Persuasive Contrastive Explanations for Bayesian Networks

Tara Koopman and Silja Renooij[0000−0003−4339−8146]

Department of Information and Computing Sciences
Utrecht University, Utrecht, The Netherlands
tara.koopman@hotmail.com, s.renooij@uu.nl

Abstract. [1] Explanation in Artificial Intelligence is often focused on providing reasons for why a model under consideration and its outcome are correct. Recently, research in explainable machine learning has initiated a shift in focus on including so-called counterfactual explanations. In this paper we propose to combine both types of explanation in the context of explaining Bayesian networks. To this end we introduce *persuasive contrastive explanations* that aim to provide an answer to the question *Why outcome t instead of t'?* posed by a user. In addition, we propose an algorithm for computing persuasive contrastive explanations. Both our definition of persuasive contrastive explanation and the proposed algorithm can be employed beyond the current scope of Bayesian networks.

Keywords: Explainable AI · Counterfactuals · Bayesian networks.

1 Introduction

Explanation of Bayesian networks has been a topic of interest ever since their introduction [15,19]. Four categories of explanation method are distinguished, depending on the focus of explanation: 1) explanation of evidence; 2) explanation of reasoning; 3) explanation of the model itself, and 4) explanation of decisions [5,11]. The last category is a recent addition to cover methods that address the question of whether or not the user can make an informed enough decision.

In the explanation of reasoning category, methods typically aim to provide justification for the obtained outcomes and the underlying inference process [11]. Approaches include those that extract reasoning chains from the Bayesian network, measure the impact of evidence, or identify supporting and conflicting evidence [10,11,12,20,22,25]. Explanation of reasoning could also address the explanation of outcomes *not* obtained [11]. In fact, upon encountering an unexpected event, people tend to request *contrastive* explanations that answer the

[1] This is the author-version of an ECSQARU 2021 paper published in Springer Lecture Notes in Computer Science, vol. 12897. The final authenticated version is available online at https://doi.org/10.1007/978-3-030-86772-0_17.

question *Why outcome t instead of t'?* [14]. Such contrastive explanations were recently adopted to explain black-box machine learning (ML) models with high-dimensional feature spaces [23] by using counterfactuals that capture the change in input required to change the outcome from t to t'. The use of counterfactuals for the purpose of explanation is popular in ML research, although the definition of what a counterfactual explanation entails varies greatly [21].

A recently identified challenge for ML research is that of unifying counterfactual explanations with more "traditional explainable AI" that focuses on justifying the original outcome [21]. This, however, is not a challenge specific to ML models only. In this paper we propose the *persuasive contrastive explanation* in the context of Bayesian networks. This explanation provides an answer to the question *Why t instead of t'?* posed by a user, where t is the outcome predicted as most likely by the Bayesian network and t' is the output expected (or desired) by the user. We will provide a contrastive explanation based upon an interpretation of counterfactuals by Wachter et al. [24]. Counterfactuals, however, do not serve to justify outcome t. To provide a more complete answer to the *Why...?* question, we will try to persuade the user into believing that in fact t is the correct outcome. To this end, we propose to include the evidence that suffices to conclude t in the explanation as well. After presenting some properties of our explanations, we propose an algorithm for their computation.

This paper is organised as follows. In Sect. 2 we introduce our new type of explanation and some of its properties. In Sect. 3 we present a search structure for explanations that is exploited by the algorithm detailed in Sect. 4. We review more related work in Sect. 5 and conclude the paper in Sect. 6.

2 Persuasive Contrastive Explanations

In this section we propose our new type of explanation in the context of Bayesian networks. A Bayesian network (BN) represents a joint probability distribution Pr over a set of discrete random variables [8]. We denote such variables V by capital letters and use $\Omega(V)$ to represent their domain. We write v as shorthand for a value assignment $V = v$, $v \in \Omega(V)$. (Sub)sets of variables are denoted by bold-face capital letters $\mathbf{V}$ and their joint value combinations, or configurations, by bold-face small letters $\mathbf{v}$; $\Omega(\mathbf{V})$ is taken to represent the domain of all configurations of $\mathbf{V}$. We write $\mathbf{v}' \subseteq \mathbf{v}$ to denote that $\mathbf{v}'$ is a configuration of $\mathbf{V}' \subseteq \mathbf{V}$ that is consistent with $\mathbf{v}$; we call $\mathbf{v}'$ a sub-configuration of $\mathbf{v}$. Moreover, for a given configuration $\mathbf{v}$, we write $\overline{\mathbf{v}}$ to indicate a configuration for $\mathbf{V}$ in which *every* $V_i \in \mathbf{V}$ takes on a value from $\Omega(V_i)$ that is different from its value in $\mathbf{v}$. Note that $\overline{\mathbf{v}}$ is unique only if all variables in $\mathbf{V}$ are binary-valued.

We are interested in the probabilities $\Pr(T \mid \mathbf{e})$ that can be computed from the Bayesian network for a target variable $T \in \mathbf{V}$ and evidence $\mathbf{e}$ for a set of variables $\mathbf{E} \subseteq \mathbf{V} \setminus \{T\}$. We assume that the most likely value of T given $\mathbf{e}$, i.e. $\arg\max_{t^* \in \Omega(T)}\{\Pr(t^* \mid \mathbf{e})\} = \arg\max_{t^* \in \Omega(T)}\{\Pr(t^* \mathbf{e})\}$, is conveyed to the user as the network's output. We refer to this as the *mode* of T given $\mathbf{e}$, written $\top(T \mid \mathbf{e})$. We now define the explanation context used throughout the paper.

Definition 1. *An* explanation context *is a tuple* $\langle \mathbf{e}, t, t' \rangle$ *where* $t = \top(T \mid \mathbf{e})$ *and* $t \neq t' \in \Omega(T)$.[2]

The explanation context describes the context for answering the question *Why* t *instead of* t'*?*. We will answer this question with a contrastive explanation that combines a sufficient explanation for t with a counterfactual explanation for t'.

