Inductive Databases and Constraint-Based Data Mining

Sašo Džeroski • Bart Goethals • Panče Panov
Editors

Inductive Databases and Constraint-Based Data Mining

Springer

Editors
Sašo Džeroski
Jožef Stefan Institute
Dept. of Knowledge Technologies
Jamova cesta 39
SI-1000 Ljubljana
Slovenia
Saso.Dzeroski@ijs.si

Panče Panov
Jožef Stefan Institute
Dept. of Knowledge Technologies
Jamova cesta 39
SI-1000 Ljubljana
Slovenia
Pance.Panov@ijs.si

Bart Goethals
University of Antwerp
Mathematics and Computer Science Dept.
Middelheimlaan 1
B-2020 Antwerpen
Belgium
Bart.Goethals@ua.ac.be

ISBN 978-1-4419-7737-3 e-ISBN 978-1-4419-7738-0
DOI 10.1007/978-1-4419-7738-0
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010938297

Preface

This book is about inductive databases and constraint-based data mining, emerging research topics lying at the intersection of data mining and database research. The aim of the book as to provide an overview of the state-of- the art in this novel and exciting research area. Of special interest are the recent methods for constraint-based mining of global models for prediction and clustering, the unification of pattern mining approaches through constraint programming, the clarification of the relationship between mining local patterns and global models, and the proposed integrative frameworks and approaches for inducive databases. On the application side, applications to practically relevant problems from bioinformatics are presented.

Inductive databases (IDBs) represent a database view on data mining and knowledge discovery. IDBs contain not only data, but also generalizations (patterns and models) valid in the data. In an IDB, ordinary queries can be used to access and manipulate data, while inductive queries can be used to generate (mine), manipulate, and apply patterns and models. In the IDB framework, patterns and models become ”first-class citizens” and KDD becomes an extended querying process in which both the data and the patterns/models that hold in the data are queried.

The IDB framework is appealing as a general framework for data mining, because it employs declarative queries instead of ad-hoc procedural constructs. As declarative queries are often formulated using constraints, inductive querying is closely related to constraint-based data mining. The IDB framework is also appealing for data mining applications, as it supports the entire KDD process, i.e., nontrivial multi-step KDD scenarios, rather than just individual data mining operations.

The interconnected ideas of inductive databases and constraint-based mining have the potential to radically change the theory and practice of data mining and knowledge discovery. The book provides a broad and unifying perspective on the field of data mining in general and inductive databases in particular. The 18 chapters in this state-of-the-art survey volume were selected to present a broad overview of the latest results in the field.

Unique content presented in the book includes constraint-based mining of global models for prediction and clustering, including predictive models for structured out-

puts and methods for bi-clustering; integration of mining local (frequent) patterns and global models (for prediction and clustering); constraint-based mining through constraint programming; integrative IDB approaches at the system and framework level; and applications to relevant problems that attract strong interest in the bioinformatics area. We hope that the volume will increase in relevance with time, as we witness the increasing trends to store patterns and models (produced by humans or learned from data) in addition to data, as well as retrieve, manipulate, and combine them with data.

This book contains sixteen chapters presenting recent research on the topics of inductive databases and queries, as well as constraint-based data, conducted within the project IQ (Inductive Queries for mining patterns and models), funded by the EU under contract number IST-2004-516169. It also contains two chapters on related topics by researchers coming from outside the project (Siebes and Puspitaningrum; Wicker et al.)

This book is divided into four parts. The first part describes the foundations of and frameworks for inductive databases and constraint-based data mining. The second part presents a variety of techniques for constraint-based data mining or inductive querying. The third part presents integration approaches to inductive databases. Finally, the fourth part is devoted to applications of inductive querying and constraint-based mining techniques in the area of bioinformatics.

The first, introductory, part of the book contains four chapters. Džeroski first introduces the topics of inductive databases and constraint-based data mining and gives a brief overview of the area, with a focus on the recent developments within the IQ project. Panov et al. then present a deep ontology of data mining. Blockeel et al. next present a practical comparative study of existing data-mining/inductive query languages. Finally, De Raedt et al. are concerned with mining under composite constraints, i.e., answering inductive queries that are Boolean combinations of primitive constraints.

The second part contains six chapters presenting constraint-based mining techniques. Besson et al. present a unified view on itemset mining under constraints within the context of constraint programming. Bringmann et al. then present a number of techniques for integrating the mining of (frequent) patterns and classification models. Struyf and Džeroski next discuss constrained induction of predictive clustering trees. Bingham then gives an overview of techniques for finding segmentations of sequences, some of these being able to handle constraints. Cerf et al. discuss constrained mining of cross-graph cliques in dynamic networks. Finally, De Raedt et al. introduce ProbLog, a probabilistic relational formalism, and discuss inductive querying in this formalism.

The third part contains four chapters discussing integration approaches to inductive databases. In the Mining Views approach (Blockeel et al.), the user can query the collection of all possible patterns as if they were stored in traditional relational tables. Wicker et al. present SINDBAD, a prototype of an inductive database system that aims to support the complete knowledge discovery process. Siebes and Puspitaningrum discuss the integration of inductive and ordinary queries (relational algebra). Finally, Vanschoren and Blockeel present experiment databases.

The fourth part of the book, contains four chapters dealing with applications in the area of bioinformatics (and chemoinformatics). Vens et al. describe the use of predictive clustering trees for predicting gene function. Slavkov and Džeroski describe several applications of predictive clustering trees for the analysis of gene expression data. Rigotti et al. describe how to use mining of frequent patterns on strings to discover putative transcription factor binding sites in gene promoter sequences. Finally, King et al. discuss a very ambitious application scenario for inductive querying in the context of a robot scientist for drug design.

