Influence-Driven Explanations for Bayesian Network Classifiers

Emanuele Albini[1], Antonio Rago[1], Pietro Baroni[2], and Francesca Toni[1]

[1] Dept. of Computing, Imperial College London, UK
[2] Dip.to di Ingegneria dell'Informazione, Università degli Studi di Brescia, Italy
{emanuele, antonio, ft}@imperial.ac.uk, pietro.baroni@unibs.it

Abstract. We propose a novel approach to building *influence-driven explanations* (IDXs) for (discrete) Bayesian network classifiers (BCs). IDXs feature two main advantages wrt other commonly adopted explanation methods. First, IDXs may be generated using the (causal) influences between *intermediate*, in addition to merely input and output, variables *within BCs*, thus providing a *deep*, rather than shallow, account of the BCs' behaviour. Second, IDXs are generated according to a configurable set of properties, specifying which influences between variables count towards explanations. Our approach is thus *flexible* and can be tailored to the requirements of particular contexts or users. Leveraging on this flexibility, we propose novel IDX instances as well as IDX instances capturing existing approaches. We demonstrate IDXs' capability to explain various forms of BCs, and assess the advantages of our proposed IDX instances with both theoretical and empirical analyses.

1 Introduction

The need for explainability has been one of the fastest growing concerns in AI of late, driven by academia, industry and governments. In response, a multitude of explanation methods have been proposed, with diverse strengths and weaknesses.

We focus on explaining the outputs of (discrete) Bayesian classifiers (BCs) of various kinds. BCs are a prominent method for classification (see [4] for an overview), popular e.g. in medical diagnosis [15,17,25], owing, in particular, to their ability to naturally extract causal influences between variables of interest.

Several bespoke explanation methods for BCs are already available in the literature, including *counterfactual* [1], *minimum cardinality* and *prime implicant* [23] explanations. Further, model-agnostic *attribution methods*, e.g. the popular *LIME* [21] and *SHAP* [16], can be deployed to explain BCs. However, these (bespoke or model-agnostic) explanation methods for BCs are predominantly *shallow*, by focusing on how inputs influence outputs, neglecting the causal influences between intermediate variables in BCs. Furthermore, most explanation methods are *rigid* wrt the users, in the sense that they are based on a single, hardwired, notion of explanation. This sort of one-size-fits-all approach may not be appropriate in all contexts: different users may need different forms of explanation and the same user may be interested in exploring alternative explanations.

To overcome these limitations, we propose the novel formalism of *influence-driven explanations* (IDXs), able to support a principled construction of various forms of explanations for a variety of BCs. The two main ingredients of IDXs are *influences* and *explanation kits*. Influences provide insights into the causal relations between variables *within BCs*, thus enabling the possibility of deep explanations, consisting of influence paths where influences are labelled with *influence types*. An explanation kit consists, of a set of influence types, each associated with a Boolean *property* specifying the condition an influence has to meet to be labelled with that type. By using different influences for the same BC and/or different explanation kits for the same BC and set of influences, a user can thus configure explanations and adjust them to different needs. Specifically, we propose four concrete instances of our general IDX approach: two amount to novel notions of deep explanations, whereas the other two are shallow, corresponding to LIME and SHAP. We evaluate the proposed instances theoretically, in particular as regards satisfaction of a desirable principle of *dialectical monotonicity*. We also conduct extensive empirical evaluation of our IDX instances.[3]

2 Related Work

There are a multitude of methods in the literature for providing explanations (e.g. see the recent surveys [6,26,9]). Many are *model-agnostic*, including: *attribution* methods such as LIME [21] and SHAP [16], which assign each feature an *attribution value* indicating its contribution towards a prediction; and methods relying upon symbolic representations, either to define explanations directly (e.g. *anchors* [22]), or to define logic-based counterparts of the underlying models from which explanations are drawn (e.g. [12,13]). Due to their model-agnosticism, all these methods restrict explanations to "correlations" between *inputs* and *outputs* and make implicit assumptions constraining the explanation [14,2]. Instead, our focus on a specific model (BCs) allows us to define *model-aware* explanations providing a deeper representation of how BCs are functioning via (selected) influences between input, output and (if present) *intermediate* model components.

Regarding BCs, [23] define *minimum cardinality* and *prime implicant* explanations to ascertain pertinent features based on a complete set of classifications, i.e. a decision function representing the BC [24]. These explanations are defined for binary variables only and again explain outputs in terms of inputs. The *counterfactual explanations* of [1] may include also intermediate model's components, but they are rigidly based on a single, hardwired notion of explanation, whereas we present a flexible method for tailoring explanations to different settings. Other works related to explaining BCs include explanation trees for causal Bayesian networks [19] and studies linking causality and explanation [10,11]. Differently from these works, influences included in our explanations represent causal behaviour *in the BC* rather than *in the world*. Finally, [27] use support graphs as explanations showing the interplay between variables (as we do) in Bayesian networks, but (differently from us) commit to a specific influence type.

