Computational Scientific Discovery and Inductive Databases

Sašo Džeroski
Department of Intelligent Systems
Jožef Stefan Institute
Jamova 39, SI-1000 Ljubljana, Slovenia

Computational scientific discovery (Langley et al. 1987; Shrager and Langley 1990) is concerned with applying computational methods to automate scientific activities. Early research on computational discovery (Langley et al. 1987) focussed on reconstructing episodes from the history of science by modeling the scientific activities and processes that led to the scientist's insight. Recent efforts in this area (for overviews see Langley 2000; Džeroski and Todorovski 2003) have focussed on individual scientific activities (such as formulating quantitative laws). Much of the work in computational scientific discovery has put emphasis on formalisms used to communicate among scientists, including numeric equations, structural models, and reaction pathways.

Over the last decade, we have developed a number of approaches to discovering quantitative laws in the form of equations or equation discovery. We have considered algebraic, ordinary differential (ODEs) and partial differential (PDEs) equations (Džeroski and Todorovski 1993; Todorovski and Džeroski 1997; Todorovski et al. 2000). The approaches developed have been applied to a number of practical modeling problems, mainly in the area of ecology (Todorovski et al. 1998; Džeroski et al. 1999). We have devoted special attention to the use of various forms of domain knowledge: we use declarative bias (Todorovski and Džeroski 1997) and background knowledge (Džeroski and Todorovski 2001), and also address the problem of revising theories that consist of quantitative laws (Todorovski and Džeroski 2001). The use of context-free grammars to define the space of equations considered (Todorovski and Džeroski 1997) allows us to treat all three forms of domain knowledge in a uniform way.

Using domain knowledge allows for a realistic approach to learning in difficult domains. Rather than trying to solve a difficult problem by starting from scratch, one can use existing domain knowledge in addition to collected observations (examples) and build upon it. Different types of domain knowledge can be taken into account, such as concepts already in common use (background knowledge), intuitions about the form of the target theory (declarative bias) and existing theories (theory revision). In this context, one can trade-off between the quantity and quality of observations and domain knowledge: high quantities of quality data may suffice to generate a good theory even with no domain knowledge, while smaller quantities of (lower quality) data may suffice if relevant domain knowledge is available.

Inductive databases (Imielinski and Mannila 1996) embody a database perspective on knowledge discovery, where knowledge discovery processes are considered as query processes. In addition to normal data, inductive databases contain patterns (either materialized or defined as views). Data mining operations looking for patterns are viewed as queries posed to the inductive database. In addition to patterns (which are of local nature), models (which are of global nature) can also be considered.

A general formulation of data mining (Mannila and Toivonen 1997) involves the specification of a language of patterns and a set of constraints that a pattern has to satisfy with respect to a given database. The constraints that a pattern has to satisfy can be divided in two parts: language constraints and evaluation constraints. The first only concern the pattern itself, the second concern the validity of the pattern with respect to a database. Constraints thus play a central role in data mining and constraint-based data mining is now a recognized research topic (Bayardo 2002).

Different types of patterns have been considered in data mining, including frequent itemsets, episodes, Datalog queries, and graphs. Designing inductive databases for these types of patterns involves the design of inductive query languages and solvers for the queries in these languages. For each type of pattern, or pattern domain, a specific solver is designed, following the philosophy of constraint logic programming (De Raedt 2002).

To bring equation discovery and inductive databases together, we consider inductive databases on the pattern domain of equations. When designing a query language for a given pattern domain, the language and evaluation constraints that are to be considered need to be specified. Language-wise, one might consider polynomial equations and search for sub-polynomials of a given polynomial which have a high correlation coefficient with a dependent variable on the data at hand. Other evaluation measures can be considered, such as maximum absolute error, mean squared error, etc. Similarity language constraints can also be used: one can search for equations that are as similar as possible to a given equation and have a correlation coefficient above a certain threshold on a given dataset. The latter is essentially a theory revision problem in equation discovery (Todorovski and Džeroski 2001).

