Explaining Robust Classification Through Prime Implicants

Hénoïk Willot, Sébastien Destercke, Khaled Belahcene

Explaining robust classification through prime implicants

Hénoïk Willot, Sébastien Destercke, and Khaled Belahcene

Heudiasyc, Université de Technologie de Compiègne, France
{henoik.willot, sebastien.destercke, khaled.belahcene}@hds.utc.fr

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.

1 Introduction

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

Regarding explainability, the notion of prime implicants corresponds to provide minimal sufficient condition to make a given prediction, 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 [12], with very efficient procedure existing for specific structures such as the Naive one [11]. In contrast with other methods such as SHAP [6] 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 if not exclusively applied to precise models, at least in the machine learning domain (this is less true, e.g., in preference modelling [4]). Yet, in some applications involving sensitive issues or where the decision maker wants to identify ambiguous cases, it may be preferable to use models that will return sets of classes in some cases where information is missing rather than always returning a point-valued prediction. Several frameworks such as conformal prediction [3], indeterminate classifiers [9] or imprecise probabilistic models [7] have been proposed to handle such issue.

The later 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. This will be done in Section 2. As the formulated problem is likely to be computationally challenging for generic models, we focus in Section 3 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 example illustrating our approach.

2 Setting and general problem formulation

In this section, we lay down our basic notations and provide necessary reminders about imprecise probabilities. We also introduce the idea of prime implicants applied to classifiers, and particular to imprecise probabilistic classifiers.

2.1 Robust classification: setting

We consider a usual discrete multi-class problem, where we must predict a variable Y taking values in $\mathcal{Y} = \{y_1, \ldots, y_m\}$ using n input variables $X_1, \ldots, X_n$ that respectively takes values in $\mathcal{X}_i = \{x_i^1, \ldots, x_i^{k_i}\}$. 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 $E \subseteq \{1, \ldots, n\}$ 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 $-E := \{1, \ldots, n\} \setminus E$ 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.

When considering precise probabilistic classifiers, a class y is said to weakly dominate[1] y', written $y \succeq_p y'$, upon observing a vector $\mathbf{x}$ when the condition[2]

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

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

[1] We work in a non cost-sensitive framework, but most of our discussion easily transfer to such cases.

[2] Using dominance expressed this way will be useful in the sequel.

$p \in \mathcal{P}$. In this case, y is said to robustly dominate y', written $y \succeq_{\mathcal{P}} y'$, upon observing a vector $\mathbf{x}$ when the condition

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

is met, or in other words when $p(y|\mathbf{x}) \geq p(y'|\mathbf{x})$ for all $p \in \mathcal{P}$. Note that the relation $\succeq_{\mathcal{P}}$ can be a partial pre-order with incomparabilities, whereas $\succeq_p$ is a pre-order.

2.2 Explaining robust classification through prime implicants

Explaining the conclusion or deduction of an algorithm, and in particular of a learning algorithm, has become an important issue. 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. When observing a vector $\mathbf{x}^o$ and making a prediction about whether y dominates y', the idea of prime implicant roughly translates as the values of $\mathbf{x}^o$ that are sufficient to know to state that y dominates y', and that are minimal with this property.

With this idea in mind, we will say that a subset $E \subseteq \{1, \ldots, n\}$ of attributes (where E are the indices of the considered attributes) is an *implicant* of $y \succeq_{\mathcal{P}} y'$ iff

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

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 $\mathbf{x}^o_E$ alone is sufficient to deduce $y \succeq_{\mathcal{P}} y'$. A set E is a *prime implicant* iff we satisfy (3) and for any $i \in E$, we have

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

that is if removing any attribute from E makes our deduction invalid, so that E is a minimal sufficient condition for $y \succeq_{\mathcal{P}} y'$ to hold. In the sequel, it will prove useful to consider the function $\phi(E)$ that associates to each possible subset the value

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

$\phi(E)$ being inclusion-monotonic (for $E \subseteq F$, $\phi(E) \leq \phi(F)$), it can be seen as a value function associated to E, and finding a prime implicant can then be seen as the task of finding a minimal "bundle of items"[3] E such that $\phi(E) \geq 1$, therefore allowing us to map the finding of robust prime implicants to an item selection problem. Also note that, in general, $\phi(E)$ will not be additive, as we will not have $\phi(E \cup \{i\}) = \phi(E) + \phi(\{i\})$.

