Proceedings of Machine Learning Research 215:461–471, 2023 ISIPTA 2023

Prime Implicants as a Versatile Tool to Explain Robust Classification

Hénoïk Willot HENOIK.WILLOT@HDS.UTC.FR
Sébastien Destercke SEBASTIEN.DESTERCKE@HDS.UTC.FR
Heudiasyc, University of Technology of Compiegne, France

Khaled Belahcène KHALED.BELAHCENE@CENTRALESUPELEC.FR
MICS, CentraleSupélec, Paris-Saclay University, France

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 subset of attributes, minimal w.r.t. inclusion, that we need to know or specify before reaching a specified conclusion (either of dominance or non-dominance between two classes). After presenting the general concepts, we derive them in the case of the well-known naive credal classifier.

Keywords: robust classifier, explainability, prime implicants, imprecise probabilities, naive credal classifier

1. Introduction

Two key aspects of trustworthy AI are the ability to provide robust and safe inferences or predictions, and to be able to provide explanations as of why those have been made.

Regarding explainability, the notion of prime implicants corresponds to providing minimal sufficient condition to make a given statement, e.g., the attributes that need to be instantiated to make a classification. They have been successfully proposed as components of explanations for large classes of models such as graphical ones [20], with very efficient procedure existing for specific structures such as the Naive one [18]. In contrast with other methods such as SHAP [22] that tries to compute the average influence of attributes, prime implicants have the advantage to be well-grounded in logic, and to provide certifiable explanation (in the sense that the identified attributes are logical, sufficient reasons).

However, explainable AI tools have been mostly applied to precise models, at least in the machine learning domain (this is less true, e.g., in preference modelling [4]). Yet, in applications involving sensitive issues or in which the decision maker wants to identify ambiguous cases, it may be preferable to use models that will return sets of classes when information is insufficient rather than always returning a point-valued prediction. Several frameworks such as conformal prediction [3], indeterminate classifiers [10] or imprecise probabilistic models [8] have been proposed to handle such issue. While some explanation methods for such models have been recently proposed [21, 26], none of them explicitly adopts a logical standpoint regarding explanations, meaning that the present work is complementary to those.

Imprecise probabilistic models in particular have the interest that they are direct extensions and generalisations of probabilistic classifiers, hence one can directly try to transport well-grounded explanation principles existing for precise probabilistic classifier to this setting. This is what we intend to do in this paper for prime implicant explanations.

We will start by introducing how the idea of prime implicants can be adapted to classifiers considering sets of probabilities as their uncertainty models. Section 2 will be a short reminder of the robust classification setting, and will introduce our notations. In Section 3, we will present the idea of prime implicant, as well as how it can answer various explanatory needs. As the formulated problems are likely to be computationally challenging for generic models, we focus in Section 4 on the naive credal classifier, that generalise the naive Bayes classifier. We show that for such a model, computing and enumerating prime implicants can be done in polynomial time, thanks to its independence assumption and decompositional properties. We also provide an experiment in Section 5 illustrating our approach. Note that this work builds upon some first results published in [24], that focused exclusively on a unique use of prime implicants and did not contain any experiment. In contrast, the study of this paper provides much more use of prime implicants, study some basic properties of their behaviour and provide some first experiments on a real-world data set.

2. Preliminaries on Robust Classification

In this section, we lay down our basic notations and provide necessary reminders about imprecise probabilities.

We consider a usual discrete multi-class problem, where we must predict a variable Y taking values in $\mathcal{Y} = \{y_1, \dots, y_m\}$ using n input variables $X_1, \dots, X_n$ that respectively takes values in $\mathcal{X}_i = \{x_i^1, \dots, x_i^{k_i}\}$. We note $\mathcal{X} = \times_{i=1}^n \mathcal{X}_i$ and $\mathbf{x} \in \mathcal{X}$ a vector in this space. When considering a subset $E \subseteq \{1, \dots, n\}$ of dimensions, we will denote by $\mathcal{X}_E = \times^{i \in E} \mathcal{X}_i$ the corresponding domain, and by $\mathbf{x}_E$ the values of a vector on this sub-domain. We will also denote by $-E := \{1, \dots, n\} \setminus E$ all dimensions not in E, with $\mathcal{X}_{-E}, \mathbf{x}_{-E}$ following the same conventions as $\mathcal{X}_E, \mathbf{x}_E$. We will also denote by $(\mathbf{x}_E, \mathbf{y}_{-E})$ the concatenation of two vectors whose values are given for different elements. Notation $(\mathbf{x}_E, \cdot)$ means that all features in $-E$ can take any value. If $E = \{1, \dots, n\}$, then we will simply ignore the subscript.

In the rest of the paper, we will often refer to partially ordered sets, their corresponding relations and sets of sufficient elements that allows to asset them. Those sufficient elements will here be composed of a vector of specified feature values and of a set of probabilities. We will denote by $y \succeq_{p,(\mathbf{x})} y'$ the fact that considering the model p and the vector $(\mathbf{x})$ is sufficient to state (or implies) $y \succeq y'$.

In the case of precise classifiers, we have $y \succeq_{p,(\mathbf{x})} y'$ when the condition[1]

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

is met, or in other words when $p(y|\mathbf{x}) \geq p(y'|\mathbf{x})$. However, probabilistic classifiers can be deceptively precise, for instance when only a small number of data are available to estimate them, or when data become imprecise.

This is why, in this paper, we consider generalised probabilistic settings, and more specifically imprecise probability theory, where one considers that the probability p belongs to some subset $\mathcal{P}$, often assumed to be convex (this will be the case here). One then needs to extend the relation $\succeq_p$ to such a case, and a common and robust way to do so is to require $\succeq_p$ to be true for all elements $p \in \mathcal{P}$. In this case, y is said to robustly dominate y' upon observing a vector $\mathbf{x}$, written $y \succ_{\mathcal{P},(\mathbf{x})} y'$, when the condition

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

is met, or in other words when $p(y|\mathbf{x}) \geq p(y'|\mathbf{x})$ for all $p \in \mathcal{P}$. Going from the precise to imprecise probabilities can introduce incomparabilities between classes, written $y \succ\prec_{\mathcal{P},(\mathbf{x})} y'$ when both

$$\inf_{p \in \mathcal{P}} \frac{p(y|\mathbf{x})}{p(y'|\mathbf{x})} < 1 \textit{ and } \inf_{p' \in \mathcal{P}} \frac{p'(y'|\mathbf{x})}{p'(y|\mathbf{x})} < 1. \qquad (3)$$

3. Explaining Robust Classification through Prime Implicants

Explaining the conclusion or deduction of an algorithm, and in particular of a learning algorithm, has become (again) an important issue [6]. A notion that can play a key role in explanation mechanisms is the one of prime implicants, i.e., which elements are sufficient before drawing a given conclusion. In this section, we detail how prime implicants can be used to answer the needs of different explanatory mechanisms, within the setting of robust, imprecise probabilistic classifiers.

3.1. Prime Implicants as Validatory Explanations

When observing a vector $\mathbf{x}^o$ and making a prediction about whether y dominates y', finding a prime implicant confirming that y dominates y' corresponds to finding the values of $\mathbf{x}^o$ that are sufficient to state that y dominates y', and that are minimal with this property.

