Machine Learning Explanations by Surrogate Causal Models (MaLESCaMo)

Alberto Termine[1],*, Alessandro Antonucci[1] and Alessandro Facchini[1]

[1]*Dalle Molle Institute for Artificial Intelligence Research (IDSIA USI-SUPSI), Lugano, Switzerland*

Abstract

Inferring causal explanations for machine learning models is a challenging task for *eXplainable Artificial Intelligence* (XAI). *Counterfactual explanations*, which are techniques to decide how to modify the model input to achieve a desired outcome, represents a possible first step in this direction. However, existing counterfactual explanation methods do not produce genuinely causal counterfactuals. These methods only exploit the correlations between features and target variables while ignoring the causal mechanisms among them. The project presented in this paper (and called MaLESCaMo) aims to develop a novel local and model-agnostic XAI procedure to generate genuine causal counterfactual explanations. Given a black-box predictor and an instance of the features, the procedure computes a counterfactual query in a *surrogate causal model* trained from a local neighbourhood of the input instance. To ease the domain expert elicitation of the causal model, we propose to adopt algorithms for *partial ancestral graphs* as a possible pre-processing step. A specialisation of the *expectation maximisation* algorithm is used instead to practically compute the causal queries.

Keywords

Counterfactual Explanations, Model-agnostic Explanations, Causal Inference, Surrogate Models, Ancestral Graphs

1. Project Overview

Consider the following example. Peter is a 32-year-old low-wage factory worker. He submits to a bank a loan request. The bank processes the request with a black-box classifier and eventually rejects the loan. Peter asks the reasons for the decision and what he can do to get the loan accepted. For such kinds of scenarios, *counterfactual explanations* (CEs) are generally considered in the Explainable AI (XAI) literature [1, 2, 3]. These correspond to synthetic instances suggesting to the user the minimal modifications of the features to get the desired outcome [4, 5]. For example, a CE might suggest that Peter increase his salary to get loan acceptance.

However, CEs have the severe limitation of ignoring the causal relations among features often present in real scenarios [6, 7]. Working in a factory, for instance, may have a direct causal impact on one's salary. Therefore, when suggesting to Peter to increase his income, it is important to

Late-breaking work, Demos and Doctoral Consortium, colocated with The 1st World Conference on eXplainable Artificial Intelligence: July 26–28, 2023, Lisbon, Portugal

*Corresponding author.

alberto.termine@idsia.ch (A. Termine); alessandro.antonucci@idsia.ch (A. Antonucci); alessandro.facchini@idsia.ch (A. Facchini)

https://sites.google.com/view/albertotermine (A. Termine); https://www.alessandroantonucci.me (A. Antonucci); https://people.idsia.ch/~facchini/ (A. Facchini)

consider that this likely entails a need for him to change his job. Nevertheless, most CE methods assume the features to be causally unrelated [8, 9, 10]. This assumption limits the explanations' plausibility and undermines their causal robustness [11]. Furthermore, the causal information provided by existing CEs is quite limited, as they neither facilitate the identification of the necessary and sufficient conditions for a desired outcome, nor provide means to measure the causal impact of each feature on that outcome. In the loan example, a CE can propose potential actions for Peter to take. However, it cannot specify the necessary and sufficient conditions he must fulfil for his request to be accepted.

The **MaLESCaMo project**, presented in this paper, aims at countering these issues by developing an XAI procedure to get causally-robust CEs based on *surrogate causal models*. The procedure is *local* and *model-agnostic*. Given a supervised learning setup with features $\boldsymbol{X}$ and target Y, the procedure takes in input a black box predictor f, a single instance of the features $\boldsymbol{x}$, and a causal query $\mathcal{Q}$, hence it answers the query via learning a surrogate causal model on a local neighbourhood of $\boldsymbol{x}$. As we want the query $\mathcal{Q}$ to be genuinely counterfactual, e.g., like Pearl's *probabilities of causation* [12], the computation requires an explicit specification of a *structural causal model* (SCM).

In the above setup, a SCM is defined as a tuple $\langle \mathcal{G}, \mathcal{F}, \{P(U)\}_{U \in \boldsymbol{U}} \rangle$ where $\mathcal{G}$ is a *causal* directed graph whose nodes are in one-to-one correspondence with a set of *endogenous* variables $(\boldsymbol{X}, Y)$ and a set of *exogenous* variables $\boldsymbol{U}$ denoting latent factors and graphically represented by the root nodes of $\mathcal{G}$ (see e.g. Fig. 1), $\mathcal{F}$ is a collection of *structural equations*, each determining the value of an endogenous variable as a deterministic function of its exogenous parents in $\mathcal{G}$, and $P(U)$ is a probability distribution ranging over $U \in \boldsymbol{U}$ [13] .