Definition 2. *Consider explanation context* $\langle \mathbf{e}, t, t' \rangle$*. A* persuasive contrastive explanation *is any pair* $[\mathbf{s}, \mathbf{c}]$ *where* $\mathbf{s} \in \Omega(\mathbf{S})$, $\mathbf{c} \in \Omega(\mathbf{C})$, $\mathbf{S}, \mathbf{C} \subseteq \mathbf{E}$*, and*

- $\mathbf{s} \subseteq \mathbf{e}$ *is a* sufficient explanation *for* t*, i.e.* $\top(T \mid \mathbf{s}\widetilde{\mathbf{e}}') = t$ *for all* $\widetilde{\mathbf{e}}' \in \Omega(\mathbf{E}')$, $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$*, and there is no* $\mathbf{s}' \subset \mathbf{s}$ *for which this property holds; and*
- $\mathbf{c} \subseteq \overline{\mathbf{e}}$ *is a* counterfactual explanation *for* t'*, i.e.* $\top(T \mid \mathbf{e}'\mathbf{c}) = t'$ *for* $\mathbf{e}' \subseteq \mathbf{e}$, $\mathbf{e}' \in \Omega(\mathbf{E}')$, $\mathbf{E}' = \mathbf{E} \setminus \mathbf{C}$*, and there is no* $\mathbf{c}' \subset \mathbf{c}$ *for which this property holds.*

Since $\mathbf{e}$ is taken to represent observations in the real world, it is common practice to assume that $\Pr(\mathbf{e}) > 0$. For computing the modes in the above definition it is required that $\Pr(\mathbf{s}\widetilde{\mathbf{e}}') > 0$ and $\Pr(\mathbf{e}'\mathbf{c}) > 0$. As it does not make sense for explanations to include impossible combinations of observations, any zero-probability configuration for $\mathbf{E}$ can be disregarded.

The sufficient explanation explains how the evidence relates to the outcome of the network by giving the user a sub-configuration of the evidence that results in the same outcome, regardless of which values are observed for the remaining evidence variables. The sufficient explanation generalizes the *PI-explanation* that was introduced for explaining naive Bayesian classifiers with binary-valued target variables [16], and more recently referred to as *sufficient reason* when used in explaining Bayesian network classifiers (BNCs) with both binary-valued target and evidence variables [4]. The word 'counterfactual' has various interpretations; in our case, a counterfactual explanation details how the evidence should be different to result in the outcome expected by the user. Our definition is a formalisation of the one by Wachter et al. [24], tailored to our specific context.

We will call a set $\mathbf{S}$ with which a sufficient explanation is associated a *sufficient set*; a *counterfactual set* is defined analogously. Sufficient sets and counterfactual sets have a number of properties that we will exploit to enable their computation. All properties assume an explanation context $\langle \mathbf{e}, t, t' \rangle$. The first property addresses the extent to which sufficient explanations are unique and follows directly from Definition 2.

Proposition 1. *A set* $\mathbf{S} \subseteq \mathbf{E}$ *has at most one associated sufficient explanation* $\mathbf{s}$*. Sets* $\mathbf{S}, \mathbf{S}' \subset \mathbf{E}$ *for which neither* $\mathbf{S} \subset \mathbf{S}'$ *nor* $\mathbf{S}' \subset \mathbf{S}$ *can both be sufficient sets.*

The next property addresses the relation between the type of evidence variables (binary or non-binary) and counterfactual explanations.

[2] In case $\top(T \mid \mathbf{e})$ is not unique, we assume all modes are output to the user and that t' is not among these.

Proposition 2. *A counterfactual set* $\mathbf{C} \subseteq \mathbf{E}$ *can have multiple associated counterfactual explanations* $\mathbf{c}$*, unless all variables in* $\mathbf{C}$ *are binary-valued. Sets* $\mathbf{C}, \mathbf{C}' \subset \mathbf{E}$ *with* $\mathbf{C} \subset \mathbf{C}'$ *can both be counterfactual sets, unless all variables in* $\mathbf{C}$ *are binary-valued.*

Proof. If all variables in $\mathbf{C}$ are binary-valued then $\mathbf{c} \subseteq \overline{\mathbf{e}}$ is unique; in this case, by Definition 2, no subset of $\mathbf{C}$ can be a counterfactual set. Otherwise, $\overline{\mathbf{e}}$ is not unique and there exist multiple configurations $\mathbf{c} \subseteq \overline{\mathbf{e}}$ for $\mathbf{C}$. Consider two such configurations $\mathbf{c}_1 \neq \mathbf{c}_2$. Then both can adhere to Definition 2 and be counterfactual explanations, but this is not necessary. Assume that $\mathbf{c}_1$ is a counterfactual explanation and that $\mathbf{c}_2$ is not, and consider a configuration $\mathbf{c}' \supset \mathbf{c}_2$ for a set $\mathbf{C}' \supset \mathbf{C}$. Then $\mathbf{c}' \not\supset \mathbf{c}_1$ so $\mathbf{c}'$ could be a counterfactual explanation, in which case both $\mathbf{C}$ and $\mathbf{C}'$ would be counterfactual sets. □

We conclude that a given explanation context can be associated with multiple persuasive contrastive explanations. Finally we establish a relation between sufficient sets and counterfactual sets.

Proposition 3. *Consider a set* $\mathbf{S} \subseteq \mathbf{E}$*, and let* $\mathbf{C} = \mathbf{E} \setminus \mathbf{S}$*. If* $\mathbf{S}$ *is a sufficient set, or is a superset of a sufficient set, then* $\mathbf{C}$ *cannot be a counterfactual set.*

Proof. If $\mathbf{S}$ is a sufficient set with sufficient explanation $\mathbf{s} \subseteq \mathbf{e}$ then by Definition 3, $\top(T \,|\, \mathbf{sc}) = t$ for all $\mathbf{c} \in \Omega(\mathbf{C})$. Since $t' \neq t$, no $\mathbf{c}$ can be a counterfactual explanation and hence $\mathbf{C}$ is not a counterfactual set. Now let $\mathbf{S}' \subset \mathbf{S}$ be a sufficient set with sufficient explanation $\mathbf{s}'$ and consider configuration $\mathbf{s} = \mathbf{s}'\mathbf{d} \subseteq \mathbf{e}$ for $\mathbf{d} \in \mathbf{S} \setminus \mathbf{S}'$. Then $\mathbf{dc} \in \Omega(\mathbf{C}')$, where $\mathbf{C}' = \mathbf{E} \setminus \mathbf{S}'$ and $\mathbf{c} \in \Omega(\mathbf{C})$. Since $\mathbf{S}'$ is a sufficient set, $\top(T \,|\, \mathbf{s}'\mathbf{c}') = t$ for all $\mathbf{c}' \in \Omega(\mathbf{C}')$. As a result, $\top(T \,|\, \mathbf{s}'\mathbf{dc}) = \top(T \,|\, \mathbf{sc}) = t$ for all $\mathbf{c} \in \Omega(\mathbf{C})$. Hence, $\mathbf{C}$ is not a counterfactual set. □

