Inductive Databases and Queries for Computational Scientific Discovery

Sašo Džeroski
Jožef Stefan Institute, Jamova 39, 1000 Ljubljana, Slovenia
Saso.Dzeroski@ijs.si, http://www-ai.ijs.si/SasoDzeroski/

Abstract

Computational scientific discovery aims to develop computer systems that automate or facilitate the various activities that humans perform in the process of scientific discovery. We propose to use the inductive databases (IDBs) framework to support computational scientific discovery. IDBs can store both data and models and support the interactive nature of knowledge discovery, as well as other aspects important for scientific discovery, such as induction under constraints and theory revision.

1. Inductive Databases and Inductive Queries

Inductive databases (IDBs, Imielinski and Mannila 1996, De Raedt 2002a) are an emerging research area at the intersection of data mining and databases. In addition to normal data, inductive databases contain patterns (either materialized or defined as views). Besides patterns (which are of local nature), models (which are of global nature) can also be considered. In the IDB framework, patterns become "first-class citizens" and can be stored and manipulated just like data in ordinary databases.

Inductive databases embody a database perspective on knowledge discovery, where knowledge discovery processes become query sessions. Ordinary queries can be used to access and manipulate data, while inductive queries (IQs) can be used to generate (mine), manipulate, and apply patterns. KDD thus becomes an extended querying process (Imielinski and Mannila 1996) in which both the data and the patterns that hold (are valid) in the data are queried. IDB research thus aims at replacing the traditional KDD process model, where steps like pre-processing, data cleaning, and model construction follow each other in succession, by a simpler model in which all data pre-processing operations, data mining operations, as well as post-processing operations are queries to an inductive database and can be interleaved in many different ways.

Given an inductive database that contains data and patterns, several different types of queries can be posed. Data retrieval queries use only the data and their results are also data: no pattern is involved in the query. In IDBs, we can also have cross-over queries that combine patterns and data in order to obtain new data, e.g., apply a predictive model to a dataset to obtain predictions for a target property. In processing patterns, the patterns are queried without access to the data: this is what is usually done in the post-processing stages of data mining. Data mining queries use the data and their results are patterns: new patterns are generated from the data and this corresponds to the traditional data mining step. When we talk about inductive queries, we most often mean data mining queries.

A general statement of the problem 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. The latter can be divided in two parts: language constraints and evaluation constraints. The first part only concerns the pattern itself, while the second part concerns the validity of the pattern with respect to a given database. Constraints thus play a central role in data mining and constraint-based data mining is now a recognized research topic (Bayardo 2002). The use of constraints enables more efficient induction and focusses the search for patterns on patterns likely to be of interest to the end user.

In the context of inductive databases, inductive queries consist of constraints. Inductive queries can involve language constraints (e.g., find association rules with item A in the head) and evaluation constraints, formed by using evaluation functions. The latter express the validity of a pattern on a given dataset. We can use these to form evaluation constraints (e.g., find all item sets with support above a threshold) or optimization constraints (e.g., find the 10 association rules with highest confidence).

Different types of data and 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, i.e., constraint-based data mining algorithms. Of central importance is the issue of defining the primitive constraints that can be applied for the chosen data and pattern

types, that can be used to compose inductive queries. For each pattern domain (type of data, type of pattern, and primitive constraints), a specific solver is designed, following the philosophy of constraint logic programming (De Raedt 2002b).

The IDB framework is an appealing approach towards a theory for data mining, because it employs declarative queries instead of ad-hoc procedural constructs. As such, it holds the promise of facilitating the formulation of an "algebra" for data mining, along the lines of Codd's relational algebra for databases (Calders et al. 2006b, Johnson et al. 2000). The IDB framework is also appealing for data mining applications, as it supports the entire KDD process (Boulicaut et al. 1999). In inductive query languages, the results of one (inductive) query can be used as input for another: nontrivial multi-step KDD scenarios can be thus supported in IDBs, rather than just single data mining operations.

