Persuasive Contrastive Explanations (Extended Abstract)

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

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

Abstract. 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 into a *persuasive contrastive explanation* that aims to provide an answer to the question *Why outcome t instead of t'?* posed by a user. In addition, we propose a model-agnostic algorithm for computing persuasive contrastive explanations from AI systems with few input variables.

1 Introduction

Upon encountering an unexpected event, people tend to request *contrastive* explanations that answer the question *Why outcome t instead of t'?* [7]. The use of such contrastive explanations, based upon counterfactuals, was recently proposed for explaining black-box machine learning models with high-dimensional feature spaces [11]. Using counterfactuals for the purpose of explanation is popular in machine learning research, although the definition of what a counterfactual explanation entails varies greatly [10]. A recently identified research challenge is that of unifying counterfactual explanations with more "traditional explainable AI" that focuses on justifying the original outcome [10]. In this paper we propose such a combination of explanations by introducing the *persuasive contrastive explanation*. A contrastive explanation, contrasting actual and expected outcomes t and t', is combined with the information that suffices to conclude t. The latter intends to persuade the user into believing that in fact t is the correct outcome. After presenting some properties of our explanations, we sketch an algorithm for their computation from AI systems with a limited number of input variables, such as e.g. Bayesian networks or decision trees. For details omitted from this extended abstract, we refer to our original ECSQARU 2021 contribution [5].

2 Persuasive Contrastive Explanations

We propose a new type of explanation for an AI system that is defined over a set of discrete variables. A variable V has domain $\Omega(V)$, and we write v as

shorthand for $V = v$, $v \in \Omega(V)$. We use bold-face letters to indicate sets of variables and their configurations (joint value assignments). We write $\mathbf{v}' \subseteq \mathbf{v}$ to indicate that $\mathbf{v}'$ is a configuration of $\mathbf{V}' \subseteq \mathbf{V}$ consistent with $\mathbf{v}$. Moreover, for a given $\mathbf{v}$, we write $\overline{\mathbf{v}}$ to denote a configuration 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 assume that our system has a target variable $T \in \mathbf{V}$ of interest and that input, or evidence, $\mathbf{e}$ is given for a set of variables $\mathbf{E} \subseteq \mathbf{V} \setminus \{T\}$. The system then computes an output value for T given $\mathbf{e}$, which we denote $\top(T \,|\, \mathbf{e})$ and call the *mode* of T given $\mathbf{e}$. An example AI system would be a Bayesian network [4] that represents a joint probability distribution over $\mathbf{V}$ and for which we take the most probable value of T given $\mathbf{e}$ as output.

Throughout the paper we assume an *explanation context* $\langle \mathbf{e}, t, t' \rangle$, where $t = \top(T \,|\, \mathbf{e})$ and $t \neq t' \in \Omega(T)$, describing the context in which we seek an answer to the user's question *Why t instead of t'?*. We will answer this question with a sufficient explanation for t and a counterfactual explanation for t'.

Definition 1. *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.* $\top(T \,|\, \mathbf{s}\widetilde{\mathbf{e}}') = t$ *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.* $\top(T \,|\, \mathbf{e}'\mathbf{c}) = t'$ *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.*

We 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 several properties (proven in [5]) that enable their computation:

Properties Consider $\mathbf{S}, \mathbf{C} \subseteq \mathbf{E}$.

- Set $\mathbf{S}$ 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.
- Set $\mathbf{C}$ 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.
- If $\mathbf{S}$ is a sufficient set, or is a superset of a sufficient set, then $\mathbf{E} \setminus \mathbf{S}$ cannot be a counterfactual set.

We note that a given explanation context can be associated with multiple persuasive contrastive explanations.