With this idea in mind, we will say that a subset $E \subseteq \{1, \dots, n\} := [\![1, n]\!]$ of attributes (where E are the indices of the considered attributes) is a *validatory implicant* of $y \succ_{\mathcal{P},(\mathbf{x}_E^o, \cdot)} y'$ iff

$$\inf_{\mathbf{x}_{-E}^v \in \mathcal{X}_{-E}} \inf_{p \in \mathcal{P}} \frac{p(y|(\mathbf{x}_E^o, \mathbf{x}_{-E}^v))}{p(y'|(\mathbf{x}_E^o, \mathbf{x}_{-E}^v))} > 1, \qquad (4)$$

that is if dominance holds for any values of attributes whose indices are outside E, and any probability $p \in \mathcal{P}$. This means that knowing $\mathbf{x}_E^o$ alone is sufficient to deduce $y \succ y'$. A set E is a *prime implicant* iff we satisfy (4) and for any $i \in E$, we have

$$\inf_{\mathbf{x}_{-E\cup\{i\}}^v \in \mathcal{X}_{-E\cup\{i\}}} \inf_{p \in \mathcal{P}} \frac{p(y|(\mathbf{x}_{E\setminus\{i\}}^o, \mathbf{x}_{-E\cup\{i\}}^v))}{p(y'|(\mathbf{x}_{E\setminus\{i\}}^o, \mathbf{x}_{-E\cup\{i\}}^v))} \leq 1, \qquad (5)$$

that is if removing any attribute from E makes our deduction invalid, so that E is a minimal sufficient condition for $y \succ_{\mathcal{P},(\mathbf{x}_E^o, \cdot)} y'$ to hold for any completion of $-E$. In the sequel, it will prove useful to consider the function that associates to each possible subset the value of the ratio between the obtained posterior probabilities. This function ϕ^v is defined by :

$$\phi^v(y, y', \mathbf{x}^o, E) := \inf_{\mathbf{x}_{-E}^v \in \mathcal{X}_{-E}} \inf_{p \in \mathcal{P}} \frac{p(y|(\mathbf{x}_E^o, \mathbf{x}_{-E}^v))}{p(y'|(\mathbf{x}_E^o, \mathbf{x}_{-E}^v))}. \qquad (6)$$

To ease the use of this function, we will omit the observed vector $\mathbf{x}^o$ when context is clear and we will write "y, y'" as a subscript meaning that class y is the numerator and y' the denominator, i.e., $\phi_{y,y'}^v(E) := \phi^v(y, y', \mathbf{x}^o, E)$

Note that when the set $\mathcal{P}$ reduces to a singleton, that is when we consider precise classifiers instead of robust ones, then our notion of prime implicant reduces to previously proposed ones [18], and our approach is therefore a formal generalisation of those.

[1] Using dominance expressed this way will be useful in the sequel. We will also restrict ourselves to 0/1 loss functions here.

Monotony with respect to imprecision. one can note that the notion of validatory prime implicant is monotonic with respect to imprecision, in the following sense

Proposition 1 *Consider two credal sets $\mathcal{P}' \subseteq \mathcal{P}$, then*

$$y \succ_{\mathcal{P},(\mathbf{x}^o_E,\cdot)} y' \implies y \succ_{\mathcal{P}',(\mathbf{x}^o_E,\cdot)} y'$$

Proof Immediate, since if Equation (4) is true for E and $\mathcal{P}$, it must be true for E and $\mathcal{P}'$, as the infimum is taken over a smaller domain. ■

This means that a validatory implicant will remain so if we consider a more precise model (obtained, e.g., by observing additional data). However, if a subset E was prime for $\mathcal{P}$, it does not need to be so for $\mathcal{P}'$, meaning that the size of validatory prime implicants should decrease as imprecision decreases. This is somehow natural, as a more informative model should need less measurements to provide a conclusion.

3.2. Prime Implicants as Contrastive Explanations

Another quite common way to audit or explain a statement *"Why X is P?"* is by answering the implicit question *"Why X is P and not Q?"* [19, 12]. This can classically be answered by finding a counter-factual, i.e., a modification of the example with sufficient changes so as to change our conclusion. Replying to this question in a minimal way can be seen as the task of finding a minimal set of attributes or features for which a modification could change our decision. We will call $E \subseteq [\![1,n]\!]$ a *contrastive prime implicant* if modifying the attributes within E is a minimal sufficient condition to change our decision, that is, if

$$\inf_{\mathbf{x}^c_E \in \mathcal{X}_E} \inf_{p \in \mathcal{P}} \frac{p(y|(\mathbf{x}^c_E, \mathbf{x}^o_{-E}))}{p(y'|(\mathbf{x}^c_E, \mathbf{x}^o_{-E}))} < 1, \tag{7}$$

and if for any $i \notin E$, we do have

$$\inf_{\mathbf{x}^c_{E\setminus\{i\}} \in \mathcal{X}_{E\setminus\{i\}}} \inf_{p \in \mathcal{P}} \frac{p(y|(\mathbf{x}^c_{E\setminus\{i\}}, \mathbf{x}^o_{-E\cup\{i\}}))}{p(y'|(\mathbf{x}^c_{E\setminus\{i\}}, \mathbf{x}^o_{-E\cup\{i\}}))} \geq 1, \tag{8}$$

that is there is at least one modification of feature values in E that lead to a different decision, and any change done within a subset of it would not change the decision. Denoting $\mathbf{x}^c_E$ the argument of Equation (7), E is a contrastive implicant if $y \not\succ_{\mathcal{P},(\mathbf{x}^c_E,\mathbf{x}^o_{-E})} y'$. We also consider the function $\phi^c_{y,y'}$ that associate to each possible subset the value

$$\phi^c_{y,y'}(E) := \inf_{\mathbf{x}^c_E \in \mathcal{X}_E} \inf_{p \in \mathcal{P}} \frac{p(y|(\mathbf{x}^c_E, \mathbf{x}^o_{-E}))}{p(y'|(\mathbf{x}^c_E, \mathbf{x}^o_{-E}))} \tag{9}$$

One of the interesting aspects of considering imprecise models is that contrastive explanations do not necessary lead to reversing the initial preference (which is the case for precise models). Indeed, modifying the conclusion $y \succ_{\mathcal{P},\mathbf{x}^o} y'$ by considering the modified vector $(\mathbf{x}^c_E, \mathbf{x}^o_{-E})$ can lead to two quite different situations, resulting either in $y' \succ_{\mathcal{P},(\mathbf{x}^c_E,\mathbf{x}^o_{-E})} y$ (reversing of preference) or $y \succ\prec_{\mathcal{P},(\mathbf{x}^c_E,\mathbf{x}^o_{-E})} y'$ (weakening of preference) and we will define two notions of contrastive explanations.

Given that E is a contrastive prime implicant, we say that it is also a *reversing prime implicant* if in addition we have[2]

$$\inf_{p \in \mathcal{P}} \frac{p(y'|(\mathbf{x}^c_E, \mathbf{x}^o_{-E}))}{p(y|(\mathbf{x}^c_E, \mathbf{x}^o_{-E}))} \geq 1, \tag{10}$$

as this contrastive prime implicant change the initial statement or conclusion into its reverse. Otherwise, if it does satisfy Equations (7) and (8), but not (10), we say that E is a *weakening prime implicant*, as it changes a preference between two classes into incomparability. The vector $(\mathbf{x}^c_E, \mathbf{x}^o_{-E})$ also provides us with a contrastive example for which the decision would change.

