Distributionally robust, skeptical binary inferences in multi-label problems

12th International Symposium on Imprecise Probabilities: Theories and Applications

CARRANZA-ALARCON Yonatan-Carlos
Ph.D. in Computer Science

DESTERCKE Sébastien
Ph.D. in Computer Science

06 to 09 July 2021

Overview

- Multi-label classification problem
- Cautious inferences in multi-label problems
  - General case for the Hamming loss
  - Experimental results
- Conclusions and Perspectives

Multi-label classification problem

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

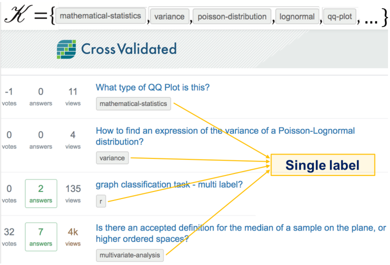
$$\text{Given a training data : } \mathscr{D} = \{\mathbf{x}^i, \mathbf{y}^i\}_{i=0}^{N} \subseteq \mathbb{R}^p \times \mathscr{Y}$$

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

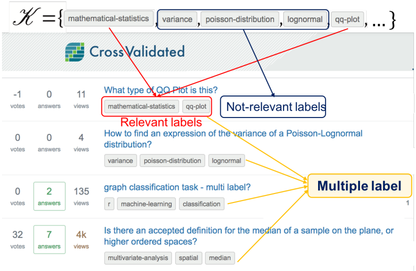
Learning a multi-label classification rule : $\varphi : \mathbb{R}^p \rightarrow \mathscr{Y}$

☞ **Example :**

Classical classification



Multi-label classification



Existing results for precise and cautious[1] inferences

- Inference in precise case difficult, but there are
  - ✓ Efficient algorithms for specific losses [**...** ; DEMBCZYŃSKI et al. 2012 ; WAEGEMAN et al. 2014]
  - ✓ Several simplified learning model : Binary relevance, Classifier chains [READ et al. 2019], ...

- This issue is poorly explored in IP [DESTERCKE 2015 ; ANTONUCCI et al. 2017], and even less in other cautious settings [NGUYEN et al. 2019 ; PILLAI et al. 2013].

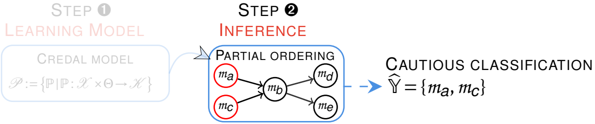
- Our contribution consists in providing :
  - More efficient, dedicated algorithm for the Hamming Loss under the maximality criterion.
  - Polynomial-time inference on restricted credal sets $\mathscr{P}_{PR}$ (imprecise Binary relevance).

1. Cautious and Skeptical are here used interchangeably.

Overview

- Multi-label classification problem
- Cautious inferences in multi-label problems
  - General case for the Hamming loss
  - Experimental results
- Conclusions and Perspectives

Cautious inferences in form of set-valued solutions



☞ **Problem setting and challenges :**

✓ Step ❶ : The uncertainty model $\mathscr{P}$ is known.

✗ Step ❷ : Under the maximality criterion and a generic loss matrix

$\Longrightarrow$ the set-valued solutions require at worst $2^m(2^m - 1) \propto 2^{2m}$ computations.

☞ **Example :** $|\mathscr{Y}| = 10$, it needs to $2^{10}(2^{10} - 1) = 1047552$ computations.

$(0,0,\ldots,0) \succ_{\ell}^{\mathscr{P}} (0,\ldots,1,1)$?

$(0,0,\ldots,0) \succ_{\ell}^{\mathscr{P}} (0,\ldots,1,0)$?

$\vdots \quad \vdots \quad \vdots \quad \vdots$

⇨

A set-valued solution

$$\widehat{\mathbb{Y}}^{M}_{\ell,\mathscr{P}} = \begin{pmatrix} 1 & 1 & 0 & \ldots & 0 \\ 0 & 1 & 1 & \ldots & 0 \\ 0 & 0 & 0 & \ldots & 0 \end{pmatrix}$$

✍ **Can we obtain cautious predictions efficiently ?**

General case for the Hamming case

Proposition 1 (Ceteris paribus comparison)

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

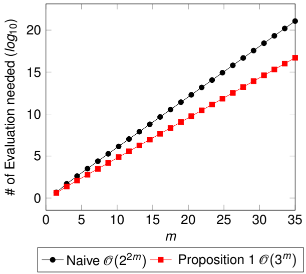
$$\boldsymbol{y}^1 \succ_{\ell}^{\mathscr{P}} \boldsymbol{y}^2 \iff \inf_{P\in\mathscr{P}} \sum_{i\in\mathscr{I}} P(Y_i = a_i) > \frac{|\mathscr{I}|}{2} \quad (1)$$

Prop. 1 amounts to focus on partial binary vector, e.g. $|\mathscr{Y}| = n+3, \boldsymbol{a} = (0,0,0,*,\dots,*)$

$$(0,0,0,\underbrace{*,\dots,*}_{\text{n labels}}) \succ_{\ell_H}^{\mathscr{P}} (1,1,1,\underbrace{*,\dots,*}_{\text{n labels}})$$

1 comparaison instead of 2^n.