The sufficient explanation explains how the evidence relates to the outcome of the system by giving the user a subset of the evidence that results in the same outcome, regardless of which values are observed for the remaining evidence variables. The sufficient explanation is similar to the *PI-explanation* that was introduced for explaining naive Bayesian classifiers with binary-valued target variables [8], and more recently referred to as *sufficient reason* when used

in explaining Bayesian network classifiers (BNCs) with both binary-valued target and evidence variables [2]. The word 'counterfactual' has various interpretations and several papers have introduced a concept of counterfactual explanation for Bayesian network classifiers, all using a different definition. A common denominator in these definitions is, however, that they determine counterfactual explanations from PI-explanations [1, 2]. Our definition of counterfactual explanation, on the contrary, is a formalisation of the one by Wachter et al. [12], and PI-explanations do not provide any information other than excluding some configurations as possible counterfactuals.

3 Computing Explanations

We can search for explanations by doing a breadth first search (BFS) on an annotated lattice that organises the search space; we will call this lattice an explanation lattice.

Definition 2. *Consider context* $\langle \mathbf{e}, t, t' \rangle$ *and lattice* $\mathcal{L} = (\mathcal{P}(\mathbf{E}), \subseteq)$ *for the 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* $\exists\, \overline{\mathbf{e}}$ *with* $\mathbf{c} \subseteq \overline{\mathbf{e}}\}$
- $l_\mathbf{S} \in \{\mathsf{true}, \mathsf{exp}, \mathsf{oth}\}$*, 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.*

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 original system outcome t and exp indicating that this is always the user's 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 Example 1.

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 $|\mathbf{S}| + 1$. Likewise we use the terms *descendant* and *child* to refer to subsets of $\mathbf{S}$.

The next proposition (proven in [5]) details exactly how to establish sufficient and counterfactual explanations from the lattice.

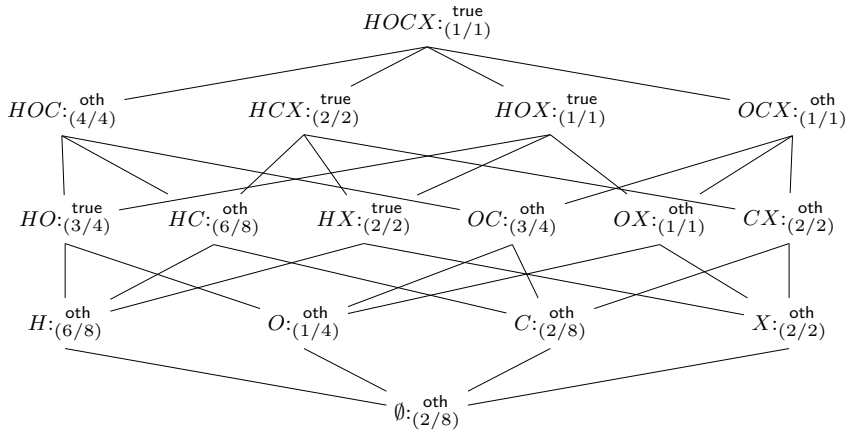

Fig. 1. A partially annotated explanation lattice for the evidence in the CHILD network: elements $\mathbf{S} \subseteq \mathbf{E} = \{H, O, C, X\}$ are annotated with label $l_{\mathbf{S}}$. Numbers between brackets indicate the fraction of modes actually computed. See Example 1 for further details.

Proposition 1. *Consider context* $\langle \mathbf{e}, t, t' \rangle$ *and lattice element* $\mathbf{S} \subseteq \mathbf{E}$ *in explanation lattice* $\mathcal{L}$*. Then*

- *Set* $\mathbf{S}$ *is a sufficient set iff 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* $\mathbf{S}^-$ *of* $\mathbf{S}$*, either* $l_{\mathbf{S}^-} \neq \mathsf{true}$*, or* $\mathbf{S}^-$ *has an ancestor* $\mathbf{S}^+$ *with label* $l_{\mathbf{S}^+} \neq \mathsf{true}$*.*
- *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* $(\mathbf{c}', t') \in \mathcal{M}_{\mathbf{S}^+}$*.*

The BFS now starts at the top lattice element $\mathbf{E}$ and returns all sufficient explanations and all counterfactual explanations for a given explanation context in a single search. During the search, the lattice is annotated *dynamically* in order to minimize the number of mode computations, since not all lattice elements will be necessarily visited. The extent of the search for sufficient explanations is independent of variable type (binary vs non-binary), but the search for counterfactual explanations can become quite extensive for non-binary variables. Pseudocode and a proof of correctness of the algorithm can be found in [5]; we will now illustrate the computation of explanations from an existing Bayesian network.