Monotony with respect to imprecision. as with validatory implicants, the notion of contrastive implicants is monotonic with respect to imprecision, but in the other direction.

Proposition 2 *Consider two credal sets $\mathcal{P} \subseteq \mathcal{P}'$, then*

$$y \not\succ_{\mathcal{P},(\mathbf{x}^c_E,\mathbf{x}^o_{-E})} y' \implies y \not\succ_{\mathcal{P}',(\mathbf{x}^c_E,\mathbf{x}^o_{-E})} y'$$

Proof Immediate, since if Equation (7) is true for E and $\mathcal{P}$, it must be true for E and $\mathcal{P}'$, as the infinimum is taken over a larger domain, and is of lower value. ■

This means that a contrastive implicant and the associated example $(\mathbf{x}^c_E, \mathbf{x}^o_{-E})$ will remain so if we consider a more imprecise model. However, if a subset E was prime for $\mathcal{P}$, it does not need to be so for $\mathcal{P}'$, meaning that the size of contrastive prime implicants should decrease as imprecision increases. Again, this is somehow intuitive, as a dominance obtained for a more imprecise model should be easier to modify than the same dominance obtained from a more precise model. It should also be noted that the argument $\mathbf{x}^c_E$ obtained for $\mathcal{P}$ in Equation (7) may actually change when considering $\mathcal{P}'$.

Remark 3 *Equation (7) corresponds to finding some minimal changes that could modify our conclusion, and can therefore be viewed as a tool to analyse the robustness of this decision. As such, it can be useful to analyse the model and its robustness. However, answering the question "what should I change to be sure to reverse the dominance" would necessitate another notion where the satisfaction of Equation (10) is enforced, that we will not consider here, leaving it for future work.*

[2] Recall that $\mathbf{x}^c_E$ is the argument of Equation (7).

3.3. Prime Implicants as Explanations of Doubt

For precise models, the statement *"Why X is P?"* that we have to explain is typically a precise assignment or a dominance relation between two classes. In the case of robust classification, the question *"Why X is neither P nor Q?"*, and for what reasons cannot I classify X precisely, also makes sense.

In this case, we say that $E \subseteq [\![1, n]\!]$ is a *prime implicant of doubt* if

$$\sup_{\mathbf{x}^d_{-E} \in \mathcal{X}_{-E}} \inf_{p \in \mathcal{P}} \frac{p(y|(\mathbf{x}^o_E, \mathbf{x}^d_{-E}))}{p(y'|(\mathbf{x}^o_E, \mathbf{x}^d_{-E}))} < 1, \tag{11a}$$

and

$$\sup_{\mathbf{x}^d_{-E} \in \mathcal{X}_{-E}} \inf_{p \in \mathcal{P}} \frac{p(y'|(\mathbf{x}^o_E, \mathbf{x}^d_{-E}))}{p(y|(\mathbf{x}^o_E, \mathbf{x}^d_{-E}))} < 1, \tag{11b}$$

that is any change performed outside of E (in particular the changes for the most favourable values for y in Equation (11a) and the most favourable for y' in Equation (11b)) will not modify the fact that the two classes are incomparable given our model and knowledge of it. It is further more minimal if for any $i \notin E$, we do have either

$$\sup_{\mathbf{x}^d_{-E\cup\{i\}} \in \mathcal{X}_{-E\cup\{i\}}} \inf_{p \in \mathcal{P}} \frac{p(y'|(\mathbf{x}^o_{E\setminus\{i\}}, \mathbf{x}^d_{-E\cup\{i\}}))}{p(y|(\mathbf{x}^o_{E\setminus\{i\}}, \mathbf{x}^d_{-E\cup\{i\}}))} \geq 1, \tag{12a}$$

or

$$\sup_{\mathbf{x}^d_{-E\cup\{i\}} \in \mathcal{X}_{-E\cup\{i\}}} \inf_{p \in \mathcal{P}} \frac{p(y|(\mathbf{x}^o_{E\setminus\{i\}}, \mathbf{x}^d_{-E\cup\{i\}}))}{p(y'|(\mathbf{x}^o_{E\setminus\{i\}}, \mathbf{x}^d_{-E\cup\{i\}}))} \geq 1, \tag{12b}$$

We also consider the function $\phi^d_{y,y'}$ that associate to each possible subset the value

$$\phi^d_{y,y'}(E) := \sup_{\mathbf{x}^d_{-E} \in \mathcal{X}_{-E}} \inf_{p \in \mathcal{P}} \frac{p(y|(\mathbf{x}^o_E, \mathbf{x}^d_{-E}))}{p(y'|(\mathbf{x}^o_E, \mathbf{x}^d_{-E}))} \tag{13}$$

$\phi^d_{y,y'}$ corresponds to Equation (11a) and $\phi^d_{y',y}$ to Equation (11b) and both will be used in the computations as we will see in Section 4 for the naive credal classifier. In general, the vectors $\mathbf{x}^d_{-E}$ for which the bounds of (11a)-(11b) are obtained will be different.

Monotony with respect to imprecision. as before, we can easily show that implicants of doubt are somehow monotonic with imprecision, in the following sense

Proposition 4 *Consider two credal sets $\mathcal{P} \subseteq \mathcal{P}'$, then*

$$y \succ\!\prec_{\mathcal{P},(\mathbf{x}^o_E,\cdot)} y' \implies y \succ\!\prec_{\mathcal{P}',(\mathbf{x}^o_E,\cdot)} y'$$

Proof Immediate, since if Equations (11a)-(11b) are true for E and $\mathcal{P}$, it must be true for E and $\mathcal{P}'$, as the infinimum is taken over a larger domain, and is of lower value. ■

It should be remarked that while those implicants are also of validatory nature, in the sense that they confirm our conclusion, their monotonicity is not in the same direction as the validatory implicants of dominance relations. This is however not surprising: as imprecision increases, it becomes easier to obtain that two classes are incomparable, hence prime implicants should decrease in size as credal sets become more imprecise.

Remark 5 *We could also have considered contrastive implicants of doubt that would transform the incomparability into a dominance relation, as done for example in [26]. Such implicants would be of the same kind as the ones mentioned in Remark 3, as they would answer the question "what should I change to be sure to have a dominance relation".*

Remark 6 *In contrast with the previous implicants trying to either verify or contradict a dominance relation between two classes, it may be that $E = \emptyset$ is the only prime implicant of $y \succ\!\prec y'$, in which case doubt is simply due to inherent imprecision of our information (think, for instance, about the case of total ignorance).*

3.4. Short Discussion about the 3 Types of Prime Implicants

We defined three functions $\phi^c_{y,y'}, \phi^d_{y,y'}, \phi^v_{y,y'}$ which are inclusion-monotonic: for $\phi^d_{y,y'}$, $\phi^v_{y,y'}$ and $E \subseteq F$, we do have $\phi^{\cdot}_{y,y'}(E) \leq \phi^{\cdot}_{y,y'}(F)$, and for $\phi^c_{y,y'}$ and $E \subseteq F$, we have $\phi^c_{y,y'}(E) \geq \phi^c_{y,y'}(F)$.

