Relation-Based Counterfactual Explanations for Bayesian Network Classifiers

Emanuele Albini[1,2] , **Antonio Rago**[2] , **Pietro Baroni**[1] and **Francesca Toni**[2]
[1]Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Brescia, Italy
[2]Department of Computing, Imperial College London, UK
{emanuele.albini19, a.rago15}@imperial.ac.uk, pietro.baroni@unibs.it, ft@imperial.ac.uk

Abstract

We propose a general method for generating *counterfactual explanations* (CFXs) for a range of Bayesian Network Classifiers (BCs), e.g. single- or multi-label, binary or multidimensional. We focus on explanations built from relations of (*critical* and *potential*) *influence* between variables, indicating the reasons for classifications, rather than any probabilistic information. We show by means of a theoretical analysis of CFXs' properties that they serve the purpose of indicating (potentially) pivotal factors in the classification process, whose absence would give rise to different classifications. We then prove empirically for various BCs that CFXs provide useful information in real world settings, e.g. when race plays a part in parole violation prediction, and show that they have inherent advantages over existing explanation methods in the literature.

1 Introduction

One of the most pressing issues in AI is the lack of explainability of many of its methods. In recent years, this has been accelerated from within academia and industry, as well as by government regulations and guidance by policy makers, e.g. the EU Ethics guidelines on trustworthy AI[1]. Indeed, there has been an influx of general-purpose, *model-agnostic* methods for generating explanations for AI systems (e.g. see [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017; Ribeiro *et al.*, 2018] for some popular proposals and [Guidotti *et al.*, 2019] for a recent survey), as well as explanation methods tailored to specific reasoning or machine learning methods (e.g. see [Bach *et al.*, 2015] for explanations for neural networks). In this paper we give a method-specific approach.

Bayesian network classifiers (BCs) [Friedman *et al.*, 1997] are probabilistic reasoning models whose underlying mechanism is a Bayesian network [Pearl, 1989]. Many different forms of BC exist in the literature (see [Bielza and Larrañaga, 2014] for an overview of the *discrete* BCs that we consider in this paper), e.g. *naive* [Maron and Kuhns, 1960] or *Markov blanket-based* [Koller and Sahami, 1996], but their classification methods are based on the same fundamental principles. It is well known that BCs are inherently *interpretable*, however, the generation of suitable explanations of their results depends on the user requirements arising in their application. Many methods for explaining Bayesian networks have been defined (see [Lacave and Díez, 2002] for an overview), including abduction methods, e.g. [Pearl, 1989], explaining evidence by making inferences about unobserved variables, and model explanation, which can be at a micro/variable or macro/model level, e.g. a graphical display of the probabilities [Lacave *et al.*, 2000]. Recently, [Shih *et al.*, 2018; Shih *et al.*, 2019] have shown how representing restricted forms of BCs as tractable *decision functions* can open up new pathways towards two novel forms of explanations. In this paper we provide novel *counterfactual explanations* (CFXs) for *generic* BCs, based on the causal reasoning underpinning them. CFXs allow us to answer the question: what would have caused the BC to determine a different classification?

Our explanation method (Section 2) relies upon mapping the *influences* between a BC's *variables*, e.g. between observations and classifications. From these influences we extract two relations between variables deemed to be relevant to the CFXs. These relations amount, respectively, to *critical* and *potential* influences, indicating (potentially) pivotal factors, whose absence would give rise to a different classification. We identify (Section 3) formal properties of CFXs, in particular pointing to the informative role of the relations we define. We then evaluate CFXs empirically (Section 5) by testing them with a range of datasets and BCs, showing that they perform well against the baselines from [Shih *et al.*, 2018] (re-defined, for the purposes of this paper, in Section 4), highlighting the advantages that CFXs hold in certain settings, and exhibit desirable behaviour in real world situations.

2 Explaining Bayesian Classifiers

In this section we define a novel notion of counterfactual explanations for (an abstract representation of) BCs. For the purposes of this paper, a BC consists of *variables*, which may be *classifications* or *observations*, *conditional dependencies* between variables, *domains* of *values* that can be ascribed to variables, and an *evaluation*, i.e. a mapping from (assignments of values to) observations to (assignments of values to) classifications. Formally:

[1] See https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

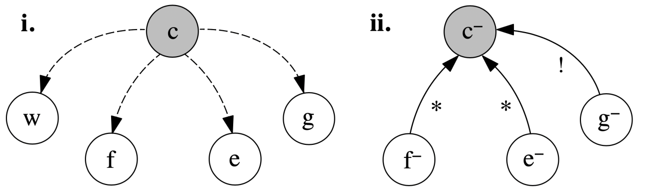

Figure 1: University admissions BC (i) with observations in white, classifications in grey and conditional dependencies as dashed arrows and (ii) corresponding CFX for the evaluation with $w\bar{e}\bar{f}\bar{g}$, with critical and potential influences indicated by ! and $*$, respectively.

Definition 1. *A* BC *is a tuple* $\langle\mathcal{O},\mathcal{C},\mathcal{D},\mathcal{V},\sigma\rangle$ *such that:*
- $\mathcal{O}$ *is a (finite) set of* observations*;*
- $\mathcal{C}$ *is a (finite) set of* classifications*; we refer to* $\mathcal{X}=\mathcal{O}\cup\mathcal{C}$ *as the set of* variables*;*
- $\mathcal{D}\subseteq\mathcal{X}\times\mathcal{X}$ *is the set of* conditional dependencies *between the variables; for any* $\alpha\in\mathcal{X}$*,* $\mathcal{D}(\alpha)=\{\beta\in\mathcal{X}|(\beta,\alpha)\in\mathcal{D}\}$ *are the conditional dependencies of* α*;*
- $\mathcal{V}$ *is a set of sets such that for any* $x\in\mathcal{X}$ *there is a unique* $v\in\mathcal{V}$ *associated to* x*, called the* domain *of* x *(*$\mathcal{V}(x)$ *for short); we refer to the elements of* $\mathcal{V}(x)$ *as* values *for* x*;*
- *let an* assignment *to a set of variables* $S\subseteq\mathcal{X}$ *be a mapping* $a_S : S\rightarrow\bigcup_{x\in S}\mathcal{V}(x)$ *such that, for all* $x\in S$*,* $a_S(x)\in\mathcal{V}(x)$*, and let* $\mathcal{A}_S$ *be the set of all assignments to variables in* S*; then the BC realises a mapping* $\sigma:\mathcal{A}_{\mathcal{O}}\rightarrow\mathcal{A}_{\mathcal{C}}$ *from (assignments to) observations to (assignments to) classifications; we denote* $\{\langle a_{\mathcal{O}},\sigma(a_{\mathcal{O}})\rangle \mid a_{\mathcal{O}}\in\mathcal{A}_{\mathcal{O}}\}$ *as* Σ *and refer to any pair* $\epsilon\in\Sigma$ *as an* evaluation *by the BC.*

For the sake of concision, given $\epsilon{=}\langle a_{\mathcal{O}},\sigma(a_{\mathcal{O}})\rangle{\in}\Sigma$, for any $o{\in}\mathcal{O},c{\in}\mathcal{C}$ we refer to $a_{\mathcal{O}}(o)$ as $\epsilon(o)$ and to $\sigma(a_{\mathcal{O}})(c)$ as $\epsilon(c)$.

