Lessons for the Computational Discovery of Scientific Knowledge

Pat Langley LANGLEY@ISLE.ORG
Institute for the Study of Learning and Expertise, 2164 Staunton Court, Palo Alto, California 94306 USA

Abstract

In this essay, I review some early analyses of machine learning applications, along with more recent treatments of successful discoveries of scientific knowledge. Although the two problem areas have much in common, we use recent work on computational discovery in Earth science and microbiology to illustrate some important differences. The lessons that emerge from these efforts run counter to some rhetorical claims and assumptions that are widespread in the machine learning and data mining communities.

1. Historical Background

Applications of machine learning date back at least to the early 1980s, when Donald Michie and colleagues championed use of decision-tree induction on industrial problems. This innovative work led to a number of fielded systems in which the knowledge base was constructed by supervised learning methods. Largely parallel developments occurred within other induction paradigms, especially with neural networks and case-based methods, throughout the 1980s.

By the early 1990s, there were enough fielded systems of this sort that I began to collect examples and talk with developers about their experiences. This led to an initial review (Langley, 1992) and a workshop on fielded applications of machine learning, which Yves Kodratoff and I organized in 1993, in conjunction with the Tenth International Conference on Machine Learning. These events occurred before the first conference on data mining and, indeed, before that term even became widely used.

The case studies and workshop led in turn to an article (Langley & Simon, 1995) that reviewed many of these applications. The paper also drew some initial lessons about the factors that influence whether such applied efforts are successful. These included:

- formulating the problem in terms amenable to existing learning algorithms;
- engineering the representation to make the learning task tractable;
- collecting data and manipulating it to aid learning;
- evaluating the learned knowledge in terms of its behavior and acceptability to domain experts;
- fielding the knowledge base in ways that lead to its acceptance and utilization.

One key hypothesis was that making appropriate decisions about these issues was more crucial to success than decisions about which learning algorithm to use. Once they are handled, many induction techniques would achieve positive results. The primary evidence for this claim was that few successful applications involved development of new algorithms; almost invariably, they relied on creative use of existing methods.

More recently, other authors (e.g., Fayyad, Piatetsky-Shapiro, & Smyth, 1996) have proposed similar analyses of stages that arise in knowledge discovery and data mining. They have also emphasized the importance of issues like problem formulation, representation engineering, and data collection. Unfortunately, most publications in both machine learning and data mining still focus on algorithm development, despite general agreement that other issues play at least as important a role in practical systems.

2. Knowledge Discovery in Science

Computational research on the discovery of scientific knowledge has existed for well over three decades (e.g., Simon, 1966). However, this area became truly active only in the late 1970s and early 1980s, along with the generally increased activity in machine learning, with which it was closely associated. Early research on computational discovery focused on replicating events from the history of science and mathematics (e.g., Langley, 1981; Lenat, 1978), which was a reasonable starting point for the field.

However, more recent work has applied similar techniques to discover new scientific knowledge. A clear criterion for success here is whether the new knowledge is published in the refereed literature of the relevant scientific field. This has occurred in enough cases that, a few years ago, I decided to analyze a number of successful applications in an effort to understand their sources of power (Langley, 2000). These included results in astronomy, biology, chemistry, ecology, graph theory, and metallurgy, and the forms of knowledge included taxonomies, qualitative laws, numeric equations, structural models, and reaction pathways.

To this end, I adapted the framework from our earlier treatment of machine learning applications. This analysis revealed evidence that many of the same factors were required to obtain positive results. In particular, successful developers spent significant time in formulating (and reformulating) their problem, in engineering the representation, and in collecting and manipulating the data. In this analysis, I added steps for manipulating the discovery algorithm (e.g., by altering parameters), as well as for filtering and interpreting the results. I also concluded that future systems should support for these activities explicitly, rather than forcing them to occur behind the scenes.

In summary, the initial lessons from scientific applications were quite similar to those from industrial applications of machine learning. This was encouraging since, in many cases, the discovered knowledge for these scientific domains often took a different form, and, in some examples, the discovery process involved abduction rather than induction. However, our own recent efforts in this arena have revealed significant differences between the two types of applications, and also suggest some quite different lessons from those already mentioned.

3. New Lessons for Scientific Discovery

Most work on computational scientific discovery has focused on modeling isolated phenomena in a constrained setting. This strategy has been as profitable for discovery research as for the natural sciences themselves. However, there is a growing interest in *systems science*, which aims to model the behavior of complex systems in terms of interacting parts. Another difference is that systems science typically has access only to observational or correlational data, rather than to data from controlled experiments. Our applied work has addressed scientific discovery in such contexts.