This means that they can be seen as value functions associated to E, and that finding a prime implicant amounts to the task of finding a minimal "bundle of items"[3] E such that $\phi^v_{y,y'}(E) \geq 1$, $\phi^c_{y,y'}(E) < 1$ or $\phi^d_{y,y'}(E) < 1$, therefore allowing us to map the finding of robust prime implicants to an item selection problem or to knapsack problems where we have to fill the sack until it reaches a certain value. Unfortunately, in general, the log-functions[4] of each problem will not be additive, as we will not have $\log {}^{\cdot}\phi_{y,y'}(E \cup \{i\}) = \log {}^{\cdot}\phi_{y,y'}(E) + \log {}^{\cdot}\phi_{y,y'}(\{i\})$. We will nevertheless show in Section 4 that it is the case for the Naive credal classifier.

While providing a minimal subset of features such that a preference/dominance is preserved or changed can be considered to some extent as satisfactory for the user [23, 7] (as long as those features have a meaning for the user),

[3] Each index of an attribute being associated to an item.

[4] As we will deal later with joint probabilities and independence, using log transform will allow us to turn products into sums.

the same cannot really be said about non-dominance or incomparability. In such a case, the user will probably not be satisfied by the mere fact that features values in E are sufficient to claim incomparability, and will request to know why this incomparability happens.

In a machine learning setting, it makes sense to differentiate between incomparability due to ambiguity, where a small change in our knowledge representation $\mathcal{P}$ would lead to a decision, from incomparability due to lack of knowledge, where it would require significantly more knowledge to obtain a decision. These two types of uncertainty sources are often referred to as epistemic and aleatoric uncertainties, and those can be quantified [13].

It seems reasonable that the complementary explanation to incomparability should differ according to the dominating source of uncertainty or indecision. In particular:

- if the indecision is mainly due to aleatoric uncertainties, it is clear that collecting more data is unlikely to solve the issue, and that it would be important to identify those features that generate the ambiguity. In this case, it would seem preferable to provide a contrastive explanation (in the sense of Section 3.2) rather than recommending the collection of further data, so as to answer the question: "which features generate my ambiguity?".
- if the indecision is mainly due to epistemic uncertainties, a possible way to answer this question is to know how many further data points would we need to collect (and which ones) in order to reach a conclusion rather than producing none. The question we would answer would then be "what data should I collect to gain knowledge?"

It is clear to us that providing formal reasons as to why an incomparability is observed, and proposing tools in this direction is a worthwile undertaking, and that our proposal could be useful to the analyst as a way to audit the model (why is my model doubting, and what could I do about it?). It is less clear that the notion proposed in this paper is instrumental to the end-user. Indeed, once $\mathbf{x}^o$ is known, letting the end-user know that we could have known earlier (i.e., with less measurement) that we could not reach a decision is not very helpful. However, our approach can also be considered to detect from partial observations, and before measuring all features, that incomparability will ensue whatever happens, therefore sending an early signal that with this model and this degree of cautiousness, taking more measurements is fruitless.

A definite goal we have in mind for future work is to go beyond the definition of prime implicants of doubts, and investigate problems such as active learning or feature acquisition in which they could offer an operational advantage.

4. The Case of the Naive Credal Classifier

We now study the specific case of the Naive credal classifier [25], and show that in this case, computing prime implicants becomes easy, as such a computation can be brought back to selecting items with an additive value functions, or equivalently to simple knapsack problems.

4.1. Generic Case

The basic idea of the Naive credal classifier (the same as its precise counterpart) is to assume that attributes are independent of each other given the class. This modelling assumption means that

$$p(y|\mathbf{x}) = \frac{\prod_{i=1}^{n} p_i(\mathbf{x}_i|y) \times p_Y(y)}{p(\mathbf{x})}$$

once we apply the Naive assumption and Bayes rule. This means in particular that

$$\frac{p(y|\mathbf{x})}{p(y'|\mathbf{x})} = \frac{p_Y(y)}{p_Y(y')} \prod_{i=1}^{n} \frac{p_i(\mathbf{x}_i|y)}{p_i(\mathbf{x}_i|y')}$$

with every $p_i(\cdot|y)$ being independent of $p_i(\cdot|y')$, and every $p_i(\cdot|y), p_j(\cdot|y)$ independent for $i \neq j$. When switching to credal models, one considers sets of conditional distributions $\mathcal{P}_{X_i}(\cdot|y)$ and a set $\mathcal{P}_{\mathcal{Y}}$ of priors rather than precise probabilities. We will abuse the notation $\mathcal{P}_{X_i}$ by $\mathcal{P}_i$ and p_{X_i} by p_i for the sake of conciseness.

In our Equations (4), (7) and (11a, 11b) we have two optimisation problems, one in $\mathcal{X}_E$ (or $\mathcal{X}_{-E}$) and one in $\mathcal{P}$. Thanks to the independence assumptions, the two problems can be solved independently. Let us now see how the common part of the three Equations, the problem in $\mathcal{P}$, transform in this case. We do have

$$\inf_{p \in \mathcal{P}} \frac{p(y|\mathbf{x})}{p(y'|\mathbf{x})} = \inf_{p_Y \in \mathcal{P}_{\mathcal{Y}}} \frac{p_Y(y)}{p_Y(y')} \prod_{i=1}^{n} \inf_{p_i \in \mathcal{P}_i} \frac{p_i(\mathbf{x}_i|y)}{p_i(\mathbf{x}_i|y')} \tag{14}$$

Once again, thanks to our Independence assumption, each term of the equation can be taken independently of the others (different variables) and inside each feature the numerator is independent of the denominator (different conditioning element). Moreover, as our probability sets $\mathcal{P}_i$ are convex, finding the minimum and maximum value is usually easy. Finally, we get that Equation (14) becomes

$$\inf_{p_Y \in \mathcal{P}_{\mathcal{Y}}} \frac{p_Y(y)}{p_Y(y')} \prod_{i=1}^{n} \frac{\underline{p}_i(\mathbf{x}_i|y)}{\overline{p}_i(\mathbf{x}_i|y')} \tag{15}$$

where $\underline{p}(\mathbf{x}) = \inf_{p \in \mathcal{P}} p(\mathbf{x})$ and $\overline{p}(\mathbf{x}) = \sup_{p \in \mathcal{P}} p(\mathbf{x})$. Let's note $\underline{p}^{y,y'} = \inf_{p_Y \in \mathcal{P}_{\mathcal{Y}}} \frac{p_Y(y)}{p_Y(y')}$. We can now rewrite our functions as :

$$\phi^{v}_{y,y'}(E) = \underline{p}^{y,y'} \prod_{i\in E} \frac{\underline{p}_i(\mathbf{x}^o_i|y)}{\overline{p}_i(\mathbf{x}^o_i|y')} \prod_{i\in -E} \inf_{\mathbf{x}^v_i\in\mathcal{X}_i} \frac{\underline{p}_i(\mathbf{x}^v_i|y)}{\overline{p}_i(\mathbf{x}^v_i|y')} \tag{16a}$$

$$\phi^{c}_{y,y'}(E) = \underline{p}^{y,y'} \prod_{i\in E} \inf_{\mathbf{x}^c_i\in\mathcal{X}_i} \frac{\underline{p}_i(\mathbf{x}^c_i|y)}{\overline{p}_i(\mathbf{x}^c_i|y')} \prod_{i\in -E} \frac{\underline{p}_i(\mathbf{x}^o_i|y)}{\overline{p}_i(\mathbf{x}^o_i|y')} \tag{16b}$$

$$\phi^{d}_{y,y'}(E) = \underline{p}^{y,y'} \prod_{i\in E} \frac{\underline{p}_i(\mathbf{x}^o_i|y)}{\overline{p}_i(\mathbf{x}^o_i|y')} \prod_{i\in -E} \sup_{\mathbf{x}^d_i\in\mathcal{X}_i} \frac{\underline{p}_i(\mathbf{x}^d_i|y)}{\overline{p}_i(\mathbf{x}^d_i|y')} \tag{16c}$$