2. Computational Scientific Discovery

Research on computational scientific discovery aims to develop computer systems which produce results that, if a human scientist did the same, we would refer to as discoveries. Of course, if we hope to develop computational methods for scientific discovery, we must be more specific about the nature of such discoveries and how they relate to the broader context of the scientific enterprise.

The term science refers both to scientific knowledge and the process of acquiring such knowledge. It includes any systematic field of study that relates to observed phenomena (as opposed to mathematics) and that involves claims which can be tested empirically (as opposed to philosophy). Science is perhaps the most complex human intellectual activity, which makes it difficult to describe. Shrager and Langley (1990) analyze it in terms of the knowledge structures that scientists consider and the processes or activities they use to transform them. Basic knowledge structures that arise in science include observations, laws, and theories, and related activities include data collection, law formation, and theory construction.

There are two primary reasons why we might want to study scientific discovery from a computational perspective:

- to understand how humans perform this intriguing activity, which belongs to the realm of cognitive science; and
- to automate or assist in facets of the scientific process, which belongs to the realm of artificial intelligence.

Science is a highly complex intellectual endeavor, and discovery is arguably the most creative part of the scientific process. Thus, efforts to automate it completely would rightfully be judged as audacious, but, as Simon (1966) noted, one can view many kinds of scientific discovery as examples of problem solving through heuristic search. Most research in automating scientific discovery has focused on small, well-defined tasks that are amenable to such treatment and that allow measurable progress.

Traditional accounts of science (Klemke et al. 1998) focus on the individual, who supposedly observes nature, hypothesizes laws or theories, and tests them against new observations. Most computational models of scientific discovery share this concern with individual behavior. However, science is almost always a collective activity that is conducted by interacting members of a scientific community. The most fundamental demonstration of this fact is the emphasize placed on communicating one's findings to other researchers in journal articles and conference presentations.

This emphasis on exchanging results makes it essential that scientific knowledge be communicable. We will not attempt to define this term, but it seems clear that contributions are more communicable if they are cast in established formalisms and if they make contact with concepts that are familiar to most researchers in the respective field of study. In the remainder of this section, we first examine more closely the scientific method and its relation to scientific discovery. After this, we discuss the components of scientific behavior, that is, the knowledge structures that arise in science and the processes that manipulate them.

2.1. The Scientific Method and Scientific Discovery

The Merriam-Webster Dictionary (2003) defines science as: "a) knowledge or a system of knowledge covering general truths or the operation of general laws, especially as obtained and tested through the scientific method, and b) such knowledge or such a system of knowledge concerned with the physical world and its phenomena". The scientific method, in turn, is defined as the "principles and procedures for the systematic pursuit of knowledge involving the recognition and formulation of a problem, the collection of data through observation and experiment, and the formulation and testing of hypotheses".

While there is consensus that science revolves around knowledge, there are different views in the philosophy of science (Klemke et al. 1998, Achinstein 2004) about the nature of its content. The 'causal realism' position is that scientific knowledge is ontological, in that it identifies entities in the world, their causal powers, and the mechanisms through which they exert influence. In contrast, 'constructive empiricism' posits that, scientific theories are objective, testable, and predictive. We believe both frameworks are correct, describing different facets of the truth.

The *scientific method* (Gower 1996), dedicated to the systematic pursuit of reliable knowledge, incorporates a

number of steps. First we ask some meaningful question or identify a significant problem. We next gather information relevant to the question, which might include existing scientific knowledge or new observations. We then formulate a hypothesis that could plausibly answer the question.

Next we must test this proposal by making observations and determining whether they are consistent with the hypothesis' predictions. When observations are consistent with the hypothesis, they lend it support and we may consider publishing it. If other scientists can reproduce our results, then the community comes to consider it as reliable knowledge. In contrast, if the observations are inconsistent, we should reject the hypothesis and either abandon it or, more typically, modify it, at which point the testing process continues. Hypotheses can take many different forms, including taxonomies, empirical laws, and explanatory theories, but all of them can be evaluated by comparing their implications to observed phenomena.

