr{R}} = \emptyset$, $\forall i \in \mathscr{I}^j_{\mathscr{I}},\ y_i = 0$,
3. (abstained labels) $\mathscr{I}^j_{\mathscr{A}} = [\![j]\!] \backslash (\mathscr{I}^j_{\mathscr{R}} \cup \mathscr{I}^j_{\mathscr{I}})$, $\forall i \in \mathscr{I}^j_{\mathscr{A}},\ y_i = \{0,1\} := *$.

---

1. $[\![j]\!] = \{1,\ldots,j\}$ set of the first j integers

(Precise) Multi-label chaining

☞ Learning a multi-label chaining

❶ Learning a binary classifier at each step of the chaining [3] :

$$\varphi_i : \mathbb{R}^p \times \{0,1\}^{i \leq m} \rightarrow \{0,1\}$$

❷ Decision step under a binary classifier $\ell(y_j, \hat{y}_j) \rightarrow$

"Optimal" decision [4] : $\varphi_i := \hat{y}_j = \begin{cases} 1 & P^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j = 1) \geq 0.5 \\ 0 & otherwise \end{cases}$

☞ An example of multi-label chaining

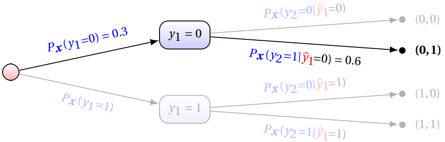


FIGURE – *Precise multi-label chaining with two labels.*

(Imprecise) Multi-label chaining

☞ **Learning a multi-label chaining using imprecise probabilities**

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

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

❷ Making a cautious decision

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

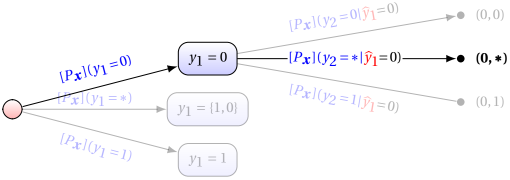


FIGURE – *An example of multi-label chaining using imprecise … babilities*

Strategy ❶ : Imprecise branching

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

$$\underline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j = 1) = \min_{\mathbf{y}\in\{0,1\}^{|\mathscr{I}_\mathscr{A}|}} \underline{P}_{\boldsymbol{x}}(Y_j = 1 | Y_{\mathscr{I}^{j-1}_\mathscr{R}} = 1, Y_{\mathscr{I}^{j-1}_\mathscr{I}} = 0, Y_{\mathscr{I}^{j-1}_\mathscr{A}} = \mathbf{y}),$$

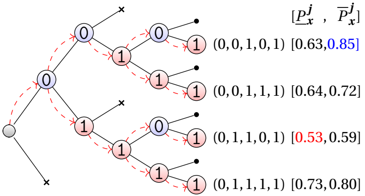
$$\overline{P}^{\boldsymbol{j}}_{\boldsymbol{x}}(Y_j = 1) = \max_{\mathbf{y}\in\{0,1\}^{|\mathscr{I}_\mathscr{A}|}} \overline{P}_{\boldsymbol{x}}(Y_j = 1 | Y_{\mathscr{I}^{j-1}_\mathscr{R}} = 1, Y_{\mathscr{I}^{j-1}_\mathscr{I}} = 0, Y_{\mathscr{I}^{j-1}_\mathscr{A}} = \mathbf{y}).$$

(IB)

Example :

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

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



Strategy ❷ : Marginalization

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

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

where $\mathscr{P}^*$ is the set of joint probability distributions described by the imprecise probabilistic tree [2].

Strategy ❷ : Marginalization

☞ An example with four labels.

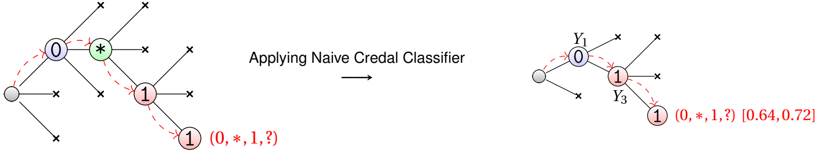
$$\underline{P}_{\boldsymbol{x}}^{4}(Y_4 = 1) = \min_{P_{\boldsymbol{x}}^{4} \in \mathscr{P}^*} P_{\boldsymbol{x}}(Y_4{=}1 | Y_1{=}0, (Y_2{=}0 \cup Y_2{=}1), Y_3{=}1)$$