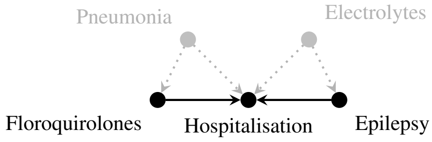

Figure 1: An SCM with two exogenous variables (in grey).

A complete SCM specification might require extensive specific domain knowledge to identify the common latent causes, determine the orientation of the edges in the graph, and specify both the exogenous distributions and the structural equations.

In the lack of background domain knowledge to elicit the structural equations, we adopt a conservative approach consisting of enumerating *all* the possible deterministic relations between an endogenous variable and its endogenous parents in the causal graph [14, 15, 16]. Consider the example in Fig. 1 and assume we do not know how *Epilepsy* impacts on *Hospitalisation*. In this case, it suffices to enumerate all the possible deterministic relations between *Epilepsy* and *Hospitalisation*, and let *Electrolytes* have the same number of states, viz. four if the variables are Boolean, in order to index all those relations.

Regarding the exogenous distributions, we consider the possibility of inferring them from a joint endogenous distribution, trained, for instance, from a dataset of endogenous observations. Following [15], it is possible to back-propagate the observational distribution through the structural equations, thus inducing constraints on the exogenous distributions. The multi-valued specification of the exogenous distributions induced by the constraints transforms the SCM into a *credal network* (CN) [17]. Therefore, a counterfactual causal query can be computed in the CN by dedicated inference algorithms [18], possibly leading to interval-valued outputs instead of sharp estimates. Those intervals reflect the well-known *partial identifiability* of counterfactual queries in SCMs [19].

Like the bounding of counterfactual queries [20], exact CN inference is a hard task [21]. As an alternative to approximate CN inference algorithms [22, 23], there are approaches proposed explicitly for the bounding of counterfactual queries based on sampling [16], polynomial programming [14], and linear programming [24]. Here we focus on the *expectation maximisation* (EM) solution presented in [20], which provides good inner approximations in relatively short execution times with credibility guarantees [25].

The last challenge that remains is the identification of the causal graph $\mathcal{G}$ with a complete specification of the confounders (i.e., common exogenous causes) [13]. For this task, we plan to use *partial ancestral graphs* (PAGs) [26, 27], which are a type of graphical model used to represent the causal connectivity among endogenous variables when the causal relations and the presence of confounders are only partially known.

In a PAG, directed edges encode so-called *ancestral relations* between endogenous variables. Specifically, an edge from X to X' marks the existence of a causal path from X to X' (possibly mediated by some unknown exogenous variables) and excludes the possibility of X' to be a cause of X. A bi-directed edge $X \leftrightarrow X'$ excludes that both X is a cause of X' and X' is a cause of X, which means that it must exist some latent common variable explaining the correlation between X and X'. Circles on the edges mark uncertainty regarding the existence of a causal dependence, i.e., $X \multimap X'$ indicates that we do not definitively know whether there is a causal path from X to X' or not. Finally, directed edges labelled by v mark a *direct* causal dependence that excludes the presence of possible confounders [28].

The significant advantage of PAGs is that they can be learned directly from data via standard structural learning algorithms [29, 30, 31]. In contrast, the learning of causal directed graphs is a much more challenging task and typically necessitates substantial assumptions about the causal-generating mechanism beyond the available data [32]. However, PAGs are less informative than standard causal diagrams since they encompass uncertainties in both edge orientations and the presence of confounders. Specifically, a PAG corresponds to an equivalence class of causal graphs, with each graph representing a distinct causal-generating mechanism compatible with the available endogenous information (see, Fig. 2). To obtain a causal graph from a PAG, we can then proceed by selecting one of the graphs in the equivalence class using domain-specific background knowledge.

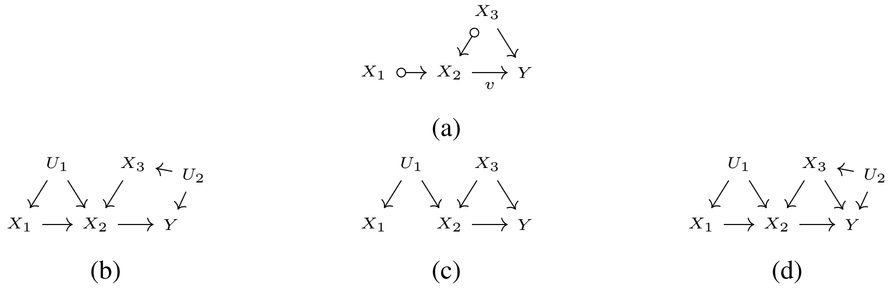

Figure 2: A sample of three causal directed graphs (b,c,d) in the equivalence class determined by a PAG (a).

