# Explications de classifications robustes a l'aide d'implicants premiers

Hénoïk Willot, Khaled Belahcene, Sébastien Destercke

# Explications de classifications robustes à l'aide d'implicants premiers

Hénoïk Willot Khaled Belahcene Sébastien Destercke
Heudiasyc, Université de Technologie de Compiègne, France

{henoik.willot, khaled.belahcene, sebastien.destercke}@hds.utc.fr

**Résumé :**

Dans ce papier, nous étudions à quel point la notion d'implicant premier peut expliquer les résultats d'une classification robuste, en s'intéressant particulièrement à expliquer des relations de dominances par paires. Nous sous-entendons par robuste des modèles imprécis qui peuvent s'abstenir de classifier ou de comparer deux classes quand l'information disponible est insuffisante. Cela se reflète en considérant des ensembles (convexes) de probabilités. Par implicants premiers nous sous-entendons un ensemble minimal d'attributs dont les valeurs doivent être connues avant de décider qu'une classe domine/est préférée à une autre.

**Mots-clés :**

Classification robuste, Explicabilité, Implicants premiers, Probabilités imprécises, Naive credal classifier

**Abstract:**

In this paper, we investigate how robust classification results can be explained by the notion of prime implicants, focusing on explaining pairwise dominance relations. By robust, we mean that we consider imprecise models that may abstain to classify or to compare two classes when information is insufficient. This will be reflected by considering (convex) sets of probabilities. By prime implicants, we understand a minimal number of attributes whose value needs to be known before stating that one class dominates/is preferred to another.

**Keywords:**

Robust classifier, Explainability, Prime implicants, Imprecise probabilities, Naive credal classifier

## 1 Introduction

Deux aspects importants de l'IA de confiance sont la capacité d'effectuer des inférences ou prédictions robustes et sûres, et la capacité à les accompagner d'une explication sur les raisons de ces dernières.

Du coté de l'explicabilité, la notion d'implicant premier revient à fournir une condition suffisante minimale permettant de faire une certaine prédiction, par exemple les attributs qui ont besoin d'être instanciés pour faire une classification. Ils ont été proposés comme composants de l'explication pour un large ensemble de modèles tels que les modèles graphiques [10], avec des procédures efficaces existantes pour des structures spécifiques telle que la structure Naïve [9]. Comparé à d'autres méthodes comme SHAP [11] qui essaye de calculer l'influence moyenne des attributs, les implicants premiers ont l'avantage d'être rattachés à la logique et de fournir un certificat d'explication (dans le sens où les attributs identifiés forment une raison logique et suffisante).

Cependant, les outils d'IA explicable ont souvent, si ce n'est exclusivement, été appliqués à des modèles précis, au moins dans le domaine du *machine learning* (c'est moins vrai par exemple en représentation des connaissances [4]). Cependant, dans certaines applications impliquant des problèmes sensibles ou quand le *decision maker* veut identifier des cas ambiguës, il serait sûrement préférable d'utiliser des modèles qui vont renvoyer un ensemble de classes dans certains cas où l'information est manquante plutôt que de toujours renvoyer une prédiction précise. Plusieurs cadres comme la prédiction conformelle [3], les *indeterminate classifiers* [7] ou les modèles de probabilités imprécises [6] ont été proposés pour s'occuper de tels cas.

Ces derniers ont l'avantage d'être des extensions et généralisations directes des classifieurs probabilistes, donc nous pouvons directement essayer de transférer des principes d'explications fondés existants pour la classification probabiliste précise dans ce cadre. C'est ce que nous avons l'intention de faire dans ce papier avec les implicants premiers.

Nous commençons par introduire comment

cette idée d'implicant premier peut s'adapter à de la classification utilisant des ensembles de probabilité comme modèle d'incertitude en Section 2. Comme le problème formulé est vraisemblablement difficile à résoudre pour des modèles génériques, nous nous intéressons en Section 3 au classifieur crédal naïf, qui généralise le classifieur bayesien naïf. Nous montrons que pour un tel modèle, calculer et énumérer les implicants premiers peut être effectué en temps polynomial, grâce à son hypothèse d'indépendance et ses propriétés de décomposition. Nous finissons avec un exemple illustratif de notre approche.