3 Explanation Lattice

To find all sufficient and counterfactual explanations for a given explanation context $\langle \mathbf{e}, t, t' \rangle$, we can typically do better than naively looping through all possible configurations $\widetilde{\mathbf{e}}$ for $\mathbf{E}$ and computing all distributions $\Pr(T \mid \widetilde{\mathbf{e}})$. In order to exploit the properties from Propositions 1–3 in our search for explanations, we propose to organise the search space using an annotated lattice.

Definition 3. *Consider context* $\langle \mathbf{e}, t, t' \rangle$ *and lattice* $\mathcal{L} = (\mathcal{P}(\mathbf{E}), \subseteq)$*, for powerset* $\mathcal{P}(\mathbf{E})$ *of* $\mathbf{E}$*. An* explanation lattice *for this context is the lattice* $\mathcal{L}$ *in which each lattice element* $\mathbf{S} \subseteq \mathbf{E}$ *is annotated with the tuple* $(\mathbf{s}, \mathcal{M}_\mathbf{S}, l_\mathbf{S})$ *such that*

- $\mathbf{s} \subseteq \mathbf{e}$ *is the configuration of* $\mathbf{S}$ *consistent with* $\mathbf{e}$*;*
- *if* $\mathbf{S} = \mathbf{E}$ *then* $\mathcal{M}_\mathbf{S} = \{(\emptyset, t)\}$*; otherwise,* $\mathcal{M}_\mathbf{S} = \{(\mathbf{c}, t^*) \mid t^* = \top(T \,|\, \mathbf{sc})$*, with* $\mathbf{c} \in \Omega(\mathbf{C}), \mathbf{C} = \mathbf{E} \setminus \mathbf{S}$ *and* $\mathbf{c} \subseteq \overline{\mathbf{e}}\}$*;*
- $l_\mathbf{S} \in \{\mathsf{true}, \mathsf{exp}, \mathsf{oth}\}$*, where* $l_\mathbf{S} = \mathsf{true}$ *if* $t^* = t$ *for all* $(\mathbf{c}, t^*) \in \mathcal{M}_\mathbf{S}$*,* $l_\mathbf{S} = \mathsf{exp}$ *if* $t^* = t'$ *for all* $(\mathbf{c}, t^*) \in \mathcal{M}_\mathbf{S}$*, and* $l_\mathbf{S} = \mathsf{oth}$*, otherwise.*

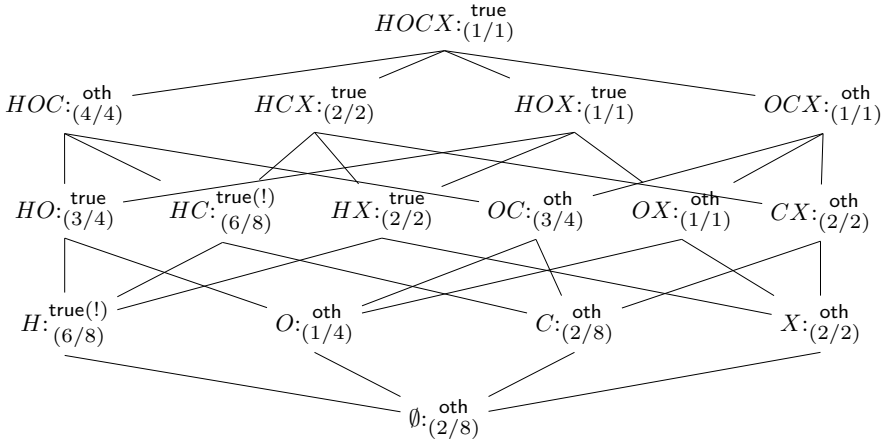

Fig. 1. A partially annotated explanation lattice for the evidence in the CHILD network: elements $\mathbf{S} \subseteq \mathbf{E} = \{H, O, C, X\}$ are annotated with label $l_{\mathbf{S}}$. Numbers between brackets indicate the fraction of modes actually computed. See Example 1 for further details.

The elements of lattice $\mathcal{L}$ are all subsets of $\mathbf{E}$ and hence represent *potential* sufficient sets $\mathbf{S}$ with associated sufficient explanation $\mathbf{s}$. For each lattice element $\mathbf{S}$, the set $\mathbf{C} = \mathbf{E} \setminus \mathbf{S}$ is a *potential* counterfactual set. To determine if $\mathbf{c}$ is a counterfactual explanation, we need to know the corresponding outcome; these pairs are stored in $\mathcal{M}$. Label l summarises whether or not all $\mathbf{sc}$ configurations associated with a lattice element result in the same outcome, with true indicating that this is always the originally predicted outcome t and exp indicating that this is always the expected output t'. Note that if all variables in $\mathbf{E}$ are binary-valued then each $\mathcal{M}_{\mathbf{S}}$ contains a *single* pair $(\mathbf{c}, t^*)$. If the target variable T is binary-valued, then $l_{\mathbf{S}} = \mathsf{oth}$ can only occur with non-binary evidence variables. Fig. 1 shows a partially annotated lattice, which is further explained in Sect. 4.3.

For a lattice element $\mathbf{S}$ we will use the term *ancestors* to refer to all supersets of $\mathbf{S}$ in the lattice, and *parents* to refer to the supersets of size $|\mathbf{S}|+1$; the parent set will be denoted $\mathbf{S}_{\Uparrow}$. Similarly, the term *descendants* is used to refer to all subsets of $\mathbf{S}$ in the lattice, and a *child* is a subset of size $|\mathbf{S}|-1$; the set of all children will be denoted $\mathbf{S}_{\Downarrow}$. The lattice now provides all information necessary for determining whether or not a lattice element represents a sufficient set.