[3] An extended version (with proofs) is available at https://arxiv.org/abs/2012.05773

3 Bayesian Network Classifiers and Influences

We first define (discrete) BCs and their *decision functions*:

Definition 1. *A* BC *is a tuple* $\langle \mathcal{O}, \mathcal{C}, \mathcal{V}, \mathcal{D}, \mathcal{A} \rangle$ *such that:*

- $\mathcal{O}$ *is a (finite) set of* observations*;*
- $\mathcal{C}$ *is a (finite) set of* classifications*; we call* $\mathcal{X} = \mathcal{O} \cup \mathcal{C}$ *the set of* variables*;*
- $\mathcal{V}$ *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* values *of* x *(*$\mathcal{V}(x)$ *for short);*
- $\mathcal{D} \subseteq \mathcal{X} \times \mathcal{X}$ *is a set of* conditional dependencies *such that* $\langle \mathcal{X}, \mathcal{D} \rangle$ *is an acyclic directed graph (we refer to this as the underlying* Bayesian network*); for any* $x \in \mathcal{X}$*,* $\mathcal{D}(x) = \{y \in \mathcal{X} | (y, x) \in \mathcal{D}\}$ *are the* parents *of* x*;*
- *For each* $x \in \mathcal{X}$*, each* $x_i \in \mathcal{V}(x)$ *is equipped with a* prior probability $P(x_i) \in [0, 1]$ *where* $\sum_{x_i \in \mathcal{V}(x)} P(x_i) = 1$*;*
- *For each* $x \in \mathcal{X}$*, each* $x_i \in \mathcal{V}(x)$ *is equipped with a set of* conditional probabilities *where if* $\mathcal{D}(x) = \{y, \ldots, z\}$*, for every* $y_m, \ldots, z_n \in \mathcal{V}(y) \times \ldots \times \mathcal{V}(z)$*, we have* $P(x_i | y_m, \ldots, z_n)$*, again with* $\sum_{x_i \in \mathcal{V}(x)} P(x_i | y_m, \ldots, z_n) = 1$*;*
- $\mathcal{A}$ *is the set of all possible* input assignments*: any* $a \in \mathcal{A}$ *is a (possibly partial) mapping* $a : \mathcal{X} \mapsto \bigcup_{x \in \mathcal{X}} \mathcal{V}(x)$ *such that, for every* $x \in \mathcal{O}$*,* a *assigns a value* $a(x) \in \mathcal{V}(x)$ *to* x*, and for every* $x \in \mathcal{X}$*, for every* $x_i \in \mathcal{V}(x)$*,* $P(x_i | a)$ *is the* posterior probability *of the value of* x *being* x_i*, given* a*.*[4]

Then, the decision function *(of the BC) is* $\sigma : \mathcal{A} \times \mathcal{X} \mapsto \bigcup_{x \in \mathcal{X}} \mathcal{V}(x)$ *where, for any* $a \in \mathcal{A}$ *and any* $x \in \mathcal{X}$*,* $\sigma(a, x) = argmax_{x_i \in \mathcal{V}(x)} P(x_i | a)$*.*[5]

We consider various concrete BCs, all special cases of Def. 1 satisfying, in addition, an *independence property* among the parents of each variable. For all these BCs, the *conditional probabilities* can be defined, for each $x \in \mathcal{X}, x_i \in \mathcal{V}(x), y \in \mathcal{D}(x), y_m \in \mathcal{V}(y)$, as $P(x_i | y_m)$ with $\sum_{x_i \in \mathcal{V}(x)} P(x_i | y_m) = 1$. For single-label classification we use Naive Bayes Classifiers (NBCs), with $\mathcal{C} = \{c\}$ and $\mathcal{D} = \{(c, x) | x \in \mathcal{O}\}$. For multi-label classification we use a variant of the Bayesian network-based Chain Classifier (BCC) [7] in which leaves of the network are observations, the other variables classifications, and every classification c is estimated with an NBC where the children of c are inputs. In the remainder, unless specified otherwise, we assume as given a generic BC $\langle \mathcal{O}, \mathcal{C}, \mathcal{V}, \mathcal{D}, \mathcal{A} \rangle$ satisfying independence.

For illustration, consider the *play-outside* BCC in Fig. 1i-ii, in which classifications *play* $\underline{o}$*utside* and $\underline{r}$*aining* are determined from observations $\underline{w}$*ind*, $\underline{t}$*emperature* and $\underline{p}$*ressure*. Here, $\mathcal{C} = \{o, r\}$, $\mathcal{O} = \{w, t, p\}$ and $\mathcal{D}$ is as in Figure 1ii. Then, let $\mathcal{V}$ be such that $\mathcal{V}(w) = \mathcal{V}(t) = \{low, medium, high\}$, $\mathcal{V}(p) = \{low, high\}$ and $\mathcal{V}(r) = \mathcal{V}(o) = \{-, +\}$, i.e. w and t are categorical while p, r and o are binary. Figure 1i gives the posterior probabilities and decision function by the BCC. Given our focus on *explaining* BCs, we ignore how they are obtained.

[4] Posterior probabilities may be estimated from prior and conditional probabilities. Note that, if $a(x) = x_i$, then we assume $P(x_i | a) = 1$ and, $\forall x_j \in \mathcal{V}(x) \setminus \{x_i\}$, $P(x_j | a) = 0$.

[5] Note that if $a(x) = x_i$ then $\sigma(a, x) = x_i$.