# 2 Formulation générale du problème

Dans cette section, nous posons nos notations et donnons les rappels nécessaires à propos des probabilités imprécises. Nous introduisons aussi l'idée des implicants premiers appliqués aux classifieurs, et tout particulièrement aux classifieurs utilisant les probabilités imprécises.

## 2.1 Classification robuste : cadre

Nous nous plaçons dans un problème multi-classes discret classique où nous devons prédire une variable $Y$ prenant ses valeurs dans $\mathcal{Y} = \{y_1, \ldots, y_m\}$ en utilisant $n$ variables d'entrée $X_1, \ldots, X_n$ qui prennent respéctivement leurs valeurs dans $\mathcal{X}_i = \{x_i^1, \ldots, x_i^{k_i}\}$. Nous notons $\mathcal{X} = \times_{i=1}^n \mathcal{X}_i$ et $\mathbf{x} \in \mathcal{X}$ un vecteur de cet espace. Quand nous considérons un sous-ensemble $E \subseteq \{1, \ldots, n\}$ des critères, nous notons par $\mathcal{X}_E = \times^{i \in E} \mathcal{X}_i$ le domaine correspondant, et par $\mathbf{x}_E$ les valeurs d'un vecteur sur ce sous-domaine. Nous notons aussi par $-E := \{1, \ldots, n\} \setminus E$ toutes les dimensions qui ne sont pas dans $E$, avec $\mathcal{X}_{-E}, \mathbf{x}_{-E}$ suivant les mêmes convensions que $\mathcal{X}_E, \mathbf{x}_E$. Nous notons aussi par $(\mathbf{x}_E, \mathbf{y}_{-E})$ la concaténation de deux vecteurs dont les valeurs sont données pour des éléments différents.

Dans le cadre des classifieurs probabilistes imprécis, une classe $y$ domine faiblement [1] $y'$, noté $y \succeq_p y'$, en observant un vecteur $\mathbf{x}$ quand la condition [2]

$$\frac{p(y|\mathbf{x})}{p(y'|\mathbf{x})} \geq 1 \tag{1}$$

est vérifiée, ou autrement dit quand $p(y|\mathbf{x}) \geq p(y'|\mathbf{x})$. Cependant, les classifieurs probabilistes peuvent être trompeurs sur leur précision, par exemple quand seulement un petit nombre de données sont disponibles pour les estimer, ou encore quand les données sont imprécises.

C'est pourquoi, dans ce papier, nous considérons un cadre de probabilités généralisées, et plus spécifiquement issu de la théorie des probabilités imprécises, où nous considérons que la probabilité $p$ appartient à un sous-ensemble $\mathcal{P}$, souvent convexe (ce sera le cas ici). Nous devons donc étendre la relation $\succeq_p$ dans un tel cas, et un moyen classique et robuste de le faire est d'exiger $\succeq_p$ d'être vraie pour tout élément $p \in \mathcal{P}$. Dans ce cas, $y$ domine de manière robuste $y'$, écrit $y \succeq_{\mathcal{P}} y'$, en observant un vecteur $\mathbf{x}$ quand la condition

$$\inf_{p \in \mathcal{P}} \frac{p(y|\mathbf{x})}{p(y'|\mathbf{x})} \geq 1 \tag{2}$$

est vérifiée, ou en d'autres termes quand $p(y|\mathbf{x}) \geq p(y'|\mathbf{x})$ pour tout $p \in \mathcal{P}$. Notons que la relation $\succeq_{\mathcal{P}}$ peut être un pre-ordre partiel avec des incomparabilitiés, là où $\succeq_p$ est un pre-ordre.