For illustration, consider an example from [Shih *et al.*, 2018], in which a university admissions decision c is predicted using a BC with 4 observations: *prior work experience*, *first-time applicant*, *passed entrance exam* and *met GPA*. The graph in Figure 1i visualises, for this BC, $\mathcal{O}=\{w,f,e,g\}$, $\mathcal{C}=\{c\}$, and $\mathcal{D}=\{(c,w),(c,f),(c,e),(c,g)\}$. This BC is *binary*, i.e. $\forall x\in\mathcal{X}$, $\mathcal{V}(x)=\{+,-\}$. The first column in Table 1 gives Σ for the BC in this example, amounting to 16 possible evaluations, with each row an evaluation in Σ (e.g. the first row amounts to ϵ with observations and classification all assigned to $-$, with e.g. $\epsilon(w)=-$ and $\epsilon(c)=-$).

Our definition of BC does not include any probabilistic information, but the evaluations result from an underlying probability distribution, obtained when the classifier is learnt, but compiled away in providing evaluations (e.g. by setting the classifier to predict the most likely classification for a set of observations). For example, in the university admissions BC, σ amounts to a simple decision function, which may result from the use of a threshold t to determine a positive result for an assignment to observations if the probability of the single classification c being positive, given the assignments, is equal to or greater than t. The representation of BCs as sets of evaluations Σ/decision functions is a popular topic of late when explaining their output (e.g. see [Shih *et al.*, 2019]). Indeed, some argue that explanations based on simple relational considerations are preferred to explanations resorting to more complex conceptual machineries (e.g. probabilistic considerations) [Lombrozo, 2007]. Note that the problem of extracting Σ is not the focus of this paper.

Σ					CFX	MCX	PIX
w	f	e	g	c	$\mathcal{R}_!(c)/\mathcal{R}_*(c)$		
$-$	$-$	$-$	$-$	$-$	$\{\}/\{\bar{w}\bar{f}\bar{e}\bar{g}\}$	$(\bar{w}\bar{f})(\bar{w}\bar{e})(\bar{w}\bar{g})(\bar{f}\bar{g})(\bar{e}\bar{g})$	$(\bar{w}\bar{f})(\bar{w}\bar{e})(\bar{w}\bar{g})(\bar{f}\bar{g})(\bar{e}\bar{g})$
$-$	$-$	$-$	$+$	$-$	$\{\bar{w}\}/\{\bar{f}\bar{e}\}$	$(\bar{w}\bar{f})(\bar{w}\bar{e})$	$(\bar{w}\bar{f})(\bar{w}\bar{e})$
$-$	$-$	$+$	$-$	$-$	$\{\}/\{\bar{w}\bar{f}\bar{g}\}$	$(\bar{w}\bar{f})(\bar{w}\bar{g})(\bar{f}\bar{g})$	$(\bar{w}\bar{f})(\bar{w}\bar{g})(\bar{f}\bar{g})$
$-$	$-$	$+$	$+$	$-$	$\{\bar{w}\bar{f}\}/\{\}$	$(\bar{w}\bar{f})$	$(\bar{w}\bar{f})$
$-$	$+$	$-$	$-$	$-$	$\{\}/\{\bar{w}\bar{e}\bar{g}\}$	$(\bar{w}\bar{e})(\bar{w}\bar{g})(\bar{e}\bar{g})$	$(\bar{w}\bar{e})(\bar{w}\bar{g})(\bar{e}\bar{g})$
$-$	$+$	$-$	$+$	$-$	$\{\bar{w}\bar{e}\}/\{\}$	$(\bar{w}\bar{e})$	$(\bar{w}\bar{e})$
$-$	$+$	$+$	$-$	$-$	$\{\bar{w}\bar{g}\}/\{\}$	$(\bar{w}\bar{g})$	$(\bar{w}\bar{g})$
$-$	$+$	$+$	$+$	$+$	$\{feg\}/\{\}$	(feg)	(feg)
$+$	$-$	$-$	$-$	$-$	$\{\bar{g}\}/\{\bar{f}\bar{e}\}$	$(\bar{f}\bar{g})(\bar{e}\bar{g})$	$(\bar{f}\bar{g})(\bar{e}\bar{g})$
$+$	$-$	$-$	$+$	$+$	$\{wg\}/\{\}$	(wg)	(wg)
$+$	$-$	$+$	$-$	$-$	$\{\bar{f}\bar{g}\}/\{\}$	$(\bar{f}\bar{g})$	$(\bar{f}\bar{g})$
$+$	$-$	$+$	$+$	$+$	$\{wg\}/\{e\}$	(wg)	(wg)
$+$	$+$	$-$	$-$	$-$	$\{\bar{e}\bar{g}\}/\{\}$	$(\bar{e}\bar{g})$	$(\bar{e}\bar{g})$
$+$	$+$	$-$	$+$	$+$	$\{wg\}/\{f\}$	(wg)	(wg)
$+$	$+$	$+$	$-$	$+$	$\{wfe\}/\{\}$	(wfe)	(wfe)
$+$	$+$	$+$	$+$	$+$	$\{\}/\{wfeg\}$	(wg)	$(wg)(wfe)(feg)$

Table 1: Evaluations and explanations for the university admissions BC, where, e.g., w signifies $\epsilon(w)=+$ and $\bar{w}$ signifies $\epsilon(w)=-$.

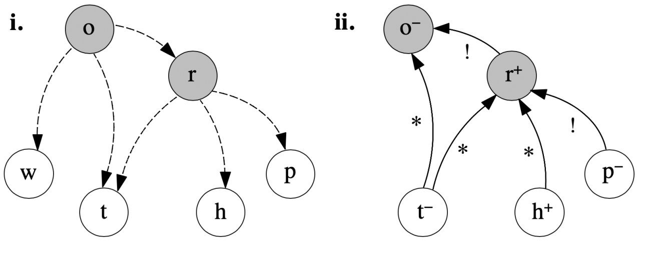

Figure 2: Play-outside BC (i) with observations in white, classifications in grey and conditional dependencies as dashed arrows and (ii) corresponding CFX for the evaluation with $\bar{w}\bar{t}h\bar{p}$, with critical and potential influences indicated by ! and $*$, respectively.