Most analyses of the scientific method come from philosophers of science, who have focused mainly on the evaluation of hypotheses and largely ignored their generation and revision. Unfortunately, what we refer to as discovery resides in just these activities. Thus, although there is a large literature on normative methods for making predictions from hypotheses, checking their consistency, and determining whether they are valid, there are remarkably few treatments of their production. Some (e.g., Popper 1959) have even suggested that rational accounts of the discovery process are impossible. A few philosophers (e.g., Darden 2006, Hanson 1958, Lakatos 1976) have gone against this trend and made important contributions to the topic, but most efforts have come from artificial intelligence and cognitive science.

Briefly, scientific discovery is the process by which a scientist creates or finds some hitherto unknown knowledge, such as a class of objects, an empirical law, or an explanatory theory. The knowledge in question may also be referred to as a scientific discovery. An important aspect of many knowledge structures, such as laws and theories, is their generality, in that they apply to many specific situations or many specific observations. We maintain that generality is an essential feature of a meaningful discovery, as will become apparent in the next section when we discuss types of scientific knowledge.

A defining aspect of discovery is that the knowledge should be new and previously unknown. Naturally, one might ask 'new to whom?'. We take the position that the knowledge should be unknown to the scientist in question with respect to the observations and background knowledge available to him when he made the discovery. This means that two or more scientists can make the same discovery independently, sometimes years apart, which has indeed often happened throughout the history of science.

2.2. The Elements of Scientific Behavior

To describe scientific behavior, we follow Shrager and Langley (1990) and use as basic components knowledge structures and the activities that transform them. The former represent the raw materials and products of science, while the latter concern the process of producing scientific knowledge. The account below mostly follows the earlier treatise, but the definitions of several knowledge structures and activities have changed, reflecting improvements in our understanding over the past 15 years.

Scientific Knowledge Structures. Science is largely about understanding the world in which we live. To this end, we gather information about the world. Observation is the primary means of collecting this information, and observations are the primary input to the process of scientific discovery.

Observations (or data) represent recordings of the environment made by sensors or measuring instruments. Typically, the state of the environment varies over time or under different conditions, and one makes recordings for these different states, where what constitutes a state depends on the object of scientific study. We will refer to each of these recordings as an observation.

We can identify three important types of scientific knowledge – taxonomies, laws, and theories – that constitute the major products of the scientific enterprise. The creation of new taxonomies, laws, and theories, as well as revising and improving existing ones, make up the bulk of scientific discovery, making them some of the key activities in science.

Taxonomies define or describe concepts for a domain, along with specialization relations among them. *Laws* summarize relations among observed variables, objects, or events. *Theories* are statements about the structures or processes that arise in the environment.

Note that all three kinds of knowledge are important and present in the body of scientific knowledge. Different types of knowledge are generated at different stages in the development of a scientific discipline. Taxonomies are generated early in a field's history, providing the basic concepts for the discipline. After this, scientists formulate empirical laws based on their observations. Eventually, these laws give rise to theories that provide a deeper understanding of the structures and processes studied in the discipline.

A knowledge structure that a scientist has proposed, but that has not yet been tested with respect to observations, is termed an hypothesis. Note that taxonomies, laws, and theories can all have this status. As mentioned earlier, hypotheses must be evaluated to determine whether they are consistent with observations (and background knowledge). If it is consistent, we say that a hypothesis has been corroborated and it comes to be viewed as scientific knowledge.

If an hypothesis is inconsistent with the evidence, then we either reject or modify it, giving rise to a new hypothesis that is further tested and evaluated.

Background knowledge is knowledge about the environment separate from that specifically under study. It typically includes previously generated scientific knowledge in the domain of study. Such knowledge differs from theories or laws at the hypothesis stage, in that the scientist regards it with relative certainty rather than as the subject of active evaluation. Scientific knowledge begins its life cycle as a hypothesis which (if corroborated) becomes background knowledge.