Lemma 1. *Consider context* $\langle \mathbf{e}, t, t' \rangle$ *and explanation lattice* $\mathcal{L}$ *with lattice element* $\mathbf{S} \subseteq \mathbf{E}$*. Then for any possible configuration* $\widetilde{\mathbf{e}}' \in \Omega(\mathbf{E}')$ *for* $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$*, output* $\top(T \,|\, \mathbf{s}\widetilde{\mathbf{e}}')$ *is available from the annotation of* $\mathbf{S}$ *or one of its ancestors.*

Proof. From Definition 3 we have that $\mathcal{M}_{\mathbf{S}}$ contains $\top(T \,|\, \mathbf{s}\mathbf{e}')$ for all $\mathbf{e}' \in \Omega(\mathbf{E}')$ with $\mathbf{e}' \subseteq \overline{\mathbf{e}}$. Now consider a configuration $\mathbf{d}\mathbf{e}^+ \in \Omega(\mathbf{E}')$ where $\mathbf{d} \subseteq \mathbf{e}$ and $\mathbf{e}^+ \subseteq \overline{\mathbf{e}}$ for some set $\mathbf{E}^+$. Then $\mathbf{S}^+ = \mathbf{E} \setminus \mathbf{E}^+$ is a superset of $\mathbf{S}$, annotated with $\mathbf{s}^+ = \mathbf{s}\mathbf{d}$ and outcome $\top(T \,|\, \mathbf{s}\mathbf{d}\mathbf{e}^+)$. We conclude that the outcomes for all

remaining $\widetilde{\mathbf{e}}' \in \Omega(\mathbf{E}')$ are found in the annotations of all supersets $\mathbf{S}^+$ of $\mathbf{S}$, which are exactly the ancestors of $\mathbf{S}$. □

The exact way to establish a sufficient set from the lattice is given by the following proposition.

Proposition 4. *Consider context* $\langle \mathbf{e}, t, t' \rangle$ *and lattice element* $\mathbf{S} \subseteq \mathbf{E}$ *in explanation lattice* $\mathcal{L}$*. Set* $\mathbf{S}$ *is a sufficient set iff all of the following hold:*

1. $\mathbf{S}$ *and each of its ancestors* $\mathbf{S}^+$ *is annotated with label* $l_{\mathbf{S}} = l_{\mathbf{S}^+} = \mathsf{true}$*, and*
2. *for each child* $\mathbf{S}^- \in \mathbf{S}_{\Downarrow}$*, either* $l_{\mathbf{S}^-} \neq \mathsf{true}$*, or* $\mathbf{S}^-$ *has an ancestor* $\mathbf{S}^+$ *with label* $l_{\mathbf{S}^+} \neq \mathsf{true}$*.*

Proof. From Definition 3 and Lemma 1 we have that the first property holds iff $\top(T|\mathbf{s}\widetilde{\mathbf{e}}') = t$ for any possible configuration $\widetilde{\mathbf{e}}' \in \Omega(\mathbf{E}')$ for $\mathbf{E}' = \mathbf{E} \setminus \mathbf{S}$. Therefore, by Definition 2 $\mathbf{S}$ is a sufficient set, unless there exists a subset $\mathbf{S}^- \subset \mathbf{S}$, i.e. a lattice descendant, that adheres to the first property. This case is covered by the second property. Consider a child $\mathbf{S}^- \in \mathbf{S}_{\Downarrow}$ of $\mathbf{S}$ in the lattice. Then $l_{\mathbf{S}^-} \neq \mathsf{true}$ iff there exists a configuration $\mathbf{e}^+ \in \Omega(\mathbf{E}^+)$ for $\mathbf{E}^+ = \mathbf{E} \setminus \mathbf{S}^-$ such that $\top(T|\mathbf{s}^-\mathbf{e}^+) \neq t$. Hence neither $\mathbf{S}^-$, nor any of its descendants, can represent a sufficient set. Now suppose that $l_{\mathbf{S}^-} = \mathsf{true}$ then $\mathbf{S}^-$ can only be sufficient if all its ancestors $\mathbf{S}^+$ have $l_{\mathbf{S}^+} = \mathsf{true}$. If there exists an ancestor with $l_{\mathbf{S}^+} \neq \mathsf{true}$, then none of the descendants of $\mathbf{S}^+$, which include $\mathbf{S}^-$ and its descendants, can represent a sufficient set. □

The next proposition provides the means for determining counterfactual explanations from the lattice.

Proposition 5. *Consider context* $\langle \mathbf{e}, t, t' \rangle$ *and lattice element* $\mathbf{S} \subseteq \mathbf{E}$ *in explanation lattice* $\mathcal{L}$*. Configuration* $\mathbf{c}$ *for set* $\mathbf{C} = \mathbf{E} \setminus \mathbf{S}$ *is a counterfactual explanation for* t' *iff* $(\mathbf{c}, t') \in \mathcal{M}_{\mathbf{S}}$ *and for none of the ancestors* $\mathbf{S}^+$ *of* $\mathbf{S}$ *there exists a* $\mathbf{c}' \subset \mathbf{c}$ *with* $(\mathbf{c}', t') \in \mathcal{M}_{\mathbf{S}^+}$*.*

Proof. From Definition 3 we have that $(\mathbf{c}, t') \in \mathcal{M}_{\mathbf{S}}$ iff $\top(T|\mathbf{sc}) = t'$. Therefore, by Definition 2, $\mathbf{c}$ is a counterfactual explanation for t', unless there exists a $\mathbf{c}' \subset \mathbf{c}$ that also results in outcome t'. Such a $\mathbf{c}'$ can only be found for a set $\mathbf{C}' = \mathbf{E} \setminus \mathbf{S}^+$ where $\mathbf{S}^+$ is a superset of $\mathbf{S}$ and hence a lattice ancestor. □

4 Computing Sufficient and Counterfactual Explanations

We will use a breadth first search on the explanation lattice to return the sufficient and counterfactual explanations. During the search, the lattice is annotated *dynamically* in order to minimize the number of mode computations, since not all lattice elements need necessarily be visited. As a result, $l_{\mathbf{S}} = \emptyset$ as long as a lattice element has not been processed during search, and the modes in $\mathcal{M}_{\mathbf{S}}$ are unknown (unkn) until actually computed. We will first present two algorithms for separately computing the two types of explanation. We will then illustrate the combined search and discuss further optimisations.