✓ We can reduce : $\mathscr{O}(2^{2m}) \longrightarrow \mathscr{O}(3^m)$



Existing approximate results for Hamming loss

- The partial vector $\widehat{\boldsymbol{y}}^* = (\widehat{y}_1^*, \ldots, \widehat{y}_m^*) \in \mathfrak{Y} = \{0, 1, *\}$

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

  is an outer-approximation of $\widehat{\mathbb{Y}}^M_{\ell_H, \mathscr{P}}$ [DESTERCKE 2015]

- Only requires to know imprecise marginal bounds $\mathscr{P}_{Y_i}$ on each label.

- Note that not all cautious multi-label predictions can be exactly represented as a partial vector

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix} \Longrightarrow \text{cannot be represented in } \mathfrak{Y}$$

$$\begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} = (*, *, 0) \in \mathfrak{Y}$$

Exact $\widehat{\mathbb{Y}}^{M}_{\ell_H,\mathscr{P}}$ vs. $\widehat{\boldsymbol{y}}^*$ outer-approximation inferences

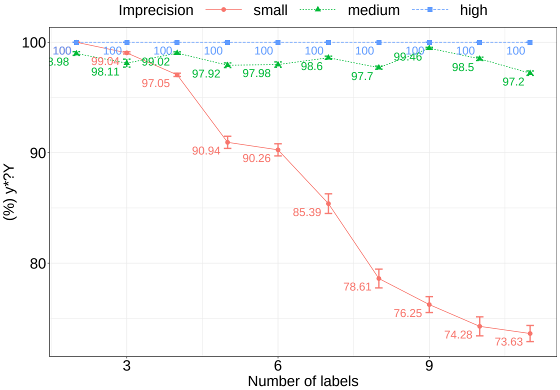


FIGURE – % of instances where $\widehat{\boldsymbol{y}}^* = \widehat{\mathbb{Y}}^{M}_{\ell_H,\mathscr{P}}$.

✓ The quality of $\widehat{\boldsymbol{y}}^*$ decreases as the number of labels increases.

✓ The quality of $\widehat{\boldsymbol{y}}^*$ seems to be the worst for moderate imprecision.

Exact $\widehat{\mathbb{Y}}^{M}_{\ell_H,\mathscr{P}}$ vs. $\widehat{y}^*$ outer-approximation inferences

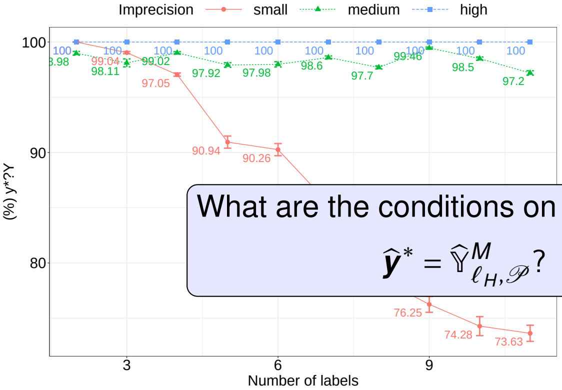


FIGURE – % of instances ... M ... $\ell_{H,\mathscr{P}}$.

What are the conditions on $\mathscr{P}$ ensuring

$$\widehat{\boldsymbol{y}}^* = \widehat{\mathbb{Y}}^{M}_{\ell_H,\mathscr{P}}?$$

- ✓ The quality of $\widehat{\boldsymbol{y}}^*$ decreases as the number of labels increases.
- ✓ The quality of $\widehat{\boldsymbol{y}}^*$ seems to be the worst for moderate imprecision.

Binary relevance and partial vectors

Under the assumption of label independence :

$$\mathscr{P}_{BR} := \left\{ \prod_{\{i|y_i=1\}} p_i \prod_{\{i|y_i=0\}} (1-p_i) \middle| p_i \in [\underline{p}_i, \overline{p}_i] \right\}.$$

Proposition 2 (Domain restriction on $\mathscr{P}$)

Given a probability set $\mathscr{P}_{BR}$ and the Hamming loss, $\widehat{\mathbb{Y}}^{M}_{\ell_H,\mathscr{P}_{BR}} \in \mathfrak{Y}$.

