PRIME IMPLICANTS AS A VERSATILE TOOL TO EXPLAIN ROBUST CLASSIFICATION

Hénoïk Willot; Sébastien Destercke; Khaled Belahcene
{henoik.willot, sebastien.destercke, khaled.belahcene}@hds.utc.fr

13th International Symposium on Imprecise Probabilities: Theories and Applications

Zoo data example

Objective : predict class $\mathbf{y}$ in $\mathcal{Y} = \{$ Mammal (M), Bird (B), Reptile (R), Fish (F) or Invertebrate (I) $\}$ from the following n features

Feature	feathers	eggs	aquatic	toothed	backbone	breathes with lungs	venomous	fins	legs	tail
Domain	{🪶, ❌}	{🥚, ❌}	{🌊, ❌}	{🦷, ❌}	{🦴, ❌}	{🫁, ❌}	{☣️, ❌}	{🐟, ❌}	{❌, 2 ∗ 🦵, 4 ∗ 🦵, 5 ∗ 🦵, 6 ∗ 🦵, 8 ∗ 🦵}	{〰️ , ❌}

If the value is left unspecified, we use the question mark symbol as for instance ❓ for feathers

Observation : 🦈 = (❌, ❌, 🌊, 🦷, 🦴, 🫁, ☣️, 🐟, ❌, ❌)

Crisp Case

A class y dominates a class $\mathbf{y}'$ given a probability distribution p

$$\mathbf{y} \succeq_{p,(\mathbf{x})} \mathbf{y}' \text{ if } p(\mathbf{y}|\mathbf{x}) \geq p(\mathbf{y}'|\mathbf{x})$$

A Validatory Prime Implicant (PI) of $\mathbf{y} \succeq_{p,(\mathbf{x})} \mathbf{y}'$ is a subset E of features whose value in the observation is sufficient to obtain dominance no matter the other values

Example : $M \succeq_{p,(🦈)} B$, $E = \{feathers, eggs\}$ is a Validatory Prime Implicant because $M \succeq_{p,(🦈_E, ❓_{-E})} B$ with $(🦈_E, ❓_{-E}) = (❌, ❌, ❓, ❓, ❓, ❓, ❓, ❓, ❓)$

A Contrastive Prime Implicant (PI) of $\mathbf{y} \succeq_{p,(\mathbf{x})} \mathbf{y}'$ is a subset E of features such that changing the observation on these values is sufficient to obtain a reversed dominance, while keeping other values untouched

Example : $M \succeq_{p,(🦈)} B$, $E = \{feathers, eggs, legs\}$ is a Constrastive Prime Implicant because $B \succeq_{p,(🦈_{-E}, (🪶, 🥚, 2 ∗ 🦵))} M$ *i.e.* (🪶, 🥚, 🌊, 🦷, 🦴, 🫁, ☣️, 🐟, 2 ∗ 🦵, ❌) is classified as a Bird

Robust prediction with credal sets

Credal set: Probability distribution p replaced by convex sets of probabilities $\mathcal{P}$

Robust classification : Necessary recommendation $\mathbf{y} \succ_{\mathcal{P},(\mathbf{x})} \mathbf{y}'$:

$$\mathbf{y} \succ_{\mathcal{P},(\mathbf{x})} \mathbf{y}' \Leftrightarrow \forall p \in \mathcal{P},\ p(\mathbf{y}|\mathbf{x}) \geq p(\mathbf{y}'|\mathbf{x}) \Leftrightarrow \inf_{p \in \mathcal{P}} \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}'|\mathbf{x})} \geq 1$$

Incomparability : When neither classes is necessarily better, incomparability $\mathbf{y} \succ\prec_{\mathcal{P},(\mathbf{x})} \mathbf{y}'$ arises :

$$\exists p \in \mathcal{P}\ \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}'|\mathbf{x})} < 1 \ \textit{and}\ \exists p' \in \mathcal{P} \frac{p'(\mathbf{y}'|\mathbf{x})}{p'(\mathbf{y}|\mathbf{x})} < 1$$
$$\Leftrightarrow \inf_{p \in \mathcal{P}} \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}'|\mathbf{x})} < 1 \ \textit{and}\ \inf_{p' \in \mathcal{P}} \frac{p'(\mathbf{y}'|\mathbf{x})}{p'(\mathbf{y}|\mathbf{x})} < 1$$