## 2.2 Expliquer des classifications robustes avec des implicants premiers

Expliquer la conclusion ou les déductions d'un algorithme, et en particulier un algorithme d'apprentissage, est devenu un problème important. Une notion qui peut jouer un role dans les mécanismes explicatifs est celles des implicants premiers, *i.e.*, quels éléments sont suffisants avant de fournir une certaine décision.

1. Nous sommes dans un cadre sans coûts, mais une grande partie de notre discussion se transfert facilement dans ce cadre

2. Utiliser la dominance exprimée de cette façon sera utile plus tard.

Quand nous observons un vecteur $\mathbf{x}^o$ et faisons la prédiction qu'un certain $y$ domine $y'$, l'idée principale des implicants premiers peut globalement se traduire par les valeurs de $\mathbf{x}^o$ qui suffisent pour savoir décider que $y$ domine $y'$, et qui est minimale pour cette propriété.

Avec cette idée en tête, nous disons qu'un sous-ensemble $E \subseteq \{1, \ldots, n\}$ d'attributs (où $E$ contient les indices des attributs sélectionnes) est un *implicant* de $y \succeq_{\mathcal{P}} y'$ ssi

$$\inf_{p \in \mathcal{P}, \mathbf{x}^a_{-E} \in X_{-E}} \frac{p(y|(\mathbf{x}^o_E, \mathbf{x}^a_{-E}))}{p(y'|(\mathbf{x}^o_E, \mathbf{x}^a_{-E}))} \geq 1, \quad (3)$$

c'est à dire si la dominance reste valide pour toutes les valeurs des attributs en dehors de $E$ et pour toutes les probabilités $p \in \mathcal{P}$. Cela signifie que savoir $\mathbf{x}^o_E$ seul est suffisant pour déduire $y \succeq_{\mathcal{P}} y'$. Un ensemble $E$ est un *implicant premier* ssi l'équation (3) est satisfaite et que pour chaque $i \in E$, nous avons

$$\inf_{p \in \mathcal{P}, \mathbf{x}^a_{-E\cup\{i\}} \in X_{-E\cup\{i\}}} \frac{p(y|(\mathbf{x}^o_{E\setminus\{i\}}, \mathbf{x}^a_{-E\cup\{i\}}))}{p(y'|(\mathbf{x}^o_{E\setminus\{i\}}, \mathbf{x}^a_{-E\cup\{i\}}))} \leq 1, \quad (4)$$

c'est à dire, retirer n'importe quel attribut de $E$ rends notre déduction invalide, donc que $E$ est une condition suffisante minimale pour vérifier $y \succeq_{\mathcal{P}} y'$. Plus tard nous aurons besoin de considerer la fonction $\phi(E)$ qui associe pour tout sous-ensemble la valeur

$$\phi(E) := \inf_{p \in \mathcal{P}, \mathbf{x}^a_{-E} \in X_{-E}} \frac{p(y|(\mathbf{x}^o_E, \mathbf{x}^a_{-E}))}{p(y'|(\mathbf{x}^o_E, \mathbf{x}^a_{-E}))}. \quad (5)$$

$\phi(E)$ est monotone vis à vis de l'inclusion (pour $E \subseteq F$, $\phi(E) \leq \phi(F)$), et peut être vue comme une fonction de valeur associée à $E$, et trouver un implicant premier peut être vu comme la tâche de trouver un "ensemble d'objets" [3] minimal $E$ tel que $\phi(E) \geq 1$, par conséquent nous permettant de caractériser la recherche d'implicants premiers robustes par un problème de sélection d'objets. Notons aussi que, en général, $\phi(E)$ n'est pas additive, vu que nous n'avons pas $\phi(E \cup \{i\}) = \phi(E) + \phi(\{i\})$.

3. Chaque index d'un attribut étant associé à un objet.

Notons que quand les ensembles $\mathcal{P}$ sont réduits à des singletons, c'est à dire quand nous considérons un classifieur précis à la place de robustes, alors notre notion d'implicant premier se réduit à celle déjà proposée [9], notre approche est donc une généralisation formelle de celle-ci.

