Computational Discovery of Communicable Knowledge: Symposium Report

Sašo Džeroski[1] and Pat Langley[2]

[1] Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
Saso.Dzeroski@ijs.si, www-ai.ijs.si/SasoDzeroski/

[2] Institute for the Study of Learning and Expertise
2164 Staunton Court, Palo Alto, CA 94306 USA
langley@isle.org, www.isle.org/~langley/

Abstract. The *Symposium on Computational Discovery of Communicable Knowledge* was held from March 24 to 25, 2001, at Stanford University. Fifteen speakers reviewed recent advances in computational approaches to scientific discovery, focusing on their discovery tasks and the generated knowledge, rather than on the discovery algorithms themselves. Despite considerable variety in both tasks and methods, the talks were unified by a concern with the discovery of knowledge cast in formalisms used to communicate among scientists and engineers.

Computational research on scientific discovery has a long history within both artificial intelligence and cognitive science. Early efforts focused on reconstructing episodes from the history of science, but the past decade has seen similar techniques produce a variety of new scientific discoveries, many of them leading to publications in the relevant scientific literatures. Work in this paradigm has emphasized formalisms used to communicate among scientists, including numeric equations, structural models, and reaction pathways.

However, in recent years, research on data mining and knowledge discovery has produced another paradigm. Even when applied to scientific domains, this framework employs formalisms developed by artificial intelligence researchers themselves, such as decision trees, rule sets, and Bayesian networks. Although such methods can produce predictive models that are highly accurate, their outputs are not stated in terms familiar to scientists, and thus typically are not very communicable.

To highlight this distinction, Pat Langley organized the *Symposium on Computational Discovery of Communicable Knowledge*, which took place at Stanford University's Center for the Study of Language and Information on March 24 and 25, 2001. The meeting's aim was to bring together researchers who are pursuing computational approaches to the discovery of communicable knowledge and to review recent advances in this area. The primary focus was on discovery in scientific and engineering disciplines, where communication of knowledge is often a central concern.

Each of the 15 presentations emphasized the discovery tasks (the problem formulation and system input, including data and background knowledge) and the generated knowledge (the system output). Although artificial intelligence and machine learning traditionally focus on differences among algorithms, the meeting addressed the results of computational discovery at a more abstract level. In particular, it explored what methods for the computational discovery of communicable knowledge have in common, rather than the great diversity of methods used to that end.

The commonalities among methods for communicable knowledge discovery were summarized best by Raul Valdés-Pérez in a presentation titled *A Recipe for Designing Discovery Programs on Human Terms.* The key step in his recipe was identifying a set of possible solutions for some discovery task, as it is here that one can adopt a formalism that humans already use to represent knowledge. Valdés-Pérez viewed computational discovery as a problem-solving activity to which one can apply heuristic-search methods. He illustrated the recipe on the problem of discovering niche statements, i.e., properties of items that make them unique or distinctive in a given set of items.

The knowledge representation formalisms considered in the different presentations were diverse and ranged from equations through qualitative rules to reaction pathways. Most talks at the symposium fell within two broad categories. The first was concerned with equation discovery in either static systems or dynamic ones that change over time. The second addressed communicable knowledge discovery in biomedicine and in the related fields of biochemistry and molecular biology.

One formalism that scientists and engineers rely on heavily is equations. The task of equation discovery involves finding numeric or quantitative laws, expressed as one or more equations, from collections of measured numeric data. Most existing approaches to this problem deal with the discovery of algebraic equations, but recent work has also addressed the task of dynamic system identification, which involves discovering differential equations.

Takashi Washio from Osaka University presented a talk about *Conditions on Law Equations as Communicable Knowledge*, in which he discussed the conditions that equations must satisfy to be considered communicable. In addition to fitting the observed data, these include generic conditions and domain-dependent conditions. The former include objectiveness, generality, and reproducibility, as well as parsimony and mathematical admissibility with respect to unit dimensions and scale type constraints.

