

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

© Springer Science+Business Media, LLC 2010

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

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 is 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 inductive 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 mining, 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 Commission 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 University, Finland
Cristoph Hema	In Silico Toxicology, Switzerland
Andreas Karwath	Albert-Ludwigs-Universität, 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

Part I Introduction

1	Inductive Databases and Constraint-based Data Mining: Introduction and Overview	3
	Sašo Džeroski	
1.1	Inductive Databases	3
1.2	Constraint-based Data Mining	7
1.3	Types of Constraints	9
1.4	Functions Used in Constraints	12
1.5	KDD Scenarios	14
1.6	A Brief Review of Literature Resources	15
1.7	The IQ (Inductive Queries for Mining Patterns and Models) Project	17
1.8	What's in this Book	22
2	Representing Entities in the OntoDM Data Mining Ontology	27
	Panče Panov, Larisa N. Soldatova, and Sašo Džeroski	
2.1	Introduction	27
2.2	Design Principles for the OntoDM ontology	29
2.3	OntoDM Structure and Implementation	33
2.4	Identification of Data Mining Entities	38
2.5	Representing Data Mining Entities in OntoDM	46
2.6	Related Work	52
2.7	Conclusion	54
3	A Practical Comparative Study Of Data Mining Query Languages	59
	Hendrik Blockeel, Toon Calders, Élisabeth Fromont, Bart Goethals, Adriana Prado, and Céline Robardet	
3.1	Introduction	60
3.2	Data Mining Tasks	61
3.3	Comparison of Data Mining Query Languages	62
3.4	Summary of the Results	74
3.5	Conclusions	76

4	A Theory of Inductive Query Answering	79
	Luc De Raedt, Manfred Jaeger, Sau Dan Lee, and Heikki Mannila	
4.1	Introduction	80
4.2	Boolean Inductive Queries	81
4.3	Generalized Version Spaces	88
4.4	Query Decomposition	90
4.5	Normal Forms	98
4.6	Conclusions	100
 Part II Constraint-based Mining: Selected Techniques		
5	Generalizing Itemset Mining in a Constraint Programming Setting	107
	Jérémy Besson, Jean-François Boulicaut, Tias Guns, and Siegfried Nijssen	
5.1	Introduction	107
5.2	General Concepts	109
5.3	Specialized Approaches	111
5.4	A Generalized Algorithm	114
5.5	A Dedicated Solver	116
5.6	Using Constraint Programming Systems	120
5.7	Conclusions	124
6	From Local Patterns to Classification Models	127
	Björn Bringmann, Siegfried Nijssen, and Albrecht Zimmermann	
6.1	Introduction	127
6.2	Preliminaries	131
6.3	Correlated Patterns	132
6.4	Finding Pattern Sets	137
6.5	Direct Predictions from Patterns	142
6.6	Integrated Pattern Mining	146
6.7	Conclusions	152
7	Constrained Predictive Clustering	155
	Jan Struyf and Sašo Džeroski	
7.1	Introduction	155
7.2	Predictive Clustering Trees	156
7.3	Constrained Predictive Clustering Trees and Constraint Types	161
7.4	A Search Space of (Predictive) Clustering Trees	165
7.5	Algorithms for Enforcing Constraints	167
7.6	Conclusion	173
8	Finding Segmentations of Sequences	177
	Ella Bingham	
8.1	Introduction	177
8.2	Efficient Algorithms for Segmentation	182
8.3	Dimensionality Reduction	183

8.4	Recurrent Models	185
8.5	Unimodal Segmentation	188
8.6	Rearranging the Input Data Points	189
8.7	Aggregate Segmentation	190
8.8	Evaluating the Quality of a Segmentation: Randomization	191
8.9	Model Selection by BIC and Cross-validation	193
8.10	Bursty Sequences	193
8.11	Conclusion	194
9	Mining Constrained Cross-Graph Cliques in Dynamic Networks . . .	199
	Loïc Cerf, Bao Tran Nhan Nguyen, and Jean-François Boulicaut	
9.1	Introduction	199
9.2	Problem Setting	201
9.3	DATA-PEELER	205
9.4	Extracting δ -Contiguous Closed 3-Sets	208
9.5	Constraining the Enumeration to Extract 3-Cliques	212
9.6	Experimental Results	217
9.7	Related Work	224
9.8	Conclusion	226
10	Probabilistic Inductive Querying Using ProbLog	229
	Luc De Raedt, Angelika Kimmig, Bernd Gutmann, Kristian Kersting, Vítor Santos Costa, and Hannu Toivonen	
10.1	Introduction	229
10.2	ProbLog: Probabilistic Prolog	233
10.3	Probabilistic Inference	234
10.4	Implementation	238
10.5	Probabilistic Explanation Based Learning	243
10.6	Local Pattern Mining	245
10.7	Theory Compression	249
10.8	Parameter Estimation	252
10.9	Application	255
10.10	Related Work in Statistical Relational Learning	258
10.11	Conclusions	259