# 3 Le cas du classifier naïf crédal

Nous étudions maintenant le cas spécifique du classifieur naïf, et montrons que dans ce cas obtenir les implicants premiers devient facile, car le calcul peut être ramené à de la sélection d'objets avec une fonction additive, ou de manière équivalente à un problème de sac à dos très simple.

## 3.1 Cas générique

L'idée de base du classifieur naïf est d'assumer que les attributs sont indépendants les uns des autres étant donné la classe. Cette hypothèse du modèle signifie que

$$p(y|\mathbf{x}) = \frac{\prod_{i=1}^n p_i(x_i|y) \times p_{\mathcal{Y}}(y)}{p(\mathbf{x})}$$

une fois que nous appliquons l'hypothèse naïve et la règle de Bayes. Cela signifie en particulier que

$$\frac{p(y|\mathbf{x})}{p(y'|\mathbf{x})} = \frac{p_{\mathcal{Y}}(y)}{p_{\mathcal{Y}}(y')} \prod_{i=1}^n \frac{p_i(x_i|y)}{p_i(x_i|y')}$$

avec chaque $p_i(|y)$ indépendant de $p_i(|y')$, et chaque $p_i(|y), p_j(|y)$ indépendants pour $i, j$. Quand nous passons aux modèles crédaux, nous avons un ensemble de distributions conditionnelles $\mathcal{P}_i(|y)$ et un ensemble $\mathcal{P}_{\mathcal{Y}}$ d'a priori.

Voyons maintenant comment l'équation (3) se

transforme dans ce cas. Nous avons

$$\inf_{\substack{p\in\mathcal{P}\\ \mathbf{x}^a_{-E}\in X_{-E}}} \frac{p(y|(\mathbf{x}^o_E,\mathbf{x}^a_{-E}))}{p(y'|(\mathbf{x}^o_E,\mathbf{x}^a_{-E}))} =$$

$$\inf_{\substack{p\in\mathcal{P}\\ \mathbf{x}^a_{-E}\in X_{-E}}} \frac{p_{\mathcal{Y}}\,(y)}{p_{\mathcal{Y}}\,(y')} \underbrace{\prod_{i\in E}\frac{p_i(x^o_i|y)}{p_i(x^o_i|y')}}_{PartieA} \underbrace{\prod_{i\notin E}\frac{p_i(x^a_i|y)}{p_i(x^a_i|y')}}_{PartieB}. \tag{6}$$

Dans l'équation 6, nous pouvons traiter le problème de minimisation des parties A et B totalement indépendamment, à cause de deux observations principales. Tout d'abord, les ensembles $\mathcal{P}_i(|y)$ sont tous indépendants quand $i$ (l'attribut) ou $y$ (le conditionnement) change. Cela implique que la partie A et B son minimisées sur des ensembles de probabilités convexes indépendants (comme elles se situent sur des $i$ distincts), mais aussi que les numérateurs et dénominateurs de chaque fraction au sein des deux parties peuvent aussi être traitées de manière séparées (étant conditionnées par des $y, y'$ différents). De plus, $E$ et $-E$ sont disjoints, ce qui signifie que la valeur $\mathbf{x}^a_{-E}$ pour laquelle la partie B est minimisée dépends uniquement de la partie B, ainsi, dans ce cas, il est donc sensé de définir un vecteur unique du "pire cas" $\mathbf{x}^{a*}$ qui minimise la partie B pour n'importe quel $E$. Aussi, puisque les lois conditionnelles avec des classes conditionnelles différentes sont indépendantes, nous obtenons que l'équation (6) devient

$$\inf_{p_{\mathcal{Y}}\in\mathcal{P}_{\mathcal{Y}}} \frac{p_{\mathcal{Y}}\,(y)}{p_{\mathcal{Y}}\,(y')} \prod_{i\in E}\frac{\underline{p}_i(x^o_i|y)}{\overline{p}_i(x^o_i|y')} \inf_{\mathbf{x}^a_{-E}\in X_{-E}} \prod_{i\notin E}\frac{\underline{p}_i(x^a_i|y)}{\overline{p}_i(x^a_i|y')}. \tag{7}$$