Kazumi Saito from Nippon Telegraph and Telephone and Mark Schwabacher from NASA Ames Research Center presented two related applications of computational equation discovery in the environmental sciences, both concerned with global models of the Earth ecosystem. Saito's talk on *Improving an Ecosystem Model Using Earth Science Data* addressed the task of revising an existing quantitative scientific model for predicting the net plant production of carbon in the light of new observations. Schwabacher's talk, *Discovering Communicable Scientific Knowledge from Spatio-Temporal Data in Earth Science*, dealt with the problem of predicting from climate variables the Normalized Difference Vegetation Index, a measure of greenness and a key component of the previous ecosystem model.

Four presentations discussed the task of dynamic system identification, which involves identifying the laws that govern behavior of systems with continuous variables that change over time. Such laws typically take the form of differential equations. Two of these talks described extensions to equation discovery methods to address system identification, whereas the other talks reported work that began with methods for system identification and incorporated artificial intelligence techniques that take advantage of domain knowledge.

Saso Džeroski from the Jožef Stefan Institute, in his talk on *Discovering Ordinary and Partial Differential Equations*, gave an overview of computational methods for discovering both ordinary and partial differential equations, the second of which describe dynamic systems that involve change over several dimensions (e.g., space and time). Ljupčo Todorovski, from the same research center, discussed an approach that uses domain knowledge to aid the discovery process in his talk, *Using Background Knowledge in Differential Equations Discovery.* He showed how knowledge in the form of context-free grammars can constrain discovery in the domain of population dynamics.

Reinhard Stolle, from Xerox PARC, spoke about *Communicable Models and System Identification.* He described a discovery system that handles both structural identification and parameter estimation by integrating qualitative reasoning, numerical simulation, geometric reasoning, constraint reasoning, abstraction, and other mechanisms. Matthew Easley from the University of Colorado, Boulder, reported extensions to Stolle's framework in his presentation, *Incorporating Engineering Formalisms into Automated Model Builders.* His approach relied on input-output modeling to plan experiments and using the resulting data, combined with knowledge at different levels of abstraction, to construct a differential equation model.

The talk by Feng Zhao from Xerox PARC, *Structure Discovery from Massive Spatial Data Sets*, described an approach to analyzing spatio-temporal data that relies on the notion of spatial aggregation. This mechanism generates summary descriptions of the raw data, which it characterizes at varying levels of detail. Zhao reported applications to several challenging problems, including the interpretation of weather data, optimization for distributed control, and the analysis of spatio-temporal diffusion-reaction patterns.

The rapid growth of biological databases, such as that for the human genome, has led to increased interest in applying computational discovery to biomedicine and related fields. Five presentations at the symposium focused on this general area. They covered a variety of discovery methods, including both propositional and first-order rule induction, genetic programming, theory revision, and abductive inference, with similar breadth in the biological discovery tasks to which they were applied.

Bruce Buchanan and Joseph Phillips, from the University of Pittsburgh, gave a presentation titled *Introducing Semantics into Machine Learning.* This focused on their incorporation of domain knowledge into rule-induction algorithms to let them find interesting and novel relations in medicine and science. They reviewed both syntactic and semantic constraints on the rule discovery process and showed that stronger forms of background knowledge increase the chances that discovered rules are understandable, interesting, and novel.

Stephen Muggleton from York University, in his talk *Knowledge Discovery in Biological and Chemical Domains*, described his application of first-order rule induction to predicting the structure of proteins, modeling the relations between a chemical's structure and its activity, and predicting a protein's function from its structure (e.g., identifying precursors of neuropeptides). Knowledge discovered in these efforts has appeared in journals for the respective scientific areas.

John Koza from Stanford University presented *Reverse Engineering and Automatic Synthesis of Metabolic Pathways from Observed Data.* His approach utilized genetic programming to carry out search through a space of metabolic pathway models, with search directed by the models' abilities to fit time-series data on observed chemical concentrations. The target model included an internal feedback loop, a bifurcation point, and an accumulation point, suggesting the method can handle complex metabolic processes.

The presentation by Pat Langley, from the Institute for the Study of Learning and Expertise, addressed *Knowledge and Data in Computational Biological Discovery.* He reported an approach that used data on gene expressions to revise a model of photosynthetic regulation in Cyanobacteria previously developed by plant biologists. The result was an improved model with altered processes that better explains the expression levels observed over time. The ultimate goal is an interactive system to support human biologists in their discovery activities.