In our definition of BCs, variables may have different domains or share the same domains. BCs can be *binary*, when the domains of variables consist of exactly two values, e.g. $\mathcal{V}(x)=\{+,-\}$ for all $x\in\mathcal{X}$ (as in the university admissions BC), or *multidimensional* (non-binary), where no restrictions are imposed on the domains of variables.

Finally, BCs may be single- or multi-label, depending on whether $\mathcal{C}$ is a singleton or not, respectively. Single-label BCs may be *naive* (as in the illustration for university admissions), when there are no conditional dependencies between observations. An example of a binary multi-label BC is shown graphically in Figure 2i. This BC determines the classifications *play outside* and *raining* based on the observations *windy*, *temperature*, *humidity* and *pressure*. We thus have $\mathcal{C}=\{o,r\}$, $\mathcal{O}=\{w,t,h,p\}$, $\mathcal{D}$ as in the figure, and Σ as in Table 2, where, e.g. when all the observations are (assigned to) + (bottom row), *raining* is (assigned to) $-$ and *play outside* is (assigned to) +.

We will define counterfactual explanations in terms of the variables that may *influence* (assignments to) classifications:

Definition 2. *The* influences *of a BC* $\langle\mathcal{O},\mathcal{C},\mathcal{D},\mathcal{V},\sigma\rangle$ *are* $\mathcal{I}\subseteq$

Σ						CFX		MCX	PIX
w	t	h	p	r	o	$\mathcal{R}_!(r)/\mathcal{R}_*(r)$	$\mathcal{R}_!(o)/\mathcal{R}_*(o)$		
−	−	−	−	+	−	$\{\bar{t}\bar{p}\}/\{\}$	$\{r\}/\{\bar{t}\}$	$(\bar{t})(\bar{p})$	$(\bar{t}\bar{p})$
−	−	−	+	−	+	$\{p\}/\{\bar{h}\}$	$\{\bar{w}\bar{r}\}/\{\}$	(p)	(p)
−	−	+	−	+	−	$\{\bar{p}\}/\{\bar{t}h\}$	$\{r\}/\{\bar{t}\}$	$(\bar{t})(\bar{p})$	$(h\bar{p})(\bar{t}\bar{p})$
−	−	+	+	−	+	$\{p\}/\{\}$	$\{\bar{w}\bar{r}\}/\{\}$	(p)	(p)
−	+	−	−	−	+	$\{t\bar{h}\}/\{\}$	$\{\bar{r}\}/\{\bar{w}t\}$	(t)	$(t\bar{h})$
−	+	−	+	−	+	$\{\}/\{t\bar{h}p\}$	$\{\bar{r}\}/\{\bar{w}t\}$	$(t)(p)$	$(t\bar{h})(p)$
−	+	+	−	+	−	$\{h\bar{p}\}/\{\}$	$\{r\}/\{\}$	$(\bar{p})$	$(h\bar{p})$
−	+	+	+	−	+	$\{p\}/\{t\}$	$\{\bar{r}\}/\{\bar{w}t\}$	$(t)(p)$	(p)
+	−	−	−	+	−	$\{\bar{t}\bar{p}\}/\{\}$	$\{\}/\{w\bar{t}r\}$	$(\bar{t})(\bar{p})$	$(w\bar{t})(\bar{t}\bar{p})$
+	−	−	+	−	−	×	$\{w\bar{t}\}/\{\}$	$(\bar{t})$	$(w\bar{t})(p)$
+	−	+	−	+	−	$\{\bar{p}\}/\{\bar{t}h\}$	$\{\}/\{w\bar{t}r\}$	$(\bar{t})(\bar{p})$	$(w\bar{t})(h\bar{p})(\bar{t}\bar{p})$
+	−	+	+	−	−	×	$\{w\bar{t}\}/\{\}$	$(\bar{t})$	$(w\bar{t})(p)$
+	+	−	−	−	+	$\{t\bar{h}\}/\{\}$	$\{t\bar{r}\}/\{\}$	(t)	$(t\bar{h})$
+	+	−	+	−	+	$\{\}/\{t\bar{h}p\}$	$\{t\bar{r}\}/\{\}$	$(t)(p)$	$(t\bar{h})(p)$
+	+	+	−	+	−	$\{h\bar{p}\}/\{\}$	$\{r\}/\{w\}$	$(\bar{p})$	$(h\bar{p})$
+	+	+	+	−	+	$\{p\}/\{t\}$	$\{t\bar{r}\}/\{\}$	$(t)(p)$	(p)

Table 2: Evaluations and explanations for the play-outside BC. We indicate with × when r is not in the relevant variables of o's CFX.

$\{(\alpha,\beta) \in \mathcal{X} \times \mathcal{C} | \beta \in \mathcal{D}(\alpha)\}$. *For any* $(\alpha,\beta) \in \mathcal{I}$, *we refer to* α *as the* influencer *and to* β *as the* influencee*; also, the* influencers *of* $\alpha \in \mathcal{X}$ *are given by* $\mathcal{I}(\alpha) = \{\beta \in \mathcal{X} | (\beta,\alpha) \in \mathcal{I}\}$.

Intuitively, (α,β) should be in $\mathcal{I}$ if the value of influencer α may have an impact on the value of influencee β given by the BC. We choose binary, rather than more complex, influences to fulfil a desideratum of simplicity. We allow $\mathcal{I}$ to be any subset of $\mathcal{X}\times\mathcal{C}$; the specific choice of $\mathcal{I}$ depends on the BC/classification task. In the university admissions BC, $\mathcal{I} = \{(w,c),(f,c),(e,c),(g,c)\} = \mathcal{O}\times\mathcal{C}$, thus each observation may influence the classification, while in the play-outside BC, $\mathcal{I}=\{(w,o),(t,o),(t,r),(h,r),(p,r),(r,o)\}$, thus classification r and observations w and t may all influence o. For more complex BCs, other forms of influences may be required, e.g. using *Markov blankets* [Pearl, 1989]. Note that naive BCs induce shallow graphs of influences, whereas multi-label BCs may induce non-shallow graphs.