- ✓ $\widehat{\mathbb{Y}}^{M}_{\ell_H,\mathscr{P}_{BR}}$ can be represented as partial vector $\mathfrak{Y}$.
- ✓ $\widehat{\mathbb{Y}}^{M}_{\ell_H,\mathscr{P}_{BR}}$ is equal to outer-approximation $\hat{\boldsymbol{y}}^*$ [Destercke 2015].
- ✓ The time complexity becomes linear on m, i.e. $\mathscr{O}(m)$!

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
scene	**294**	**6**	**2407**	**1.07**	**0.18**

☞ Naive credal classifier (NCC) [ZAFFALON 2002] for each marginal credal $\mathscr{P}_{Y_i}$.

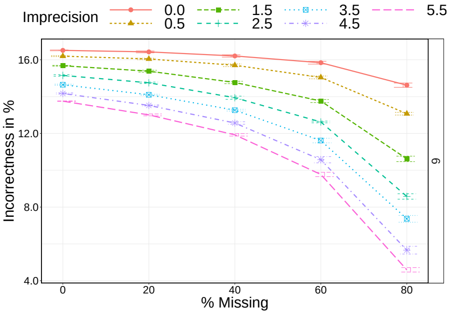
☞ Metric evaluations : (Q denotes the set of predicted label s.t. $\widehat{y}_i = 1$ or $\widehat{y}_i = 0$)

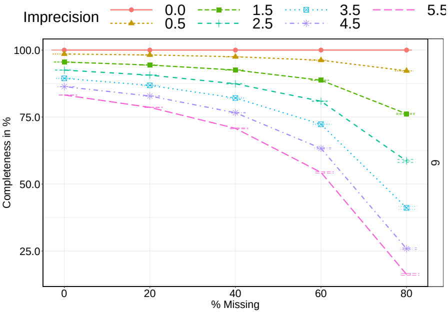
$$IC(\widehat{\mathbb{Y}}, \boldsymbol{y}) = \frac{1}{|Q|} \sum_{\widehat{y}_i \in Q} 1_{(\widehat{y}_i \neq y_i)} \quad \text{and} \quad CP(\widehat{\mathbb{Y}}, \boldsymbol{y}) = \frac{|Q|}{m}$$

Missing labels

We uniformly pick at random a percentage of missing labels $Y_{i,j}$ (the jth label of the ith instance) which are then removed from the training data, i.e. $Y_{i,j} = 1 \wedge 0 \longrightarrow Y_{i,j} = *$

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
…	…	…	…	…	…	…	…





Evolution of the incorrectness (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 ($s = 0.0$) is not really affected by randomly missing labels.
- ✌ Our proposal, however, becomes more cautious as the number of missing labels increases.

Overview

- Multi-label classification problem
- Cautious inferences in multi-label problems
  - General case for the Hamming loss
  - Experimental results
- Conclusions and Perspectives

Conclusions and Perspective

❶ Works done in this paper :

- Provide efficient algorithmic procedures to solve the maximality criterion under Hamming loss and generic probability sets.
- When considering sets of distributions and cautious inferences, it is not sufficient to consider marginal probabilities to get exact set-valued predictions, as opposed to the case of precise distributions.

❷ What remains to do

- ✘ Compare our proposal against those rejecting and abstaining approaches.
- ✘ Solve the maximality criterion using other loss functions, e.g. ; ranking loss, Jaccard loss, F-measure, and so on.


Thank You!

References

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

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

PILLAI, Ignazio, Giorgio FUMERA et Fabio ROLI (2013). "Multi-label classification with a reject option". In : *Pattern Recognition* 46.8, p. 2256-2266.

WAEGEMAN, Willem et al. (2014). "On the bayes-optimality of f-measure maximizers". In : *Journal of Machine Learning Research* 15, p. 3333-3388.

DESTERCKE, Sébastien (2015). "Multilabel predictions with sets of probabilities : the Hamming and ranking loss cases". In : *Pattern Recognition* 48.11, p. 3757-3765.

ANTONUCCI, Alessandro et Giorgio CORANI (2017). "The multilabel naive credal classifier". In : *International Journal of Approximate Reasoning* 83, p. 320-336.

NGUYEN, Vu-Linh et Eyke HÜLLERMEIER (2019). "Reliable Multi-label Classification : Prediction with Partial Abstention". In : *CoRR* abs/1904.09235. arXiv : `1904.09235`. URL : `http://arxiv.org/abs/1904.09235`.

READ, Jesse et al. (2019). "Classifier Chains : A Review and Perspectives". In : *arXiv preprint arXiv :1912.13405*.