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

Note that when sets $\mathcal{P}$ reduce to singletons, that is when we consider precise classifiers instead of robust ones, then our notion of prime implicant reduces to previously proposed ones [11], and our approach is therefore a formal generalisation of those.

3 The case of the Naive credal classifier

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

3.1 Generic case

The basic idea of the Naive classifier is to assume that attributes are independent of each other given the class. This modelling assumption means that

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

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

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

with every $p_i(|y)$ independent of $p_i(|y')$, and every $p_i(|y), p_j(|y)$ independent for i, j. When switching to credal models, one has sets of conditional distributions $\mathcal{P}_i(|y)$ and a set $\mathcal{P}_{\mathcal{Y}}$ of priors.

Let us now see how Equation (3) transform in this case. We do have

$$\begin{aligned} &\inf_{\substack{p\in\mathcal{P}\\ \mathbf{x}^a_{-E}\in X_{-E}}} \frac{p(y|(\mathbf{x}^o_E, \mathbf{x}^a_{-E}))}{p(y'|(\mathbf{x}^o_E, \mathbf{x}^a_{-E}))} = \\ &\inf_{\substack{p\in\mathcal{P}\\ \mathbf{x}^a_{-E}\in X_{-E}}} \frac{p_{\mathcal{Y}}(y)}{p_{\mathcal{Y}}(y')} \underbrace{\prod_{i\in E} \frac{p_i(x^o_i|y)}{p_i(x^o_i|y')}}_{PartA} \underbrace{\prod_{i\notin E} \frac{p_i(x^a_i|y)}{p_i(x^a_i|y')}}_{PartB}. \end{aligned} \tag{6}$$

In Equation (6), we can treat the minimization problem of parts A and B completely separately, due to two principal observations. First, the sets $\mathcal{P}_i(|y)$ are all independent when i (the attribute) or y (the conditioning element) changes. This implies that part A and B are minimised over independent convex sets of probabilities (as they are over distinct i's), and that the numerator and denominators of each fraction within the two parts can also be treated separately (being conditioned on different y, y'). Second, E and $-E$ are disjoint, which means that the value $\mathbf{x}^a_{-E}$ for which part B is minimised only depends on part B, hence in

this case it makes sense to define a unique "worst case" vector $\mathbf{x}^{a*}$ which minimises part B for any E. Also, since conditional laws with different conditional classes are independent, we get that Equation (6) becomes

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

where $\underline{p}(x) = \inf_{p \in \mathcal{P}} p(x)$ and $\overline{p}(x) = \sup_{p \in \mathcal{P}} p(x)$. If we consider the vector $\mathbf{x}^{a*}_{-E}$, we finally have

$$\inf_{\substack{p \in \mathcal{P} \\ \mathbf{x}^a_{-E} \in X_{-E}}} \frac{p(y|(\mathbf{x}^o_E, \mathbf{x}^a_{-E}))}{p(y'|(\mathbf{x}^o_E, \mathbf{x}^a_{-E}))} = \inf_{p_{\mathcal{Y}} \in \mathcal{P}_{\mathcal{Y}}} \frac{p_{\mathcal{Y}}(y)}{p_{\mathcal{Y}}(y')} \prod_{i \in E} \frac{\underline{p}_i(x_i^o|y)}{\overline{p}_i(x_i^o|y')} \prod_{i \notin E} \frac{\underline{p}_i(x_i^{a*}|y)}{\overline{p}_i(x_i^{a*}|y')} \tag{8}$$

Let us now go back to our idea of selecting minimal bundle of items (or attribute) making $\phi(E) > 1$ or equivalently $\log \phi(E) > 0$. Let us first note by

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

the value of $\log \phi(\emptyset)$, and by

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

the positive[4] gain obtained by adding element i to E. Developing Equation (7), one can check that

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

has an additive form. Finding a smallest prime implicant is then computationally easy, as it amounts to order the $G_i's$ in decreasing order, and add them until $\sum_{i \in E} G_i \geq -C$. The whole procedure is summarised in Algorithm 1.