où $\underline{p}(x) = \inf_{p\in\mathcal{P}} p(x)$ et $\overline{p}(x) = \sup_{p\in\mathcal{P}} p(x)$. Si nous considérons le vecteur $\mathbf{x}^{a*}_{-E}$, nous obtenons finalement

$$\inf_{\substack{p\in\mathcal{P}\\ \mathbf{x}^a_{-E}\in X_{-E}}} \frac{p(y|(\mathbf{x}^o_E,\mathbf{x}^a_{-E}))}{p(y'|(\mathbf{x}^o_E,\mathbf{x}^a_{-E}))} =$$

$$\inf_{p_{\mathcal{Y}}\in\mathcal{P}_{\mathcal{Y}}} \frac{p_{\mathcal{Y}}\,(y)}{p_{\mathcal{Y}}\,(y')} \prod_{i\in E}\frac{\underline{p}_i(x^o_i|y)}{\overline{p}_i(x^o_i|y')} \prod_{i\notin E}\frac{\underline{p}_i(x^{a*}_i|y)}{\overline{p}_i(x^{a*}_i|y')} \tag{8}$$

Revenons maintenant à notre idée de sélection d'un ensemble d'objets (ou attributs) minimal tel que $\phi(E) > 1$ ou de manière équivalente $\log\phi(E) > 0$. Notons d'abord par

$$C = \log \inf_{p_{\mathcal{Y}}\in\mathcal{P}_{\mathcal{Y}}} \frac{p_{\mathcal{Y}}\,(y)}{p_{\mathcal{Y}}\,(y')} \prod_{i\in\{1,\dots,n\}} \frac{\underline{p}_i(x^{a*}_i|y)}{\overline{p}_i(x^{a*}_i|y')} \tag{9}$$

la valeur de $\log\phi(\emptyset)$, et par

$$G_i = (\log\underline{p}_i(x^o_i|y) - \log\overline{p}_i(x^o_i|y')) - (\log\underline{p}_i(x^{a*}_i|y) - \log\overline{p}_i(x^{a*}_i|y')) \tag{10}$$

le gain positif[4] obtenu en ajoutant l'élément $i$ à $E$. En développant l'équation (7), nous pouvons vérifier que

$$\log\phi(E) = C + \sum_{i\in E} G_i$$

a une forme additive. Trouver le premier implicant premier est très facile, étant donné qu'il suffit d'ordonner les $G_i$ dans l'ordre décroissant, et les ajouter jusqu'à ce que $\sum_{i\in E} G_i \geq -C$. La procédure est synthétisée dans l'algorithme 1.

**Algorithme 1 :** Calcul de l'implicant premier immédiatement disponible

**Entrées :** $C : log(\phi(\emptyset))$ ; $G$ : Contributions des critères dans l'ordre décroissant ;
**Output :** $Xpl = (E, \mathbf{x}_E)$ : explication en termes d'attributs

```
Ordonner G de manière décroissante, avec σ la permutation associée
i ← 1
tant que φ(E) + C < 0 faire
    i ← i + 1
    E ← E ∪ {σ^{-1}(i)}
    φ(E) ← φ(E) + G_{σ(i)}
Xpl ← (E, x^o_E)
retourner (Xpl)
```

4. Vu que $\log\underline{p}_i(x^{a*}_i|y) - \log\overline{p}_i(x^{a*}_i|y') < \log\underline{p}_i(x^o_i|y) - \log\overline{p}_i(x^o_i|y')$ par définition.

