Data-Driven and Knowledge-Based Causal Network Discovery for Identifying Differential Equations

Mitsuhiro Odaka[1,2,3,4], **Morgan Magnin**[3,2], **Katsumi Inoue**[2,1,3]
[1]The Graduate University for Advanced Studies, SOKENDAI, [2]National Institute of Informatics, [3]École Centrale Nantes, LS2N, France, [4]Japan Society for the Promotion of Science

Background

Discovering Ordinary Differential Equations (ODEs) governing the dynamics of a system is significant for our better understanding of the world. Various techniques have been developed to distill ODEs from observations, such as constraining hypothesis space with context-free grammar [1], constraint system identification with human-computer communication [2], process model induction from data and knowledge about the observed behavior [3], genetic learning of free-form natural laws [4], sparse regression to learn equations [5], physics-informed learning from small data [6], learning of symmetry and separability [7]. These methods render the dynamics interpretable as ODEs, addressing the following omnipresent pitfalls. Specifically, unlike synthetic data, noise-free data are rarely available (*noise*). Additionally, frequent longitudinal time courses of a complete set of variables without hidden variables are seldom observable (*partial observability*). Moreover, the acquired equations should balance accuracy and complexity (*overfitting*). Uncovering underlying dynamics requires addressing these issues, but simultaneously solving all of them remains challenging. To solve them effectively, we should unify data and knowledge on symbolic structures instead of handling either data or knowledge.

Data-Driven and Knowledge-Based Integrated Approach

The above background has motivated us to implement **an integrated framework of data-driven and knowledge-based approaches for parsimonious equation discovery from noisy and partial observations**. In the framework, ODEs are yielded from state variables of interest. First, once the user inputs desired state variables, multivariate temporal data and knowledge represented by causal networks are retrieved online from databases and domain-specific knowledge bases. A causal network can also be built purely from data. For example, we have developed a methodology to combine network inference driven by gene expression data and model validation based on multiple knowledge bases [8]. An undirected partial correlation network was created from data, and causal directions were assigned according to knowledge. Meanwhile, knowledge is also used for the abduction of latent variables in causal networks. In this case, new variables and their associated temporal data and knowledge are fetched via an API and automatically added to the model. Second, causal networks inferred from data or derived from knowledge bases result in building candidate ODEs. Third, the ODEs are verified by investigating their predictive accuracy and stability and simplified by parameter pruning with sensitivity indices. For instance, we have constructed different ODEs of viral dynamics by querying the state variables, collecting the viral load time series and state transition diagrams, with each transition corroborated by primary sources, and verifying the ODEs with sensitivity and stability analyses [9]. While iteratively expanding and reducing the model, the identified ODEs are evaluated by model selection criteria to avoid overfitting. Finally, the ODEs are obtained with a causal network by integrating data and knowledge.

Parsimonious Equation Learning with Causality

Following our previous studies that provided knowledge-based causal directions for data-driven undirected networks, our current efforts are dedicated to **learning causal network topology from temporal data based on Generative Adversarial Networks (GANs) for robustness to adversarial noise and incorporating it into equation discovery for less model complexity**. Predicting all edge weights of a fully-connected bipartite graph whose nodes are time series of endogenous and higher-order candidate variables is necessary to select amenable variables and function bases for dynamics and to reduce model complexity and computational cost. Existing equation discovery techniques [4–7] have not explicitly discovered causality. We now conjecture that observational causal discovery can limit the search space in equation discovery and prevent models from overfitting due to those terms that are irrelevant to the causal model. In the proposed framework, GANs learn coefficient matrices of linear structural equations with time delays to detect causality between time series. Subsequently, the coefficient matrix to be estimated in equation discovery is sparsified by injecting this causal network into the hypothesis space as an inductive bias and replacing the non-causal elements in the matrix with zeros. These procedures formulate the problem as a continuous constrained problem with penalty term. Besides, GANs accelerate the mining of robust patterns from noisy data using continuous algebraic computing. The above framework for data-driven and knowledge-based learning of causal networks and ODEs is realized as *Parsimonious Equation Learning with Causality (PELC)*, which can be a novel attempt to open an avenue for linking causality and equation discovery. In the symposium, we expect to report on PELC, its experimental results, and its future.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP21J22938 and JP21H04905, and JST, CREST Grant Number JPMJCR22D3, Japan. This research was also supported by "Strategic Research Projects" grant from ROIS (Research Organization of Information and Systems).

References

1. Todorovski, L. and Džeroski, S. Declarative Bias in Equation Discovery. *Proceedings of the Fourteenth International Conference on Machine Learning*. pp.376–384. 1997. doi:10.5555/645526.657279.
2. Stolle, R. and Bradley, E. Communicable Knowledge in Automated System Identification. *Computational Discovery of Scientific Knowledge. Lecture Notes in Computer Science*. vol. 4660. pp.17–43. 2007. doi:10.1007/978-3-540-73920-3_2.
3. Bridewell, W. and Langley, P. and Todorovski, L. Inductive process modeling. *Mach. Learn.* 71, 1–32. 2008. doi:10.1007/s10994-007-5042-6.
4. Schmidt, M. and Lipson, H. Distilling Free-Form Natural Laws from Experimental Data. *Science*. 324(5923), 81–85. 2009. doi:10.1126/science.1165893.
5. Brunton, S. L. and Proctor, J. L. and Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *PNAS*. 113(15):3932–3937. 2016. doi:10.1073/pnas.1517384113.
6. Raissi, M. and Perdikaris, P. and Karniadakis, G.E. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *J. Comput. Phys.* vol. 378. pp.686–707. 2019. doi:10.1016/j.jcp.2018.10.045.
7. Udrescu S.M. and Tegmark M. AI Feynman: A physics-inspired method for symbolic regression. *Sci. Adv.* 6(16)eaay2631. 2020. doi:10.1126/sciadv.aay2631.
8. Odaka, M. and Magnin, M. and Inoue, K. Gene Network Inference from Single-Cell Omics Data and Domain Knowledge for Constructing COVID-19-Specific *ICAM1*-Associated Pathways. Preprint *Research Square*. 2022. doi:10.21203/rs.3.rs-1300133/v1.
9. Odaka, M. and Inoue, K. Modeling viral dynamics in SARS-CoV-2 infection based on differential equations and numerical analysis. *Heliyon*. 7(10) e08207. 2021. doi:10.1016/j.heliyon.2021.e08207.