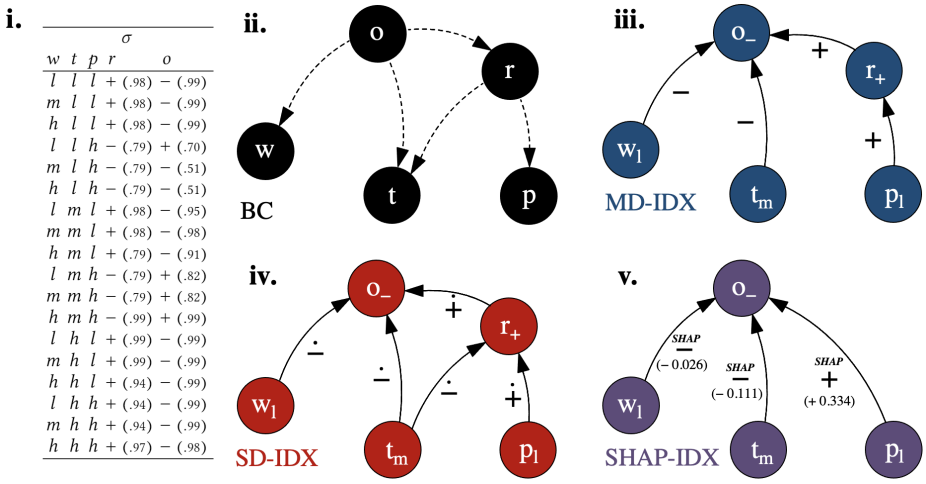

Fig. 1: (i) Decision function (with posterior probabilities explicitly indicated) and (ii) Bayesian network for the play-outside BC, with conditional dependencies as dashed arrows. (iii-v) Corresponding MD-IDX, SD-IDX and SHAP-IDX (shown as graphs, with influences given by edges labelled with their type) for input *low wind* (w_l), *medium temperature* (t_m), and *low pressure* (p_l) and output *not play outside* (o_-) (for the SHAP-IDX we also show the attribution values).

Our method for generating explanations relies on modelling how the variables within a BC *influence* one another. For this, we use two alternative sets of influences. First, similarly to [1], we use *deep influences*, defined as the (acyclic) relation $\mathcal{I}_d = \{(x,c) \in \mathcal{X} \times \mathcal{C} | (c,x) \in \mathcal{D}\}$. Second, we use *input-output influences*, defined as the (acyclic) relation $\mathcal{I}_{io} = \mathcal{O} \times \mathcal{C}_o$, where $\mathcal{C}_o \subseteq \mathcal{C}$ are designated *outputs*. Obviously, $\mathcal{I}_{io}$ ignore the inner structure of BCs. Note that deep influences indicate the direction of the inferences in determining classifications' values, neglecting dependencies between observations as considered in the BCs of [8].

For illustration, in Figure 1i-ii, $\mathcal{I}_d = \{(w,o),(t,o),(r,o),(t,r),(p,r)\}$ and $\mathcal{I}_{io} = \{(w,o),(t,o),(p,o)\}$ for $\mathcal{C}_o = \{o\}$, while $\mathcal{I}_{io} = \{(w,r),(t,r),(p,r),(w,o),(t,o),(p,o)\}$ for $\mathcal{C}_o = \{o,r\}$. Note that in the former $\mathcal{I}_{io}$ case, r is neglected, while in the latter, the influence (w,r) is extracted even though *wind* cannot influence *raining* in this BC, highlighting that using $\mathcal{I}_{io}$, instead of $\mathcal{I}_d$, may have drawbacks for non-naive BCs, except when the notions coincide, i.e. when $\mathcal{D} = \mathcal{C}_o \times \mathcal{O}$.

4 Influence-Driven Explanations

Our explanations are drawn from (deep or input-output) influences by categorising (some of) them as being of different *types*, depending on the satisfaction of *properties*. The choice of types and properties is captured in *explanation kits*:

Definition 2. *Given influences $\mathcal{I}$, an* explanation kit for $\mathcal{I}$ *is a finite set of pairs $\{\langle t_1, \pi_1\rangle, \ldots \langle t_n, \pi_n\rangle\}$ with $\pi_i : \mathcal{I} \times \mathcal{A} \to \{true, false\}$, for $i \in \{1, \ldots, n\}$: we say that t_i is an* influence type characterised by influence property π_i, *and that π_i is* satisfied *for $(x, y) \in \mathcal{I}$ and $a \in \mathcal{A}$ iff $\pi_i((x, y), a) = true$.*

We will focus on explanation kits $\{\langle t_1, \pi_1\rangle, \langle t_2, \pi_2\rangle\}$ with two *mutually exclusive* "dialectical" influence types, of "attack" (t_1) and "support" (t_2): intuitively an influence (x, y) is of type attack (support) if x is a "reason" against (for, resp.) y; mutual exclusion is guaranteed for t_1 and t_2 iff $\pi_i((x, y), a) = true$ implies $\pi_j((x, y), a) = false$ (for $i, j = 1, 2$, $i \neq j$). We will show that these influence types may be characterised by different influence properties, leading to explanations which can all be deemed "dialectical", while differing in other respects.

In general, explanations are obtained from explanation kits as follows:

Definition 3. *Given influences $\mathcal{I}$ and explanation kit $EK = \{\langle t_1, \pi_1\rangle, \ldots \langle t_n, \pi_n\rangle\}$ for $\mathcal{I}$, an* influence-driven explanation *(IDX) drawn from EK for* explanandum *$e \in \mathcal{C}$ with input assignment $a \in \mathcal{A}$ is a tuple $\langle \mathcal{X}_r, \mathcal{I}_{t_1}, \ldots, \mathcal{I}_{t_n}\rangle$ with:*
- *$\mathcal{X}_r \subseteq \mathcal{X}$ such that $e \in \mathcal{X}_r$ (we call $\mathcal{X}_r$ the set of* relevant variables*);*
- *$\mathcal{I}_{t_1}, \ldots \mathcal{I}_{t_n} \subseteq \mathcal{I} \cap (\mathcal{X}_r \times \mathcal{X}_r)$ such that for any $i \in \{1 \ldots n\}$, for every $(x, y) \in \mathcal{I}_{t_i}$, $\pi_i((x, y), a) = true$;*
- *$\forall x \in \mathcal{X}_r$ there is a sequence $x_1, \ldots, x_k$, $k \geq 1$, such that $x_1 = x$, $x_k = e$, and $\forall 1 \leq i < k$ $(x_i, x_{i+1}) \in \mathcal{I}_{t_1} \cup \ldots \cup \mathcal{I}_{t_n}$.*

An IDX thus consists of a set of *relevant variables* ($\mathcal{X}_r$), including the explanandum, connected to one another by influences satisfying the influence properties specified in the explanation kit. Several choices of $\mathcal{X}_r$ may be possible and useful: in the remainder we will restrict attention to *maximal IDXs*, i.e. IDXs with $\subseteq$-maximal $\mathcal{X}_r$ satisfying the conditions set in the second and third bullets of Def. 3. These may be deemed to convey in full the workings of the underlying BC, shaped by the chosen explanation kit. We leave the study of non-maximal IDXs to future work. Note that maximal IDXs, for mutually exclusive influence types, are guaranteed to be unique for a given explanandum and input assignment, due to the "connectedness" requirement in the third bullet of Def. 3.