One effort concerns global models of the Earth ecosystem. Existing models match the observed behavior to some extent, but Earth scientists would like to improve their predictive ability. Our collaborators want to account for changes in the global production of carbon through vegetative growth, as well as the production and absorption of biogenic trace gases in the atmosphere. Hypothesized predictive variables include surface temperature, moisture levels, available sunlight, and properties of soil. Data for these variables come partly from measurements collected at ground stations and partly from estimations based on satellite images. The resulting model should explain observed differences in the behavioral variables as a function of the predictive variables.

The second application involves models of gene regulation in microorganisms. Biologists understand the basic mechanisms through which DNA produces biochemical behavior, but they have not yet mastered the regulatory networks that control the degree to which each gene is expressed. Our collaborators are concerned especially with determining how gene regulation controls the photosynthesis process in Cyanobacteria, an important type of phytoplankton. The available data come from experiments with wild type and mutant organisms grown in a chemostat and exposed to environmental stress such as bright light. cDNA microarrays measure the expression levels for culture samples at different points in time for approximately 300 genes believed to play a role in photosynthesis. The resulting model should explain both the observed expression levels and high-level behavior, such as the fact that Cyanobacteria bleaches when exposed to bright light.

Our experience with these two domains has suggested five lessons about factors that influence the success of applied computational work on scientific discovery. We will see that some of the claims run counter to prevailing wisdom in machine learning and data mining.

- *Lesson 1. Traditional machine learning notations are not easily communicated to scientists.*

Most sciences differ from industries in that they have a long history of representing their knowledge formally. Different fields have developed different notations that suit their needs, including structural models in organic chemistry, numeric equations in physics, and reaction networks in nuclear astrophysics. However, few scientific notations bear much relation to the formalisms popular in machine learning and data mining, which have mainly been invented by the learning researchers themselves.

We encountered this issue in both our scientific applications. We found that Earth scientists state their

models as sets of algebraic and difference equations, and that our collaborators were most comfortable with methods that produced knowledge in this format. We had some success with regression rules (Schwabacher & Langley, 2001), since they were familiar with piecewise linear models, but this approach does not support the theoretical terms that, as noted below, often appear in Earth science models. A similar problem arose with our microbiology colleagues, whose models of gene regulation took the form of qualitative causal diagrams. Here we were able to adapt techniques for inducing linear causal models (Langley, Shrager, & Saito, in press), but we had to introduce a number of additional features before our results made biological sense.

- *Lesson 2. Scientists often have initial models that should influence the discovery process.*

Science is an inherently incremental activity. Although scientists sometimes discover isolated laws from data, development of coherent models for complex systems takes place gradually and involves revising knowledge rather than discovering it from scratch. However, few techniques from machine learning and data mining support knowledge revision. Moreover, the rhetoric of these fields often discourages the incorporation of existing knowledge, since this would presumably bias the discovery system.

In our two application efforts, we found that our collaborators had initial models they hoped to improve upon, but not replace entirely, using computational techniques. For instance, the Earth scientists had an extensive model, stated as sets of equations, that partially accounted for the production of vegative carbon as a function of climate and other variables. Similarly, the microbiologists had a qualitative causal model that hypothesized which genes regulated others and how this activity influenced photosynthesis. Both groups felt their models might contain inaccurate assumptions and might omit important variables, but neither was interested in developing an entirely new model that did not build on the existing one.

- *Lesson 3. Scientific data are often rare and difficult to obtain rather than plentiful.*

The data mining community assures us that data are abundant in both business and scientific domains. Indeed, we have so much data that our primary goals should be developing algorithms that can process them efficiently and that can extract the last bit of knowledge from them. Such claims about the quantity of data may well hold for many business arenas, and there are some scientific contexts in which many observations are available, but the scientific applications we have examined (Langley, 2000) suggest this is the exception rather than the rule.

Our personal experience with scientific domains is consistent with this view. For ecosystem modeling, we had access to some 14,000 observations for a few variables that could be extracted from satellite images, but we had only 303 data points for the other variables, which we needed to determine the relations of interest. For modeling gene regulation, we had available thousands of measurements, since with DNA microarrays one can estimate expression levels for many genes at the same time. However, we had only 20 distinct samples measured over five time steps, which provided very few constraints on candidate models. Thus, the frequently heard rhetoric about the massive data sets generated by satellite imagery and microarray technology is misleading at best.