2. The MaLESCaMo XAI Procedure

The ideas sketched in the previous section are the basis for defining a complete XAI procedure to extract genuinely causal CEs based on a query $\mathcal{Q}$, an instance of the features $\boldsymbol{x}$, and minimal access to expert domain knowledge. The procedure is articulated in several distinct steps, graphically outlined in Fig. 3 and described below.

- The procedure is **model-agnostic** and the training of f is obtained from the supervised data $\mathcal{D}^{\boldsymbol{X},Y}$ through any (supervised) **machine learning** (ML) algorithm.
- The procedure is **local** aiming to explain a single test instance $\boldsymbol{x}$ of the features. A *synthetic neighbourhood* $\mathcal{D}_{\boldsymbol{x}}^{\boldsymbol{X}}$ of $\boldsymbol{x}$ is generated by perturbing the input space locally on $\boldsymbol{x}$ and then selecting the instances whose **distance** δ from $\boldsymbol{x}$ is lower than a threshold δ^*. The distance δ can be a standard metric (e.g., Hamming) or some *ad-hoc* distance functions considered suitable for the application. Similarly, the choice of δ^* is up to the user, depending on the specific application context.
- The instances of the features in the neighbourhood $\mathcal{D}_{\boldsymbol{x}}^{\boldsymbol{X}}$ are **automatically annotated** by using f as an oracle. The resulting annotated dataset is denoted as $\mathcal{D}_{\boldsymbol{x}}^{\boldsymbol{X},Y}$.
- If the number of features is too high to be tractable and explainable in our setup, the dimensionality of the instances in $\mathcal{D}_{\boldsymbol{x}}^{\boldsymbol{X},Y}$ can be reduced by applying standard **feature selection** (FS) methods. The resulting dataset is denoted as $\tilde{\mathcal{D}}_{\boldsymbol{x}}^{\boldsymbol{X},Y}$.
- **Structural learning** (SL) algorithms, such as [31], infer the PAG $\mathcal{G}_{\mathsf{PAG}}$ over $(\boldsymbol{X}, Y)$ from the dataset $\tilde{\mathcal{D}}'_{X,Y}$.
- **Domain knowledge** (DK) in the form of a knowledge base or a human expert is required to select a causal directed graph $\mathcal{G}$ in the equivalence class of $\mathcal{G}_{\mathsf{PAG}}$.
- The **structural equations** (SEs) of the *surrogate* SCM based on $\mathcal{G}$ can be also based on DK. Alternatively, the conservative approach above described should be considered.
- Finally, the EM procedure of [20], already embedded in a software library for causal inference [33], can be used to answer the **causal query** $\mathcal{Q}$.

$$\begin{tikzcd}
\mathcal{D}^{\boldsymbol{X},Y} \arrow[r, "\mathsf{ML}"] & f \arrow[dr] & & & & & \mathsf{DK} \arrow[dl] \arrow[d, "\mathsf{SEs}"] & \\
\boldsymbol{x} \arrow[r, "\delta,\,\delta^*"] & \mathcal{D}_{\boldsymbol{x}}^{\boldsymbol{X}} \arrow[r] & \mathcal{D}_{\boldsymbol{x}}^{\boldsymbol{X},Y} \arrow[r, "\mathsf{FS}"] & \tilde{\mathcal{D}}_{\boldsymbol{x}}^{\boldsymbol{X},Y} \arrow[r, "\mathsf{SL}"] \arrow[rrrr, bend right] & \mathcal{G}_{\mathsf{PAG}} \arrow[r] & \mathcal{G} \arrow[r] & \mathsf{SCM} \arrow[r, "\mathsf{EM}"] & \mathsf{CE} \\
\mathcal{Q} \arrow[urrrrrrr, bend right] & & & & & & &
\end{tikzcd}$$

Figure 3: The pipeline of the MaLESCaMo XAI procedure for causal CEs.

Acknowledgments

The research in this paper has been supported by the Hasler Foundation grant No. 22050.

References

[1] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

[2] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115.

[3] R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, F. Giannotti, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.

[4] R. Guidotti, Counterfactual explanations and how to find them: literature review and benchmarking, Data Mining and Knowledge Discovery (2022) 1–55.

[5] S. Verma, V. Boonsanong, M. Hoang, K. E. Hines, J. P. Dickerson, C. Shah, Counterfactual explanations and algorithmic recourses for machine learning: A review, arXiv:2010.10596 (2020).

[6] S. Beckers, Causal explanations and xai, in: First Conference on Causal Learning and Reasoning, 2021.

[7] C. Molnar, Interpretable machine learning, Lulu. com, 2020.

[8] R. K. Mothilal, A. Sharma, C. Tan, Explaining machine learning classifiers through diverse counterfactual explanations, in: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 2020, pp. 607–617.