Validatory PI

$$\forall x_{-E} \in \mathcal{X}^{-E}\ \inf_{p \in \mathcal{P}} \frac{p(\mathbf{y}|(\mathbf{x}_E, x_{-E}))}{p(\mathbf{y}'|(\mathbf{x}_E, x_{-E}))} \geq 1$$
$$\Leftrightarrow \phi^v_{\mathbf{y},\mathbf{y}'}(E) = \inf_{\substack{p \in \mathcal{P} \\ x_{-E} \in \mathcal{X}^{-E}}} \frac{p(\mathbf{y}|(\mathbf{x}_E, x_{-E}))}{p(\mathbf{y}'|(\mathbf{x}_E, x_{-E}))} \geq 1$$

Contrastive PI

$$\exists x_{-E} \in \mathcal{X}^{-E}\ \inf_{p \in \mathcal{P}} \frac{p(\mathbf{y}|(x_E, \mathbf{x}_{-E}))}{p(\mathbf{y}'|(x_E, \mathbf{x}_{-E}))} < 1$$
$$\Leftrightarrow \phi^c_{\mathbf{y},\mathbf{y}'}(E) = \inf_{\substack{p \in \mathcal{P} \\ x_E \in \mathcal{X}^E}} \frac{p(\mathbf{y}|(x_E, \mathbf{x}_{-E}))}{p(\mathbf{y}'|(x_E, \mathbf{x}_{-E}))} < 1$$

Doubt PI

$$\forall x_{-E} \in \mathcal{X}^{-E}\ \inf_{p \in \mathcal{P}} \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}'|\mathbf{x})} < 1 \ \textit{and}\ \inf_{p' \in \mathcal{P}} \frac{p'(\mathbf{y}'|\mathbf{x})}{p'(\mathbf{y}|\mathbf{x})} < 1$$
$$\Leftrightarrow \phi^d_{\mathbf{y},\mathbf{y}'}(E) = \sup_{x_{-E} \in \mathcal{X}^{-E}} \inf_{p \in \mathcal{P}} \frac{p(\mathbf{y}|(\mathbf{x}_E, x_{-E}))}{p(\mathbf{y}'|(\mathbf{x}_E, x_{-E}))} < 1$$
$$\textit{and}\ \phi^d_{\mathbf{y}',\mathbf{y}}(E) = \sup_{x_{-E} \in \mathcal{X}^{-E}} \inf_{p' \in \mathcal{P}} \frac{p'(\mathbf{y}'|(\mathbf{x}_E, x_{-E}))}{p'(\mathbf{y}|(\mathbf{x}_E, x_{-E}))} < 1$$

Application with NCC [1, 2]

Naive Bayes leads to $\frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}'|\mathbf{x})} = \frac{p_\mathcal{Y}(\mathbf{y})}{p_\mathcal{Y}(\mathbf{y}')} \prod_{i \in N} \frac{p_i(\mathbf{x}_i|\mathbf{y})}{p_i(\mathbf{x}_i|\mathbf{y}')}$ and with **credal sets** for all $p_i(\cdot|\mathbf{y}')$ distributions we obtain $\inf_{p \in \mathcal{P}} \frac{p(\mathbf{y}|\mathbf{x})}{p(\mathbf{y}'|\mathbf{x})} = \frac{\underline{p}_\mathcal{Y}(\mathbf{y})}{\overline{p}_\mathcal{Y}(\mathbf{y}')} \prod_{i \in N} \frac{\underline{p}_i(\mathbf{x}_i|\mathbf{y})}{\overline{p}_i(\mathbf{x}_i|\mathbf{y}')}$ with $\underline{p}$ and $\overline{p}$ lower/upper bounds of $p \in \mathcal{P}$

Validatory + Contrastive PI

$$\mathbf{x}^{v:\mathbf{y},\mathbf{y}'} = \mathbf{x}^{c:\mathbf{y},\mathbf{y}'} = \times_{i=1}^n \arg \inf_{x_i^a \in \mathcal{X}_i} \frac{\underline{p}_i(x_i^a|\mathbf{y})}{\overline{p}_i(x_i^a|\mathbf{y}')}$$

$$\forall E \subset N\ \log \phi^v_{\mathbf{y},\mathbf{y}'}(E \cup \{i\}) - \log \phi^v_{\mathbf{y},\mathbf{y}'}(E) = G^v_{\mathbf{y},\mathbf{y}'}(i)$$
$$\Rightarrow \log\ \phi^v_{\mathbf{y},\mathbf{y}'}(E) = \log\ \phi^v_{\mathbf{y},\mathbf{y}'}(\emptyset) + \sum_{i \in E} G^v_{\mathbf{y},\mathbf{y}'}(i) \geq 0$$

