Data Mining and Knowledge Discovery Handbook

Second Edition

Oded Maimon · Lior Rokach
Editors

Data Mining and Knowledge Discovery Handbook

Second Edition

Springer

Editors
Prof. Oded Maimon
Tel Aviv University
Dept. Industrial Engineering
69978 Ramat Aviv
Israel
maimon@eng.tau.ac.il

Dr. Lior Rokach
Ben-Gurion University of the Negev
Dept. Information Systems
Engineering
84105 Beer-Sheva
Israel
liorrk@bgu.ac.il

ISBN 978-0-387-09822-7 e-ISBN 978-0-387-09823-4
DOI 10.1007/978-0-387-09823-4
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010931143

To my family
– Oded Maimon

To my parents Ines and Avraham
– Lior Rokach

Preface

Knowledge Discovery demonstrates intelligent computing at its best, and is the most desirable and interesting end-product of Information Technology. To be able to discover and to extract knowledge from data is a task that many researchers and practitioners are endeavoring to accomplish. There is a lot of hidden knowledge waiting to be discovered – this is the challenge created by today's abundance of data.

Knowledge Discovery in Databases (KDD) is the process of identifying valid, novel, useful, and understandable patterns from large datasets. Data Mining (DM) is the mathematical core of the KDD process, involving the inferring algorithms that explore the data, develop mathematical models and discover significant patterns (implicit or explicit) -which are the essence of useful knowledge. This detailed guide book covers in a succinct and orderly manner the methods one needs to master in order to pursue this complex and fascinating area.

Given the fast growing interest in the field, it is not surprising that a variety of methods are now available to researchers and practitioners. This handbook aims to organize all major concepts, theories, methodologies, trends, challenges and applications of Data Mining into a coherent and unified repository. This handbook provides researchers, scholars, students and professionals with a comprehensive, yet concise source of reference to Data Mining (and additional selected references for further studies).

The handbook consists of eight parts, each part consists of several chapters. The first seven parts present a complete description of different methods used throughout the KDD process. Each part describes the classic methods, as well as the extensions and novel methods developed recently. Along with the algorithmic description of each method, the reader is provided with an explanation of the circumstances in which this method is applicable, and the consequences and trade-offs incurred by using that method. The last part surveys software and tools available today.

The first part describes preprocessing methods, such as cleansing, dimension reduction, and discretization. The second part covers supervised methods, such as regression, decision trees, Bayesian networks, rule induction and support vector machines. The third part discusses unsupervised methods, such as clustering, association rules, link analysis and visualization. The fourth part covers soft computing

methods and their application to Data Mining. This part includes chapters about fuzzy logic, neural networks, and evolutionary algorithms.

Parts five and six present supporting and advanced methods in Data Mining, such as statistical methods for Data Mining, logics for Data Mining, DM query languages, text mining, web mining, causal discovery, ensemble methods, and a great deal more. Part seven provides an in-depth description of Data Mining applications in various interdisciplinary industries, such as finance, marketing, medicine, biology, engineering, telecommunications, software, and security.

The motivation: Over the past few years we have presented and written several scientific papers and research books in this fascinating field. We have also developed successful methods for very large complex applications in industry, which are in operation in several enterprises. Thus, we have first hand experience in the needs of the KDD/DM community in research and practice. This handbook evolved from these experiences.

The first edition of the handbook, which was published five years ago, was extremely well received by the data mining research and development communities. The field of data mining has evolved in several aspects since the first edition. Advances occurred in areas, such as Multimedia Data Mining, Data Stream Mining, Spatio-temporal Data Mining, Sequences Analysis, Swarm Intelligence, Multi-label classification and privacy in data mining. In addition new applications and software tools become available. We received many requests to include the new advances in the field in a second edition of the handbook. About half of the book is new in this edition. This second edition aims to refresh the previous material in the fundamental areas, and to present new findings in the field. The new advances occurred mainly in three dimensions: new methods, new applications and new data types, which can be handled by new and modified advanced data mining methods.

We would like to thank all authors for their valuable contributions. We would like to express our special thanks to Susan Lagerstrom-Fife of Springer for working closely with us during the production of this book.