The complexity of Algorithm 1 is obviously 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 $\mathbf{x}^{a*}$, whose components $\mathbf{x}_i^{a*}$ 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 [8] (induced, e.g., by the classical Imprecise Dirichlet Model [5]), this can even be done in linear time. Therefore, the overall method is clearly 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] As $\log \underline{p}_i(x_i^{a*}|y) - \log \overline{p}_i(x_i^{a*}|y') < \log \underline{p}_i(x_i^o|y) - \log \overline{p}_i(x_i^o|y')$ by definition.

Algorithm 1: Compute first available prime implicants explanation

Input: $C : log(\phi(\emptyset))$; G : Contributions of criteria in decreasing order;
Output: $Xpl = (E, \mathbf{x}_E)$: explanation in terms of attribute

Order G in decreasing order, with σ the associated permutation
$i \leftarrow 1$
while $\phi(E) + C < 0$ **do**
  $i \leftarrow i + 1$
  $E \leftarrow E \cup \{\sigma^{-1}(i)\}$
  $\phi(E) \leftarrow \phi(E) + G_{\sigma(i)}$
$Xpl \leftarrow (E, \mathbf{x}_E^o)$
return (Xpl)

3.2 Illustrative case

We will present a small illustrative example using categorical data and probability intervals. Those later could, for instance, be obtained through the use of the classical Imprecise Dirichlet Model [5], possibly with some regularisation to avoid zero probabilities, or in the case of continuous variable, by parametric [1] or non-parametric models [10].

In this example we want to predict the class of animal from its physical characteristics. We have data concerning the set $\mathcal{Y} = \{$D(og), C(at), H(orse), B(unny)$\}$ of animals and observe the length of their $\mathcal{X} =\{$E(ars), T(ail), H(air)$\}$. Each of these criteria can have a value in $\{L(ong), A(verage), S(hort)\}$. To identify easily variables in the example, we will use the notation LE for long ears, and similarly for all other attribute combinations. The prior probabilities are presented in table 1 and the conditioned probabilities in table 2, 3 and 4.

Dog	Cat	Horse	Bunny
$[0.25, 0.26]$	$[0.29, 0.31]$	$[0.20, 0.22]$	$[0.25, 0.26]$

Table 1: Probability intervals of each animal class

Assume that we observe the vector $\mathbf{x}^o =$ (Long Ear, Short Tail, Long Hair) or (LE,ST,LH) for short. As we are using an imprecise classification model, the predicted classes will correspond to the non dominated classes, and our explanations will mostly be used to understand why we rejected the other classes. For every pair (y,y') of animals we compare $\inf_{p \in \mathcal{P}} \frac{p(y|\mathbf{x})}{p(y'|\mathbf{x})}$ to 1 to build the partial order between them. In our specific case, this comes down to compare

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

		Animal			
		Dog	Cat	Horse	Bunny
Length	Long	[0.33,0.40]	[0.02,0.08]	[0.10,0.19]	[0.58,0.65]
	Average	[0.30,0.37]	[0.55,0.61]	[0.66,0.75]	[0.26,0.33]
	Short	[0.30,0.37]	[0.37,0.43]	[0.15,0.23]	[0.09,0.16]

Table 2: Conditional probabilities of the length of the ears knowing the animal

		Animal			
		Dog	Cat	Horse	Bunny
Length	Long	[0.54,0.61]	[0.31,0.37]	[0.66,0.75]	[0.02,0.09]
	Average	[0.23,0.30]	[0.61,0.67]	[0.23,0.32]	[0.30,0.37]
	Short	[0.16,0.23]	[0.02,0.08]	[0.02,0.10]	[0.61,0.69]

Table 3: Conditional probabilities of the length of the tail knowing the animal

		Animal			
		Dog	Cat	Horse	Bunny
Length	Long	[0.40,0.47]	[0.46,0.52]	[0.23,0.32]	[0.02,0.09]
	Average	[0.26,0.33]	[0.17,0.22]	[0.10,0.19]	[0.19,0.26]
	Short	[0.26,0.33]	[0.31,0.37]	[0.58,0.66]	[0.72,0.79]

Table 4: Conditional probabilities of the length of hair knowing the animal

to 0. As we have probability intervals, the bound $\underline{p}$ (resp. $\overline{p}$) can be read directly from the tables. Taking the pair (Dog, Horse) or (D,H) as an example, we have

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

We then have that $D \succeq_{\mathcal{P}} H$. Repeating this for all pairs, we obtain the partial order in Figure 1. The cautious prediction will be $\{D, B\}$, and each arc of Figure 1 can be explained with prime implicants.