We have exactly the same properties for $\phi^c_{\mathbf{y},\mathbf{y}'}$.
It is *item selection* problem [3]

$$G^v_{\mathbf{y},\mathbf{y}'}(i) = \left(\log \frac{\underline{p}_i(\mathbf{x}_i|\mathbf{y})}{\overline{p}_i(\mathbf{x}_i|\mathbf{y}')}\right) - \left(\log \frac{\underline{p}_i(\mathbf{x}_i^{v:\mathbf{y},\mathbf{y}'}|\mathbf{y})}{\overline{p}_i(\mathbf{x}_i^{v:\mathbf{y},\mathbf{y}'}|\mathbf{y}')}\right)$$

$$G^c_{\mathbf{y},\mathbf{y}'}(i) = \left(\log \frac{\underline{p}_i(\mathbf{x}_i^{c:\mathbf{y},\mathbf{y}'}|\mathbf{y})}{\overline{p}_i(\mathbf{x}_i^{c:\mathbf{y},\mathbf{y}'}|\mathbf{y}')}\right) - \left(\log \frac{\underline{p}_i(\mathbf{x}_i|\mathbf{y})}{\overline{p}_i(\mathbf{x}_i|\mathbf{y}')}\right)$$

Example

Worst opponent for $M \succ_{\mathcal{P},(🦈)} B$ and $M \succ\prec_{\mathcal{P},(🦈)} R$

$\mathbf{x}^{v:M,B} = (🪶, 🥚, 🌊, ❌, ❌, ❌, ☣️, 🐟, 5 ∗ 🦵, 〰️)$
$\mathbf{x}^{d:M,R} = (❌, 🥚, 🌊, 🦷, 🦴, ❌, ☣️, 🐟, 4 ∗ 🦵, 〰️)$
$\mathbf{x}^{d:R,M} = (❌, 🥚, 🌊, 🦷, 🦴, ❌, ☣️, 🐟, ❌, 〰️)$

	🪶	🥚	🌊	🦷	❓	∅
$G^v_{M,B}$	2.64	2.29	0	2.29	.515	-6.77 | 3.27
$G^d_{M,R}$	0	0	-0.472	0	0	1.58
$G^d_{R,M}$	0	-1.49	0	-0.61	0	2.45
	🫁	☣️	🐟	❓	❓	
$G^v_{M,B}$	.515	.515	.187	.91	.187	
$G^d_{M,R}$	0	0	-0.219	-1.31	-0.219	
$G^d_{R,M}$	-0.809	-1.25	-0.707	0	-0.707	

Doubt PI

(α, β) is either $(\mathbf{y}, \mathbf{y}')$ either $(\mathbf{y}', \mathbf{y})$

$$\mathbf{x}^{d:\alpha,\beta} = \times_{i=1}^n \arg \sup_{x_i^a \in \mathcal{X}_i} \frac{\underline{p}_i(x_i^a|\alpha)}{\overline{p}_i(x_i^a|\beta)}$$

Like Validatory or Contrastive PI, we obtain :

$$\log\ \phi^d_{\alpha,\beta}(E) = \log\ \phi^d_{\alpha,\beta}(\emptyset) + \sum_{i \in E} G^d_{\alpha,\beta}(i) < 0$$

$$G^d_{\alpha,\beta}(i) = \left(\log \frac{\underline{p}_i(\mathbf{x}_i|\alpha)}{\overline{p}_i(\mathbf{x}_i|\beta)}\right) - \left(\log \frac{\underline{p}_i(\mathbf{x}_i^{d:\alpha,\beta}|\alpha)}{\overline{p}_i(\mathbf{x}_i^{d:\alpha,\beta}|\beta)}\right)$$

Adding feature i in E changes both functions $\phi^d_{\mathbf{y},\mathbf{y}'}$ and $\phi^d_{\mathbf{y}',\mathbf{y}}$ by adding in both a positive value, while both functions are bounded by 0.
It is a *2-dimentional knapsack* problem !

Conclusion

Summary :

- Prime implicants for robust preferences
- Application to the NCC with convex domains
- Polynomial calculation of implicants

Perspectives :

- Pairwise or holistic explanations ?
- Implications on complexity to remove the independence hypothesis ?
- Refine explanations for incomparability (random or epistemic) ?

References

[1] Marco Zaffalon. "The naive credal classifier". In: *Journal of Statistical Planning and Inference* 105.1 (2002), pp. 5–21.

[2] Jean-Marc Bernard. "An introduction to the imprecise Dirichlet model for multinomial data". In: *International Journal of Approximate Reasoning* 39.2-3 (2005), pp. 123–150.

[3] João Marques-Silva et al. "Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay". In: *NeurIPS 2020, December 6-12, 2020, virtual*. 2020.