Our counterfactual explanations are defined to answer the question "Why does the classification $e \in \mathcal{C}$ (the *explanandum*) have the value $\epsilon(e)$ in the context of the evaluation $\epsilon \in \Sigma$?" They rely upon a selection of variables $\mathcal{X}_r \subseteq \mathcal{X}$, including the explanandum e, deemed to be *relevant* to e, as well as (two, mutually exclusive) *relations* between these variables, drawn from the restriction of the influences to the relevant variables and "chaining" (via "paths") the relevant variables $\mathcal{X}_r$ to e, where, as conventional, a *path* from A to B by a relation R is a sequence of distinct variables $\alpha_1, \ldots, \alpha_k$, $k \geq 1$, such that $\alpha_1 = A$, $\alpha_k = B$, and for each i, $1 \leq i < k$, $(\alpha_i, \alpha_{i+1}) \in R$. The motivation for $\mathcal{X}_r$ is that only some of the variables in $\mathcal{X}$ (potentially very few [Miller, 2019]) may suffice to explain the value assigned to e by ϵ. Formally:

Definition 3. *Given a BC* $\langle \mathcal{O}, \mathcal{C}, \mathcal{D}, \mathcal{V}, \sigma \rangle$ *with influences* $\mathcal{I}$, *a* counterfactual explanation *(CFX) for an explanandum* $e \in \mathcal{C}$ *in an evaluation* $\epsilon \in \Sigma$ *is a triple* $\langle \mathcal{X}_r, \mathcal{R}_!, \mathcal{R}_* \rangle$ *such that:*
• $\mathcal{R}_! \subseteq \mathcal{I}$ *is a binary relation of* critical influence *where* $\forall(\alpha,\beta) \in \mathcal{R}_!$, $\forall \gamma \in \mathcal{I}(\beta)\backslash\{\alpha\}$,
– $\exists_{\geq 1} \epsilon' \in \Sigma$ *such that* $\epsilon'(\gamma) = \epsilon(\gamma)$ *and* $\epsilon'(\alpha) \neq \epsilon(\alpha)$ *and*
– $\forall \epsilon' \in \Sigma$ *if* $\epsilon'(\gamma) = \epsilon(\gamma)$ *and* $\epsilon'(\alpha) \neq \epsilon(\alpha)$ *then* $\epsilon'(\beta) \neq \epsilon(\beta)$*;*
• $\mathcal{R}_* \subseteq \mathcal{I}$ *is a binary relation of* potential influence *where* $\forall(\alpha,\beta) \in \mathcal{R}_*$*:*
– $(\alpha,\beta) \notin \mathcal{R}_!$*;*
– $\exists \epsilon', \epsilon'' \in \Sigma$ *such that* $\epsilon(\alpha) = \epsilon'(\alpha) \neq \epsilon''(\alpha), \epsilon(\beta) = \epsilon'(\beta) \neq \epsilon''(\beta)$ *and* $\forall \gamma \in \mathcal{I}(\beta)\backslash\{\alpha\}, \epsilon'(\gamma) = \epsilon''(\gamma)$*;*
• $\{e\} \subseteq \mathcal{X}_r \subseteq \mathcal{X}$ *such that* $\forall \alpha \in \mathcal{X}_r\backslash\{e\}$ *there exists a path from* α *to* e *by* $\mathcal{R}_! \cup \mathcal{R}_*$.
The critical influencers *(CIs) and* potential influencers *(PIs) of any* $\alpha \in \mathcal{X}$ *are defined as* $\mathcal{R}_!(\alpha) = \{\beta \in \mathcal{X}_r | (\beta,\alpha) \in \mathcal{R}_!\}$ *and* $\mathcal{R}_*(\alpha) = \{\beta \in \mathcal{X}_r | (\beta,\alpha) \in \mathcal{R}_*\}$, *respectively.*

The relations of critical and potential influence are based on the notion that if a change away from the current evaluation of an influencer may see a change in the evaluation of its influencee (when all its other influencers' evaluations remain unchanged), the influencee's current state somewhat *depends on* that of the influencer. A critical influence is a relation where *any* change in the evaluation of the influencer (with other influencers unchanged) will have this effect, thus it represents an immediate change which can be guaranteed. Meanwhile, a potential influence is not "critical" in that there exists at least one such possible change when other influencees remain in *some* constant (not necessarily the current) state, and thus it represents the possibility of such a change.

For illustration, the CFXs for the university admissions BC are shown in Table 1. Consider the evaluation with $w\bar{f}\bar{e}\bar{g}$, whose CFX, shown in Figure 1ii, includes a CI relation from $\bar{g}$. If a change in the evaluation of g from − to + were seen with all other influencers remaining unchanged, the classification would also change from − to +. This CFX also shows that the relations from $\bar{f}$ and $\bar{e}$ are PIs, indicating the presence of some other case where a change in their evaluations from − to + would lead to a change in the classification from − to + (e.g. $w\bar{f}e\bar{g}$ to $wfe\bar{g}$ for $\bar{f}$ and $\bar{w}f\bar{e}g$ to $\bar{w}feg$ for $\bar{e}$).

Note that CFXs can be understood as sub-graphs of the graph of influences, including (an assignment to) a classification (the explanandum) and with edges drawn from the two relations $\mathcal{R}_!$ and $\mathcal{R}_*$. If the given BC is naive, as for university admissions, then these sub-graphs (the CFXs) are shallow. Consider instead the multi-label play-outside BC, with CFXs for the classification o shown in Table 2. These CFXs now make use of other classifications and are thus non-shallow (e.g. the evaluation with $\bar{w}\bar{t}h\bar{p}$ admits the CFX in Figure 2ii, including the intermediate classification r explained in turn).

3 Formal Properties of CFXs

Here we prove that CFXs provide not only insights into the pivotal factors which lead to a classification but also potential changes which may result in different classifications. First we show uniqueness and (conditional) non-emptiness of CFXs.

Theorem 1. *Given a BC* $\langle \mathcal{O}, \mathcal{C}, \mathcal{D}, \mathcal{V}, \sigma \rangle$ *and its influences* $\mathcal{I}$, *let* $e \in \mathcal{C}$ *be an explanandum in an evaluation* $\epsilon \in \Sigma$. *Then,* $\exists_1 \langle \mathcal{X}_r, \mathcal{R}_!, \mathcal{R}_* \rangle$ *for* e *in* ϵ, *and if* $\mathcal{R}_! \cup \mathcal{R}_* = \emptyset$ *then* $\nexists \epsilon', \epsilon'' \in \Sigma$ *such that* $\exists_1 \alpha \in \mathcal{I}(e)$ *where* $\epsilon''(\alpha) \neq \epsilon'(\alpha) = \epsilon(\alpha)$, $\epsilon''(e) \neq \epsilon'(e) = \epsilon(e)$ *and* $\forall \beta \in \mathcal{I}(e)\backslash\{\alpha\}$, $\epsilon''(\beta) = \epsilon'(\beta)$.

Proof. Follows from Definition 3. □

We posit that the guarantee of a single, non-empty explanation is desirable, as it increases the chance of a user comprehending it and not being overwhelmed by multiple distinct explanations. We guarantee that the CFX (of an explanandum) is non-empty in the likely case that there are two distinct evaluations where changing the evaluation of a single influencer changes that of the influencee (the explanandum).