Example 1. We consider the CHILD Bayesian network [9] 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 $\mathbf{e}$: $H =$ `yes`, $O =$ `5-12`, $C =$ `<7.5`, and $X =$ `oligaemic`. We find $\top(D|\mathbf{e}) =$ `PAIVS`, whereas the user instead expected outcome `TGA` $\in \Omega(D)$. Fig. 1 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 configurations associated with the element.

Starting at the top of the lattice, BFS first searches for potential sufficient sets. After computing modes for $HOCX$, HOC, HCX, HOX, OCX, and HX (in total: 12), the algorithm has found all sufficient sets, resulting in a single sufficient explanation '$H = \texttt{yes}$ and $X = \texttt{oligaemic}$'. In the process, also counterfactual explanation '$X = \texttt{plethoric}$' is found; had all variables been binary-valued then we would have been done. Instead, the search for counterfactuals is continued and results in another three counterfactual explanations: '$X = \texttt{normal}$ & $H = \texttt{no}$', '$X = \texttt{grd\_glass}$ & $H = \texttt{no}$', and '$H = \texttt{no}$ & $O = \texttt{<5}$ & $X = \texttt{asy/patchy}$'. The persuasive contrastive explanations for PAIVS are now given by all four pairs $[\mathbf{s}, \mathbf{c}]$ of the sufficient ($\mathbf{s}$) and counterfactual ($\mathbf{c}$) explanations. □

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 a lot of potential counterfactual explanations need to be considered. In worst case, therefore, BFS will visit and process all of the $2^{|\mathbf{E}|}$ lattice elements. The cost of a single visit is determined by the complexity of mode computations for the AI system under consideration. By exploiting monotonicity properties in the domain [3] we can effectively prune the explanation lattice and hence reduce the number of required mode computations (see [6] for details).

4 Conclusions

In this paper we introduced persuasive contrastive explanations for AI systems with a limited number of discrete input variables. We sketched an algorithm for the computation of these explanations and illustrated this on an example Bayesian network. Our definitions and algorithm are in essence model-agnostic, albeit that the required mode computations need to be performed by the specific AI system under consideration. The complexity of our algorithm is therefore determined by the size of the lattice ($2^{|\mathbf{E}|}$), the number of evidence configurations ($|\Omega(\mathbf{E})|$) and the cost of computing modes from the AI system. The computation of explanations will be feasible as long as the number of different configurations for a typical set of evidence variables is limited enough to process.

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, but other criteria could be more suitable. This is a topic for further research.

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. 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)
3. 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)
4. Jensen, F.V., Nielsen, T.D.: Bayesian networks and decision graphs. Springer Science & Business Media, 2 edn. (2007)
5. Koopman, T., Renooij, S.: Persuasive contrastive explanations for Bayesian networks. In: Vejnarová, J., Wilson, N. (eds.) Proceedings of the 16th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty. Lecture Notes in Artificial Intelligence, Springer (2021)
6. 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
7. Miller, T.: Explanation in Artificial Intelligence: Insights from the social sciences. Artificial Intelligence **267**, 1–38 (2019)
8. 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)
9. Spiegelhalter, D.J., Dawid, A.P., Lauritzen, S.L., G.Cowell, R.: Bayesian analysis in expert systems. Statistical Science **8**(3), 219–247 (1993)
10. Verma, S., Dickerson, J.P., Hines, K.E.: Counterfactual explanations for machine learning: A review. ArXiv **abs/2010.10596** (2020)
11. 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)
12. 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)