$$= \min_{P_{\boldsymbol{x}}^{4} \in \mathscr{P}^*} \frac{\sum_{y_2 \in \{0,1\}} P_{\boldsymbol{x}}(Y_4{=}1, Y_1{=}0, Y_2{=}y_2, Y_3{=}1)}{\sum_{y_2 \in \{0,1\}} P_{\boldsymbol{x}}(Y_1{=}0, Y_2{=}y_2, Y_3{=}1)}$$

$$= \min_{P_{\boldsymbol{x}}^{4} \in \mathscr{P}^*} P_{\boldsymbol{x}}(Y_4{=}1 | Y_1{=}0, Y_3{=}1).$$

✗ The optimization problem can be tricky.

✓ But, we propose to use NCC classifier to compute $P_{\boldsymbol{x}}$



Overview

- Introduction to multi-label classification
- Multi-label chaining with imprecise probabilities
- Evaluation
  - Settings and Datasets
  - Experimental results
- Conclusions and Perspectives

Datasets and experimental setting

Material and method

- 3 data sets issued from MULAN repository.

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

- 10×10-fold cross-validation procedure.
- Naive imprecise classifier (NCC) [1]
- Applying minimax strategy to compare precise approaches.

e.g. $(0,*,1) \overset{minimax}{\longrightarrow} (0,1,1)$

Missing and Noise labels

1. **Missing (miss)** $Y_{j,i} = 0 \wedge 1 \longrightarrow Y_{j,i} = *$.
2. **Noise**
   - 2.1 **Reversing (rev)** $Y_{j,i} = 1 \longrightarrow Y_{j,i} = 0$ or $Y_{j,i} = 0 \longrightarrow Y_{j,i} = 1$).
   - 2.2 **Flipping (flip)** $Y_{j,i} \sim \mathscr{B}er(\beta), \beta := P(Y_{j,i} = 1),\ \beta \in \{0.2, 0.8\}$.

Experimental results

TABLE – 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**

Experimental results

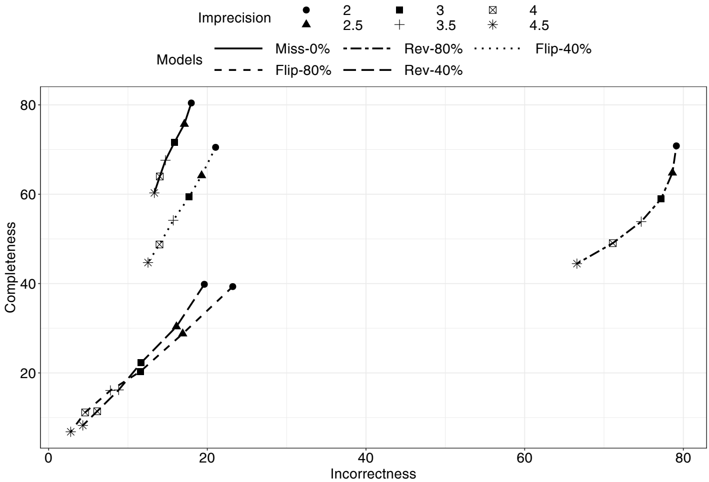


FIGURE – Evolution of the incorrectness and completeness for the imprecise branching strategy and Emotion dataset.

Overview

- Introduction to multi-label classification
- Multi-label chaining with imprecise probabilities
- Evaluation
  - Settings and Datasets
  - Experimental results
- Conclusions and Perspectives

Conclusions and Perspectives

- ✓ We propose two new innovative strategies to treat the multi-label chaining under uncertainty.
- ✓ Our proposal overcomes those precise ones in the noise setting.
- ✗ How to come up with general but efficient optimisation methods to solve Equations (IB) and (MAR).
- ✗ Investigating the performance of our proposed strategies on other imprecise classifier (eg. continuous classifier).


Thank
You!

References

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

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

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

Krzysztof DEMBCZYŃSKI et al. "On label dependence and loss minimization in multi-label classification". In : *Machine Learning* 88.1-2 (2012), p. 5-45.

Thomas AUGUSTIN et al. *Introduction to imprecise probabilities*. John Wiley & Sons, 2014.