Algorithm 1: BFS-SFX for computing sufficient explanations.

Input : BN $\mathcal{B}$, context $\langle \mathbf{e}, t, t' \rangle$ and explanation lattice $\mathcal{L}$
Output: Set $\mathcal{S}$ with all sufficient explanations

```
1  SQ ← E; PotS ← ∅
2  while SQ not empty do
3  |  S ← Dequeue(SQ);
4  |  if l_S = ∅ and ∀S⁺ ∈ S_⇑: l_S⁺ = true then
5  |  |  ComputeModesAndLabel(L, S);
6  |  |  if l_S = true then
7  |  |  |  PotS ← PotS ∪{S};
8  |  |  |  for all S⁻ ∈ S_⇓ do Enqueue(SQ, S⁻);
9  |  |  |
10 |  |  end
11 |  end
12 end
13 S ← {s ⊆ e | S ∈ PotS, s ∈ Ω(S), ∀S⁻ ∈ S_⇓ : l_S⁻ ≠ true};
14
15 return S
```

4.1 Searching the Lattice for Sufficient Explanations

The breadth-first search for sufficient explanations (BFS-SFX) is described in pseudo code in Algorithm 1. Since sufficient explanations are set-minimal, it may seem most optimal to start the search at the bottom of the explanation lattice. However, to decide whether a lattice element represents a sufficient set, we require the labels of its lattice ancestors (see Proposition 4). Hence we start the search at the top of the explanation lattice. We add an unvisited set $\mathbf{S}$ to the set PotS of potential sufficient sets if all its lattice parents (if any) are in PotS and $\top(T \,|\, \mathbf{sc}) = t$ for all configurations $\mathbf{c}$ for $\mathbf{C} = \mathbf{E} \setminus \mathbf{S}$ with $\mathbf{c} \subseteq \overline{\mathbf{e}}$. Line 5 of the algorithm (ComputeModesAndLabel) serves for computing modes from the Bayesian network and for recording them, together with the summarising label, in the explanation lattice. For a given explanation context, the algorithm returns sufficient explanations for all sufficient sets that adhere to the properties stated in Proposition 4.

Proposition 6. *Consider context* $\langle \mathbf{e}, t, t' \rangle$ *and explanation lattice* $\mathcal{L}$*. Algorithm 1 returns all sufficient explanations for* t*.*

Proof. Queue SQ is initially filled with the top lattice element $\mathbf{S} = \mathbf{E}$, which is subsequently processed since it wasn't visited and does not have any parents. ComputeModesAndLabel gives $\mathcal{M}_{\mathbf{E}} = \{(\emptyset, t)\}$ and $l_{\mathbf{E}} = \mathsf{true}$. Since all modes associated with this lattice element equal t, it is added to the set PotS of potential sufficient sets and its lattice children are enqueued in SQ. Subsequently, labels are only computed for lattice elements $\mathbf{S}$ for which all parents are potential sufficient sets, and only if $l_{\mathbf{S}} = \mathsf{true}$ will its children be enqueued in SQ. As a result, a lattice element is in PotS iff it adheres to the first property in Proposition 4.

Algorithm 2: BFS-CFX for computing counterfactual explanations.

Input : BN $\mathcal{B}$, context $\langle \mathbf{e}, t, t' \rangle$, explanation lattice $\mathcal{L}$
Output: Set $\mathcal{C}$ with all counterfactual explanations

```
1  CQ ← E; C ← ∅
2  while CQ not empty do
3  |  S ← Dequeue(CQ);
4  |  if l_S = ∅ and ∀S+ ∈ S_⇑: l_S+ ≠ exp then
5  |  |  potc ← {c | (c, unkn) ∈ M_S, ¬∃c' ∈ C : c' ⊂ c};
6  |  |  if potc ≠ ∅ then
7  |  |  |  ComputeModesAndLabel(L, S, potc);
8  |  |  |  if l_S ≠ true then
9  |  |  |  |  C ← C ∪ {c ∈ potc | (c, t') ∈ M_S};
10 |  |  |  end
11 |  |  |  if l_S ≠ exp then
12 |  |  |  |  for all S− ∈ S_⇓ do Enqueue(CQ, S−);
13 |  |  |  end
14 |  |  end
15 |  end
16 end
17 return C
```

The algorithm now returns (line 14) only explanations for the sets in PotS for which the children $\mathbf{S}^-$ in the lattice are labelled $l_{\mathbf{S}^-} = \mathsf{exp}$ or $l_{\mathbf{S}^-} = \mathsf{oth}$, or for which $l_{\mathbf{S}^-} = \emptyset$. The former two cases clearly adhere to the second property from Proposition 4. If $l_{\mathbf{S}^-} = \emptyset$, we have not computed the modes and label for $\mathbf{S}^-$, so we could still have that in fact $l_{\mathbf{S}^-}$ should be true. However, during search the label for a set $\mathbf{S}^-$ remains undetermined if it has a parent that is not in PotS; in that case it has an ancestor $\mathbf{S}^+$ with label $l_{\mathbf{S}^+} = \mathsf{exp}$ or $l_{\mathbf{S}^+} = \mathsf{oth}$. So with $l_{\mathbf{S}^-} = \emptyset$, $\mathbf{S}^-$ also adheres to property 2 of Proposition 4. Hence set $\mathcal{S}$ contains all sufficient explanations for t. □

4.2 Searching the Lattice for Counterfactual Explanations

The breadth-first search for counterfactual explanations (BFS-CFX) is described in pseudo code in Algorithm 2. The search again starts at the top of the explanation lattice, processing an unvisited set $\mathbf{S}$ if it can potentially have counterfactual explanations associated with it. Whereas the extent of the search for sufficient explanations is independent of variable type (binary vs non-binary), the search for counterfactual explanations can become quite more extensive for non-binary variables. For a given explanation context, Algorithm 2 returns all counterfactual explanations that adhere to the properties stated in Proposition 5.