La complexité de l'algorithme 1 est manifestement linéaire sur les contributions ordonnées, donc du nombre d'attributs. Calculer les contributions reste facile vu que la complexité provient du calcul du "pire cas" $\mathbf{x}^{a*}$, dont les composantes $\mathbf{x}_i^{a*}$ nécessitent $|X_i| = k_i$ évaluations sur chaque dimension. Comme les ensembles $\mathcal{P}$ sont des polytopes définis par des contraintes linéaires, trouver les valeurs $\underline{p}$ et $\overline{p}$ revient à résoudre des programmes linéaires, ce qui peut être fait en temps polynomial. Pour quelques cas particuliers comme les intervalles de probabilités [13] (induits, par exemple, par le classique modèle de Dirichlet imprécis [5]) le calcul peut même être fait en temps linéaire. C'est pourquoi, la méthode complète est linéaire, avec un pré-traitement linéaire sur les sommes des $k_i$, suivis par un algorithme de tri, après lequel s'applique l'algorithme 1 qui est linéaire du nombre d'attributs.

### 3.2 Exemple illustratif

Nous allons présenter un petit exemple illustratif utilisant des données catégorielles et des intervalles de probabilités. Ces derniers peuvent, par exemple, être obtenus à l'aide du modèle imprécis de Dirichlet [5], possiblement avec quelques régularisations pour éviter d'avoir des probabilités valant zéro, ou dans le cas de variables continues des modèles paramétriques [1] ou non-paramétriques [8].

Dans cet exemple nous souhaitons prédire une classe d'animal à partir de ses caractéristiques physique. Nous avons des données sur l'ensemble d'animaux $\mathcal{Y}$ = {Chien(D), Chat(C), Cheval(H), Lapin(B)} et nous observons la longueur des $\mathcal{X}$ ={Oreilles(E), Queue(T), Poils(H)}. Chacun de ces critères prends ses valeurs dans l'ensemble $\{Longue(L), Moyenne(A), Courte(S)\}$.
Pour identifier plus facilement les variables dans l'exemple, nous utiliserons la notation $LE$ pour désigner des longues oreilles et de manière similaire pour toute autre combinaison d'attributs. Les probabilités a priori sont données dans le tableau 1 et les probabilités conditionnelles dans les tableaux 2, 3 and 4.

TABLEAU 1 – Intervalles de probabilités pour chaque animal

| Chien | Chat | Cheval | Lapin |
|---|---|---|---|
| $[0.25, 0.26]$ | $[0.29, 0.31]$ | $[0.20, 0.22]$ | $[0.25, 0.26]$ |

TABLEAU 2 – Probabilités conditionnelles de la longueur des oreilles sachant l'animal

| | Chien | Chat | Cheval | Lapin |
|---|---|---|---|---|
| L | [0.33,0.40] | [0.02,0.08] | [0.10,0.19] | [0.58,0.65] |
| A | [0.30,0.37] | [0.55,0.61] | [0.66,0.75] | [0.26,0.33] |
| S | [0.30,0.37] | [0.37,0.43] | [0.15,0.23] | [0.09,0.16] |

TABLEAU 3 – Probabilités conditionnelles de la longueur de la queue sachant l'animal

| | Chien | Chat | Cheval | Lapin |
|---|---|---|---|---|
| L | [0.54,0.61] | [0.31,0.37] | [0.66,0.75] | [0.02,0.09] |
| A | [0.23,0.30] | [0.61,0.67] | [0.23,0.32] | [0.30,0.37] |
| S | [0.16,0.23] | [0.02,0.08] | [0.02,0.10] | [0.61,0.69] |

A partir de maintenant, nous observons le vecteur $\mathbf{x}^o$ = (Longues Oreilles, Queue Courte, Longs Poils) ou (LE,ST,LH) abrégé. Comme nous utilisons un modèle de classification imprécise, les classes prédites seront celles qui sont non dominées, et nos explications servirons principalement à comprendre pourquoi nous avons rejeté les autres. Pour chaque paire (y,y') d'animaux, nous comparons $\inf_{p\in\mathcal{P}} \frac{p(y|\mathbf{x})}{p(y'|\mathbf{x})}$ à 1 pour construire l'ordre partiel entre elles. Dans notre cas spécifique, cela revient à comparer

$$log\ \underline{p}(y) - log\ \overline{p}(y') + \sum_{i=1}^{3} log\ \underline{p}(\mathbf{x}_i^o \mid y) - \sum_{i=1}^{3} log\ \overline{p}(\mathbf{x}_i^o \mid y')$$

avec 0. Comme nous avons des intervalles de probabilités, la borne $\underline{p}$ (resp. $\overline{p}$) peut se lire directement des tableaux. Si nous prenons la paire