Inductive databases and constraint-based data mining open the door to more intensive use of domain knowledge in data mining and thus bring it closer to computational scientific discovery. In the pattern domain of equations, inductive queries that would allow for a combination of data-driven modeling and modeling from first principles would be possible. These would include queries that perform theory revision on models consisting of sets of equations.

In a given application domain, an inductive database would contain not only data about the domain but also models or model components (patterns). Alternative models of different aspects of the domain can be stored. Inductive queries would generate models from data only, from the data and model components, or revise models in light of the data. Computational scientific discovery can then be supported through inductive query sessions, which allows for a much more active role of the user as compared to traditional data mining.

Note that this calls for more effort on documenting and storing the results of the modeling process. Within he inductive database paradigm, this could provide strong support for further modeling activities (computational scientific discovery). While it is usual to have datasets in databases, models are typically not stored in databases. Efforts to create databases of models in different fields of science, such as the ECOBAS initiative in the field of ecological modeling, should thus be encouraged and supported. Hopefully this would facilitate computational scientific discovery by synthesizing new data and existing knowledge into new knowledge through the interaction of scientists with inductive databases.

References

[1] Bayardo, R., editor (2002). Constraint-based data mining. Special issue. *SIGKDD Explorations*.

[2] De Raedt, L. (2002). Data mining as constraint logic programming. In *Computational Logic: From Logic Programming into the Future (In honor of Bob Kowalski)*. Springer, Berlin.

[3] Džeroski, S., & Todorovski, L., editors (2003). *Computational Discovery of Communicable Knowledge*. Springer, Berlin. Forthcoming.

[4] Džeroski, S., & Todorovski, L. (2002). Encoding and using domain knowledge on population dynamics in equation discovery. In L. Magnani, N. J. Nersessian, and C. Pizzi, (editors), *Logical and Computational Aspects of Model-Based Reasoning*. Kluwer, Dordrecht.

[5] Džeroski, S., Todorovski, L., Bratko, I., Kompare, B., & Križman, V. (1999). Equation discovery with ecological applications. In A.H. Fielding, editor, *Machine Learning Methods for Ecological Applications* (pp. 185–207). Kluwer, Dordrecht.

[6] Džeroski, S., & Todorovski, L. (1993). Discovering dynamics. In *Proc. 10th International Conference on Machine Learning* (pp. 97–103). Morgan Kaufmann, San Mateo, CA.

[7] Imielinski, T., and Mannila, H. (1996). A database perspective on knowledge discovery. *Communications of the ACM*, 39(11): 58–64.

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

[9] Langley, P., Simon, H.A., Bradshaw, G.L., & Zytkow, J. (1987). *Scientific Discovery*. MIT Press, Cambridge, MA.

[10] Mannila, H., and Toivonen, H. (1997). Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3): 241–258.

[11] Shrager, J., & Langley, P., editors (1990). *Computational Models of Scientific Discovery and Theory Formation*. Morgan Kaufmann, San Mateo, CA.

[12] Todorovski, L., & Džeroski, S. (2001). Theory revision in equation discovery. In *Proc. 4th International Conference on Discovery Science* (pp. 390–400). Springer, Berlin.

[13] Todorovski, L., Džeroski, S., Srinivasan, A., Whiteley, J., & Gavaghan, D. (2000). Discovering the structure of partial differential equations from example behavior. In *Proc. 17th International Conference on Machine Learning* (pp. 991–998). Morgan Kaufmann, San Francisco, CA.

[14] Todorovski, L., Džeroski, S., & Kompare, B. (1998). Modeling and prediction of phytoplankton growth with equation discovery. *Ecological Modelling* 113: 71–81.

[15] Todorovski, L., & Džeroski, S. (1997). Declarative bias in equation discovery. In *Proc. 14th International Conference on Machine Learning* (pp. 376–384). Morgan Kaufmann, San Francisco, CA.