Proposition 7. *Consider context $\langle \mathbf{e}, t, t' \rangle$ and explanation lattice $\mathcal{L}$. Algorithm 2 returns all counterfactual explanations for t'.*

Proof. First note that if a set $\mathbf{S} \subseteq \mathbf{E}$ is a potential sufficient set (PotS) according to Algorithm 1, then set $\mathbf{C} = \mathbf{E} \setminus \mathbf{S}$ cannot be a counterfactual set (see Proposition 3). Therefore, only if we encounter a set $\mathbf{S}$ with $l_{\mathbf{S}} \neq \mathsf{true}$ can $\mathbf{C}$ possibly be a counterfactual set. Since the search starts at the top of the lattice, any potential counterfactual set is encountered earlier in the search than any of its supersets. Therefore, the first $\mathbf{c}$ with $(\mathbf{c}, t') \in \mathcal{M}_{\mathbf{S}}$ found is in fact a counterfactual explanation and added to the set $\mathcal{C}$ of counterfactual explanations. As as result, no $\mathbf{c}' \supset \mathbf{c}$ can be a counterfactual explanation (see Proposition 5) and set potc prevents such $\mathbf{c}'$ from being added to $\mathcal{C}$ (line 5). If potc is empty then all configurations $\mathbf{c}$ associated with the current lattice element $\mathbf{S}$ are already covered by $\mathcal{C}$ so neither $\mathbf{S}$ nor its descendants can have an associated true counterfactual set $\mathbf{C}$. If potc is non-empty, then the computation of modes and the resulting label can be restricted to configurations in potc (see ComputeModesAndLabel's optional argument). Now any $\mathbf{c} \in$ potc with $(\mathbf{c}, t') \in \mathcal{M}_{\mathbf{S}}$ is a counterfactual explanation. If there also exists a $\mathbf{c} \in$ potc with $(\mathbf{c}, t^*) \in \mathcal{M}_{\mathbf{S}}$ such that $t^* \neq t'$, then this $\mathbf{c}$ could be part of a counterfactual explanation associated with a superset of $\mathbf{C}$ in one of the descendants of $\mathbf{S}$. Hence the children of $\mathbf{S}$ are enqueued in CQ. If such a child has a parent $\mathbf{S}^+$ for which $l_{\mathbf{S}^+} = \mathsf{exp}$ then all possible configurations for $\mathbf{C}^- = \mathbf{E} \setminus \mathbf{S}^+$ are counterfactual explanations or are covered by counterfactual explanations in ancestors; hence, the child is not processed further. We conclude that once queue CQ is empty, set $\mathcal{C}$ contains all counterfactual explanations. □

4.3 Combining the Search for Explanations

Both algorithms BFS-SFX and BFS-CFX do a breadth first search through the explanation lattice and can easily be combined to compute both types of explanation in a single search. Recall that Algorithm 1 goes through the lattice until it encounters lattice elements that cannot be sufficient sets. It is not until this point that the search for counterfactual explanations needs to start (Proposition 3). Rather than starting BFS-CFX at the top of the lattice, we could therefore have Algorithm 1 initialize set $\mathcal{C}$ and queue CQ. The following additions to Algorithm 1 serve for checking for counterfactual explanations upon encountering an element $\mathbf{S}$ that cannot be sufficient, adding those to $\mathcal{C}$ and, if necessary, enqueueing the children of $\mathbf{S}$ in queue CQ:

▷ line 1, add: $CQ \leftarrow \emptyset$; $\mathcal{C} \leftarrow \emptyset$
▷ for $l_{\mathbf{S}} \neq \mathsf{true}$, fill in blank line 9 with:

```
9    else C ← C ∪ {c | (c, t') ∈ M_S};
9a       if l_S ≠ exp then
9b        | for all S⁻ ∈ S_⇓ do Enqueue(CQ, S⁻);
9c       end
```

If all variables are binary-valued then any lattice element $\mathbf{S}$ has only a single associated $(\mathbf{c}, t^*) \in \mathcal{M}_{\mathbf{S}}$; if $t^* = t'$ and this $\mathbf{c}$ is added to $\mathcal{C}$, then none of the descendants of $\mathbf{S}$ can contain counterfactual explanations. In this case, therefore, the above adaptation leaves queue CQ empty. As a result, once all sufficient

sets are found, set $\mathcal{C}$ contains all counterfactual explanations and we are done. An example of computing sufficient and counterfactual explanations from the well-known ASIA network[3] with only binary-valued variables is given in the first author's MSc. thesis [9]. In case the target variable and/or the evidence variables are non-binary, queue CQ will probably be non-empty and the search can now fully focus on finding any remaining counterfactual explanations by continuing the search with the already partially filled queue CQ and set $\mathcal{C}$. That is, we add the following to Algorithm 1:

- ▷ fill in blank line 14 with: GetRemainingCounterfactuals($\mathcal{L}, \mathcal{C}$,CQ), which executes lines 2–16 of BFS-CFX;
- ▷ line 15, add $\mathcal{C}$.

We will use BFS-SFX-CFX to refer to Algorithm 1 with the four above changes to include the computation of both types of explanation. We now illustrate their computation with an example.

Example 1. We consider the CHILD Bayesian network [18] with 6-valued target variable Disease(D) and four of its evidence variables: LVH Report (H) with 2 values, Lower Body O2 (O) with 3 values, CO2 Report (C) with 2 values and X-ray Report (X) with 5 values. We enter the evidence $\mathbf{e} \equiv$ '$H =$ `yes` $\wedge$ $O =$ `5-12` $\wedge$ $C =$ `<7.5` $\wedge$ $X =$ `Oligaemic`'. We find $\top(D|\mathbf{e}) =$ `PAIVS`, whereas the user instead expected outcome `TGA` $\in \Omega(D)$. Fig. 1 now shows the elements $\mathbf{S} \subseteq \mathbf{E}$ in the explanation lattice for context $\langle \mathbf{e}, \texttt{PAIVS}, \texttt{TGA} \rangle$. In addition, the figure shows the labels $l_{\mathbf{S}}$ computed for each element and, between brackets, the number of computed modes versus the total number of associated configurations.