TABLEAU 4 – Probabilités conditionnelles de la longueur des poils sachant l'animal

| | Chien | Chat | Cheval | Lapin |
|---|---|---|---|---|
| L | [0.40,0.47] | [0.46,0.52] | [0.23,0.32] | [0.02,0.09] |
| A | [0.26,0.33] | [0.17,0.22] | [0.10,0.19] | [0.19,0.26] |
| S | [0.26,0.33] | [0.31,0.37] | [0.58,0.66] | [0.72,0.79] |

(Chien, Cheval) ou (D,H) comme exemple, nous avons

$$log\ \underline{p}(D) - log\ \overline{p}(H) + \sum_{i=1}^{3} log\ \underline{p}(\mathbf{x}_i^o \mid D) - \sum_{i=1}^{3} log\ \overline{p}(\mathbf{x}_i^o \mid H) = 0.58 > 0$$

Nous avons donc que $D \succeq_{\mathcal{P}} H$. En répétant cette méthode pour toutes les paires, nous obtenons l'ordre partiel de la figure 1. Le prédiction prudente sera $\{D, B\}$, et chaque arc de la figure 1 peut être expliquée par des implicants premiers.

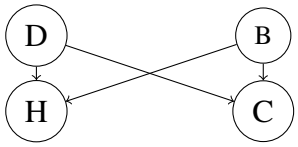


FIGURE 1 – Dominances entre les classes pour la prédiction de $\mathbf{x}^o$ = (LE, ST, LH)

Nous détaillons le calcul seulement pour $D \succeq_{\mathcal{P}} H$. Tout d'abord, nous devons calculer le pire adversaire $\mathbf{x}^a$ qui minimise $log\ \underline{p}(x_i^{a*} \mid D) - log\ \overline{p}(x_i^a \mid H)$ pour chaque variable $i$. Nous obtenons $\mathbf{x}^{a*} = (AE, AT, SH)$. En appliquant l'équation (9), nous avons

$$C = log\ \underline{p}(D) - log\ \overline{p}(H) + \sum_{i=1}^{3} log\ \underline{p}(\mathbf{x}_i^{a*} \mid D) - log\ \overline{p}(\mathbf{x}_i^{a*} \mid H) = -0.90$$

Les contributions des critères requis par l'algorithme 1 sont :

$$\begin{aligned} G_i &= log\ \underline{p}(\mathbf{x}_i^o \mid D) - log\ \overline{p}(\mathbf{x}_i^o \mid H) \\ &\quad - (log\ \underline{p}(\mathbf{x}_i^{a*} \mid D) - log\ \overline{p}(\mathbf{x}_i^{a*} \mid H)) \\ G_{Oreilles} &= log(0.33) - log(0.19) \\ &\quad - (log(0.30) - log(0.75)) = 0.65 \\ G_{Queue} &= log(0.16) - log(0.10) \\ &\quad - (log(0.23) - log(0.32)) = 0.33 \\ G_{Poils} &= log(0.40) - log(0.32) \\ &\quad - (log(0.26) - log(0.66)) = 0.50 \end{aligned}$$