[9] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: automated decisions and the GDPR, Harvard Journal of Law & Technology 31 (2017) 841.

[10] X. Zhang, A. Solar-Lezama, R. Singh, Interpreting neural network judgments via minimal, stable, and symbolic corrections, in: Advances in Neural Information Processing Systems, volume 31, Curran Associates, Inc., 2018.

[11] A. Karimi, G. Barthe, B. Schölkopf, I. Valera, A survey of algorithmic recourse: definitions, formulations, solutions, and prospects, arXiv:2010.04050 (2020).

[12] J. Tian, J. Pearl, Probabilities of causation: Bounds and identification, Annals of Mathematics and Artificial Intelligence 28 (2000) 287–313.

[13] J. Pearl, Causality, Cambridge University Press, 2009.

[14] G. Duarte, F. N., D. Knox, J. Mummolo, I. Shpitser, An automated approach to causal inference in discrete settings, arXiv:2109.13471 (2021).

[15] M. Zaffalon, A. Antonucci, R. Cabañas, Structural causal models are (solvable by) credal networks, in: International Conference on Probabilistic Graphical Models, PMLR, 2020, pp. 581–592.

[16] J. Zhang, J. Tian, E. Bareinboim, Partial counterfactual identification from observational and experimental data, in: Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 26548–26558.

[17] F. G. Cozman, Credal networks, Artificial intelligence 120 (2000) 199–233.

[18] D. Huber, R. Cabañas, A. Antonucci, M. Zaffalon, Crema: A Java library for credal network inference, in: Proceedings of the tenth International Conference on Probabilistic Graphical

Models, volume 138 of *PMLR*, JMLR.org, 2020, pp. 613–616.

[19] I. Shpitser, J. Pearl, What counterfactuals can be tested, in: Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, 2007, pp. 352–359.

[20] M. Zaffalon, A. Antonucci, R. Cabañas, Causal expectation-maximisation, in: Proceeding of the WHY-21 NeurIPS Workshop, 2021.

[21] D. D. Mauá, C. P. De Campos, A. Benavoli, A. Antonucci, Probabilistic inference in credal networks: new complexity results, Journal of Artificial Intelligence Research 50 (2014) 603–637.

[22] A. Antonucci, Y. Sun, C. de Campos, M. Zaffalon, Generalized loopy 2U: A new algorithm for approximate inference in credal networks, International Journal of Approximate Reasoning 51 (2010) 474–484.

[23] A. Antonucci, C. P. de Campos, D. Huber, M. Zaffalon, Approximate credal network updating by linear programming with applications to decision making, International Journal of Approximate Reasoning 58 (2015) 25–38.

[24] A. Balke, J. Pearl, Counterfactual probabilities: Computational methods, bounds and applications, in: Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann, 1994, pp. 46–54.

[25] M. Zaffalon, A. Antonucci, R. Cabañas, D. Huber, D. Azzimonti, Bounding counterfactuals under selection bias, in: Proceedings of The 11th International Conference on Probabilistic Graphical Models, volume 186 of *Proceedings of Machine Learning Research*, 2022, pp. 289–300.

[26] J. Zhang, Generalized do-calculus with testable causal assumptions, in: Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, volume 2 of *Proceedings of Machine Learning Research*, PMLR, 2007, pp. 667–674.

[27] J. Zhang, Causal reasoning with ancestral graphs, Journal of Machine Learning Research 9 (2008) 1437–1474.

[28] A. Jaber, A. Ribeiro, J. Zhang, E. Bareinboim, Causal identification under markov equivalence: calculus, algorithm, and completeness, Advances in Neural Information Processing Systems 35 (2022) 3679–3690.

[29] B. Huang, K. Zhang, J. Zhang, J. Ramsey, R. Sanchez-Romero, C. Glymour, B. Schölkopf, Causal discovery from heterogeneous/nonstationary data, The Journal of Machine Learning Research 21 (2020) 3482–3534.

[30] J. M. Mooij, T. Claassen, Constraint-based causal discovery using partial ancestral graphs in the presence of cycles, in: Conference on Uncertainty in Artificial Intelligence, PMLR, 2020, pp. 1159–1168.

[31] J. Zhang, On the completeness of orientation rules for causal discovery in the presence of latent confounders and selection bias, Artificial Intelligence 172 (2008) 1873–1896.

[32] P. Spirtes, An anytime algorithm for causal inference, in: International Workshop on Artificial Intelligence and Statistics, PMLR, 2001, pp. 278–285.

[33] R. Cabañas, A. Antonucci, D. Huber, M. Zaffalon, CREDICI: a Java library for causal inference by credal networks, in: International Conference on Probabilistic Graphical Models, PMLR, 2020, pp. 597–600.