Causal Inference in ProM (Extended Abstract)

Mahnaz Sadat Qafari, Wil M. P. van der Aalst
Chair of Process and Data Science, RWTH Aachen University, Aachen, Germany
{m.s.qafari, wvdaalst}@pads.rwth-aachen.de

***Abstract*—Process mining is widely used to turn the stored data by the information systems of companies into actionable information. Companies are not just interested in discovering their processes but also want to know how to enhance them. Thus, they are interested not just in detecting the performance and conformance problems in their processes, but also in designing specific action steps to reengineer their processes. Knowing the causal relationships among the process features is vital information that may help to improve a process. In this paper, we present a ProM plug-in that helps process owners discover the causal relationships among the features of their processes and also provide them with the possibility of foreseeing the effect of interventions on their processes.**

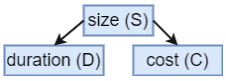
***Index Terms*—Process mining, Structural equation model, Process enhancement**

I. Introduction

Process enhancement is one of the applications of *process mining* which gain more and more attention from both academic and industrial communities. Process enhancement provides process managers and stockholders with insights on the friction points of the process and actionable suggestions on how to resolve each issue. Providing actionable insight towards process reengineering requires a deep understanding of the process, including the causal relationships among the process features. Today, there are several robust techniques for process monitoring and finding their friction points, but little work is done on discovering the causal relationships. In the presented tool, we focus on uncovering causal relationships among process features and investigating the impact of interventions.

The structure of the causal relationships among the process features can be encoded and visualized using a graph which is called the *Causal Structure (CS)*. In a CS, each vertex is corresponding to a process feature also the existence of a directed edge (v_1, v_2) between two vertex v_1 and v_2 means that the corresponding feature of v_1 is a direct cause of the corresponding feature of v_2. The CS can be further used to discover the *Structural Equation Model (SEM)* of the data which is a set of equations encoding the observational and interventional distribution of the data [5]. Having CS and the data, discovering the SEM of the features, is a statistical estimation problem. The SEM can be used to foresee the effect of the intervention on any of the process features. An example of a CS is shown in Figure 1a and a possible SEM with the same CS is shown in Figure 1b.

Determining the CS and the SEM of a set of features require incorporating both data-driven methods and domain knowledge. In this paper, we present a ProM plug-in, [10], that provides the process managers and stockholders with an easy and interactive way of discovering the causal relationships and their qualities among the process features.



(a) An example of a CS with three features.

$$S = N_S, \quad N_S = \mathcal{N}(5,4)$$
$$D = 5D + N_D, \quad N_D = \mathcal{N}(0,1)$$
$$C = 7S + N_C, \quad N_C = \mathcal{N}(0,2)$$

(b) A causal equation model with the same CS as the one in 1a.

Fig. 1: Suppose in a delivery process, the duration (D) and the cost (C) of delivering items are correlated. If the CS of these features is as in Figure 1a, where S indicates the size of the item, then there is no causal relationship between C and D. The correlation between C and D is explainable by their common cause, S. This CS indicates that intervention on D (e.g., by increasing the resources such that the delivery takes a shorter time) does not have any effect on C.

There exist relevant work on discovering the causal relationships among process features. For example, in [3], [6], the goal is to uncover the causal equation model of the process features at the process level. Moreover, in [2], an approach based on time series analysis is used to discover the cause-effect relations. Also, causal reasoning in the case level has been investigated in [1], [8].

The rest of the paper is organized as follows. In Section II we explain the method used in the implemented plug-in. In Section III, we mention some of the applications of the tool. Finally, in IV, we depart with the conclusion.

II. Method

The inputs of our plug-in are the event log of the process, the process model, and the conformance checking results of replaying the given event log on the given model. An overview of our approach is shown in Figure 2.

As a preprocessing step, we enrich the event log by adding several derived attributes; e.g. conformance diagnostics. Then the user determines the target and the descriptive features. A tabular data is extracted from the enriched event log such that all the data related to each occurrence of the target feature are gathered from the part of the trace that happens before. In this plug-in, we focus on three types of target features, which we call them *situations*: 1) choice situation, e.g., which factors influence the decision made in a choice place, 2) trace situation, e.g., why deviations happen in some cases, and 3) event situation, e.g., what causes the bottleneck in an event.

We thank the Alexander von Humboldt (AvH) Stiftung for supporting our research.

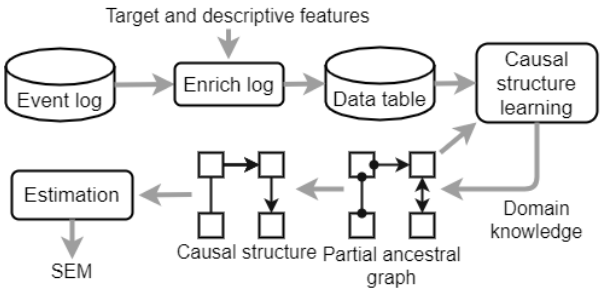


Fig. 2: The general overview of the implemented causal inference method.

