Lecture Notes in Artificial Intelligence 4660

Edited by J. G. Carbonell and J. Siekmann

Subseries of Lecture Notes in Computer Science

Sašo Džeroski   Ljupčo Todorovski (Eds.)

Computational Discovery of Scientific Knowledge

Introduction, Techniques, and Applications in Environmental and Life Sciences

Springer

Volume Editors

Sašo Džeroski
Department of Knowledge Technologies
Jožef Stefan Institute
Ljubljana, Slovenia
E-mail: Saso.Dzeroski@ijs.si

Ljupčo Todorovski
Department of Knowledge technologies
Jožef Stefan Institute
Jjubliana, Slovenia
E-mail: Ljupco.Todorovski@ijs.si

Library of Congress Control Number: 2007931366

CR Subject Classification (1998): I.2, H.2.8, H.2-3, J.2-3, F.1

LNCS Sublibrary: SL 7 – Artificial Intelligence

ISSN 0302-9743
ISBN-10 3-540-73919-X Springer Berlin Heidelberg New York
ISBN-13 978-3-540-73919-7 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springer.com

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 12099696 06/3180 5 4 3 2 1 0

Preface

Advances in technology have enabled the collection of data from scientific observations, simulations, and experiments at an ever-increasing pace. For the scientist and engineer to benefit from these enhanced data-collecting capabilities, it is becoming clear that semi-automated data analysis techniques must be applied to find the useful information in the data. Techniques from both data mining and computational discovery can be used to that end.

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 scientist's 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. The discoveries made using computational tools 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.

However, in recent years, research on data mining and knowledge discovery has produced another paradigm. Data mining is 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. Although such methods can produce predictive models that are highly accurate, their outputs are not cast in terms familiar to scientists, and thus typically are not very communicable.

Mining scientific data focuses on building highly predictive models, rather than producing knowledge in any standard scientific notation. In contrast, much of the work in computational scientific discovery has put a strong emphasis on formalisms used by scientists to communicate scientific knowledge, such as numeric equations, structural models, and reaction pathways. In this sense, computational scientific discovery is complementary to mining scientific data.

The book provides an introduction to computational approaches to the discovery of communicable scientific knowledge and gives an overview of recent advances in this area. The primary focus is on discovery in scientific and engineering disciplines, where communication of knowledge is often a central concern.

This volume has its origins in the symposium "Computational Discovery of Communicable Knowledge," organized by Pat Langley, held March 24-25, 2001 at Stanford University. A detailed report on the symposium can be found in the Proceedings of the DS-2001 Conference (S. Džeroski, and P. Langley. Computational discovery of communicable knowledge: Symposium report. In *Proceedings of the Fourth International Conference on Discovery Science*, pages 45-49. Springer, Berlin, 2001). Many of the presentations from that symposium have a corresponding chapter in the book. To achieve a more representative coverage of recent research in computational discovery, we have invited a number of additional contributions as well.

The book is organized as follows. The first chapter introduces the field of computational scientific discovery and provides a brief overview thereof. It also provides a more detailed overview of the contents of the book. The majority of the contributed chapters fall within two broad categories, which correspond to Parts I and II of the book, respectively. The first describes a number of computational discovery methods for system identification and automated modelling, while the second discusses a number of methods for computational discovery developed for biomedical and bioinformatics applications.

In the first part of the book, the focus is on establishing models of dynamic systems, i.e., systems that change their state over time. The models are mostly based on equations, in particular ordinary differential equations that represent a standard formalism for modelling dynamic systems in many engineering and scientific areas. This is in contrast to the bulk of previous research on equation discovery, which focuses on algebraic equations. Topics covered in this part include a reasoning tool for nonlinear system identification, the use of different forms of domain knowledge when inducing models of dynamic systems (including the use of existing models for theory revision, partial knowledge of the target model, knowledge on basic processes, and knowledge on measurement scales of the system variables), and applications to Earth sciences.

While the first part of the book focuses on a class of methods and covers a variety of scientific fields and areas, the second focuses on the field of biomedicine. The first three chapters are in line with the first part of the book and continue with the theme of model formation, but use representation formalisms specific to the biomedical field, such as chemical reaction networks and genetic pathways. The last two chapters present approaches to forming scientific hypotheses by connecting disconnected scientific literatures on the same topic. This part also includes a chapter on using learning in logic for predicting gene function.

We would like to conclude with some words of thanks. Pat Langley organized the symposium that motivated this volume and encouraged us to edit it. More importantly, he has pioneered research on computational scientific discovery and provided unrelenting support to our research in this area. We would also like to thank the participants of the symposium. Finally, we would like to thank all the contributors to this volume for their excellent contributions and their patience with the editors.

May 2007

Sašo Džeroski
Ljupčo Todorovski

Table of Contents

I Equation Discovery and Dynamic Systems Identification

II Computational Scientific Discovery in Biomedicine