Besides the basic data and knowledge types considered above, several other types of structures play important roles in science. These include models, predictions, and explanations. These occupy an intermediate position, as they are derived from laws and theories and, as such, they are not primary products of the scientific process.

Models are special cases of laws and theories that apply to particular situations in the environment and only hold under certain environmental conditions. *Predictions* represent expectations about the behavior of the environment under specific conditions. *Explanations* are narratives that connect a theory to a law (or a model to a prediction) by a chain of inferences appropriate to the field.

Scientific Activities. Scientific processes and activities are concerned with generating and manipulating scientific data and knowledge structures. Here we consider the processes and activities in the same order as we discussed the structures that they generate in the previous subsection.

The process of observation involves inspecting the environmental setting by focusing an instrument, sometimes simply the agent's senses, on that setting. The result is a concrete description of the setting, expressed in terms from the agent's taxonomy and guided by the model of the setting. Since one can observe many things in any given situation, the observer must select some aspects to record and some to ignore.

As we have noted, scientific discovery is concerned with generating scientific knowledge in the form of taxonomies, laws and theories. These can be generated directly from observations (and possibly background knowledge), but, quite often, scientists modify an existing taxonomy, law, or theory to take into account anomalous observations.

Taxonomy formation (and revision) involves the organization of observations into classes and subclasses, along with the definition of those classes. *Inductive law formation (and revision)* involves the generation of empirical laws that cover observed data. *Theory formation (and revision)* stands in the same relation to empirical laws as does law formation to data. Given one or more laws, this activity generates a theory from which one can derive the laws for a given model by explanation.

While some scientific activities revolve around inductive reasoning, others instead rely on deduction. Scientists typically derive predictions from laws or models, and sometimes they even deduce laws from theoretical principles.

In contrast to inductive law discovery from observations, *deductive law formation* starts with a theory and uses an explanatory framework to deduce both a law and an explanation of how that law follows from the theory. The *prediction* process takes a law, along with a particular setting, and produces a prediction about what one will observe in the setting. The analogous process of *postdiction* takes place in cases where the scientist must account for existing observations. The process of *explanation* connects a theory to a law (or a law to a prediction) by specifying the deductive reasoning that derives the law from the theory.

To assess the validity of theories or laws, scientists compare their predictions or postdictions with observations. This produces either consistent results or anomalies, which may serve to stimulate further theory or law formation or revision. This process is called *evaluation* and generally follows experimentation and observation.

Experimentation involves experimental design and manipulation. *Experimental design* specifies settings in which the scientist will collect measurements. Typically, he varies selected aspects of the environment (the independent variables) to determine their effect on other aspects (the dependent variables). He then constructs a physical setting (this is called *manipulation*) that corresponds to the desired environmental conditions and carries out the experiment.

Observation will typically follow or will be interleaved with systematic experimentation, in which case we call it active observation. However, there are fields and phenomena where experimental control is difficult, and sometimes impossible. In such cases the scientist can still collect data to test his hypotheses through passive observation.

2.3. Computational Discovery and Data Mining

Computational scientific discovery focuses on applying computational methods to automate scientific activities, such as finding laws from observational data. It has emerged from the view that science is a problem-solving activity and that problem solving can be cast as search through a space of possible solutions. Early research on computational discovery within the fields of artificial intelligence and cognitive science focused on reconstructing episodes from the history of science. This typically included identifying data and knowledge available at the time and implementing a computer program that models the scientific activities and processes that led to the scientists insight.

Recent efforts in this area have focused on individual scientific activities (such as formulating quantitative laws) and have produced a number of new scientific discoveries,

many of them leading to publications in the relevant scientific literatures. These discoveries include qualitative laws of metallic behavior, quantitative conjectures in graph theory, and temporal laws of ecological behavior. Work in this paradigm has emphasized formalisms used to communicate among scientists, including numeric equations, structural models, and reaction pathways.