We now show CFXs are "compositional", in that they include CFXs of other classifications. Formally:

Theorem 2. *Given a BC $\langle \mathcal{O}, \mathcal{C}, \mathcal{D}, \mathcal{V}, \sigma \rangle$, its influences $\mathcal{I}$, and a CFX $\langle \mathcal{X}_r, \mathcal{R}_!, \mathcal{R}_* \rangle$ for an explanandum $e \in \mathcal{C}$ in an evaluation $\epsilon \in \Sigma$, a CFX $\langle \mathcal{X}'_r, \mathcal{R}'_!, \mathcal{R}'_* \rangle$ for any explanandum $e' \in \mathcal{C} \cap \mathcal{X}_r$ in ϵ is such that $\mathcal{X}'_r \subseteq \mathcal{X}_r$, $\mathcal{R}'_! \subseteq \mathcal{R}_!$ and $\mathcal{R}'_* \subseteq \mathcal{R}_*$.*

Proof. (Sketch) Whether in the context of $\langle \mathcal{X}_r, \mathcal{R}_!, \mathcal{R}_* \rangle$ or $\langle \mathcal{X}'_r, \mathcal{R}'_!, \mathcal{R}'_* \rangle$, Definition 3 does not differ in determining $\mathcal{R}_!(e')$ or $\mathcal{R}'_!(e')$, $\mathcal{R}_*(e')$ or $\mathcal{R}'_*(e')$ and, thus, the relevant variables linked by these relations $\mathcal{X}_r(e')$ or $\mathcal{X}'_r(e')$, and thus their relations, and so on. Therefore, the theorem holds. □

We thus simplify explanations of multi-label BCs since an explanation of a *set* of classifications may be represented by a single CFX. For example, for the play-outside BC, in all of the CFXs for o where r is present, the critical and potential influences towards r (and thus the relevant variables linked by these relations) are as in the corresponding CFX for r.

Our next result concerns the effect of changing the evaluation of an explanandum's CIs.

Theorem 3. *Given a BC $\langle \mathcal{O}, \mathcal{C}, \mathcal{D}, \mathcal{V}, \sigma \rangle$, its influences $\mathcal{I}$, and a CFX $\langle \mathcal{X}_r, \mathcal{R}_!, \mathcal{R}_* \rangle$ for $e \in \mathcal{C}$ in $\epsilon \in \Sigma$, let $\alpha, \beta \in \mathcal{X}$ be such that there is a unique path $\alpha_1, \ldots, \alpha_k$ ($k \geq 1$) from α to β by $\mathcal{I}$. Then, $\forall \epsilon' \in \Sigma$ such that $\forall \gamma \in \mathcal{I}(\alpha_1) \cup \ldots \cup \mathcal{I}(\alpha_k) \backslash \{\alpha_1, \ldots, \alpha_{k-1}\}$, $\epsilon'(\gamma) = \epsilon(\gamma)$ and $\epsilon'(\alpha) \neq \epsilon(\alpha)$, if there exists the same unique path by $\mathcal{R}_!$, then $\epsilon'(\beta) \neq \epsilon(\beta)$.*

Proof. (Sketch) By induction on the length k of the path: the property trivially holds if $k = 1$ and thus $\alpha = \beta$; also, if the property holds for $k - 1$ ($k > 1$) then it holds for k, since $\epsilon(\alpha_k)$ is influenced exclusively by $\mathcal{I}(\alpha_k)$=α_{k-1}. Indeed $\forall \gamma \in \mathcal{I}(\alpha_k) \backslash \{\alpha_{k-1}\}$, $\epsilon'(\gamma) = \epsilon(\gamma)$ since there is no path from α_1 to γ by $\mathcal{I}$. Then, if we assume that $\epsilon'(\alpha_{k-1}) \neq \epsilon(\alpha_{k-1})$, by Definition 3, if $(\alpha_{k-1}, \alpha_k) \in \mathcal{R}_!$, then $\epsilon'(\alpha_k) \neq \epsilon(\alpha_k)$. □

Since a change in the evaluation of any CI will see a change in the evaluation of its influencee, which will in turn see a change in the evaluation of *its* (critical) influencee and so on, it becomes apparent that one single change sees a guaranteed chain reaction of changes along a path by the CIs. For example, in the play-outside BC, if we consider the CFX for $\bar{w}\bar{t}h\bar{p}$ in Figure 2ii, there is a single, critical path from $\bar{p}$ to $\bar{o}$ via r. This not only shows the important factors in the classifications but also that: if the *pressure* were high (positive), it would not be *raining* and *play outside* would be positive.

Our final result concerns changing the evaluation of an explanandum's PI when its evaluation does not change.

Theorem 4. *Given a BC $\langle \mathcal{O}, \mathcal{C}, \mathcal{D}, \mathcal{V}, \sigma \rangle$, its influences $\mathcal{I}$, and a CFX $\langle \mathcal{X}_r, \mathcal{R}_!, \mathcal{R}_* \rangle$ for $e \in \mathcal{C}$ in $\epsilon \in \Sigma$ such that $\mathcal{R}_!(e) = \varnothing$ and $\mathcal{R}_*(e) \neq \varnothing$, let $\alpha \in \mathcal{R}_*(e)$. Then, $\forall \epsilon' \in \Sigma$ such that $\epsilon'(e) = \epsilon(e)$, $\epsilon'(\alpha) \neq \epsilon(\alpha)$ and $\epsilon'(\beta) = \epsilon(\beta)$ $\forall \beta \in \mathcal{I}(e) \backslash \{\alpha\}$, given a CFX $\langle \mathcal{X}'_r, \mathcal{R}'_!, \mathcal{R}'_* \rangle$ for e in ϵ', it must hold that $|\mathcal{R}'_*(e)| \leq |\mathcal{R}_*(e)|$ and (trivially) $|\mathcal{R}'_!(e)| \geq |\mathcal{R}_!(e)|$.*

Proof. (Sketch) We consider the states of e's incoming influences from $\epsilon = \langle \mathcal{X}_r, \mathcal{R}_!, \mathcal{R}_* \rangle$ to $\epsilon' = \langle \mathcal{X}'_r, \mathcal{R}'_!, \mathcal{R}'_* \rangle$. For (α, e), by Definition 3, α either loses its status as a PI if $\epsilon'(e) = \epsilon(e)$ is no longer potentially influenced by $\epsilon'(\alpha) \neq \epsilon(\alpha)$, or it remains a PI, as it cannot become a CI in ϵ' as this would require a change from $\epsilon(e)$ to $\epsilon'(e)$. For $\{(\gamma, e) \in \mathcal{R}_* | \gamma \in \mathcal{X} \backslash \{\alpha\}\}$, the only modified evaluation is $\epsilon'(\alpha)$, and thus, by Definition 3, these PIs either remain so or become CIs. For $\{(\delta, e) \in \mathcal{I} \backslash \mathcal{R}_* | \delta \in \mathcal{X}\}$, any influence in $\mathcal{R}'_! \cup \mathcal{R}'_*$ must have already been in $\mathcal{R}_*$ by Definition 3. Thus the theorem holds. □