Tel-Aviv, Israel *Oded Maimon*
Beer-Sheva, Israel *Lior Rokach*

April 2010

Contents

List of Contributors

Maria M. Abad
Software Engineering Department,
University of Granada, Spain

Ajith Abraham
Center of Excellence for Quantifiable Quality of Service
Norwegian University of Science and Technology,
Trondheim, Norway

Bruno Agard
Département de Mathmatiques et de Génie Industriel,
École Polytechnique de Montréal, Canada

Daniel Barbara
Department of Information and Software Engineering,
George Mason University, USA

Christopher D. Barko
Customer Analytics, Inc.

Irad Ben-Gal
Department of Industrial Engineering,
Tel-Aviv University, Israel

Moty Ben-Dov
School of Computing Science,
MDX University, London, UK.

Yoav Benjamini
Department of Statistics, Sackler Faculty for Exact Sciences
Tel Aviv University, Israel

Richard A. Berk
Department of Statistics
UCLA, USA

Jean-Francois Boulicaut
INSA Lyon, France

Pavel Brazdil
Faculty of Economics,
University of Porto, Portugal

Abel Browarnik
Department of Industrial Engineering,
Tel-Aviv University, Israel

Christopher J.C. Burges
Microsoft Research, USA

Cory J. Butz
Department of Computer Science,
University of Regina, Canada

Nitesh V. Chawla
Department of Computer Science and Engineering,
University of Notre Dame, USA

Ping Chen
Department of Computer and Mathematics Science,
University of Houston-Downtown, USA

Barak Chizi
Department of Industrial Engineering,
Tel-Aviv University, Israel

Shahar Cohen
Department of Industrial Engineering,
Tel-Aviv University, Israel

Antonio Congiusta
Dipartimento di Elettronica, Informatica e Sistemistica,
University of Calabria, Italy

Gautam Das
Computer Science and Engineering Department,
University of Texas, Arlington, USA

Swagatam Das
Department of Electronics and Telecommunication Engineering,
Jadavpur University, India.

Steve Donoho
Mantas, Inc. USA

Sašo Džeroski
Jožef Stefan Institute, Slovenia

Ronen Feldman
Department of Mathematics and Computer Science,
Bar-Ilan university, Israel

Eibe Frank
Department of Computer Science,
University of Waikato, New Zealand

Alex A. Freitas
Computing Laboratory,
University of Kent, UK

Johannes Fürnkranz
TU Darmstadt, Knowledge Engineering Group, Germany

Mohamed Medhat Gaber
Centre for Distributed Systems and Software Engineering
Monash University

Pierre Geurts
Department of Electrical Engineering and Computer Science,
University of Liège, Belgium

Christophe Giraud-Carrier
Department of Computer Science,
Brigham Young University, Utah, USA

Paolo Giudici
Faculty of Economics,
University of Pavia, Italy

Bart Goethals
Departement of Mathemati1cs and Computer Science,
University of Antwerp, Belgium

Jerzy W. Grzymala-Busse
Department of Electrical Engineering and Computer Science,
University of Kansas, USA

Witold J. Grzymala-Busse
FilterLogix Inc., USA

Dimitrios Gunopulos
Department of Computer Science and Engineering,
University of California at Riverside, USA

Petr Hájek
Institute of Computer Science,
Academy of Sciences of the Czech Republic

Maria Halkidi
Department of Computer Science and Engineering,
University of California at Riverside, USA

Mark Hall
Department of Computer Science,
University of Waikato, New Zealand

Howard J. Hamilton
Department of Computer Science,
University of Regina, Canada

Jiawei Han
Department of Computer Science,
University of Illinois, Urbana Champaign, USA

Geoffrey Holmes
Department of Computer Science,
University of Waikato, New Zealand

Frank Höppner
Department of Information Systems,
University of Applied Sciences Braunschweig/Wolfenbüttel, Germany

Yan Huang
Department of Computer Science,
University of Minnesota, USA

Sushil Jajodia
Center for Secure Information Systems,
George Mason University, USA

Ioannis Katakis
Dept. of Informatics, Aristotle University of Thessaloniki, 54124 Greece

Eamonn Keogh
Computer Science and Engineering Department,
University of California at Riverside, USA

Richard Kirkby
Department of Computer Science,
University of Waikato, New Zealand

Slava Kisilevich
University of Konstanz, Germany

Boris Kovalerchuk
Department of Computer Science,
Central Washington University, USA

Shonali Krishnaswamy
Centre for Distributed Systems and Software Engineering
Monash University