We will define four instances of our notion of IDX: the first two use $\mathcal{I}_d$, whereas the others use $\mathcal{I}_{io}$. In doing so, we will make use of the following notion.

Definition 4. *Given influences $\mathcal{I}$, a variable $x \in \mathcal{X}$ and an input $a \in \mathcal{A}$, the* modified input *$a'_{x_k} \in \mathcal{A}$ by $x_k \in \mathcal{V}(x)$ is such that, for any $z \in \mathcal{X}$: $a'_{x_k}(z) = x_k$ if $z = x$, and $a'_{x_k}(z) = a(z)$ otherwise.*

A modified input thus assigns a desired value (x_k) to a specified variable (x), keeping the preexisting input assignments unchanged. For example, if $a \in \mathcal{A}$ amounts to *low wind*, *medium temperature* and *low pressure* in the running example, then $a'_{w_h} \in \mathcal{A}$ refers to *high wind*, *medium temperature* and *low pressure*.

4.1 Monotonically Dialectical IDXs

Our first IDX instance draws inspiration from work in bipolar argumentation [3] to define an instance of the explanation kit notion so as to fulfil a form of

dialectical monotonicity: intuitively, this requires that attacks (supports) have a negative (positive, resp.) effect on influenced variables. Concretely, we require that an influencer is an attacker (a supporter) if its assigned value minimises (maximises, resp.) the posterior probability of the influencee's current value.

Definition 5. *An explanation kit* $\{\langle t_1, \pi_1\rangle, \langle t_2, \pi_2\rangle\}$ *for* $\mathcal{I}_d$ *is* monotonically dialectical *iff* $t_1 = -$ *(called* monotonic attack*),* $t_2 = +$ *(*monotonic support*) and for any* $(x, y) \in \mathcal{I}_d$*,* $a \in \mathcal{A}$*, the influence properties* $\pi_1 = \pi_-, \pi_2 = \pi_+$ *are defined as:*
- $\pi_-((x, y), a) = true$ *iff* $\forall x_k \in \mathcal{V}(x) \setminus \{\sigma(a, x)\}$ $P(\sigma(a, y)|a) < P(\sigma(a, y)|a'_{x_k})$*;*
- $\pi_+((x, y), a) = true$ *iff* $\forall x_k \in \mathcal{V}(x) \setminus \{\sigma(a, x)\}$ $P(\sigma(a, y)|a) > P(\sigma(a, y)|a'_{x_k})$.

A monotonically dialectical IDX (MD-IDX) *(for given explanandum and input assignment) is an IDX drawn from a monotonically dialectical explanation kit.*

For illustration, consider the MD-IDX in Figure 1iii (for explanandum o and input assignment a such that $a(w) = l$, $a(t) = m$, $a(p) = l$): here, for example, p_l monotonically supports r_+ because $\sigma(a, r) = +$, $P(\sigma(a, r)|a) = 0.94$ whereas for a' such that $a'(p) = h$ (the only other possible value for p), $P(\sigma(a, r)|a') = 0.01$.

Even though dialectical monotonicity is a natural property, it is a strong requirement that may lead to very few influences, if any, in MD-IDXs. For contexts where this is undesirable, we introduce a weaker form of IDX next.

4.2 Stochastically Dialectical IDXs

Our second IDX instance relaxes the requirement of dialectical monotonicity while still imposing that attacks/supports have a negative/positive, resp., effect on their targets. Concretely, an influencer is an attacker (supporter) if the posterior probability of the influencee's current value is lower (higher, resp.) than the average of those resulting from the influencer's other values, weighted by their prior probabilities (with all other influencers' values unchanged). Formally:

Definition 6. *An explanation kit* $\{\langle t_1, \pi_1\rangle, \langle t_2, \pi_2\rangle\}$ *for* $\mathcal{I}_d$ *is* stochastically dialectical *iff* $t_1 = \dot{-}$ *(called* stochastic attack*),* $t_2 = \dot{+}$ *(*stochastic support*) and for any* $(x, y) \in \mathcal{I}_d$*,* $a \in \mathcal{A}$*, the influence properties* $\pi_1 = \pi_{\dot{-}}, \pi_2 = \pi_{\dot{+}}$ *are defined as:*
- $\pi_{\dot{-}}((x, y), a) = true$ *iff* $P(\sigma(a, y)|a) < \frac{\sum_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} \left[P(x_k) \cdot P(\sigma(a,y)|a'_{x_k})\right]}{\sum_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} P(x_k)}$*;*
- $\pi_{\dot{+}}((x, y), a) = true$ *iff* $P(\sigma(a, y)|a) > \frac{\sum_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} \left[P(x_k) \cdot P(\sigma(a,y)|a'_{x_k})\right]}{\sum_{x_k \in \mathcal{V}(x) \setminus \{\sigma(a,x)\}} P(x_k)}$.

