Empowering Causal Machine Learning for Large-scale Manufacturing Pipelines with Knowledge Graphs

Yuxin Zi[1,*,†], Cory Henson[2] and Amit Sheth[1]

[1]Artificial Intelligence Institute, University of South Carolina, Columbia, USA
[2]Bosch Center for Artificial Intelligence, Pittsburgh, USA

Abstract

Understanding causal relations within manufacturing pipelines is crucial for key manufacturing tasks such as anomaly detection and root cause analysis. However, existing causal machine learning (causal ML) approaches struggle to scale effectively to the vast number of variables present in manufacturing settings. We advocate for incorporating domain knowledge within the manufacturing pipelines, represented as knowledge graphs (KGs), for designing causal ML methods for large-scale manufacturing problems. Knowledge graphs can encode rich contextual information about the interactions and dependencies between different components and stages of the manufacturing pipeline, providing a structured framework to guide the discovery of causal relationships. By incorporating KGs, causal ML models can leverage both data-driven approaches and domain knowledge, enhancing scalability and improving the accuracy of causal learning in large scale manufacturing settings.

1. Introduction

Causal machine learning (causal ML) encompasses ML methods aimed at identifying cause-and-effect relationships among variables, primarily using observational data. Manufacturing pipelines involve tens of thousands of observed and unobserved variables, including physical sensor readings, material properties, machine parameters, and environmental factors. Understanding the causal relations between these variables is critical for downstream tasks such as anomaly detection, root cause analysis and process optimization [1]. Existing causal ML methods have demonstrated effectiveness with relatively small numbers of variables; however, they either cannot scale to manufacturing problem or do so with significant inefficiency. The challenge with scaling arises from the combinatorial complexity of evaluating possible causal relationships among variables [2], compounded by noise, confounders, and unobserved variables. We argue that integrating expert-curated knowledge graphs (KGs) of manufacturing processes can enable the development of scalable approaches that discern meaningful causal relationships amidst the complexity of data. Specifically, this knowledge can clarify existing causal relationships and specify additional constraints over the search space, drastically improving the computational tractability and learning stability of causal ML methods. Due to their symbolic form, KGs can significantly enhance the interpretability of causal ML model outputs. Such knowledge-guided approaches have the potential to enhance both the scalability and accuracy of causal ML techniques, ultimately supporting more informed decision-making and process optimization in real-world manufacturing environments.

2. Constructing KGs for Manufacturing Pipelines

The observational data from manufacturing pipelines typically includes sensor readings and details about the manufactured parts. Beyond this, however, there is abundant structured and unstructured knowledge available from most manufacturing pipelines. Organizing this knowledge into KGs requires close collaboration between manufacturing experts, who understand the intricacies of the individual

Posters, Demos, and Industry Tracks at ISWC 2024, November 13–15, 2024, Baltimore, USA
*Corresponding author.
†This work was completed during an internship at Bosch Center for Artificial Intelligence.
yzi@email.sc.edu (Y. Zi); cory.henson@us.bosch.com (C. Henson); amit@sc.edu (A. Sheth)

manufacturing processes, and KG construction experts, who can effectively structure and integrate the data [3] [4]. Specifically, we are exploring the integration of the following information into KGs to develop large-scale causal machine learning methods:

1. ***Direct and indirect causal influences***: expert knowledge on known causal relationships, non-causal relationships, and conditional independence between variables.
2. ***Manufacturing pipeline workflow***: a structured representation of the production process, including temporal and logical dependencies between stages and variables.
3. ***Noise characteristics of variables***: information regarding which variables are subject to measurement noise and the nature of that noise (e.g., Gaussian, uniform, etc.).
4. ***Latent variables and confounders***: identification of important unobserved variables and potential confounding factors in the production process.

3. KGs Enables Large-Scale Causal ML for Manufacturing

The scalability issue of causal learning from observational data lies in determining whether a causal relationship exists between each pair of variables (nodes) [2]. Traditional causal discovery methods are the least scalable due to the need to compute conditional independence between pairs of nodes, making it a combinatorial optimization problem. A more scalable approach involves determining the topological ordering of nodes by iteratively identifying the leaf nodes of the causal graph and then applying feature selection techniques [5]. Recently, approximation methods using deep learning have achieved significantly higher scalability (up to 500 nodes [2]). However, no existing method can scale up to the manufacturing setting due to limitations in assumptions and complexity of the problem. We argue that scalable causal ML methods for manufacturing demands explicit relational descriptions beyond ground-level sensor measurements. Below, we analyse how knowledge can facilitate large-scale causal ML methods:

- Knowledge on ***direct and indirect causal influences*** provides constraints on the causal structure, prunes invalid edges early in the discovery process, and thereby effectively narrows down candidate causal graphs and reduces the search space for learning algorithms.
- Knowledge on ***manufacturing pipeline workflow*** can guide the node selection process to focus on the most relevant nodes, reduce the candidates of possible causes and thus improve scalability. This knowledge also impose temporal constrains on the topological ordering of nodes, reducing the number of candidate graphs in a Markov equivalent class. Temporal constrains can also facilitate efficient time-series based causal ML methods. Note that temporal constrains are not sufficient conditions for identifying causal relations.
- Knowledge on ***noise characteristics of variables*** provides valuable insights into causal processes in additive noise models (ANM) [6] or post-nonlinear (PNL) models [7] for discovering nonlinear causal relationships. By incorporating this knowledge, researchers can better tailor causal ML methods to account for specific types and levels of noise present in the data. This knowledge can enhance the effectiveness of causal discovery by mitigating the impact of noise and uncovering more accurate causal relationships amidst complex, real-world data scenarios.
- Knowledge on ***latent variables and confounders*** allows for more precise identification of causal relationships and better decision-making in complex systems. Current causal ML methods often overlook the presence of unobserved variables. However, in the real world, these latent factors can significantly influence the observed data and lead to erroneous causal learning if not properly accounted for.

We are researching the development of neuro-symbolic or hybrid causal ML methods to incorporate domain-specific KGs. Neuro-symbolic methods combine the strengths of symbolic reasoning with neural network approaches [8], allowing for more scalable, interpretable and robust causal modeling frameworks [9]. Furthermore, KG-guided large-scale causal ML methods should also consider leveraging

parallel computing, efficient data structures, and advanced statistical techniques to manage the scale. In conclusion, by addressing scalability issues in causal ML with domain KGs, we can make significant strides in solving problems such as anomaly detection and root cause analysis, leading to higher product quality, more efficient operations and improved process control, ultimately benefiting the entire manufacturing industry.

4. Ackowledgements

This work is supported by Bosch Center for Artificial Intelligence and NSF Award 2335967 EAGER: Knowledge-guided neurosymbolic AI with guardrails for safe virtual health assistants[1]. The views expressed here are those of the authors, not those of the sponsors.

References

[1] K. Budhathoki, L. Minorics, P. Bloebaum, D. Janzing, Causal structure-based root cause analysis of outliers, in: K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, S. Sabato (Eds.), Proceedings of the 39th International Conference on Machine Learning, volume 162 of *Proceedings of Machine Learning Research*, PMLR, 2022, pp. 2357–2369. URL: https://proceedings.mlr.press/v162/budhathoki22a.html.

[2] P. Sanchez, X. Liu, A. Q. O'Neil, S. A. Tsaftaris, Diffusion models for causal discovery via topological ordering, 2023. URL: https://arxiv.org/abs/2210.06201. arXiv:2210.06201.

[3] A. Sheth, C. Henson, S. S. Sahoo, Semantic sensor web, IEEE Internet Computing 12 (2008) 78–83. doi:10.1109/MIC.2008.87.

[4] H. Cheng, P. Zeng, L. Xue, Z. Shi, P. Wang, H. Yu, Manufacturing ontology development based on industry 4.0 demonstration production line, in: 2016 Third International Conference on Trustworthy Systems and their Applications (TSA), 2016, pp. 42–47. doi:10.1109/TSA.2016.17.

[5] P. Rolland, V. Cevher, M. Kleindessner, C. Russel, B. Schölkopf, D. Janzing, F. Locatello, Score matching enables causal discovery of nonlinear additive noise models, 2022. URL: https://arxiv.org/abs/2203.04413. arXiv:2203.04413.

[6] J. Peters, J. Mooij, D. Janzing, B. Schölkopf, Causal discovery with continuous additive noise models, 2014. URL: https://arxiv.org/abs/1309.6779. arXiv:1309.6779.

[7] K. Zhang, A. Hyvarinen, On the identifiability of the post-nonlinear causal model, 2012. URL: https://arxiv.org/abs/1205.2599. arXiv:1205.2599.

[8] A. Sheth, K. Roy, M. Gaur, Neurosymbolic artificial intelligence (why, what, and how), IEEE Intelligent Systems 38 (2023) 56–62. doi:10.1109/MIS.2023.3268724.

[9] U. Jaimini, C. Henson, A. Sheth, Causal neurosymbolic ai: A synergy between causality and neurosymbolic methods, IEEE Intelligent Systems 39 (2024) 13–19. doi:10.1109/MIS.2024.3395936.

[1] https://www.nsf.gov/awardsearch/showAward?AWD_ID=2335967