Computational Revision of Ecological Process Models

Nima Asgharbeygi,[1] Pat Langley,[1] Stephen Bay,[1] and Kevin Arrigo[2]
[1]Computational Learning Laboratory, CSLI, Stanford University, Stanford, CA 94305
[2]Department of Geophysics, Stanford University, Stanford, CA 94305

Most ecological models are developed manually by scientists, who decide on their basic structure, tune their parameters, compare them against available data, and refine them in response. In contrast, most work on computational scientific discovery has emphasized the automated generation of models from data and background knowledge. In this abstract, we describe an approach to model revision that incorporates ideas from both traditions. We believe that computational tools for model revision offer great practical value to scientists by decreasing the time required to search for models while letting them retain control over the search space.

Our approach involves the modification of *quantitative process models*, a representation of knowledge that constrains search while remaining interpretable. In this framework, a model consists of a set of processes, each of which specifies one or more differential or algebraic equations that represent causal relationships among variables. Processes can include threshold conditions on variables that characterize when they are active. A variable may be labeled as observable, meaning it is present in the data, or play the role of a theoretical term that serves to link processes. Each variable is also marked as either exogenous, in that it influences other variables but is not influenced in return, or as endogenous, which means it is causally dependent on other variables.

For example, we have developed a process model for the aquatic ecosystem of the Ross Sea based on Arrigo et al.'s (2003) earlier model, which is cast as a set of differential equations. The new version incorporates four observable variables: the available light, the amount of ice, and the concentrations of phytoplankton and nitrate. The model includes processes for the loss of phytoplankton from natural causes and for its growth as a function of current concentration and growth rate. A third process specifies the decrease in nitrate associated with its update by phytoplankton, and another indicates that the growth rate is a product of the unconstrained rate and the minimum of two theoretical terms, nitrate-rate and light-rate, which determine the fraction of the unconstrained rate achievable when the available nitrate or light are limited. Two final processes specify parameters that occur across processes and describe the variable light as a function of time. We can utilize a process model of this sort, together with initial values, to simulate its behavior over time and thus predict values for each endogenous variable.

In previous work (Langley et al., 2003), we developed an initial algorithm, called IPM, to address the task of inducing process models like the one described above from time-series data and from knowledge about the domain. We cast this background knowledge as a set of *generic* processes which are distinguished from specific processes in that they do not commit to particular variables or parameter values. However, they can contain constraints, such as types for generic variables and intervals for parameter values. Although IPM produced encouraging results, it had drawbacks that limited its applicability: the space of explored models could still be large, and it provided no way to guide the search

In response, we have developed a new system, IPM/R, which adopts a revision approach to process model induction. This algorithm requires the user to specify four inputs. These include an initial model that encodes the user's beliefs about the processes that are most likely involved, a set of possible changes that specify which initial processes can be removed or have their parameters altered, a set of generic processes that may be added to the initial model, and observations to which the revised model should be fit. The possible changes, combined with the candidate processes for addition, guide IPM/R's search toward parts of the model space that are consistent with the user's knowledge about the domain.

IPM/R generates a set of revised models that are sorted by their distance from the initial model and presented to the user with their mean squared errors on the data. This output format lets the user observe the trade-off between the performance of the revised models and their similarity to the initial model, in order to determine the best compromise when selecting among the candidate revisions. We applied both IPM and IPM/R to the problem of modeling phytoplankton population dynamics in the Ross Sea, using the initial model described above and alternative generic processes that included mechanisms for zooplankton grazing on phytoplankton, nitrate remineralization, and residue loss. Our input data consisted of 188 daily measurements of sunlight, ice amount, phytoplankton concentration, and nitrate concentration in the Ross Sea.

Our runs revealed that IPM/R found revised models which reduced error substantially by making only a few ecologically plausible revisions to the original model, including the addition of processes for nutrient remineralization and zooplankton grazing. In contrast, IPM generated models with comparable error after a much longer execution time, and these were very different from the initial model and less comprehensible. These results demonstrate that IPM/R can produce accurate and comprehensible models that make contact with existing domain knowledge. Although some earlier work has utilized machine learning to revise quantitative models (Todorovski et al., 2003; Whigham & Recknagel, 2001), we have adapted this approach to the improvement of dynamical process models, which seem especially appropriate for fields like ecology.

References

Arrigo, K. R., Worthen, D. L. & Robinson, D. H. (2003). A coupled ocean-ecosystem model of the Ross Sea: 2. Iron regulation of phytoplankton taxonomic variability and primary production. *Journal of Geophysical Research*, *108*, C7, 3231.

Langley, P., George, D., Bay, S., & Saito, K. (2003). Robust induction of process models from time-series data. *Proceedings of the Twentieth International Conference on Machine Learning* (pp. 432–439). Washington, DC: AAAI Press.

Todorovski, L., Dzeroski, S., Langley, P., & Potter, C. (2003). Using equation discovery to revise an Earth ecosystem model of carbon net production. *Ecological Modeling*, *170*, 141–154.

Whigham, P. A. & Recknagel, F. (2001). Predicting Chlorophyll-a in freshwater lakes by hybridising process-based models and genetic algorithms. *Ecological Modelling*, *146*, 243–251.