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

http://www.springer.com/978-1-4419-7737-3

Inductive Databases and Constraint-Based Data Mining
Dzeroski, S.; Goethals, B.; Panov, P. (Eds.)
2010, XVII, 456 p., Hardcover
ISBN: 978-1-4419-7737-3