A stochastically dialectical IDX (SD-IDX) *(for given explanandum and input assignment) is an IDX drawn from a stochastically dialectical explanation kit.*

For illustration, Figure 1iv gives the SD-IDX for our running example (using uniform prior probabilities on the domains $\mathcal{V}(w)$, $\mathcal{V}(t)$, and $\mathcal{V}(p)$ and $P(r_+) = .67$, $P(o_+) = 0.22$). Note that this SD-IDX extends the MD-IDX in Figure 1iii by including the negative (stochastic) effect which t_m has on r_+.

SD-IDXs are *stochastic* in that they take into account the prior probabilities of the possible changes of the influencers. This implies that attacks and supports in SD-IDXs will not be empty except in special cases.

4.3 Attribution Method Based Dialectical IDXs

We further show the versatility of the notion of IDX by instantiating it to integrate attribution methods, notably LIME and SHAP. For our purposes, attribution methods can be thought of as mappings $\alpha : \mathcal{O} \times \mathcal{A} \times \mathcal{C}_o \mapsto \mathbb{R}$, basically assigning real values to input-output influences, given input assignments. These values represent the importance of input features towards outputs, and are computed differently by different attribution methods (we will use α_{LIME} and α_{SHAP}, omitting the computation details). To reflect attribution methods' focus on input-output variables, these instances are defined in terms of $\mathcal{I}_{io}$, as follows:

Definition 7. *Given an attribution method α, an* α-explanation kit $\{\langle t_1, \pi_1\rangle, \langle t_2, \pi_2\rangle\}$ *for $\mathcal{I}_{io}$ is such that $t_1 \overset{\alpha}{=\!\!-} t_2$ (*α-attack*), $t_2 \overset{\alpha}{=\!\!+} t_1$ (*α-support*) and for any $(x, y) \in \mathcal{I}_{io}$, $a \in \mathcal{A}$, the influence properties $\pi_1 = \pi_{\overset{\alpha}{-}}$ and $\pi_2 = \pi_{\overset{\alpha}{+}}$, are defined as:*
- $\pi_{\overset{\alpha}{-}}((x, y), a) = true$ *iff* $\alpha(x, a, y) < 0$*;*
- $\pi_{\overset{\alpha}{+}}((x, y), a) = true$ *iff* $\alpha(x, a, y) > 0$.

An α-IDX *is an IDX drawn from an α-explanation kit.*

LIME- and SHAP-explanation kits are instances of α-explanation kits for choices, resp., of $\alpha{=}\alpha_{LIME}$ and $\alpha{=}\alpha_{SHAP}$. Then, *LIME-IDXs* and *SHAP-IDXs* are drawn, resp., from LIME- and SHAP-explanation kits. For illustration, Figure 1v shows a SHAP-IDX for our running example. Here, the restriction to input-output influences implies that the intermediate variable *raining* is not considered in the IDX. Thus, IDXs based on attribution methods are suitable only when the users prefer explanations with a simpler structure. However, in real world applications such as medical diagnosis, where BCs are particularly prevalent, the inclusion of intermediate information could be beneficial: we will illustrate this in Sect. 5.2.

5 Evaluation

We evaluate IDXs theoretically (by showing how different IDX instances relate and how they differ in satisfying a desirable *principle of dialectical monotonicity*) and empirically (for several BCs /datasets). Proofs are omitted for lack of space.

5.1 Theoretical Analysis

Our first two results show the relation/equivalence between MD- and SD-IDXs.[6]

Proposition 1. *Given MD-IDX $\langle \mathcal{X}_r, \mathcal{I}_-, \mathcal{I}_+\rangle$ and SD-IDX $\langle \mathcal{X}'_r, \mathcal{I}_{\dot{-}}, \mathcal{I}_{\dot{+}}\rangle$, both for $e \in \mathcal{X}_r \cap \mathcal{X}'_r$ and $a \in \mathcal{A}$, it holds that $\mathcal{X}_r \subseteq \mathcal{X}'_r$, $\mathcal{I}_- \subseteq \mathcal{I}_{\dot{-}}$ and $\mathcal{I}_+ \subseteq \mathcal{I}_{\dot{+}}$.*

Thus, an MD-IDX, for given explanandum/input assignment, is always (element-wise) a subset of the SD-IDX for the same explanandum/input assignment.

When all variables are binary, MD-IDXs and SD-IDXs are equivalent:

[6] From now on the subscript *io* and *d* of influences for instantiated IDXs will be left implicit, as it is univocally determined by the IDX instance being considered.

Proposition 2. *Given MD-IDX* $\langle \mathcal{X}_r, \mathcal{I}_-, \mathcal{I}_+ \rangle$ *and SD-IDX* $\langle \mathcal{X}'_r, \mathcal{I}_{\dot{-}}, \mathcal{I}_{\dot{+}} \rangle$*, both for explanandum* $e \in \mathcal{X}_r \cap \mathcal{X}'_r$ *and input assignment* $a \in \mathcal{A}$*, if, for all* $x \in \mathcal{X}'_r \setminus \{e\}$*,* $|\mathcal{V}(x)| = 2$*, then* $\mathcal{X}_r = \mathcal{X}'_r$*,* $\mathcal{I}_- = \mathcal{I}_{\dot{-}}$ *and* $\mathcal{I}_+ = \mathcal{I}_{\dot{+}}$*.*

In general, as discussed in Sect. 4.1, MD-IDXs may be much smaller (element-wise) than SD-IDXs, due to the strong requirements imposed by the principle of *dialectical monotonicity*, defined formally as follows, for generic dialectical IDXs:

Principle 1 *An explanation kit* $\{\langle a, \pi_a \rangle, \langle s, \pi_s \rangle\}$[7] *for* $\mathcal{I}$ *satisfies* dialectical monotonicity *iff for any IDX* $\langle \mathcal{X}_r, \mathcal{I}_a, \mathcal{I}_s \rangle$ *drawn from the kit (for any explanandum* $e \in \mathcal{X}_r$*, input assignment* $a \in \mathcal{A}$*), it holds that, for any* $(x, y) \in \mathcal{I}_a \cup \mathcal{I}_s$*, if* $a' \in \mathcal{A}$ *is such that* $\sigma(a', x) \neq \sigma(a, x)$ *and* $\sigma(a', z) = \sigma(a, z)$ $\forall z \in \mathcal{X} \setminus \{x\}$ *such that* $(z, y) \in \mathcal{I}$*, then:*
- *if* $(x, y) \in \mathcal{I}_a$ *then* $P(\sigma(a, y)|a') > P(\sigma(a, y)|a)$*;*
- *if* $(x, y) \in \mathcal{I}_s$ *then* $P(\sigma(a, y)|a') < P(\sigma(a, y)|a)$*.*

Monotonically dialectical explanation kits satisfy this principle by design, while it is worth noting that this does not hold for the other explanations kits:

Proposition 3. *Monotonically dialectical explanation kits satisfy dialectical monotonicity; stochastically dialectical, LIME and SHAP explanation kits do not.*

5.2 Empirical Analysis

For an empirical comparison of the proposed IDX instances, we used several datasets/Bayesian networks (see Table 1),[8] for each of which we deployed an NBC (for single-label classification dataset) or a BCC (for multi-label classification datasets and non-shallow Bayesian networks). Two illustrative IDXs for the same input assignment and explanandum (amounting to the output computed by a model built from the *Child* dataset) are shown in Fig. 2. Note that the MD-IDX provides a deeper account of the influences within the BC than the SHAP-IDX, while also being selective on observations included in the explanations (with two observations playing no role in the MD-IDX), to better reflect the inner workings (Bayesian network) of the model.

The comparison is carried out by analysing the computational viability of IDXs and two aspects linked to their effectiveness, i.e. the size of the produced explanations and the actual amount of violations of dialectical monotonicity.

Computational cost. MD-IDXs and SD-IDXs can be computed efficiently, in linear time in the number of variables' values. Formally, let t_p be the time to compute a prediction and its associated posterior probabilities by the BC (in our experiments, t_p ranged from $3\mu s$ for the simplest NBC to $40ms$ for the most complex BCC).[9] The time complexity to compute whether an influence $(x, y) \in \mathcal{I}$

[7] Here a and s are some form of attack and support, resp., depending on the specific explanation kit; e.g. for *stochastically* dialectical explanation kits $a = \dot{-}$ and $s = \dot{+}$.

[8] *Votes/German*: ML Repo [28]; *COMPAS*: ProRepublica Data Store [20]; *Emotions*: Multi-Label Classification Dataset Repo [18]; *Asia/Child*: Bayesian Net Repo [5].

[9] We used a machine with *Intel i9-9900X* at $3.5Ghz$ and $32GB$ of RAM with no GPU acceleration. For BCCs, we did not use optimised production-ready code.

Dataset	BC†	Size	Variables		Types‡		Performance§	
			$\lvert\mathcal{O}\rvert$	$\lvert\mathcal{C}\rvert$	$\mathcal{O}$	$\mathcal{C}$	Accuracy	F1
Votes	NBC	435	16	1	B	B	90.8%	0.90
German	NBC	750	20	1	C	B	76.4%	0.72
COMPAS	NBC	6951	12	1	C	B	70.5%	0.71
Emotions	BCC	593	72	6	C	B	80.2%	0.70
Asia	BCC	4	2	6	B	B	100%	1.00
Child	BCC	1080	7	13	C	C	80.6%	0.66

Table 1: Characteristics of datasets/BCs used in the empirical analysis. (†) NBC (Naive BC) or BCC (Bayesian Chain Classifier); (‡) **B**inary or **C**ategorical; (§) accuracy and macro F1 score on the test set, averaged for multi-label settings.

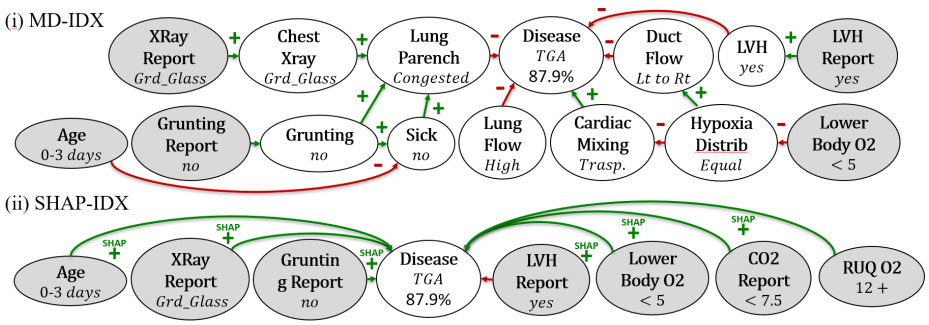

Fig. 2: Example MD-IDX (i) and SHAP-IDX (ii), in graphical form, for explanandum *Disease* for the *Child* BCC (predicting value *TGA* for *Disease* with posterior probability 87.9%). Each node represents a variable with the assigned/estimated value in italics. Grey/white nodes indicate, resp., observations/classifications. $+/\overset{\text{SHAP}}{+}$ and $-/\overset{\text{SHAP}}{-}$ indicate, resp., supports (green arrows) and attacks (red arrows).