Starting at the top of the lattice, BFS-SFX-CFX first searches for potential sufficient sets. After computing modes for $HOCX$, HOC, HCX, HOX, OCX, and HX (in total: 11), the algorithm has found all sufficient sets, resulting in a single sufficient explanation: $\mathcal{S} = \{$'$H =$ `yes` $\wedge$ $X =$ `Oligaemic`'$\}$. In the process, set $\mathcal{C}$ is initialised to $\mathcal{C} = \{$'$X =$*Plethoric*'$\}$; had all variables been binary-valued then we would have been done. Instead, queue CQ is initialised with CQ$= [HO, HC, OC, OX, CX]$ so the search for counterfactuals continues, finally resulting in four counterfactual explanations: $\mathcal{C} = \{$'$X =$ `Plethoric`', '$X =$ `Normal` $\wedge$ $H =$ `no`', '$X =$ `Grd_Glass` $\wedge$ $H =$ `no`', '$H =$ `no` $\wedge$ $O =$ `<5` $\wedge$ $X =$ `Asy/Patchy`'$\}$. The persuasive contrastive explanations for `PAIVS` are now given by all four pairs $[\mathbf{s}, \mathbf{c}]$ such that $\mathbf{s} \in \mathcal{S}$ and $\mathbf{c} \in \mathcal{C}$. □

We note that in the example we ultimately computed modes for 39 out of the 60 represented evidence configurations. The search for counterfactual explanations continued all the way to the bottom of the lattice: since the target variable has a large state-space, the majority of elements is labelled with `oth`, indicating that possibly another counterfactual explanation is to be found. Two of the labels in Fig. 1 have an exclamation mark (for HC and H). Here the computed labels

[3] All mentioned networks are available from `https://www.bnlearn.com/bnrepository/`.

are in fact different from what they should be according to Definition 3. These labels should be `oth`, since both modes `PAIVS` and `TGA` are found. However, the configurations that would result in mode `TGA` in both cases are excluded from potc due to '$X =$ `Plethoric`' $\in \mathcal{C}$, leaving their modes `unkn`. As a result, the computed labels are based only on configurations that result in mode `PAIVS`. Note that this does not affect the outcome or correctness of the algorithm.

4.4 Complexity and Further Optimisations

The search for sufficient and counterfactual explanations is aborted as soon as all explanations are guaranteed to be found. In worst case, however, BFS-SFX-CFX will visit and process all of the $2^{|\mathbf{E}|}$ lattice elements. In processing a lattice element $\mathbf{S}$, at most $\prod_{C_i \in \mathbf{E} \setminus \mathbf{S}} (|\Omega(C_i)| - 1)$ modes are computed. For Bayesian networks, these computations can be time-consuming, since in general probabilistic inference in Bayesian networks is NP-hard [3], even if we prune computationally irrelevant variables from the network [2]. The overall computational burden can be reduced in at least two ways:

- We can do Bayesian network inference using so-called saturated junction trees: a $\mathbf{C}$-saturated junction tree will allow for computing all probabilities $\Pr(T\mathbf{Cs})$ through efficient local operations only [8].
- We can further reduce the number of mode computations by exploiting monotonicity properties in the domain, such as monotonicity in distribution or in mode [6]; this effectively serves for pruning the explanation lattice.

Assuming an ordering on the values of each variable, inducing a partial order on configurations, the Bayesian network is for example monotonic in mode if higher values of $\mathbf{e}$ result in a higher mode. In such a case, if we have observed the highest value $\widehat{e}$ for some individual evidence variable E_i, then lower values for E_i will never result in a higher mode. If $t' > t$, then we can disregard the values of E_i in our search for counterfactuals. Exploiting such properties can greatly reduce the number of configurations BFS-SFX-CFX needs to consider. Further details on how monotonicity can be exploited to this end, including example computations on an adapted version of the INSURANCE network, can again be found in [9].

5 Related work

As discussed in Sect. 2, our notion of sufficient explanation is similar to the concept of PI-explanation introduced for explaining Bayesian network classifiers. In the related research the Bayesian networks are assumed to be restricted in topology (naive vs general) [13,16] and/or restricted in variable type (binary vs non-binary) [4,16]. The algorithms used for computing the PI-explanations all assume that the Bayesian network is used purely as a classifier and rely on transforming the classifier into a tractable model, such as an Ordered (Binary) Decision Diagram (O(B)DD) [4,16], an Extended Linear Classifier (ELC) [13],

or a representation in First Order Logic [7]. Some transformations either apply only to naive Bayesian networks [13] or require NP-hard compilations [16].

In the past year, several papers have introduced a concept of counterfactual explanation for Bayesian network classifiers, all using different definitions. A common denominator in these definitions is that they determine counterfactual explanations from PI-explanations or vice-versa. Examples include taking the evidence that is common to all PI-explanations (critical influences) together with taking the combined non-critical evidence from the PI-explanations (potential influences) [1], or using PI-explanations to explain for which changes in evidence the current mode will be left unchanged ('even-if-because') [4]. Ignatiev et al.[7] prove a formal relationship between PI-explanations and counterfactual explanations for ML models. Their definition of counterfactual explanation is also based on Wachter et al.[24], but assumes binary-valued evidence variables. In contrast, our counterfactual explanation is defined for discrete variables in general and different counterfactual explanations can include the same evidence variables with different counterfactual values. Moreover, PI-explanations do not provide any information about our counterfactual explanations other than excluding some configurations as possible counterfactuals.

6 Conclusions and further research

In this paper we introduced persuasive contrastive explanations for Bayesian networks, detailed an algorithm for their computation and proved its correctness. The new type of explanation combines a sufficient explanation for the current most likely outcome of the network with a counterfactual explanation that explains the changes in evidence that would result in the outcome expected by the user. Sufficient explanations were introduced before as PI-explanations and efficient algorithms for their computation exist for special cases. Counterfactual explanations such as we define have, to the best of our knowledge, not been used in this context before. We have demonstrated that for special cases the counterfactual explanations are available as soon as the search for sufficient explanations finishes; in general the search for counterfactuals then starts.

Our definitions and basic algorithm are in essence model-agnostic, albeit that the required modes are computed from the Bayesian network. The modes, however, could represent the output predicted by other types of model over the same variables, since we do not exploit properties specific to the Bayesian network. We can therefore employ the same concepts and algorithm for other types of underlying model, as long as the number of different configurations for a typical set of evidence variables is limited enough to process.