The content of the book is described in more detail in the last two sections of the introductory chapter by Džeroski.

We would like to conclude with a word of thanks to those that helped bring this volume to life: This includes (but is not limited to) the contributing authors, the referees who reviewed the contributions, the members of the IQ project and the various funding agencies. A more complete listing of acknowledgements is given in the Acknowledgements section of the book.

September 2010

Sašo Džeroski
Bart Goethals
Panče Panov

Acknowledgements

Heartfelt thanks to all the people and institutions that made this volume possible and helped bring it to life.

First and foremost, we would like to thank the contributing authors. They did a great job, some of them at short notice. Also, most of them showed extraordinary patience with the editors.

We would then like to thank the reviewers of the contributed chapters, whose names are listed in a separate section. Each chapter was reviewed by at least two (on average three) referees. The comments they provided greatly helped in improving the quality of the contributions.

Most of the research presented in this volume was conducted within the project IQ (Inductive Queries for mining patterns and models). We would like to thank everybody that contributed to the success of the project: This includes the members of the project, both the contributing authors and the broader research teams at each of the six participating institutions, the project reviewers and the EU officials handling the project. The IQ project was funded by the European Comission of the EU within FP6-IST, FET branch, under contract number FP6-IST-2004-516169.

In addition, we want to acknowledge the following funding agencies:

- Sašo Džeroski is currently supported by the Slovenian Research Agency (through the research program *Knowledge Technologies* under grant P2-0103 and the research projects *Advanced machine learning methods for automated modelling of dynamic systems* under grant J2-0734 and *Data Mining for Integrative Data Analysis in Systems Biology* under grant J2-2285) and the European Commission (through the FP7 project PHAGOSYS *Systems biology of phagosome formation and maturation - modulation by intracellular pathogens* under grant number HEALTH-F4-2008-223451). He is also supported by the Centre of Excellence for Integrated Approaches in Chemistry and Biology of Proteins (operation no. OP13.1.1.2.02.0005 financed by the European Regional Development Fund (85%) and the Slovenian Ministry of Higher Education, Science and Technology (15%)), as well as the Jozef Stefan International Postgraduate School in Ljubljana.

- Bart Goethals wishes to acknowledge the support of FWO-Flanders through the project ”Foundations for inductive databases”.
- Panče Panov is supported by the Slovenian Research Agency through the research projects *Advanced machine learning methods for automated modelling of dynamic systems* (under grant J2-0734) and *Data Mining for Integrative Data Analysis in Systems Biology* (under grant J2-2285).

Finally, many thanks to our Springer editors, Jennifer Maurer and Melissa Fearon, for all the support and encouragement.

September 2010

Sašo Džeroski
Bart Goethals
Panče Panov

List of Reviewers

Hendrik Blockeel	Katholieke Universiteit Leuven, Belgium
Marko Bohanec	Jožef Stefan Institute, Slovenia
Jean-Francois Boulicaut	University of Lyon, INSA Lyon, France
Mario Boley	University of Bonn and Fraunhofer IAIS, Germany
Toon Calders	Eindhoven Technical University, Netherlands
Vineet Chaoji	Yahoo! Labs, Bangalore, India
Amanda Clare	Aberystwyth University, United Kingdom
James Cussens	University of York, United Kingdom
Tomaž Curk	University of Ljubljana, Ljubljana, Slovenia
Ian Davidson	University of California - Davis, USA
Luc Dehaspe	Katholieke Universiteit Leuven, Belgium
Luc De Raedt	Katholieke Universiteit Leuven, Belgium
Jeroen De Knijf	University of Antwerp, Belgium
Tijl De Bie	University of Bristol, United Kingdom
Sašo Džeroski	Jožef Stefan Institute, Slovenia
Elisa Fromont	University of Jean Monnet, France
Gemma C. Garriga	University of Paris VI, France
Christophe Giraud-Carrier	Brigham Young University, USA
Jiawei Han	University of Illinois at Urbana-Champaign, USA
Hannes Heikinheimo	Aalto Universit, Finland
Cristoph Hema	In Silico Toxicology, Switzerland
Andreas Karwath	Albert-Ludwigs-Universitat, Germany
Jörg-Uwe Kietz	University of Zurich, Switzerland
Arno Knobbe	University of Leiden, Netherlands
Petra Kralj Novak	Jožef Stefan Institute, Slovenia
Stefan Kramer	Technische Universität München, Germany
Rosa Meo	University of Torino, Italy
Pauli Miettinen	Max-Planck-Institut für Informatik, Germany
Siegfried Nijssen	Katholieke Universiteit Leuven, Belgium
Markus Ojala	Aalto University, Finland
Themis Palpanas	University of Trento, Italy

Panče Panov	Jožef Stefan Institute, Ljubljana, Slovenia
Juho Rousu	University of Helsinki, Finland
Nikolaj Tatti	University of Antwerp, Belgium
Grigorios Tsoumakas	Aristotle University of Thessaloniki, Greece
Giorgio Valentini	University of Milano, Italy
Jan Van den Bussche	Universiteit Hasselt, Belgium
Jilles Vreeken	University of Utrecht, Netherlands
Kiri Wagstaff	California Institute of Technology, USA
Joerg Wicker	Technische Universität München, Germany
Gerson Zaverucha	Federal University of Rio de Janeiro, Brazil
Albrecht Zimmermann	Katholieke Universiteit Leuven, Belgium
Bernard Ženko	Jožef Stefan Institute, Slovenia

Contents