When a change in the evaluation of a PI does not result in a change in the evaluation of its influencee, this theorem shows that it may no longer be a PI, while other PIs of the influencee may become CIs. Thus, PIs may highlight an iterative path towards CIs. This is an important finding demonstrating the power of CFXs as it shows how the two relations work in symphony to allow a user to assess the counterfactual factors of a BC's predictions. For example, in the BC in Table 1, CFXs may provide an admissions officer with reasons for why an applicant was accepted or not, in the form of factors which could have reversed that decision. Let us take the case $\bar{w}\bar{f}\bar{e}\bar{g}$ where the applicant was rejected (-). In the corresponding CFX, none of the observations are CIs, i.e. a change in any observation would not change the classification for the applicant (Theorem 3). However, all observations are PIs, i.e. they may provide an iterative path towards a change by increasing the number of CIs (Theorem 4). If we change the evaluation of the PI $\bar{w}$ (so now the applicant has prior work experience) we obtain $w\bar{f}\bar{e}\bar{g}$, whose CFX in Figure 1ii shows that the change generates the CI $\bar{g}$. Thus, if the applicant were to meet the GPA, the classification would also change (Theorem 3). At this point the admissions officer can see that had the applicant had prior work experience and met the GPA, the classification would be reversed.

4 Related Work

In this section we discuss related work from the literature, making formal comparisons with our work. We do not review methods for explaining general Bayesian networks such as those aiming to obtain the *most probable explanation* [Pearl, 1989] as these approaches seek to determine approximations of the evaluations, which is outside the scope of this paper.

Two types of explanations for naive BCs are defined in [Shih *et al.*, 2018]. We re-define them here so that they can be formally compared with our CFXs. The first form of explanation gives, for a positive (negative) classification, the minimal subset of the positive (negative, respectively) observations sufficient for the classification. The second form comprises the smallest set of variables that renders the other variables' evaluations irrelevant to the classification.

Definition 4. *[Shih* et al.*, 2018] Given a naive BC $\langle \mathcal{O}, \{c\}, \mathcal{D}, \mathcal{V}, \sigma \rangle$ and an explanandum $e = c$ in $\epsilon \in \Sigma$:*

- *a* minimum cardinality explanation *(MCX) for e in ϵ is an evaluation $m \in \Sigma$ such that $m(e) = \epsilon(e)$ and $\nexists \epsilon' \in \Sigma$ where $\epsilon'(e) = \epsilon(e)$ and $|\{o \in \mathcal{O} | \epsilon'(o) = \epsilon(e)\}| < |\{o \in \mathcal{O} | m(o) = \epsilon(e)\}|$;*

• *a* prime implicant explanation *(PIX) for e in ϵ is a* partial evaluation[2] $p \in 2^{\epsilon}$ *which is minimal (wrt set inclusion) and $\nexists \epsilon' \in \Sigma$ where $\epsilon' \supseteq p$ and $\epsilon'(e) \neq \epsilon(e)$.*
Let $\mathcal{M}$ be the set of all MCXs and $\mathcal{P}$ be the set of all PIXs.

The formal relationship between PIXs and CFXs follows.

Theorem 5. *Given a binary, naive BC $\langle \mathcal{O}, \{c\}, \mathcal{D}, \mathcal{V}, \sigma \rangle$, its influences $\mathcal{I}$, and a CFX explanation $\langle \mathcal{X}_r, \mathcal{R}_!, \mathcal{R}_* \rangle$ for an explanandum $e = c$ in $\epsilon \in \Sigma$: $\bigcap_{p \in \mathcal{P}} p = \mathcal{R}_!(e)$.*

Proof. For a variable x, let $x \in \bigcap_{p \in \mathcal{P}} p$ and suppose by contradiction $x \notin \mathcal{R}_!(e)$. This implies that $\exists \epsilon' \in \Sigma$ such that $\forall y \in \mathcal{X} \setminus \{x\} \epsilon'(y) = \epsilon(y)$ while $\epsilon'(x) \neq \epsilon(x)$. It follows that the partial evaluation p' obtained restricting ϵ' to $\mathcal{X} \setminus \{x\}$ is either a PIX for e or there is a partial evaluation $p'' \subsetneq p'$ which is a PIX for e. Then there is a PIX for e not including x which contradicts the hypothesis $x \in \bigcap_{p \in \mathcal{P}} p$. Conversely assume that $x \in \mathcal{R}_!(e)$ but $x \notin \bigcap_{p \in \mathcal{P}} p$. Then $\exists p_j \in \mathcal{P}$ such that $x \notin p_j$, and thus the partial evaluation p_j can be extended to an evaluation ϵ' such that $\epsilon'(x) \neq \epsilon(x)$ while $\forall y \in \mathcal{X} \setminus \{x\} \epsilon'(y) = \epsilon(y)$ (in particular $\epsilon'(e) = \epsilon(e)$) which contradicts $x \in \mathcal{R}_!(e)$. □

Since PIXs and MCXs are the most similar in the literature to our CFXs, we have provided illustrations thereof throughout the paper, and will use them as baselines in Section 5.

The explanations of [Timmer *et al.*, 2015] also show the interplay between variables, utilising *argumentation frameworks* [Simari and Rahwan, 2009] as the reasoning model. A notion of *support*, similar to our influence, is used to derive evidential (rather than counterfactual) *support graphs* based on the Markov blanket of each variable (rather than only classifications in our method). Support graphs may contain multiple instances of a variable, whereas our CFXs do not.

Explanation trees for causal Bayesian networks [Nielsen *et al.*, 2008] also use the variables *relevant* to explanations for explananda, but only for observations. The explanations are causal wrt the *dataset*, i.e. variable α causing variable β may affect the explanation, whereas ours are causal only wrt the *BC itself*, answering "what caused the BC to predict $\epsilon(e)$?".

Various model-agnostic explanation methods exist, e.g. [Ribeiro *et al.*, 2016; Lundberg and Lee, 2017], including some that generate counterfactuals, e.g. [Schwab and Karlen, 2019], but they generate *flat* explanations, ignoring factors between inputs and outputs that CFXs may use.