belongs to MD-/SD-IDXs, denoted as $T_{1\text{-}IDX}$, is a function of $|\mathcal{V}(x)|$ because determining membership of (x, y) in MD-/SD-IDXs requires checking how the posterior probability of y changes when changing x. Specifically: $T_{1\text{-}IDX}((x,y))$= $\Theta\left(t_p \cdot [1+|\mathcal{V}(x)|-1]\right)$=$\Theta\left(t_p \cdot |\mathcal{V}(x)|\right)$. Then, assuming that the cost for checking the inequalities of Defs. 5 and 6 is negligible wrt the cost of a BC call, it turns out that the cost to compute a full MD-/SD-IDX, denoted as T_{IDX}, corresponds to iterating $T_{1\text{-}IDX}((x,y))$ over all variables $x \in \mathcal{X}$: $T_{IDX}(\mathcal{V}) = \Theta\left(t_p \cdot \sum_{x\in\mathcal{X}} |\mathcal{V}(x)|\right)$, showing linearity. Thus, MD-/SD-IDXs are competitive wrt attribution methods, which rely on costly sampling of the input space. For illustration, the time taken to generate MD-IDXs for the *Child BC* is at most $60 \cdot t_p$ while the time taken to generate LIME explanations with default parameters is $5000 \cdot t_p$.

Size of the explanations. In order to understand *how many influences contribute to IDXs*, we calculated the percentage of influences (per type) in each of the instantiated IDXs from Sect. 4: the results are reported on the left in Table 2. We note that: **(1)** when non-naive BCs are used, MD- and SD-IDXs

Dataset	% Influences in Explanations										% Violating Influences		
	SD-IDX			MD-IDX			LIME-IDX		SHAP-IDX		SD-IDX	LIME-IDX	SHAP-IDX
	$\mathcal{I}_{\dot{+}}$	$\mathcal{I}_{\dot{-}}$	$\mathcal{I}^{\mathcal{C}}_{-\dot{+}}$	$\mathcal{I}_{+}$	$\mathcal{I}_{-}$	$\mathcal{I}^{\mathcal{C}}_{-+}$	$\mathcal{I}^{\text{LIME}}_{+}$	$\mathcal{I}^{\text{LIME}}_{-}$	$\mathcal{I}^{\text{SHAP}}_{+}$	$\mathcal{I}^{\text{SHAP}}_{-}$			
Votes	77.1	22.9	×	77.1	22.9	×	77.1	22.9	73.2	7.3	0.0	0.2	0.1
German	59.3	40.7	×	29.6	22.0	×	55.9	44.1	46.9	36.4	18.5	20.8	19.8
COMPAS	67.0	33.0	×	45.4	20.3	×	65.7	34.3	35.6	19.1	12.3	12.5	22.7
Emotions	56.9	24.0	1.1	10.3	5.4	1.1	60.6	39.4	56.8	10.3	12.0	11.9	8.9
Child	77.5	22.5	64.0	65.4	15.1	64.0	54.0	41.3	24.4	9.7	7.1	2.5	5.6
Asia	87.5	12.5	62.5	87.5	12.5	62.5	70.8	29.2	54.2	20.8	0.0	0.0	0.0

Table 2: Average percentages of influences that are part of IDXs (on the left, with types as shown and where, for types t, t', $\mathcal{I}^{\mathcal{C}}_{t\,t'} = \{(x, y) \in \mathcal{I}_t \cup \mathcal{I}_{t'} | x, y \in \mathcal{C}\}$) and (on the right) of influences in IDXs violating dialectical monotonicity (all percentages are drawn from a sample of 25,000 influences for 250 data-points). Here, × indicates percentages that must be 0 due to the BC type. On the left, percentages may not sum to 100 as some influences may not be part of IDXs.

include influences between classifications (see $\mathcal{I}^{\mathcal{C}}_{-+}$ and $\mathcal{I}^{\mathcal{C}}_{\dot{-}\dot{+}}$ in Table 2), as a consequence of using $\mathcal{I}_d$ and thus being non-shallow; this suggests that our deep IDXs can provide better insights into models than shallow IDXs drawn from *input-output influences*; **(2)** SD- and LIME-IDXs tend to behave similarly, and MD-IDXs tend to include fewer influences than SD-IDXs (in line with Prop. 1); **(3)** in some settings, SHAP-IDXs fail to capture the majority of attacks captured by the other IDX instances (e.g. for *Votes* and *Emotions*).

Satisfaction of Dialectical Monotonicity. We calculated the percentage of influences in SD-/LIME-/SHAP-IDXs which do not satisfy *dialectical monotonicity*: the results are reported in Table 2 (right). We note that: **(1)** All three forms of IDXs may violate the principle for deep and shallow BCs; **(2)** SM-IDXs violate the principle significantly ($p < 0.05$) less for all NBCs, but the percentage of violations by SM-IDXs increases for BCCs, possibly due to SM-IDXs being non-shallow for BCCs (differently from LIME-/SHAP-IDXs, which are always shallow). Note that the violation of dialectical monotonicity may give rise to counter-intuitive results from a dialectical perspective. For illustration, consider the (shallow) SHAP-IDX in Fig. 2ii: one would expect that for values of *Age* for which this is no longer a supporter the diagnosis that *Disease* is *TGA* becomes less likely, but this is not so here. Instead, in the MD-IDX of Fig. 2i, *Age* is an attacker of the inner *Sick* and no misunderstandings may arise.

6 Conclusions

IDXs offer a new perspective on explanation for BCs and open numerous directions for future work, including investigating other instances and other principles, exploring IDXs for other AI methods, as well as conducting user studies to assess how best IDXs can be delivered to users.