Research on data mining and knowledge discovery, however, has produced another paradigm, concerned with finding patterns (regularities) in data. Even when applied to scientific domains, such as astronomy, biology, and chemistry, this framework employs formalisms developed by artificial intelligence researchers themselves, such as decision trees, rule sets, and Bayesian networks. Mining scientific data focuses on building highly predictive models from large datasets, rather than producing knowledge in any standard scientific notation.

The differences between the two paradigms are emphasized if we take a look at the lessons learned that have emerged from work in scientific domains (Langley 2002): (1) The output of a discovery system should be communicated easily to domain scientists. (2) Discovery systems should take advantage of background knowledge to constrain their search. (3) Computational methods for scientific discovery should be able to infer knowledge from small data sets. (4) Discovery systems should produce models that move beyond description to provide explanations of data. (5) Computational discovery systems should support interaction with domain scientists.

3. Towards Inductive Scientific Databases

We propose to use the inductive databases framework to support the process of scientific discovery. This would allow the development of an interactive environment that supports the process of establishing models of real world systems from measurements and observations in various scientific disciplines. The framework would allow us to integrate computational scientific discovery and data mining methods for the induction of new models and revision of existing ones from data on one side with tools for data storage and visualization, as well as storage, application/simulation, and evaluation of models on the other.

While many data analysis methods have been developed that are capable of inducing models from data, most of these focus on the task of building the model from data only, without taking into account previously developed models. In practice, scientists and engineers build models of complex systems gradually by revising existing (i.e., previously developed) models, rather than building them from scratch. In this sense, model storage is necessary in a proper modeling environment. Furthermore, scientists and engineers rarely need complete automation of the modeling process; they rather need an interactive environment that supports model development as well as simulation, evaluation, and comparison of different models.

The above requirements are well met in the framework of inductive (scientific) databases. In addition to the data, collected through scientific experiments and observations, inductive databases would support the storage of models, where the models could be developed by scientists or automated modelling methods. Similarly, in addition to usual data retrieval queries that operate on data only, inductive databases provide queries that operate on both data and models. One type of queries can be used to combine models and data: An example query would apply an existing model to newly collected data and evaluate its performance.

Inductive databases allow inductive queries that are used to induce models from data. These queries can be composed of constraints on the space of candidate models that are based on background knowledge from the domain of use and constraints on the accuracy or validity of the model on data. The result of an inductive query is a set of models that can be stored in the database and can be re-used later in further modeling efforts. These can include model revision, where an existing model is used as a starting point for induction and is revised using recently collected data.

An inductive database would typically support different data mining operations, including clustering and predictive modelling. As such it would provide support for the different scientific activities: Clustering could be used for taxonomy formation/revision, while predictive modelling (e.g., equation discovery) In sum, scientific inductive databases can provide scientists and engineers with an interactive environment for assisting the process of modeling. Scientists can interact with the database using a series of queries. These queries allow simple data storage and manipulation of data and models, but also more complex operations with models, such as model induction from data, as well as simulation, evaluation and revision of stored models. Thus, scientific IDBs are a worthwile goal to pursue.

Acknowledgements. I would like to thank Ljupčo Todorovski, Pat Langley and the members of the IQ project (IST FET 516169, Inductive Queries for Mining Patterns and Models) for supporting this work.

References. The references that appear in the text can be found in the two articles listed below.

1. S. Džeroski. Towards a general framework for data mining. In S. Džeroski and J. Struyf, editors, *Knowledge Discovery in Inductive Databases, 5th International Workshop: Revised Selected and Invited Papers*, pages 259–300. Springer, Berlin, 2007.

2. S. Džeroski, P. Langley, and L. Todorovski. Computational Discovery of Scientific Knowledge. In S. Džeroski and L. Todorovski, editors, *Computational Discovery of Scientific Knowledge*, pages 1–14. Springer, Berlin, 2007.