As we see in Equations (16a), (16b) and (16c), in the case of the NCC the optimisation on $\mathcal{X}_E$ (or $\mathcal{X}_{-E}$) is independent of the computation of $\phi_{y,y'}(E)$. It follows that the results are unique and can be computed before choosing the items in E. We can represent them by unique "worst opponent" vectors, depending only on classes y and y' (the former in the numerator and the later at the denominator) :

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

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

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

When we solve the problem of selecting E, in the case of validatory and contrastive prime implicants, we will only use the "worst opponent" vectors with "y, y'", whereas we also need the converse "y', y" for prime implicants of doubt. We will then refer to $\mathbf{x}^v$ and $\mathbf{x}^c$ instead of $\mathbf{x}^{v:y,y'}$ and $\mathbf{x}^{c:y,y'}$. We can also note that $\mathbf{x}^v$ and $\mathbf{x}^c$ are equal in the case of NCC, the two problems of finding validatory and contrastive prime implicants only differing by the fact that in the validatory case, elements of E are fixed, while they are modified in the contrastive case. It should be noted that this uniqueness of "worst opponent" is not true for more generic models, in the sense that the arguments of Equations (4), (7) and (11a, 11b) in $\mathcal{X}_E$ will typically depend on E.

Coming back to the NCC, we can also see that each of our function is additive on their log form. Indeed, for instance for $\phi^v_{y,y'}$ we have :

$$\begin{aligned} \log \phi^v_{y,y'}(E\cup\{i\}) - \log \phi^v_{y,y'}(E) = \\ (\log\ \underline{p}_i(\mathbf{x}^o_i|y) - \log\ \overline{p}_i(\mathbf{x}^o_i|y')) \\ - (\log\ \underline{p}_i(\mathbf{x}^v_i|y) - \log\ \overline{p}_i(\mathbf{x}^v_i|y')) \end{aligned} \tag{17}$$

As this value is independent of any feature (inside or outside E) different from i. We can therefore define contribution functions G^v, G^c and $G^d_{y,y'}$ mapping each feature i to the contribution of adding i to E :

$$\begin{aligned} G^v(i) = (\log\ \underline{p}_i(\mathbf{x}^o_i|y) - \log\ \overline{p}_i(\mathbf{x}^o_i|y')) \\ - (\log\ \underline{p}_i(\mathbf{x}^v_i|y) - \log\ \overline{p}_i(\mathbf{x}^v_i|y')) \end{aligned} \tag{18a}$$

$$\begin{aligned} G^c(i) = (\log\ \underline{p}_i(\mathbf{x}^c_i|y) - \log\ \overline{p}_i(\mathbf{x}^c_i|y')) \\ - (\log\ \underline{p}_i(\mathbf{x}^o_i|y) - \log\ \overline{p}_i(\mathbf{x}^o_i|y')) \end{aligned} \tag{18b}$$

$$\begin{aligned} G^d_{y,y'}(i) = (\log\ \underline{p}_i(\mathbf{x}^o_i|y) - \log\ \overline{p}_i(\mathbf{x}^o_i|y')) \\ - (\log\ \underline{p}_i(\mathbf{x}^{d:y,y'}_i|y) - \log\ \overline{p}_i(\mathbf{x}^{d:y,y'}_i|y')) \end{aligned} \tag{18c}$$

We see that, by definition, values of functions G^v's are at least zero because we replace the worst opponent value with a better value (the observed one)[5] and is at most 0 for G^c and $G^d_{y,y'}$. It follows that

$$\log \phi^v_{y,y'}(E) = \log \phi^v_{y,y'}(\emptyset) + \sum_{i\in E} G^v(i) \tag{19a}$$

$$\log \phi^c_{y,y'}(E) = \log \phi^c_{y,y'}(\emptyset) + \sum_{i\in E} G^c(i) \tag{19b}$$

$$\log \phi^d_{y,y'}(E) = \log \phi^d_{y,y'}(\emptyset) + \sum_{i\in E} G^d_{y,y'}(i) \tag{19c}$$

We will note the log-contributions of the empty set by C^v, C^c and $C^d_{y,y'}$.

Using these additive rewriting, we will now investigate how to compute our three types of prime implicants (validatory, contrastive and doubt), and the associated complexity.

4.2. Validatory Prime Implicants

From Equation (19a), we have that $\log \phi^v_{y,y'}(E) = C^v + \sum_{i\in E} G^v(i)$ and our goal (4) is to find subsets $E \subseteq [\![1, n]\!]$ such that $\log \phi^v_{y,y'}(E) \geq 0$.

It follows that we want to optimise E so that the sum of positive contributions is greater than C^v. Finding a smallest prime implicant is then computationally easy, as it amounts to order the $G^v(i)'s$ in decreasing order, and add them until $\sum_{i\in E} G^v(i) \geq -C^v$. The whole procedure is summarised in Algorithm 1.

The complexity of Algorithm 1 is linear over the ordered contributions, in number of attributes. Computing the contributions remains easy as the only complexity is to compute the "worst case" vector $\mathbf{x}^v$, whose components $\mathbf{x}^v_i$ requires $|X_i| = k_i$ evaluations on each dimensions. As sets $\mathcal{P}$ are typically polytopes defined by linear constraints, finding the values $\underline{p}$ and $\overline{p}$ amounts to solve linear programs, something that can be done in polynomial time. For some specific cases such as probability intervals [9] (induced, e.g., by the classical Imprecise Dirichlet Model [5]), this can even

[5] Indeed, $\log\ \underline{p}_i(\mathbf{x}^v_i|y) - \log\ \overline{p}_i(\mathbf{x}^v_i|y') < \log\ \underline{p}_i(\mathbf{x}^o_i|y) - \log\ \overline{p}_i(\mathbf{x}^o_i|y')$ by definition.

```
Input: $C^v$; $G^v$
Output: $Xpl = (E, \mathbf{x}^o_E)$: explanation in terms of attribute
Order $G^v$ in decreasing order, with $\sigma$ the associated permutation
  $i \leftarrow 1$
  while $\phi^v_{y,y'}(E) + C^v < 0$ do
    $i \leftarrow i + 1$
    $E \leftarrow E \cup \{\sigma^{-1}(i)\}$
    $\phi^v_{y,y'}(E) \leftarrow \phi^v_{y,y'}(E) + G^v(\sigma(i))$
end
$Xpl \leftarrow (E, \mathbf{x}^o_E)$
  return $(Xpl)$
```

Algorithm 1: Compute first available prime implicants explanation

be done in linear time. Therefore, the overall method is polynomial, with a linear pre-treatment over the sum of k_i's, followed by a sorting algorithm, after which Algorithm 1 is linear over the number of attributes.