- *Lesson 4. Scientists want models that move beyond description to provide explanations of data.*

Descriptive laws play an essential role in science, most especially in the early stages of a field. However, as a discipline advances, its scientists desire increasingly to explain phenomena in terms of theoretical variables, entities, or processes. This holds especially for systems science, which attempts to account for observations in terms of interactions among hypothesized components. Explanations often make use of general concepts or relations that occur in different models, and thus rely on domain knowledge for their generation.

Explanatory accounts were important in both our applications. The ecosystem model we hoped to improve upon contained more theoretical terms than observable variables, many of them conceptually relevant to our Earth science collaborators. The initial photosynthesis model incorporated only one theoretical variable but many unobservable regulatory processes, some believed in strongly by our microbiology colleagues. Thus, we had to develop discovery algorithms that dealt directly with these explanatory terms, rather than drawing on the more common methods for finding descriptive knowledge that dominate the literatures on both scientific discovery and data mining.

- *Lesson 5. Scientists want computational assistance rather than automated discovery systems.*

Throughout their history, machine learning and data mining have emphasized automated methods for extracting knowledge from data. Business consumers of this discovered knowledge would be perfectly happy if such methods existed; they care about the effectiveness of the knowledge, not its source. In contrast, scientists'

careers revolve around making their own discoveries. Naturally, they have little desire to see the process automated, though many would welcome computational tools that would make their own data analyses more productive. This suggests the need for interactive discovery environments that assist the scientist in understanding data while letting him remain in control.[1]

In response, we are developing two such interactive systems to support our collaborations in Earth science and microbiology. The first will provide tools for specifying, viewing, editing, evaluating, and revising quantitative models composed of numeric equations, whereas the second will offer analogous abilities for qualitative causal models. We are endeavoring to make each environment as general as possible, but the types of scientific models that arise in these two domains seem different enough to justify separate systems. Whether scientists find these environments usable is an empirical question, but we are optimistic that, with user feedback, they can become useful aides for our collaborators and other domain experts.

4. Concluding Remarks

The above lessons point to some obvious conclusions about directions for additional work in computational approaches to scientific discovery. First, researchers should continue to focus on methods that generate knowledge in established scientific formalisms rather than those popular in the data mining movement. However, we need more concern with model revision as opposed to model construction, since this is more relevant to the incremental nature of science. We also need increased concern with methods that produce good models from small data sets rather than large ones, whether through incorporation of domain knowledge or statistical techniques for variance reduction, and with methods that generate explanatory models with theoretical terms to complement existing work on descriptive discovery. Finally, the field should expand its efforts on interactive environments for computational scientific discovery, rather than continuing its emphasis on automated methods.

These recommendations do not contradict earlier lessons drawn from applications of machine learning and discovery methods. Developers should still think carefully about how to formulate their problems, engineer the representations, manipulate their data and algorithms, and interpret their results. But they do suggest that, despite some impressive successes, we still require research that will produce a broader base of computational methods for the discovery of scientific knowledge. This research should address issues like revising existing models, handling sparse data, generating explanations, and supporting interaction with human scientists that appear crucial for the next generation of applications in this promising field.

[1] The data mining community has also developed such interactive environments, but they are designed for use by professional data miners, not those who use the knowledge.

Acknowledgements

The research reported in this paper was supported by NTT Communication Science Laboratories, Nippon Telegraph and Telephone Corporation, as well as by Grants NCC 2-5462 and NCC 2-5471 from NASA Ames Research Center.

References

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*, 37–54.

Langley, P. (1981). Data-driven discovery of physical laws. *Cognitive Science*, *5*, 31–54.

Langley, P. (1992). Areas of application for machine learning. *Proceedings of the Fifth International Symposium on Knowledge Engineering*. Seville, Spain.

Langley, P. (2000). The computational support of scientific discovery. *International Journal of Human-Computer Studies*, *53*, 393–410.

Langley, P., Shrager, J., & Saito, K. (in press). Computational discovery of communicable scientific knowledge. In L. Magnani, N. J. Nersessian, & C. Pizzi (Eds), *Logical and computational aspects of model-based reasoning*. Dordrecht: Kluwer Academic Publishers.

Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, *38*, November, 55–64.

Lenat, D. B. (1978). The ubiquity of discovery. *Artificial Intelligence*, *9*, 257–285.

Schwabacher, M., & Langley, P. (2001). Discovering communicable scientific knowledge from spatio-temporal data. *Proceedings of the Eighteenth International Conference on Machine Learning* (pp. 489–496). Williamstown, MA: Morgan Kaufmann.

Simon, H. A. (1966). Scientific discovery and the psychology of human problem solving. In R. G. Colodny (Ed.), *Mind and cosmos: Essays in contemporary science and philosophy*. University of Pittsburgh Press, Pittsburgh, PA.