Marc Weeber from the U.S. National Library of Medicine reported on a quite different approach in his talk on *Literature-based Discovery in Biomedicine.* The main idea relies on utilizing bibliographic databases to uncover indirect but plausible connections between disconnected bodies of scientific knowledge. He illustrated this method with a successful example of finding potentially new therapeutic applications for an existing drug, thalidomide.

Sakir Kocabas, from Istanbul Technical University, talked about *The Role of Completeness in Particle Physics Discoveries*, which dealt with a completely different domain. He described a computational model of historical discovery in particle physics that relies on two main criteria – consistency and completeness – to postulate new quantum properties, determine those properties' values, propose new particles, and predict reactions among particles. Kocabas' system successfully simulated an extended period in the history of this field, including discovery of the neutrino and postulation of the baryon number.

At the close of the symposium, Lorenzo Magnani from the University of Pavia commented on the presentations from a philosophical viewpoint. In particular, he cast the various efforts in terms of his general framework for abduction, which incorporates different types of explanatory reasoning. The gathering also spent time honoring the memory of Herbert Simon and Jan Żytkow, both of whom played seminal roles in the field of computational scientific discovery.

Further information on the symposium is available at the World Wide Web page `http://www.isle.org/symposia/comdisc.html`. This includes information about the speakers, abstracts of the presentations, and pointers to publications related to their talks. Slides from the presentations can be found at the Web page `http://math.nist.gov/~JDevaney/CommKnow/`. Sašo Džeroski and Ljupčo Todorovski are currently editing a book based on the talks given at the symposium. Information on the book will appear at the symposium page and the first author's Web page as it becomes available.

Acknowledgements

The Symposium on Computational Discovery of Communicable Knowledge was supported by Grant NAG 2-1335 from NASA Ames Research Center and by the Nippon Telegraph and Telephone Corporation.

References

Bradley, E., Easley, M., & Stolle, R. (in press). Reasoning about nonlinear system identification. *Artificial Intelligence*.

Kocabas, S., & Langley, P. (in press). An integrated framework for extended discovery in particle physics. *Proceedings of the Fourth International Conference on Discovery Science*. Washington, D.C.: Springer.

Koza, J. R., Mydlowec, W., Lanza, G., Yu, J., & Keane, M. A. (2001). Reverse engineering and automatic synthesis of metabolic pathways from observed data using genetic programming. *Pacific Symposium on Biocomputing*, *6*, 434–445.

Lee, Y., Buchanan, B. G., & Aronis, J. M. (1998). Knowledge-based learning in exploratory science: Learning rules to predict rodent carcinogenicity. *Machine Learning*, *30*, 217–240.

Muggleton, S. (1999). Scientific knowledge discovery using inductive logic programming. *Communications of the ACM*, *42*, 42–46.

Saito, K., Langley, P., Grenager, T., Potter, C., Torregrosa, A., & Klooster, S. A. (in press). Computational revision of quantitative scientific models. *Proceedings of the Fourth International Conference on Discovery Science*. Washington, D.C.: Springer.

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.

Shrager, J., Langley, P., & Pohorille, A. (2001). *Guiding revision of regulatory models with expression data*. Unpublished manuscript, Institute for the Study of Learning and Expertise, Palo Alto, CA.

Todorovski, L., & Džeroski, S. (2000). Discovering the structure of partial differential equations from example behavior. *Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 991–998). San Francisco: Morgan Kaufmann.

Valdés-Pérez, R. E. (1999). Principles of human-computer collaboration for knowledge discovery in science. *Artificial Intelligence*, *107*, 335–346.

Washio, T., Motoda, H., & Niwa, Y. (2000). Enhancing the plausibility of law equation discovery. *Proceedings of the Seventeenth International Conference on Machine Learning* (pp. 1127–1134). San Francisco: Morgan Kaufmann.

Yip, K., & Zhao, F. (1996). Spatial aggregation: Theory and applications. *Journal of Artificial Intelligence Research*, *5*, 1–26.