Processes and Constraints in Explanatory Scientific Discovery

Pat Langley (langley@csli.stanford.edu)
Will Bridewell (willb@csli.stanford.edu)
Computational Learning Laboratory, CSLI
Stanford University, Stanford, CA 94305 USA

Constructing Process Models

In previous publications, we have reported a computational approach to constructing explanatory process models of dynamic systems from time-series data and background knowledge. We have not aimed to mimic the detailed behavior of human researchers, but we maintain that our systems address the same tasks as ecologists, biologists, and other theory-guided scientists, and that they carry out search through similar problem spaces.

Our initial research (Langley et al., 2002) introduced an approach to model discovery that uses background knowledge about generic processes in a scientific domain to generate candidate model structures that relate a set of continuous variables. For each model structure, the method carries out a gradient descent search through the parameter space, with random restarts, to fit the structure to observations. Generic processes serve as building blocks from which to construct explanatory models.

We have applied this framework successfully to infer plausible models of dynamic systems observed by ecologists (Asgharbeygi et al., 2006) and biologists (Langley et al., 2006). These have included phytoplankton growth in the Ross Sea, predator-prey interactions in protists, gene regulation of photosynthetic activity, and water dynamics in a Danish fjord. Extensions have included making the approach robust with respect to noise (Bridewell et al., 2005) and handling data sets with missing observations, both of which reduced variance across data sets and lowered the squared error on novel test cases.

Two Forms of Scientific Knowledge

Unlike most early computational models of scientific discovery, which emphasized knowledge-lean induction of descriptive laws, our recent research has emphasized knowledge-laden construction of explanatory models. However, our initial forays produced an important insight: adding generic processes to the background knowledge increases the search space, since it supports the creation of more model structures. Clearly, the adage about knowledge reducing search does not always hold; it depends on the type of knowledge involved.

In response, we imposed a hierarchical organization on process knowledge (Todorovski et al., 2005). This reduced search, improved accuracy, and produced more plausible models, but the hierarchy itself was cumbersome and unfamiliar to the scientists we interviewed. In recent work, we have had more success by introducing a variety of constraints among processes and the entities they relate. Unlike the hierarchical structures, these constraints combine the benefits of search reduction with modularity, and one can imagine how a scientific community might alter them as it matures. In ongoing work, we are exploring methods for inducing these constraints from experience with successful and unsuccessful model structures, which may help explain the origin of such principles in a variety of disciplines.

The distinction between process and constraint knowledge has been notably absent from the cognitive science and philosophy of science literatures, yet it seems crucial to understanding the generation of scientific explanations. Processes provide the content from which scientists construct models, whereas constraints correspond to theoretical principles about how to combine processes. Knowledge about such constraints is often implicit, and our work provides a formalism for making them explicit, which in turn supports their controlled use in directing search through the space of explanatory models.

Acknowledgements

This research was funded by NSF Grant No. IIS-0326059. We thank Kevin Arrigo, 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., Bani Asadi, N., Langley, P., & Todorovski, L. (2005). Reducing overfitting in process model induction. *Proceedings of the Twenty-Second International Conference on Machine Learning* (pp. 81–88).

Langley, P., Sanchez, 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).