4.3. Contrastive Prime Implicants

The case of the contrastive prime implicants is straightforward once we solved the validatory prime implicants. Indeed, as suggested by the similarity between the definitions of $\phi^v_{y,y'}$ and $\phi^c_{y,y'}$ in Equations (16a) and (16b) and the definitions of $\mathbf{x}^v$ and $\mathbf{x}^c$, we almost compute the same thing, the difference being that the role of E for $\phi^v_{y,y'}$ is fulfilled by $-E$ for $\phi^c_{y,y'}$. We obtain that $C^c > 0$ whereas we had $C^v < 0$, as C^c is obtained when we observe the full vector $\mathbf{x}^o$, and that $G^c(i) \leq 0$. To use Algorithm 1, we only need to change the while condition to $\phi^c_{y,y'}(E) + C^c > 0$, and the vector G^v to be ordered in ascending order.

That such strong duality relations hold in general is unlikely, even if validatory and contrastive explanations are known to be linked in general [14].

4.4. Prime Implicants of Doubt

From the definition of prime implicant of doubt in Equations (11a) and (11b) we have to investigate simultaneously two problems, one in favour of y against y' and one in favour of y' against y. To do so we have two functions $\phi^d_{y,y'}$ and $\phi^d_{y',y}$, which in the case of the NCC are additive:

$$\log \phi^d_{y,y'}(E) = C^d_{y,y'} + \sum_{i \in E} G^d_{y,y'}(i),$$
$$\log \phi^d_{y',y}(E) = C^d_{y',y} + \sum_{i \in E} G^d_{y',y}(i).$$

$C^d_{y,y'}$ and $C^d_{y',y}$ are obtained when we assume observing the "worst case opponents" $\mathbf{x}^{d:y,y'}$ and $\mathbf{x}^{d:y',y}$, the two vectors the most in favour of y against y' and of y' against y. In practice, we then want to find which features of $\mathbf{x}^o$ are sufficient to observe so that both dominance relationships (if they hold for some vectors $\mathbf{x}^{d:y,y'}$, $\mathbf{x}^{d:y',y}$, which may not be the case as hinted by Remark 6) are broken, *i.e.*, both $y \not\succ_{\mathcal{P},(\mathbf{x}^o_E,\mathbf{x}^{d:y,y'}_{-E})} y'$ and $y' \not\succ_{\mathcal{P},(\mathbf{x}^o_E,\mathbf{x}^{d:y',y}_{-E})} y$. This problem can be represented as a 2-dimensional Knapsack where the objects are the features and the two Knapsacks corresponds to the dominance of y over y' and the converse. We obtain the following formulation

$$\begin{aligned} &\min \sum_{i=1}^{n} x_i \\ &\text{subject to} \\ &\sum_{i=1}^{n} x_i * G^d_{y,y'}(i) \leq -C^d_{y,y'}, \\ &\sum_{i=1}^{n} x_i * G^d_{y',y}(i) \leq -C^d_{y',y}, \\ &\forall i \in \{1,\ldots,n\}\ x_i \in \{0,1\}. \end{aligned} \tag{20}$$

This problem can be solved by using efficient MILP solver, which may provide fast solutions for the average case, even if the problem worst complexity remains NP-hard. The indexes with a non zero associated x_i are the components of E, *i.e.* are our prime implicants.

4.5. NCC with the Imprecise Dirichlet Model

In this section we will present the Imprecise Dirichlet Model [5], which is a classical model of representation of domain of probabilities, and study how the prime implicants will behave in this case.

The main idea of the IDM is to build a cautious interval around a precise probability distribution. Let's note the number of observation of an event X by n_X, same notations for a conditional event $X|Y$ by $n_{X|Y}$ and N the total number of observation. We obtain that the probability of witnessing X is $\frac{n_X}{N}$. We introduce the meta-parameter s of the IDM which can be interpreted as a number of "unwitnessed" observations. As these could be or not witnessed for X the probability of X belongs to the interval $\left[\frac{n_X}{N+s}, \frac{n_X+s}{N+s}\right]$.

We can easily see that, in the case of the IDM, the s hyper-parameter allows us to go from fully precise ($s = 0$) to fully imprecise ($s = \infty$), meaning that the monotonicity properties we mentioned so far can easily be checked by modifying its value.

5. First Experiments

This Section will present an illustrative case based on the data from the Zoo dataset from UCI repository [11] using the

NCC alongside the IDM. To avoid probabilities of 0 we will regularize them by mixing them with a uniform distribution (using a coefficient $\epsilon = 0.05$ to weight this uniform). The experiment will be separated in two parts. The first one will focus on trying to answer with a quantitative study to the questions "How do the different implicants behave, in size, absence, number, based on the kind of implicants we search, and on whether we justify a relation consistent with the ground truth ?" and "Is the size of pairwise explanations dependent of the imprecision and the number of predicted classes ?". Second one will illustrate how a discussion with a user could occur based on this data for the different types of explanations.

The Zoo dataset is a classification dataset containing 101 samples of animals with 16 input features and the class. The classes are numbers from 1 to 7 corresponding to Mammal, Bird, Reptile, Fish, Amphibian, Bug and Invertebrate. We used 14 features for classification ('feathers': $\{fe, \neg fe\}$, 'eggs': $\{e, \neg e\}$, 'airborne': $\{ai, \neg ai\}$, 'aquatic': $\{aq, \neg aq\}$, 'predator': $\{p, \neg p\}$, 'toothed': $\{to, \neg to\}$, 'backbone': $\{b, \neg b\}$, 'breathes with lungs': $\{l, \neg l\}$, 'venomous': $\{v, \neg v\}$, 'fins': $\{fi, \neg fi\}$, 'legs': $\{0, 2, 4, 5, 6, 8\}$, 'tail': $\{ta, \neg ta\}$, 'domestic': $\{d, \neg d\}$, 'at least catsize': $\{c, \neg c\}$), all binary except for the number of legs. To have sufficient classification errors, we removed 2 features ('hair' and 'milk') from the original features.

5.1. Quantitative Study

We performed this study using a 4-Fold cross-validation with a stratified data separation, due to the samples by class being very unbalanced, e.g. 41 for Mammals and 4 for Amphibians.

Are Validatory and Contrastive explanations size dependent of miss-classification ? The idea behind this question is to verify the shape of explanations when the observation is well classified against when it is miss-classified. If y is the true class, we could expect explanation for an observed dominance $y \succ_{\mathcal{P},\mathbf{x}^o} y'$ that are "true" to differ from observed dominance $y' \succ_{\mathcal{P},\mathbf{x}^o} y$ that are false. In Figure 1, we plot the size of such explanations.

First note that the monotonicity in terms of imprecision are well observed: as s increases, the size of validatory and contrastive explanations respectively increases and decreases.

