Influence-Driven Explanations for Bayesian Network Classifiers (Extended Abstract)*

Emanuele Albini[1] , **Antonio Rago**[1] , **Pietro Baroni**[2] and **Francesca Toni**[1]
[1]Department of Computing, Imperial College London, UK
[2]Dipartimento di Ingegneria dell'Informazione, Università degli Studi di Brescia, Italy
{emanuele, a.rago, 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.

Overview

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 (Bielza and Larrañaga 2014) for an overview), popular, for example, in medical diagnosis (Lipovetsky 2020; McLachlan et al. 2020; Stähli, Frenz, and Jaeger 2021), 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* (Albini et al. 2020), *minimum cardinality* (Shih, Choi, and Darwiche 2018) and *prime implicant* (Shih, Choi, and Darwiche 2018) explanations. Further, model-agnostic *attribution methods*, e.g. the popular *LIME* (Ribeiro, Singh, and Guestrin 2016) and *SHAP* (Lundberg and Lee 2017), 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. IDXs are based on two main knowledge representation components, namely *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. Informally, an explanation kit can be regarded as a set of basic explanatory patterns which can be combined together to form an actual explanation. Such patterns hence correspond to the atomic elements of explanatory knowledge in a given context. 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, namely *monotonically dialectical IDXs* (MD-IDXs) and *stochastically dialectical IDXs* (SD-IDXs), whereas the other two (LIME-IDXs and SHAP-IDXs) are shallow, corresponding to the attribution methods LIME (Ribeiro, Singh, and Guestrin 2016) and SHAP (Lundberg and Lee 2017). 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.

Illustration

All the IDX notions considered in the paper are *dialectical*, meaning that they consider two types of influences, namely *attacks* and *supports*. The relations considered as influences and the meaning of attack and support vary across different

*For the full paper see https://link.springer.com/chapter/10.1007%2F978-3-030-89188-6_7

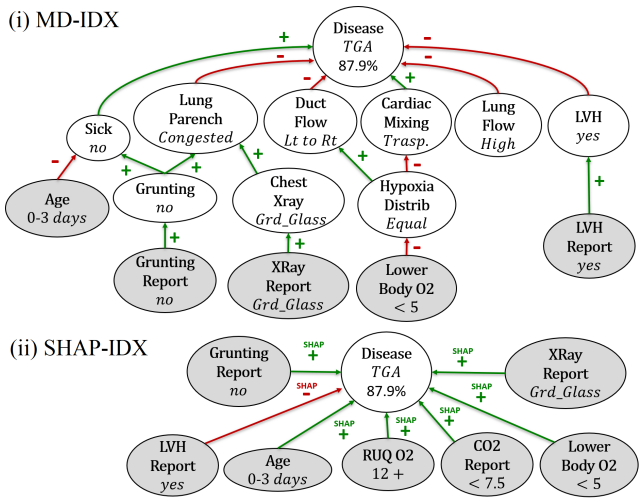

Figure 1: Example MD-IDX (i) and SHAP-IDX (ii), in graphical form, for explanandum *Disease* for the *Child* BC (predicting value *TGA* for *Disease* with posterior probability 87.9%). Each node represents a variable with the assigned/estimated value in italics. Grey/white nodes indicate, respectively, observations/classifications. $+/\overset{\text{SHAP}}{+}$ and $-/\overset{\text{SHAP}}{-}$ indicate, respectively, supports (green arrows) and attacks (red arrows).

instances, allowing to capture shallow vs. deep variants and stronger vs. weaker explanatory notions. For each instance, the meaning of attack and support is specified by the relevant explanation kit.

In particular MD-IDXs and SD-IDXs are oriented to deep explanations, as they consider as influences all the causal relations inherent to the structure of the BC, while LIME-IDXs and SHAP-IDXs are shallow, as they consider only influences from the input variables to the output variables of the BC.

Concerning stronger vs. weaker notions, in MD-IDXs an influence from a variable x to a variable y is considered a support if the current value of x (i.e. the value of x determined by the classifier's input in the explained instance) maximises the probability that y is assigned its current value. Dually, the influence is considered as an attack if the current value of x minimises the probability that y is assigned its current value. SD-IDXs correspond instead to weaker conditions on explanatory roles. For support, it requires that the current value of x raises the probability that y is assigned its current value with respect to the average (over all the possible values of x). Dually, for attack it requires that the current value of x lowers the probability with respect to the average.

LIME-IDXs and SHAP-IDXs do not lend themselves to such a distinction: both LIME and SHAP produce a real number that can be associated with an input/output influence; the positive or negative sign of this number determines whether the influence is regarded as an attack or a support.

To exemplify, MD-IDXs and SHAP-IDXs (corresponding to SHAP explanations), are illustrated in Figure 1, demonstrating the additional information which can be provided via the depth of MD-IDXs, with respect to shallow SHAP-IDXs in a case of medical diagnosis taken from the *Child* dataset (BNlearn 2020). Indeed, 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 in the explained instance.

Conclusion

In summary, our contribution is threefold: we give

- a systematic approach for generating IDXs from BCs, generalising existing work and offering great flexibility with regards to the BC model being explained and the nature of the explanation;
- various instantiations of IDXs, including two based on the cardinal principle of dialectical monotonicity; and
- theoretical and empirical analyses, showing the strengths of IDXs with respect to existing methods, along with illustrations of real world cases where the exploitation of these benefits may be particularly advantageous.

References

Albini, E.; Rago, A.; Baroni, P.; and Toni, F. 2020. Relation-based counterfactual explanations for bayesian network classifiers. In *Proc. of the 29th Int. Joint Conf. on Artificial Intelligence, IJCAI*, 451–457.

Bielza, C., and Larrañaga, P. 2014. Discrete bayesian network classifiers: A survey. *ACM Comput. Surv.* 47(1):5:1–5:43.

BNlearn. 2020. Bayesian network repository - an r package for bayesian network learning and inference.

Lipovetsky, S. 2020. Let the evidence speak - using bayesian thinking in law, medicine, ecology and other areas. *Technometrics* 62(1):137–138.

Lundberg, S. M., and Lee, S. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conf. on Neural Information Processing Systems 2017*, 4765–4774.

McLachlan, S.; Dube, K.; Hitman, G. A.; Fenton, N. E.; and Kyrimi, E. 2020. Bayesian networks in healthcare: Distribution by medical condition. *Artif. Intell. Medicine* 107:101912.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. "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*, 1135–1144.

Shih, A.; Choi, A.; and Darwiche, A. 2018. A symbolic approach to explaining bayesian network classifiers. In *Proc. of the 27th Int. Joint Conf. on Artificial Intelligence, IJCAI*, 5103–5111.

Stähli, P.; Frenz, M.; and Jaeger, M. 2021. Bayesian approach for a robust speed-of-sound reconstruction using pulse-echo ultrasound. *IEEE Trans. Medical Imaging* 40(2):457–467.