Inducing design biases that characterize successful experimentation in weak-theory domains: TIPS

Vanathi Gopalakrishnan
Intelligent Systems Laboratory, University of Pittsburgh, Pittsburgh, PA 15260
vanathi@cs.pitt.edu

Experiment design in domains with weak theories is largely a trial-and-error process. In such domains, the effects of actions are unpredictable due to insufficient knowledge about the causal relationships among entities involved in an experiment. Thus, experiments are designed based on heuristics obtained from prior experience. Assuming that past experiment designs leading to success or failure can be recorded electronically, this thesis research proposes one method for analyzing these designs to yield hints regarding effective operator application sequences. This work assumes that the order in which operators are applied matters to the overall success of experiments. Experiment design can also be thought of as a form of planning, since it involves generation of a sequence of steps comprising of one or more operations that can change the environment by changing values of some of the parameters that describe the environment. Experiment design operators can therefore be thought of as plan operators at higher levels of abstraction. This thesis proposes a method for learning contexts within which applying certain sequences of operators has favored successful experimentation in the past.

The motivation for this thesis arose from study of experiment design in macromolecular crystallography. The goal of experiments in this domain is to obtain a good quality crystal of a macromolecule (protein/DNA/complex) so that it can be X-ray diffracted to yield the 3-D molecular structure. There is little understanding about how the large number of parameters (> 25) such as temperature, pH, and macromolecular concentration, interact and influence the growth of a particular macromolecular crystal. Yet, it is observed that some people are able to produce good quality crystals faster than others (crystal growing can take anywhere from a week to a few years). This observation seems to indicate that it is quite likely that these "good" crystallographers are using some *pet* solutions or other methodology that is helping them search this vast parameter space effectively and efficiently. We can refer to the different ways that crystallographers search the parameter space (by setting up appropriate experiments) as *design biases.*

It would be desirable to capture these design biases or preferences through knowledge engineering, via interviews with the experts. The caveat is that such knowledge tends to be very hard to convey verbally without reference to a specific protein or DNA molecule. Since the experiments are done over a long period of time, the most accurate record of the precise methodology used is found in the laboratory notebooks of individual crystallographers. We have found a way to capture the information found in laboratory notebooks electronically via an easy-to-use interface [1].

The Temporal Induction of Plan Sequence (TIPS) framework proposed in this thesis aims to provide a set of hints regarding operator application sequence for a particular problem posed by an experiment designer. The major component of TIPS is a symbolic inductive Rule Learner (Temporal-RL) that uses the method of *temporal specialization* to learn design biases from past cases of successful and failed experiment designs. Design biases are characterized by a context, followed by a set of temporal relationships (e.g., before, during) expressed among operators. Contexts refer to characteristics of the macromolecule (e.g., protein name, class, weight). The learned contexts are used for matching with the particular problem characteristics given by the user to TIPS.

Learning structured attributes is an open problem in machine learning. This thesis provides one method to deal with structured attributes such as steps in a plan. New symbolic learning representations are introduced for instance and concept description languages. These representations exploit the structured aspects of steps in experiment design. A symbolic inductive learner is augmented with capabilities to learn statistically significant relations between normal attributes (not structured) and structured attributes (steps and sequence of steps), with respect to the class of successful or failed experiment designs. The thesis also demonstrates that learning at different levels of abstraction by exploiting structures inherent in a domain can be efficient and meaningful.

Acknowledgements

I am extremely grateful to Prof. Bruce Buchanan for his guidance and contribution to this thesis idea. Thanks also to Prof. John Rosenberg and Dr. Patricia Wilkosz for being most helpful domain experts. This research is supported in part by funds from the W.M.Keck Center for Advanced Training in Computational Biology at the University of Pittsburgh, Carnegie Mellon University and the Pittsburgh Supercomputing Center, and also by NIH.

References

[1] Hennessy, D., Gopalakrishnan, V., Buchanan, B. G., Subramanian, D. Induction of rules for biological macromolecule crystallization. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, August 1994, pp. 179-187.