Acknowledgements

This research was funded in part by J.P. Morgan and by the Royal Academy of Engineering under the Research Chairs and Senior Research Fellowships scheme. Any views or opinions expressed herein are solely those of the authors listed, and may differ from the views and opinions expressed by J.P. Morgan or its affiliates. This material is not a product of the Research Department of J.P. Morgan Securities LLC. This material should not be construed as an individual recommendation for any particular client and is not intended as a recommendation of particular securities, financial instruments or strategies for a particular client. This material does not constitute a solicitation or offer in any jurisdiction.

References

1. Albini, E., Rago, A., Baroni, P., Toni, F.: Relation-based counterfactual explanations for bayesian network classifiers. In: Proc. of the 29th Int. Joint Conf. on Artificial Intelligence, IJCAI. pp. 451–457 (2020)
2. Barocas, S., Selbst, A.D., Raghavan, M.: The Hidden Assumptions Behind Counterfactual Explanations and Principal Reasons. In: FAT* '20: Proc. of the 2020 Conf. on Fairness, Accountability, and Transparency. pp. 80–89 (2020)
3. Baroni, P., Rago, A., Toni, F.: How many properties do we need for gradual argumentation? In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. pp. 1736–1743 (2018)
4. Bielza, C., Larrañaga, P.: Discrete bayesian network classifiers: A survey. ACM Comput. Surv. **47**(1), 5:1–5:43 (2014)
5. BNlearn: Bayesian network repository - an r package for bayesian network learning and inference (2020), https://www.bnlearn.com/bnrepository
6. Burkart, N., Huber, M.F.: A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research **70**, 245–317 (2021)
7. Enrique Sucar, L., Bielza, C., Morales, E.F., Hernandez-Leal, P., Zaragoza, J.H., Larrañaga, P.: Multi-label classification with bayesian network-based chain classifiers. Pattern Recognition Letters **41**, 14 – 22 (2014)
8. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Mach. Learn. **29**(2-3), 131–163 (1997)
9. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5), 93:1–93:42 (2019)
10. Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach - part II: explanations. In: Proc. of the 17th Int. Joint Conf. on Artificial Intelligence, IJCAI. pp. 27–34 (2001)
11. Halpern, J.Y., Pearl, J.: Causes and explanations: A structural-model approach: Part 1: Causes. In: UAI '01: Proc. of the 17th Conf. in Uncertainty in Artificial Intelligence. pp. 194–202 (2001)
12. Ignatiev, A., Narodytska, N., Marques-Silva, J.: Abduction-based explanations for machine learning models. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. pp. 1511–1519 (2019)
13. Ignatiev, A., Narodytska, N., Marques-Silva, J.: On relating explanations and adversarial examples. In: Advances in Neural Information Processing Systems 32:

Annual Conf. on Neural Information Processing Systems 2019, NeurIPS 2019. pp. 15857–15867 (2019)
14. Kumar, I.E., Venkatasubramanian, S., Scheidegger, C., Friedler, S.A.: Problems with Shapley-value-based explanations as feature importance measures. Machine Learning Research **119**, 5491–5500 (2020)
15. Lipovetsky, S.: Let the evidence speak - using bayesian thinking in law, medicine, ecology and other areas. Technometrics **62**(1), 137–138 (2020)
16. Lundberg, S.M., Lee, S.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30: Annual Conf. on Neural Information Processing Systems 2017. pp. 4765–4774 (2017)
17. McLachlan, S., Dube, K., Hitman, G.A., Fenton, N.E., Kyrimi, E.: Bayesian networks in healthcare: Distribution by medical condition. Artif. Intell. Medicine **107**, 101912 (2020)
18. Moyano, J.M.: Multi-label classification dataset repository (2020), http://www.uco.es/kdis/mllresources/
19. Nielsen, U.H., Pellet, J., Elisseeff, A.: Explanation trees for causal bayesian networks. In: UAI 2008, Proc. of the 24th Conf. in Uncertainty in Artificial Intelligence. pp. 427–434 (2008)
20. ProPublica Data Store: Compas recidivism risk score data and analysis (2016), https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
21. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": Explaining the predictions of any classifier. In: Proc. of the 22nd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining. pp. 1135–1144 (2016)
22. Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. In: Proc. of the 32nd AAAI Conf. on Artificial Intelligence. pp. 1527–1535 (2018)
23. Shih, A., Choi, A., Darwiche, A.: A symbolic approach to explaining bayesian network classifiers. In: Proc. of the 27th Int. Joint Conf. on Artificial Intelligence, IJCAI. pp. 5103–5111 (2018)
24. Shih, A., Choi, A., Darwiche, A.: Compiling bayesian network classifiers into decision graphs. In: Proc. of the 33rd AAAI Conf. on Artificial Intelligence. pp. 7966–7974 (2019)
25. Stähli, P., Frenz, M., Jaeger, M.: Bayesian approach for a robust speed-of-sound reconstruction using pulse-echo ultrasound. IEEE Trans. Medical Imaging **40**(2), 457–467 (2021)
26. Stepin, I., Alonso, J.M., Catala, A., Pereira-Farina, M.: A Survey of Contrastive and Counterfactual Explanation Generation Methods for Explainable Artificial Intelligence. IEEE Access **9**, 11974–12001 (2021)
27. Timmer, S.T., Meyer, J.C., Prakken, H., Renooij, S., Verheij, B.: Explaining bayesian networks using argumentation. In: Symbolic and Quantitative Approaches to Reasoning with Uncertainty - 13th European Conf., ECSQARU 2015. pp. 83–92 (2015)
28. UCI Center for Machine Learning and Intelligent Systems: Machine Learning Repository (2020), https://archive.ics.uci.edu/ml/datasets.php