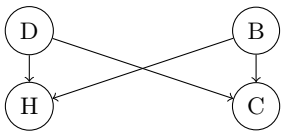

Fig. 1: Class dominance for prediction of $\mathbf{x}^o$ = (LE, ST, LH)

We detail the computation only for $D \succeq_{\mathcal{P}} H$. First we need to compute the worst opponent $\mathbf{x}^a$ that minimises $log\ \underline{p}(x_i^{a*} \mid D) - log\ \overline{p}(x_i^a \mid H)$ for each variable i. We obtain $\mathbf{x}^{a*} = (AE, AT, SH)$. Applying Equation (9), we obtain

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

The contributions of the criteria required by Algorithm 1 are :

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

We can now apply Algorithm 1 and we obtain the explanation {(Ears, Long), (Hair, Long)} as (0.65 + 0.50) - 0.90 > 0, but with an enumeration algorithm we would find a second prime implicant explanation with {(Ears, Long), (Tail, Short)}, as (0.65+0.33) -0.90 > 0, that is less important in terms of gain, but maybe intuitively more satisfying. Similarly we can compute explanations for other dominances, like {(Ears, Long), (Tail, Short)} for Dog $\succeq_{\mathcal{P}}$ Cat, {(Ears, Long), (Tail, Short)} for Bunny $\succeq_{\mathcal{P}}$ Cat and {(Ears, Long), (Tail, Short)} for Bunny $\succeq_{\mathcal{P}}$ Horse.

4 Conclusion

This paper proposes to explain robust classification by prime implicants, extending notions proposed so far in the precise setting. We show that, as for the precise

case, this task is easy for the Naive classifier. To our knowledge, this is the first attempt to combine imprecise probabilistic classification with explanation.

In the future, we would like to focus on various questions not investigated here, such as: does enumerating all prime implicants remain easy for the naive credal classifier? For which robust models (e.g., including some dependence statements) do computations remain tractable? What happens with interaction between attributes ? Can we explain incomparabilities with similar notions? 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? There are also several other explanation mechanisms we could consider [2].

References

1. Alarcón, Y.C.C., Destercke, S.: Imprecise gaussian discriminant classification. Pattern Recognition **112**, 107739 (2019)
2. Audemard, G., Koriche, F., Marquis, P.: On tractable xai queries based on compiled representations. In: Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning. vol. 17, pp. 838–849 (2020)
3. Balasubramanian, V., Ho, S.S., Vovk, V.: Conformal prediction for reliable machine learning: theory, adaptations and applications. Newnes (2014)
4. Belahcene, K., Labreuche, C., Maudet, N., Mousseau, V., Ouerdane, W.: Explaining robust additive utility models by sequences of preference swaps. Theory and Decision **82**(2), 151–183 (2017)
5. Bernard, J.M.: An introduction to the imprecise dirichlet model for multinomial data. International Journal of Approximate Reasoning **39**(2-3), 123–150 (2005)
6. Van den Broeck, G., Lykov, A., Schleich, M., Suciu, D.: On the tractability of shap explanations. In: Proceedings of the 35th AAAI (2021)
7. Corani, G., Antonucci, A., Zaffalon, M.: Bayesian networks with imprecise probabilities: Theory and application to classification. In: Data Mining: Foundations and Intelligent Paradigms, pp. 49–93. Springer (2012)
8. De Campos, L.M., Huete, J.F., Moral, S.: Probability intervals: a tool for uncertain reasoning. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems **2**(02), 167–196 (1994)
9. Del Coz, J.J., Díez, J., Bahamonde, A.: Learning nondeterministic classifiers. Journal of Machine Learning Research **10**(10) (2009)
10. Dendievel, G., Destercke, S., Wachalski, P.: Density estimation with imprecise kernels: application to classification. In: SMPS proceedings. pp. 59–67. Springer (2018)
11. Marques-Silva, J., Gerspacher, T., Cooper, M.C., Ignatiev, A., Narodytska, N.: Explaining Naive Bayes and Other Linear Classifiers with Polynomial Time and Delay. In: NeurIPS 2020, December 6-12, 2020, virtual (2020)
12. Shih, A., Choi, A., Darwiche, A.: A symbolic approach to explaining bayesian network classifiers. arXiv preprint arXiv:1805.03364 (2018)