In the next step, the user who possesses the domain knowledge can provide the CS of the process features in the form of a directed acyclic graph. However, usually, the process owner does not have such knowledge, so we use a causal structure learning algorithm, also called *search algorithm*, to uncover the CS in a data-driven manner. The input of a search algorithm is a data table (and possibly domain knowledge) and its output is a *partial ancestral graph* which is a graphical object that encodes the set of CSs that have been statistically supported by the data. A partial ancestral graph generated by our plug-in for the data extracted from a real event log is shown in Figure 3. This graph reveals valuable information about the possible causal relationships in the process and can be used as initial insight into the CS of the features. The user modifies this graph further by editing the graph or adding domain knowledge to the search algorithm and turn this graph into the CS of the process features. In this plug-in, we assume the linear dependencies. Also, we used the Tetrad, [9], implementation of *greedy fast causal inference algorithm*, [4], as the search algorithm. The final step involves estimating the strength of the causal relationships in the CS which results in the SEM of the data. The output of this approach can be used to predict the effect of an intervention on the process features which is crucial for process enhancement planning.

III. Maturity of the Tool

The implemented plug-in is available in the nightly build of ProM under the name *root-cause analysis using structural equation model*. Also, the source code of our tool[1] and a video tutorial[2] are publicly available. The tool has been used in multiple academic projects to discover the causal relationships among the process features. For example, to generate the experimental results of [7] and [6], this plug-in has been used on several synthetic and real event logs. Moreover, for providing case-level counterfactual explanations using the method proposed in [8], this plug-in has been used as a preprocessing step to discover the causal equation model of the features extracted from an event log. The results of these papers show the validity of the proposed method.

[1]https://svn.win.tue.nl/repos/prom/Packages/CausalityInference
[2]https://youtu.be/jcBqExtJRO8

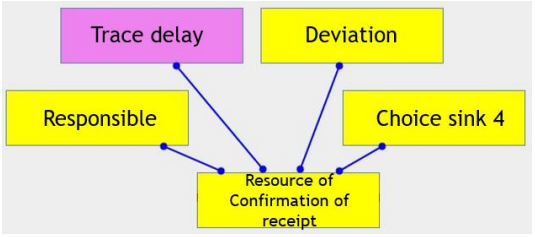


Fig. 3: In the above partial ancestral graph, “trace delay” is the target and the corresponding features of the yellow vertices are the descriptive features. This graph shows the existence of four statistically supported causal relationships among the features. If we use the CS in which the “resource of confirmation of receipt” is the cause of other features, and do the estimation, in the resulting SEM we can see that the intervention “resource of confirmation of receipt = resource23” results in “delay” with probability 0.29 and “on-time” with probability 0.71.

IV. Conclusion

The structure of the causal relationship among process features provides indispensable information for process enhancement. The CS can further be used to discover the causal equation model of the process features which provides the user the possibility of investigating the effect of interventions on the process in a data-driven manner.

In this paper, we have introduced a plug-in in ProM, which provides the user a simple yet sophisticated interactive method to discover not just the CS, but also the SEM of the process features.

References

[1] Z. D. Bozorgi, I. Teinemaa, M. Dumas, M. La Rosa, and A. Polyvyanyy. Process mining meets causal machine learning: Discovering causal rules from event logs. In *2020 2nd International Conference on Process Mining (ICPM)*, pages 129–136. IEEE, 2020.

[2] B. F. Hompes, A. Maaradji, M. La Rosa, M. Dumas, J. C. Buijs, and W. M. van der Aalst. Discovering causal factors explaining business process performance variation. In *International Conference on Advanced Information Systems Engineering*, pages 177–192. Springer, 2017.

[3] T. Narendra, P. Agarwal, M. Gupta, and S. Dechu. Counterfactual reasoning for process optimization using structural causal models. In *Proceedings of Business Process Management Forum*, volume 360, pages 91–106. Springer, 2019.

[4] J. M. Ogarrio, P. Spirtes, and J. Ramsey. A hybrid causal search algorithm for latent variable models. In *Proceedings of Probabilistic Graphical Models - Eighth International Conference*, pages 368–379, 2016.

[5] J. Pearl. *Causality*. Cambridge university press, 2009.

[6] M. S. Qafari and W. van der Aalst. Root cause analysis in process mining using structural equation models. In *International Conference on Business Process Management*, pages 155–167. Springer, 2020.

[7] M. S. Qafari and W. van der Aalst. Feature recommendation for structural equation model discovery in process mining, 2021.

[8] M. S. Qafari and W. M. van der Aalst. Case level counterfactual reasoning in process mining. In *International Conference on Advanced Information Systems Engineering*, pages 55–63. Springer, 2021.

[9] R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson. The tetrad project: Constraint based aids to causal model specification. *Multivariate Behavioral Research*, 33(1):65–117, 1998.

[10] H. Verbeek, J. Buijs, B. Van Dongen, and W. M. van der Aalst. Prom 6: The process mining toolkit. *Proc. of BPM Demonstration Track*, 615:34–39, 2010.