Nous appliquons maintenant l'algorithme 1 et obtenons l'explication {(Oreilles, Longues), (Poils, Longs)} vu que $(0.65+0.50)-0.90 > 0$, mais avec un algorithme d'énumération nous pouvons trouver un second implicant premier avec {(Oreilles, Longues), (Queue, Courte)}, vu que $(0.65 + 0.33) - 0.90 > 0$, qui est moins important en terme de gain, mais qui est peut-être intuitivement plus satisfaisant. De manière similaire, nous pouvons calculer des explications pour d'autre dominances, comme {(Oreilles, Longues), (Queue, Courte)} pour Chien $\succeq_{\mathcal{P}}$ Chat, {(Oeilles, Longues), (Queue, Courte)} pour Lapin $\succeq_{\mathcal{P}}$ Chat et {(Oreilles, Longues), (Queue, Courte)} pour Lapin $\succeq_{\mathcal{P}}$ Cheval.

## 4 Conclusion

Ce papier propose d'expliquer des classifications robustes par des implicants premiers, cette notion étant proposée actuellement uniquement dans le cadre précis. Nous montrons que, comme pour le cas précis, cette tâche est facile pour le classifier naïf. A notre connaissance, c'est la première tentative de combiner une classification imprécise avec des explications.

Dans le futur, nous souhaiterions nous concentrer sur certaines question non explorées ici, comme par exemple : est-ce qu'énumérer tous les implicants premiers reste facile pour le classifieur naïf crédal ? Pour quels modèles robustes les calculs restent faisables facilement ? Que

se passe-t-il si nous ajoutons de l'interaction entre critères ? Peut-on expliquer des incomparabilités avec des notions similaires ? Quand nous voulons expliquer un ordre partiel en entier, devons nous utiliser des comparaison par paires ou holistique (*i.e.*, les implicants premiers expliqueraient les classes non-dominées en un seul coup) ? Il y a aussi plusieurs autres mécanismes d'explication que nous pourrions considérer [2].

## Références


[1] Y. C. C. Alarcón et S. Destercke. Imprecise Gaussian Discriminant Classification. *Pattern Recognition*, vol. 112, p. 107739, 2019.

[2] G. Audemard, F. Koriche, et P. Marquis. On tractable XAI queries based on compiled representations. *International Conference on Principles of Knowledge Representation and Reasoning, 2020*, vol. 17, n° 1, p. 838-849.

[3] V. Balasubramanian, S.-S. Ho, et V. Vovk. *Conformal prediction for reliable machine learning : theory, adaptations and applications*. Newnes, 2014.

[4] K. Belahcene, C. Labreuche, N. Maudet, V. Mousseau, et W. Ouerdane. Explaining robust additive utility models by sequences of preference swaps. *Theory and Decision*, vol. 82, n° 2, Art. n° 2, 2017.

[5] J.-M. Bernard. An introduction to the imprecise Dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, vol. 39, n° 2-3, p. 123-150, 2005.

[6] G. Corani, A. Antonucci, et M. Zaffalon. Bayesian networks with imprecise probabilities : Theory and application to classification. *Data Mining : Foundations and Intelligent Paradigms*, Springer, 2012, p. 49-93.

[7] J. J. Del Coz, J. Díez, et A. Bahamonde. Learning Nondeterministic Classifiers. *Journal of Machine Learning Research*, vol. 10, n° 10, 2009.

[8] G. Dendievel, S. Destercke, et P. Wachalski. Density estimation with imprecise kernels : application to classification. *SMPS, 2018*, p. 59-67.

[9] J. Marques-Silva, T. Gerspacher, M. C. Cooper, A. Ignatiev, et N. Narodytska. Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay, *NeurIPS, 2020*, vol. 33, p. 20590-20600.

[10] A. Shih, A. Choi, et A. Darwiche. A symbolic approach to explaining bayesian network classifiers. *arXiv preprint arXiv :1805.03364, 2018*.

[11] G. Van den Broeck, A. Lykov, M. Schleich, et D. Suciu. On the tractability of SHAP explanations, *35th AAAI, 2021*.

[12] M. Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, vol. 105, n° 1, p. 5-21, 2002.

[13] L. M. De Campos, J. F. Huete, et S. Moral. Probability intervals : a tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 2, n° 02, p. 167-196, 1994.