Then, we can see that there is no significant difference between the distributions when the prediction is correct or not, except maybe for a bigger variability in the case of wrong prediction. While further experiments would be needed to confirm this, it seems that the length of explanation is not related to whether they explain a correct or incorrect prediction, suggesting that one would have to check their plausibility.

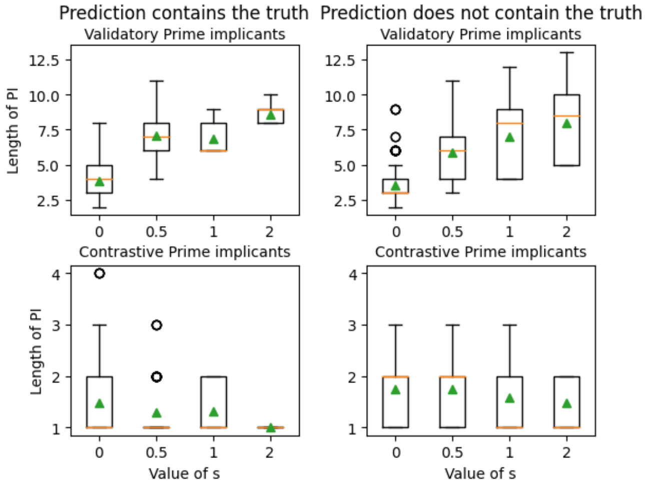

Figure 1: Explanation length according to prediction truth. Green triangles are mean values.

Are Doubt explanation dependent of the number of undominated items? We will now focus on prime implicants of doubt explanations. As said in Section 3.4, incomparabilities may arise from lack of knowledge or from ambiguity about the observed element. A question is then to know whether this affects the length of our explanation.

As a proxy, we plotted in Figure 2 the length of implicants explaining incomparabilities against the number of undominated classes, with the idea that this is a reasonable proxy of ambiguity versus lack of knowledge (the more the number of undominated, the more incomparabilities are likely to be due to lack of knowledge). Again, while the Figure 2 does show the expected monotonicity, it seems that the size of pairwise explanation is not especially affected by the final number of classes in the prediction. As we used a proxy, this independence would however have to be confirmed by more precise assessment of whether our incomparability is mainly due to epistemic or aleatoric uncertainty.

5.2. Illustrative Explanations

Let us now present some results we can get from the experiments. We will focus on values $s \in \{0.5, 1, 2\}$ and on 3 animals: Giraffe, Seal and Tortoise. We will denote NCCs the corresponding classifier.

Giraffe as a non-ambiguous problem. Described by $(\neg fe, \neg e, \neg ai, \neg aq, \neg p, to, b, l, \neg v, \neg fi, 4, ta, \neg d, c)$, the Giraffe is a prototypical example of Mammals, as all NCC0.5, NCC1 and NCC2 classifies it as a Mammal only. To illustrate the validatory prime implicants explanations, we will take a look into the preference "Mammal $\succ_{\mathcal{P},Giraffe}$ Bird". For NCC0.5, a sufficient reason to classify the Giraffe

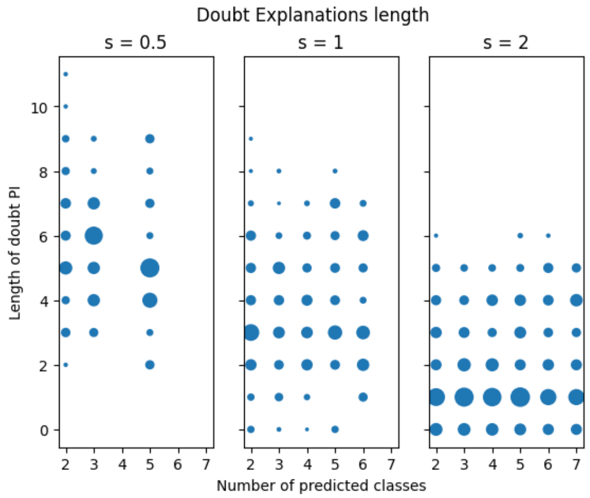

Figure 2: Length of pairwise doubt prime implicants by size of prediction and values of IDM (ball size is normalized with respect to the number of examples having the same number of undominated classes)

as Mammal and not Bird is that it has **no feathers** ($\neg fe$), does **not** produce **eggs** ($\neg e$) and is **toothed** (to). With the increasing cautiousness of NCC1 we need to add the fact that the Giraffe has **4 legs** to the explanation and for NCC2 we then add that it is **not airborne**. All advanced reasons correspond to attributes of Mammals and not of Birds.

A contrastive explanation showing how "robust" our classification is for NCC0.5 and NCC1 that we change the values of the features **feathers**, **eggs** and **toothed**. So, an animal like the giraffe, but with feathers, laying eggs and no teeth could be either a mammal or a bird.

Seal as an ambiguous animal. Described by $(\neg fe, \neg e, \neg ai, aq, p, to, b, l, \neg v, fi, 0, \neg ta, \neg d, c)$, it is classified as a Mammal for NCC0.5, but NCC1 and NCC2 are more cautious by predicting the set $\{Mammal, Reptile, Fish, Amphibian\}$. Let us now investigate the comparison between Mammal (the true class) with Fishes.

For NCC0.5 a sufficient reason to classify as a Mammal is that the Seal **has lungs**, does **not** produce **eggs**, is (at least) **catsized**, is **not venomous**, has **no feathers** and has a **backbone**. Note that this time explanations contain element that support Mammal but can nevertheless be met in fishes as well (e.g., has no feathers). The decision is also less robust, as contrastive explanation shows that flipping one of the features ['lungs', 'eggs', 'catsized', 'venomous'] is enough to make Mammal and Fish incomparable.

When going to NCC1, Mammal $\succ\prec_{\mathcal{P},Seal}$ Fish can be explained by the fact that the Seal does **not** produce **eggs**, is **aquatic**, **breathes**, has **fins**, has **no legs**. Interestingly, we can see that the explanation shows that the seal is somehow ambiguous, having some typical features of fishes (aq, fi) as well as of Mammals $(\neg e, l)$.

Tortoise as a mistaken animal. Described by $(\neg fe, e, \neg ai, \neg aq, \neg p, \neg to, b, l, \neg v, \neg fi, 4, ta, \neg d, c)$, it is wrongly labelled by NCC0.5 and NCC1 as a Mammal rather than a Reptile. NCC2 is much lesser precise and predicts that a Tortoise can be every class except for Fish.

If we investigate the reasons why NCC0.5 believes the Tortoise is a Mammal we obtain the validatory prime implicant **not venomous**, has **4 legs**, is **catsized**, **breathes**, is **not** a **predator**, has a **backbone**, is **not airborne** (for NCC1 we add that it has **no feathers**, is **not aquatic** and has **no fins**.)

The explanation is reasonable but quite long, and does not use the fact that a Tortoise lay eggs (the Platypus being one of the mammal, it is possible for mammals to lay eggs). Also, the Reptile class is poorly represented (4 examples) and most by "serpent like" animals with **no legs**, pretty **venomous**, small (so not **catsized**) and **predators**.