Part III Inductive Databases: Integration Approaches

11	Inductive Querying with Virtual Mining Views	265
	Hendrik Blockeel, Toon Calders, Élisabeth Fromont, Bart Goethals, Adriana Prado, and Céline Robardet	
11.1	Introduction	266
11.2	The Mining Views Framework	267
11.3	An Illustrative Scenario	277
11.4	Conclusions and Future Work	285

12	SINDBAD and SiQL: Overview, Applications and Future Developments	289
	Jörg Wicker, Lothar Richter, and Stefan Kramer	
12.1	Introduction	289
12.2	SiQL	291
12.3	Example Applications	296
12.4	A Web Service Interface for SINDBAD	303
12.5	Future Developments	305
12.6	Conclusion	307
13	Patterns on Queries	311
	Arno Siebes and Diyah Puspitaningrum	
13.1	Introduction	311
13.2	Preliminaries	313
13.3	Frequent Item Set Mining	319
13.4	Transforming KRIMP	323
13.5	Comparing the two Approaches	331
13.6	Conclusions and Prospects for Further Research	333
14	Experiment Databases	335
	Joaquin Vanschoren and Hendrik Blockeel	
14.1	Introduction	336
14.2	Motivation	337
14.3	Related Work	341
14.4	A Pilot Experiment Database	343
14.5	Learning from the Past	350
14.6	Conclusions	358
Part IV Applications		
15	Predicting Gene Function using Predictive Clustering Trees	365
	Celine Vens, Leander Schietgat, Jan Struyf, Hendrik Blockeel, Dragi Koccev, and Sašo Džeroski	
15.1	Introduction	366
15.2	Related Work	367
15.3	Predictive Clustering Tree Approaches for HMC	369
15.4	Evaluation Measure	374
15.5	Datasets	375
15.6	Comparison of Clus-HMC/SC/HSC	378
15.7	Comparison of (Ensembles of) CLUS-HMC to State-of-the-art Methods	380
15.8	Conclusions	384

16	Analyzing Gene Expression Data with Predictive Clustering Trees . .	389
	Ivica Slavkov and Sašo Džeroski	
16.1	Introduction	389
16.2	Datasets	391
16.3	Predicting Multiple Clinical Parameters	392
16.4	Evaluating Gene Importance with Ensembles of PCTs	394
16.5	Constrained Clustering of Gene Expression Data	397
16.6	Clustering gene expression time series data	400
16.7	Conclusions	403
17	Using a Solver Over the String Pattern Domain to Analyze Gene Promoter Sequences	407
	Christophe Rigotti, Ieva Mitašiūnaitė, Jérémy Besson, Laurène Meyniel, Jean-François Boulicaut, and Olivier Gandrillon	
17.1	Introduction	407
17.2	A Promoter Sequence Analysis Scenario	409
17.3	The <i>Marguerite</i> Solver	412
17.4	Tuning the Extraction Parameters	413
17.5	An Objective Interestingness Measure	415
17.6	Execution of the Scenario	418
17.7	Conclusion	422
18	Inductive Queries for a Drug Designing Robot Scientist	425
	Ross D. King, Amanda Schierz, Amanda Clare, Jem Rowland, Andrew Sparkes, Siegfried Nijssen, and Jan Ramon	
18.1	Introduction	425
18.2	The Robot Scientist Eve	427
18.3	Representations of Molecular Data	430
18.4	Selecting Compounds for a Drug Screening Library	444
18.5	Active learning	446
18.6	Conclusions	448
	Appendix	452
	Author index	455