Processes and Constraints in Scientific Model Construction

Will Bridewell WILLB@CSLI.STANFORD.EDU

Pat Langley LANGLEY@CSLI.STANFORD.EDU

Computational Learning Laboratory, Center for the Study of Language and Information
Stanford University, Stanford, CA 94305 USA

Abstract

In previous publications, we have reported a computational approach to constructing quantitative process models of dynamic systems from time-series data and background knowledge. However, our experience with these systems suggests that process knowledge is insufficient to avoid the consideration of implausible models. To this end, we have identified and introduced constraints that specify which processes can or must occur together, which in turn limit search to candidates that scientists will consider acceptable. We have also developed methods for inducing such constraints from the results of search through the model space. We maintain that the ability to specify, utilize, and induce constraints on quantitative process models will support deeper understanding of complex systems and constitutes an important addition to eScience.

For the past few years, we have been developing computational tools to aid scientists in constructing process models of complex systems. This work integrates ideas from artificial intelligence, simulation environments, machine learning, and human–computer interaction to represent, simulate, and discover models that explain observations while remaining consistent with knowledge about a domain. We believe this research fills an important need within the eScience movement.

We have focused on constructing *quantitative process models* (Langley et al., 2002), which explain relations in time-series data in terms of domain-specific processes. These models connect knowledge cast as differential and algebraic equations to conceptual knowledge about the processes that underlie them. To explain the dynamics in a data set, our systems instantiate generic processes such as exponential and logistic growth and various forms of predation to fit the specific scenario and use the resulting components to build candidate model structures. After this, they use gradient descent search with random restarts to estimate each structure's parameters and evalute its fit to the data. For example, a population dynamics model could include an exponential growth process for rabbits, a mortality process for foxes, and a particular predation process that relates the two populations. We have used this approach to produce accurate and plausible models in ecological (Asgharbeygi et al., 2006) and biological domains (Langley et al., 2006).

However, our previous process-modeling systems sometimes generated explanations that violated knowledge about a domain. Although the generic processes place considerable limits on the possible models, they do not provide limits on component assembly that would avoid incomplete or otherwise implausible structures. To illustrate, these constraints might ensure that if a population has a growth process, then it must have a corresponding one for mortality. This knowledge defines the structure of an acceptable solution and rules out candidates that lack any of the required components. The HIPM system (Todorovski et al., 2005) incorporated such constraints in the form of a process hierarchy, but interactions with scientific users indicated that they found this approach rigid and unwieldy.

In more recent work, we have developed SC-IPM, a system that uses modular constraints to direct search through the space of process models. Importantly, these constraints form a part of scientific knowledge that is often overlooked (Langley & Bridewell, 2008). We hold that constraints correspond to general theoretical principles that delineate the form of explanations within a scientific domain. Moreover, their application substantially reduces search and improves accuracy (Todorovski et al., 2005). However, scientists rarely discuss these constraints, as they constitute implicit background knowledge and do not appear explicitly in models. Nevertheless, experts can recognize violations and state the underlying rule, so representing them explicitly seems a reasonable objective.

Recent evidence suggests that one can induce these structural constraints from past experience (Bridewell & Todorovski, 2007). While searching for an explanation, a process-modeling system considers several candidates of varying quality. Labeling each solution as a relatively good or bad fit to the data and establishing features based on model structure produces a training set for an inductive logic programming system. The resulting rules describe the characteristics of good and bad models in a way that corresponds to scientific constraints. We have used this approach not only to recover known constraints but also to find plausible ones supported by the domain literature (Bridewell et al., 2007). Our current research is investigating how well these constraints transfer to other problems.

Both quantitative process modeling and constraint induction open the door to new scientific methodology and offer an expanded vision for eScience. Modeling environments that incorporate such capabilities will let researchers consider vast numbers of alternative explanations, rather than focusing on only a few, as they do currently. Moreover, constraint induction lets one analyze this throng of information by extracting general principles from modeling experience. This new scientific knowledge contributes to a field's general understanding of a broad class of problems.

Acknowledgements

This research was funded by NSF Grant No. IIS-0326059. We thank Kevin Arrigo, Stuart Borrett, Matt Bravo, and Ljupčo Todorovski for their contributions to the ideas presented here.

References

Asgharbeygi, N., Bay, S., Langley, P., & Arrigo, K. (2006). Inductive revision of quantitative process models. *Ecological Modelling*, *194*, 70–79.

Bridewell, W., Borrett, S., & Todorovski, L. (2007). Extracting constraints for process modeling. *Proceedings of the Fourth International Conference on Knowledge Capture*, 87–94.

Bridewell, W., & Todorovski, L. (2007). Learning declarative bias. *Proceedings of the Seventeenth International Conference on Inductive Logic Programming* (pp. 63–77).

Langley, P., & Bridewell, W. (2008). Processes and constraints in explanatory scientific discovery. *Proceedings of the Thirtieth Annual Meeting of the Cognitive Science Society.*

Langley, P., Sánchez, J., Todorovski, L., & Džeroski, S. (2002). Inducing process models from continuous data. *Proceedings of the Nineteenth International Conference on Machine Learning* (pp. 347–354).

Langley, P., Shiran, O., Shrager, J., Todorovski, L., & Pohorille, A. (2006). Constructing explanatory process models from biological data and knowledge. *Artificial Intelligence in Medicine*, *37*, 191–201.

Todorovski, L., Shiran, O., Bridewell, W., & Langley, P. (2005). Inducing hierarchical process models in dynamic domains. *Proceedings of the Twentieth National Conference on Artificial Intelligence* (pp. 892–897).