Finally, when increasing to NCC2, we obtain that the doubt between Reptile and Amphibian is not caused by any feature (empty prime implicant of doubt). This clearly shows that Reptile and Amphibian are indistinguishable "by default" and are underrepresented, as the Amphibian and reptile classes have respectively 3 and 4 learning observations which is too little compared to s=2.

6. Discussion and Perspectives

Considering explanations for imprecise classifier opens up many questions, for instance in relation with the possibility of observing incomparability, or of increasing/decreasing the imprecision of a model. In this paper, we focused on prime implicants, extending notions proposed so far in the precise setting. We introduced three notions of prime implicants in the case of pairwise comparison, answering the questions *"Why X is P?"*, *"Why X is P and not Q?"* and *"Why is X neither P nor Q?"*. When applying them to the Naïve credal classifier, we obtain that the computations are computationally easy (at least for validatory and contrastive explanations).

Complexity beyond the NCC case While Section 4 showed that the notions of Section 3 could be computed relatively easily for some case, it is clear that applying them in general will present many challenges, as the problem of finding prime implicants with minimal cardinality or of enumerating prime implicants is known to be NP-hard. Coming up with methods to extract such prime implicants from other credal classification methods therefore constitute an important avenue for future research. We are however relatively confident that the notion of (prime) implicants could be used in numerous cases, notably for the following reasons:

- while considering NCC results in an additive structure making the problme of finding prime implicants linear in the number of features, other structural assumptions may also make this task easy. For instance, monotone classifiers [17] and decision trees [16] also exhibit a structure that allows for efficient algorithms in the precise case;
- while finding prime implicants for other classifiers such as random forest can turn out to be quite complex [15], relaxing the notion of prime implicants to weaker notions such as being a prime implicant for a majority of trees may allow for efficient, polynomial search methods [2];
- many classification problems do not include that many features, meaning that even if the problem of finding or enumerating prime implicants is computationally challenging, one could still use, e.g., ILP formulations of the problem with powerful solver to obtain a solution.

Some perspectives In the future, we would like to focus on various questions not investigated here, such as for which robust models (e.g., including some dependence statements) do computations remain tractable? What happens with interaction between attributes ? When trying to explain the complete partial order, should we use pairwise or holistic (i.e., prime implicants explaining the non-dominated classes at once) explanations? How do we select what dominance to explain in such a case ? There are also several other explanation mechanisms we could consider [1]. We also want to investigate the case with prediction costs. Indeed, we can associate costs to making a prediction y' when the truth is y. During this paper we used the usual 0/1 cost for all types of miss-classification.

Finally, we also feel that we have only skimmed the surface of the role of incomparability explanations, in the sense that the operational role and advantage of such explanations still remain to be explored.

References

[1] Gilles Audemard, Frédéric Koriche, and Pierre Marquis. On tractable xai queries based on compiled representations. In *Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning*, volume 17, pages 838–849, 2020.

[2] Gilles Audemard, Steve Bellart, Louenas Bounia, Frédéric Koriche, Jean-Marie Lagniez, and Pierre Marquis. Trading complexity for sparsity in random forest explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5461–5469, 2022.

[3] Vineeth Balasubramanian, Shen-Shyang Ho, and Vladimir Vovk. *Conformal prediction for reliable machine learning: theory, adaptations and applications*. Newnes, 2014.

[4] Khaled Belahcene, Christophe Labreuche, Nicolas Maudet, Vincent Mousseau, and Wassila Ouerdane. Explaining robust additive utility models by sequences of preference swaps. *Theory and Decision*, 82(2):151–183, 2017.

[5] Jean-Marc Bernard. An introduction to the imprecise dirichlet model for multinomial data. *International Journal of Approximate Reasoning*, 39(2-3):123–150, 2005.

[6] Nadia Burkart and Marco F Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.

[7] Seth Chin-Parker and Julie Cantelon. Contrastive constraints guide explanation-based category learning. *Cognitive Science*, 41(6):1645–1655, 2017. doi:https://doi.org/10.1111/cogs.12405.

[8] Giorgio Corani, Alessandro Antonucci, and Marco Zaffalon. Bayesian networks with imprecise probabilities: Theory and application to classification. In *Data Mining: Foundations and Intelligent Paradigms*, pages 49–93. Springer, 2012.

[9] Luis M De Campos, Juan F Huete, and Serafin Moral. Probability intervals: a tool for uncertain reasoning. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 2(02):167–196, 1994.

[10] Juan José Del Coz, Jorge Díez, and Antonio Bahamonde. Learning nondeterministic classifiers. *Journal of Machine Learning Research*, 10(10), 2009.

[11] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[12] Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Data Mining and Knowledge Discovery*, pages 1–55, 2022.

[13] Eyke Hüllermeier, Sébastien Destercke, and Mohammad Hossein Shaker. Quantification of credal uncertainty in machine learning: A critical analysis and empirical comparison. In *Uncertainty in Artificial Intelligence*, pages 548–557. PMLR, 2022.

[14] Alexey Ignatiev, Nina Narodytska, Nicholas Asher, and Joao Marques-Silva. On relating'why?'and'why not?'explanations. *arXiv preprint arXiv:2012.11067*, 2020.

[15] Yacine Izza and João Marques Silva. On explaining random forests with sat. In *30th International Joint Conference on Artificial Intelligence (IJCAI 2021)*. International Joint Conferences on Artifical Intelligence (IJCAI), 2021.

[16] Yacine Izza, Alexey Ignatiev, and Joao Marques-Silva. On explaining decision trees. *arXiv preprint arXiv:2010.11034*, 2020.

[17] Joao Marques-Silva, Thomas Gerspacher, Martin C Cooper, Alexey Ignatiev, and Nina Narodytska. Explanations for monotonic classifiers. In *International Conference on Machine Learning*, pages 7469–7479. PMLR, 2021.

[18] João Marques-Silva, Thomas Gerspacher, Martin C. Cooper, Alexey Ignatiev, and Nina Narodytska. Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay. In *NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[19] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artif. Intell.*, 267: 1–38, 2019.

[20] Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining bayesian network classifiers. *arXiv preprint arXiv:1805.03364*, 2018.

[21] Lev V Utkin, Andrei V Konstantinov, and Kirill A Vishniakov. An imprecise shap as a tool for explaining the class probability distributions under limited training data. *arXiv preprint arXiv:2106.09111*, 2021.

[22] Guy Van den Broeck, Anton Lykov, Maximilian Schleich, and Dan Suciu. On the tractability of shap explanations. In *Proceedings of the 35th AAAI*, 2021.

[23] Joseph Jay Williams and Tania Lombrozo. The role of explanation in discovery and generalization: evidence from category learning. *Cognitive science*, 34 5:776–806, 2010.

[24] Hénoïk Willot, Sébastien Destercke, and Khaled Belahcene. Explaining robust classification through prime implicants. In *Scalable Uncertainty Management: 15th International Conference, SUM 2022, Paris, France, October 17–19, 2022, Proceedings*, pages 361–369. Springer, 2022.

[25] Marco Zaffalon. The naive credal classifier. *Journal of Statistical Planning and Inference*, 105(1):5–21, 2002.

[26] Haifei Zhang, Benjamin Quost, and Marie-Hélène Masson. Explaining cautious random forests via counterfactuals. In *International Conference on Soft Methods in Probability and Statistics*, pages 390–397. Springer, 2023.