Since Bayesian network classifiers of arbitrary topology allowing non-binary evidence and target variables can now be compiled into a tractable ODD [17], it is worth investigating the suitability of ODDs for more efficiently computing persuasive contrastive explanations in Bayesian network classifiers. When many different explanations are found, it is necessary to make a selection to be presented to the user. Such a selection can for example be based on the cardinality

of the explanation. A benefit of directly using Bayesian networks rather than compiled structures such as ODDs is that the computed probabilities can also be exploited for selecting explanations to present to the user. In future we aim to further study the use of probabilistic information for explanation selection. In addition, we aim to further exploit the direct use of a Bayesian network by introducing intermediate variables into the explanation.

Acknowledgements This research was partially funded by the Hybrid Intelligence Center, a 10-year programme funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research, `https://hybrid-intelligence-centre.nl`.

References

1. Albini, E., Rago, A., Baroni, P., Toni, F.: Relation-based counterfactual explanations for Bayesian network classifiers. In: Bessiere, C. (ed.) Proceedings of the 29th International Joint Conference on Artificial Intelligence. pp. 451–457 (2020)
2. Baker, M., Boult, T.E.: Pruning Bayesian networks for efficient computation. In: Bonissone, P., Henrion, M., Kanal, L., Lemmer, J. (eds.) Uncertainty in Artificial Intelligence 6. Elsevier Science, Amsterdam (1991)
3. Cooper, G.F.: The computational complexity of probabilistic inference using Bayesian belief networks. Artificial Intelligence **42**, 393–405 (1990)
4. Darwiche, A., Hirth, A.: On the reasons behind decisions. In: De Giacomo, G., Catala, A., Dilkina, B., Milano, M., Barro, S., Bugarín, A., Lang, J. (eds.) Proceedings of 24th European Conference on Artificial Intelligence. pp. 712–720. IOS Press (2020)
5. Derks, I.P., de Waal, A.: A taxonomy of explainable Bayesian networks. In: Gerber, A. (ed.) Artificial Intelligence Research. pp. 220–235. Springer, Cham (2020)
6. van der Gaag, L.C., Bodlaender, H.L., Feelders, A.: Monotonicity in Bayesian networks. In: Chickering, M., Halpern, J. (eds.) Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. pp. 569–576 (2004)
7. Ignatiev, A., Narodytska, N., Asher, N., Marques-Silva, J.: From contrastive to abductive explanations and back again. In: Baldoni, M., Bandini, S. (eds.) AIxIA 2020 – Advances in Artificial Intelligence. Lecture Notes in Artificial Intelligence, vol. 12414, pp. 335–355. Springer Nature (2021)
8. Jensen, F.V., Nielsen, T.D.: Bayesian networks and decision graphs. Springer Science & Business Media, 2 edn. (2007)
9. Koopman, T.: Computing Contrastive, Counterfactual Explanations for Bayesian Networks. Master's thesis, Universiteit Utrecht, The Netherlands (2020), `https://dspace.library.uu.nl/handle/1874/398728`
10. Kyrimi, E., Marsh, W.: A progressive explanation of inference in 'hybrid' Bayesian networks for supporting clinical decision making. In: Antonucci, A., Corani, G., de Campos, C.P. (eds.) Proceedings of the 8th Conference on Probabilistic Graphical Models. pp. 275–286 (2016)
11. Lacave, C., Díez, F.J.: A review of explanation methods for Bayesian networks. The Knowledge Engineering Review **17**(2), 107–127 (2002)
12. van Leersum, J.: Explaining the reasoning of Bayesian networks with intermediate nodes and clusters. Master's thesis, Utrecht University (2015), `http://dspace.library.uu.nl/handle/1874/313520`

13. Marques-Silva, J., Gerspacher, T., Cooper, M., Ignatiev, A., Narodytska, N.: Explaining naive Bayes and other linear classifiers with polynomial time and delay. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H. (eds.) Advances in Neural Information Processing Systems. vol. 33, pp. 20590–20600. Curran Associates, Inc. (2020)
14. Miller, T.: Explanation in Artificial Intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
15. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers (1988)
16. Shih, A., Choi, A., Darwiche, A.: A symbolic approach to explaining Bayesian network classifiers. Proceedings of the 27th International Joint Conference on Artificial Intelligence p. 5103–5111 (2018)
17. Shih, A., Choi, A., Darwiche, A.: Compiling Bayesian network classifiers into decision graphs. In: Proceedings of the 33rd AAAI Conference on Artificial Intelligence. pp. 7966–7974 (2019)
18. Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., G.Cowell, R.: Bayesian analysis in expert systems. Statistical Science **8**(3), 219–247 (1993)
19. Suermondt, H.J.: Explanation in Bayesian belief networks. Phd thesis, Stanford University (1992)
20. Timmer, S., Meyer, J.J., Prakken, H., Renooij, S., Verheij, B.: Explaining Bayesian networks using argumentation. In: Destercke, S., Denoeux, T. (eds.) Proceedings of the 13th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty. Lecture Notes in Artificial Intelligence, vol. 9161, pp. 83–92. Springer (2015)
21. Verma, S., Dickerson, J.P., Hines, K.E.: Counterfactual explanations for machine learning: A review. ArXiv **abs/2010.10596** (2020)
22. Vlek, C., Prakken, H., Renooij, S., Verheij, B.: A method for explaining Bayesian networks for legal evidence with scenarios. Artificial Intelligence and Law **24**(3), 285–324 (2016)
23. van der Waa, J., Robeer, M., van Diggelen, J., Brinkhuis, M., Neerincx, M.: Contrastive explanations with local foil trees. In: Proceedings of the Workshop on Human Interpretability in Machine Learning. pp. 41–47 (2018)
24. Wachter, S., Mittelstadt, B., Russell, C.: Counterfactual explanations without opening the black box: Automated decisions and the GPDR. Harvard Journal of Law & Technology **31**, 841 (2017)
25. Yap, G.E., Tan, A.H., Pang, H.H.: Explaining inferences in Bayesian networks. Applied Intelligence **29**(3), 263–278 (2008)