5 Experiments

We now evaluate CFXs empirically with various datasets and BCs. Our experiments indicate that CFXs (i) are of appropriate cardinality and length (compared with PIXs and MCXs from Section 4); (ii) highlight paths via PIs towards CIs; and (iii) give meaningful information about BCs' predictions. The algorithm we use to generate $\mathcal{R}_!$ for an explanandum $e \in \mathcal{C}$ in $\epsilon \in \Sigma$ has time complexity $O\left(\mu \cdot \sum_{x \in \mathcal{I}(e)} |\mathcal{V}(x)|\right)$, while that for $\mathcal{R}_*$ has time complexity $O\left(\mu \cdot \prod_{x \in \mathcal{I}(e)} |\mathcal{V}(x)|\right)$, where μ is the constant time required to compute a classification. For a binary, naive BC that can be compiled into an OBDD [Shih *et al.*, 2018], to find a CI for e we generate the sub-graph induced by ϵ, checking for all $x \in \mathcal{I}(e)$ that there is an outgoing edge leading into the opposite sink to that of the classification. This has time complexity $O\left(|\mathcal{I}(e)|\right)$, where μ was removed as the result can be determined from the OBDD. To find a PI for e, for each $x \in \mathcal{I}(e)$, we search for paths to − and + classifications, keeping the subsequent variables constant. In the worst case, the whole OBDD may be searched.

For the experiments we consider three settings: the Congressional **Voting Records** dataset[3], the **Parole Violation** dataset[4] and the **Child** Bayesian network[5]. The Voting Records dataset consists of 435 voting records, amounting to 16 key votes by (Republican or Democrat) Congressmen in the USA. We understand votes as observations and party membership of Congressmen as classifications. We binarise the dataset such that votes are either *yes* or *not yes*. The Parole Violation dataset consists of 675 records, amounting to cases of prisoners who either violated parole or not, understood as classifications. Numeric features of the dataset were mapped into categorical ones using uniform length intervals, giving 8 categorical features for each prisoner, understood as observations. Child is a multi-label Bayesian network that describes the incidence of 6 possible diseases in a baby. We understand six clinical reports and *age* (the BN's leaves) as observations that generate 1080 possible evaluations with presence/absence of one of the diseases as the main classification and all other variables as intermediate classifications. We built a binary, naive BC with test set accuracy of 89.3% for Voting Records and a non-binary, naive BC with test set accuracy of 87.3% for Parole Violation. For Child, we applied inference rules from Bayes' theorem to obtain the classifications. In each BC, classifications' values in evaluations are set to those with the highest probability.

Table 3 shows our results concerning the explanations' cardinality and length.[6] All explanations here are non-empty, though CIs only appear in a subset of the results, which is expected as the more observations and values involved, the less likely it is that any change in a single variable will guarantee a different classification. While MCXs include every observation and are usually singular, many PIXs are generated for each evaluation, despite their more compact length, which may prevent easy comprehension by users in some settings. This point is accentuated in the datasets, particularly for the PIXs for the Voting Records, as shown in Table 3, where a single evaluation may generate *hundreds* of PIXs and only 4% have a single PIX. Our CFXs, meanwhile, are unique (per evaluation) and their length is between the two baselines. In the results for Child, the lengths of the (still numerous) PIXs decrease due to the intermediate classifications that they have no way of representing, whereas the CFXs capture them, with a notable increase in CIs. Overall, we posit that our CFXs perform well against the baselines in these experiments.

[2] Here, we treat any $\epsilon \in \Sigma$ as a set.

[3] http://archive.ics.uci.edu/ml/index.php

[4] https://www.icpsr.umich.edu/icpsrweb/NACJD/studies/26521

[5] https://www.bnlearn.com/bnrepository/discrete-medium.html

[6] Note that we consider MCXs for Voting Records only as the two others are non-binary and [Shih *et al.*, 2018] define MCXs formally for binary BCs only; we consider PIXs for all BCs given that their definition directly extends to the non-binary case [Shih *et al.*, 2018].

Dataset	Exp.	Mean	Number of explanations of length:				
(size, $\|\mathcal{X}\backslash\{e\}\|$)		length	0%	1-25%	26-50%	51-75%	76-100%
Voting Records (435, 16)	MCX	100%	0	0	0	0	485
	PIX	42%	0	10627	108023	16828	0
	CFX	76%	0	0	40	161	234
	CFX!	3%	375	41	14	5	0
Parole Violation (675, 8)	PIX	50%	0	141	1132	691	10
	CFX	75%	0	0	35	438	202
	CFX!	3%	547	126	2	0	0
Child (1080, 18)	PIX	13%	0	3264	84	0	0
	CFX	86%	0	0	0	108	972
	CFX!	29%	0	330	749	1	0

Table 3: Length of explanations (as a percentage of $|\mathcal{X}\backslash\{e\}|$) for the three BCs, where CFX! counts only the CIs in the CFXs.

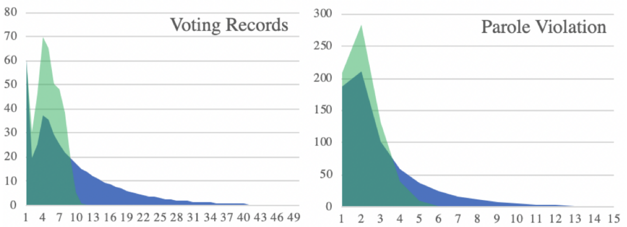

Figure 3: Number of evaluations (y-axis) vs. number of modifications required (x-axis) by selecting a random influence (solid blue) vs. an optimised PI (translucent green) in the path walk simulations.

To show the usefulness of PIs in CFXs, we demonstrate that PIs can highlight paths towards CIs by performing *path walk simulations*. For the two naive BCs, for each evaluation we count the steps it takes to achieve a change in the classification. At each step, we select a CI to change if one is available, and if not, we select a PI and change it to the value which generates the lowest number of PIs in the modified evaluation's CFX. As a baseline, instead of selecting random PIs, we choose a random influencer. For the first case we also do not change any variable twice whenever possible, a strategy which causes *dead ends* in the baseline. Figure 3 shows the results when averaged over 1000 random seeds. We see much quicker convergence to a change with PIs, reducing the mean required steps from 9.5 to 4.7 and from 3.0 to 2.1 for Voting Records and Parole Violation, respectively.

Our final assessment pinpoints single CFXs to demonstrate their usefulness in real world situations. Consider an evaluation from the Voting Records BC where only *water project cost sharing*, *physician fee freeze*, *El Salvador aid*, *religious groups in school*, *crime* are positive, and this record is predicted to be from a Republican; a user may wish to know why this is the case. The CFX shows 13 PIs and 0 CIs, and changing one of these PIs, e.g. *adoption of the budget resolution*, does not change the prediction but means that the CFX now contains a CI, *physician fee freeze* (Theorem 4). Changing this CI changes the class to Democrat (Theorem 3) and the user is informed of two observations which were pivotal in the original classification. For this evaluation, 115 PIXs are generated (with a mean length of 8.0), a potentially overwhelming amount of information for a user, while the MCX gives *physician fee freeze* as its only (positive) variable. Clearly, other variables with a negative evaluation contributed towards this classification (in fact, changing this variable to - does not change the classification) but this information is lost in MCXs. This is due to the fact that they (by definition) do not discriminate between observations with a different evaluation to that of the classification. Another example can be seen in Table 2, where the observations *windy* and *humidity* are clearly factors in determining *play outside* (see the PIXs and CFXs), but they are not used by the MCXs as they affect it non-monotonically, e.g. when they are +, *o* is likely to be -.

In the context of the Parole Violation BC, we show the ability of CFXs to highlight problematic factors in classifications, e.g. to an expert user checking for bias. Consider the evaluation with observations *female*, *white*, *39-46y/o*, *other state*, *5-6 prison*, *11.8-13.4 sentence*, *multi-offender*, *larceny* and classification *non-violator*. The CFX shows two CIs of *white* and *female*: a change in either one of these alone would change the classification, thus highlighting potential racial or sexist bias in the BC. Note that since Theorem 5 does not hold for non-binary BCs, this bias is not highlighted by PIXs.

Finally, in the Child BC, evaluation $\{$ *LowerBodyO2=12+*, *LVHreport=no*, *XrayReport=Asy/Patchy*, *RUQO2=5-12*, *CO2Report=<7.5*, *GruntingReport=Yes*, *Age=0-3 days*$\}$ has classification *Disease=Lung*, and the corresponding CFX shows a CI of *LungParench=Abnormal*, itself with a CI of *Grunting=Yes*. Each CI is a pivotal factor towards the evaluation (Theorem 2). Consideration of the intermediate classifications (i.e. BCs' internal reasoning) when explaining a classification is clearly useful in settings such as health but is absent in PIXs/MCXs, e.g. in Table 2, *raining* plays a huge role towards *play outside* and is included in all but two CFXs, but cannot be handled by PIXs and MCXs. Note that this would also be the case for the flat explanations generated by the model-agnostic methods mentioned in Section 4.

6 Conclusions

We have introduced a novel method for explaining the predictions of BCs via CFXs and have proven theoretically and provided some empirical evidence that the explanations are informative and appropriate for various settings. Our empirical results show for multiple datasets and BCs that CFXs provide important counterfactual information regarding the pivotal factors that could change a classification, opening up many applications for users ranging from non-experts (e.g counterfactually explaining a medical diagnosis) to experts (e.g. detecting bias in a BC in parole violation prediction).

For future work we will explore, from a HCI perspective, how CFXs should be delivered to users. We also plan to develop more efficient algorithms for the generation of $\mathcal{R}_!$ and $\mathcal{R}_*$. We will consider how other AI methods represented by decision functions could be explained by CFXs, e.g. (discretised) neural networks may be well suited to our method, with influences from inputs to intermediate features (e.g. see [Bau *et al.*, 2017]) to outputs. Finally, we will compare with other (possibly probabilistic) notions for evaluating Bayesian networks, e.g. *value of information* measures [Chen *et al.*, 2015] such as *same-decision-probability* [Choi *et al.*, 2012].

References

[Bach *et al.*, 2015] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015.

[Bau *et al.*, 2017] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 3319–3327, 2017.

[Bielza and Larrañaga, 2014] Concha Bielza and Pedro Larrañaga. Discrete Bayesian network classifiers: A survey. *ACM Computing Surveys*, 47(1):5:1–5:43, 2014.

[Chen *et al.*, 2015] Suming Jeremiah Chen, Arthur Choi, and Adnan Darwiche. Value of information based on decision robustness. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 3503–3510, 2015.

[Choi *et al.*, 2012] Arthur Choi, Yexiang Xue, and Adnan Darwiche. Same-decision probability: A confidence measure for threshold-based decisions. *International Journal of Approximate Reasoning*, 53(9):1415–1428, 2012.

[Friedman *et al.*, 1997] Nir Friedman, Dan Geiger, and Moisés Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.

[Guidotti *et al.*, 2019] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5):93:1–93:42, 2019.

[Koller and Sahami, 1996] Daphne Koller and Mehran Sahami. Toward optimal feature selection. In *Machine Learning, Proceedings of the Thirteenth International Conference (ICML '96), Bari, Italy, July 3-6, 1996*, pages 284–292, 1996.

[Lacave and Díez, 2002] Carmen Lacave and Francisco Javier Díez. A review of explanation methods for Bayesian networks. *Knowledge Engineering Review*, 17(2):107–127, 2002.

[Lacave *et al.*, 2000] Carmen Lacave, Roberto Atienza, and Francisco Javier Díez. Graphical explanation in Bayesian networks. In *Medical Data Analysis, First International Symposium, ISMDA 2000, Frankfurt, Germany, September 29-30, 2000, Proceedings*, pages 122–129, 2000.

[Lombrozo, 2007] Tania Lombrozo. Simplicity and probability in causal explanation. *Cognitive Psychology*, 55:232–257, 2007.

[Lundberg and Lee, 2017] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4765–4774, 2017.

[Maron and Kuhns, 1960] M. E. Maron and J. L. Kuhns. On relevance, probabilistic indexing and information retrieval. *Journal of the ACM*, 7(3):216–244, 1960.

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

[Nielsen *et al.*, 2008] Ulf H. Nielsen, Jean-Philippe Pellet, and André Elisseeff. Explanation trees for causal Bayesian networks. In *UAI 2008, Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence, Helsinki, Finland, July 9-12, 2008*, pages 427–434, 2008.

[Pearl, 1989] Judea Pearl. *Probabilistic reasoning in intelligent systems - networks of plausible inference*. Morgan Kaufmann series in representation and reasoning. Morgan Kaufmann, 1989.

[Ribeiro *et al.*, 2016] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144, 2016.

[Ribeiro *et al.*, 2018] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. Anchors: High-precision model-agnostic explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, AAAI-18, New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1527–1535, 2018.

[Schwab and Karlen, 2019] Patrick Schwab and Walter Karlen. CXPlain: Causal explanations for model interpretation under uncertainty. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 10220–10230, 2019.

[Shih *et al.*, 2018] Andy Shih, Arthur Choi, and Adnan Darwiche. A symbolic approach to explaining Bayesian network classifiers. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 5103–5111, 2018.

[Shih *et al.*, 2019] Andy Shih, Arthur Choi, and Adnan Darwiche. Compiling Bayesian network classifiers into decision graphs. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7966–7974, 2019.

[Simari and Rahwan, 2009] Guillermo Ricardo Simari and Iyad Rahwan, editors. *Argumentation in Artificial Intelligence*. Springer, 2009.

[Timmer *et al.*, 2015] Sjoerd T. Timmer, John-Jules Ch. Meyer, Henry Prakken, Silja Renooij, and Bart Verheij. Explaining Bayesian networks using argumentation. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 13th European Conference, ECSQARU 2015, Compiègne, France, July 15-17, 2015. Proceedings*, pages 83–92, 2015.