Andrew Kusiak
Department of Mechanical and Industrial Engineering,
The University of Iowa, USA

Nada Lavrač
Jožef Stefan Institute, Ljubljana, Slovenia
Nova Gorica Polytechnic, Nova Gorica, Slovenia

Moshe Leshno
Faculty of Management and Sackler Faculty of Medicine,
Tel Aviv University, Israel

Nissan Levin
Q-Ware Software Company, Israel

Tao Li
School of Computer Science,
Florida International University, USA

Churn-Jung Liau
Institute of Information Science,
Academia Sinica, Taiwan

Jessica Lin
Department of Computer Science and Engineering,
University of California at Riverside, USA

Tsau Y. Lin
Department of Computer Science,
San Jose State University, USA

Sheng Ma
Machine Learning for Systems
IBM T.J. Watson Research Center, USA

Oded Maimon
Department of Industrial Engineering,
Tel-Aviv University, Israel

Jonathan I. Maletic
Department of Computer Science,
Kent State University, USA

Florian Mansmann
University of Konstanz, Germany

Andrian Marcus
Department of Computer Science,
Wayne State University, USA

Cyrille Masson
INSA Lyon, France

Steve Moyle
Computing Laboratory,
Oxford University, UK

Mirco Nanni
University of Pisa,
Italy

Hamid R. Nemati
Information Systems and Operations Management Department
Bryan School of Business and Economics
The University of North Carolina at Greensboro, USA

Mitsunori Ogihara
Computer Science Department,
University of Rochester, USA

Nora Oikonomakou
Department of Informatics,
Athens University of Economics and Business (AUEB), Greece

Bernhard Pfahringer
Department of Computer Science,
University of Waikato, New Zealand

Marco F. Ramoni
Departments of Pediatrics and Medicine
Harvard University, USA

Chotirat Ann Ratanamahatana
Department of Computer Science and Engineering,
University of California at Riverside, USA

Yoram Reich
Center for Design Research,
Stanford University, Stanford, CA, USA

Salvatore Rinzivillo
Institute of Information Science and Technologies,
Italy

Lior Rokach
Department of Information Systems Engineering
Ben-Gurion University of the Negev, Israel

Noa Ruschin Rimini
Department of Industrial Engineering,
Tel-Aviv University, Israel

Sigal Sahar
Department of Computer Science,
Tel-Aviv University, Israel

Paola Sebastiani
Department of Biostatistics,
Boston University, USA

Richard S. Segall
Arkansas State University,
Department of Computer and Info. Tech., Jonesboro, AR
72467-0130,USA

Shashi Shekhar
Institute of Technology,
University of Minnesota, USA

Armin Shmilovici
Department of Information Systems Engineering,
Ben-Gurion University of the Negev, Israel

Gautam B. Singh
Department of Computer Science and Engineering,
Center for Bioinformatics, Oakland University, USA

Anoop Singhal
Center for Secure Information Systems,
George Mason University, USA

Domenico Talia
Dipartimento di Elettronica, Informatica e Sistemistica,
University of Calabria, Italy

Kurt Thearling
Vertex Business Services
Richardson, Texas, USA

Vicenç Torra
Institut d'Investigació en Intel·ligència Artificial, Spain

Paolo Trunfio
Dipartimento di Elettronica, Informatica e Sistemistica,
University of Calabria, Italy

Grigorios Tsoumakas
Dept. of Informatics, Aristotle University of Thessaloniki, 54124 Greece

Jiong Yang
Department of Electronic Engineering and Computer Science,
Case Western Reserve University, USA

Ying Yang
School of Computer Science and Software Engineering,
Monash University, Melbourne, Australia

Hong Yao
Department of Computer Science,
University of Regina, Canada

Philip S. Yu
IBM T. J. Watson Research Center,
USA

Michalis Vazirgiannis
Department of Informatics,
Athens University of Economics and Business, Greece

Ricardo Vilalta
Department of Computer Science,
University of Houston, USA

Evgenii Vityaev
Institute of Mathematics,
Russian Academy of Sciences, Russia

Michail Vlachos
IBM T. J. Watson Research Center,
USA

Ioannis Vlahavas
Dept. of Informatics, Aristotle University of Thessaloniki, 54124 Greece

Haixun Wang
IBM T. J. Watson Research Center,
USA

Wei Wang
Department of Computer Science,
University of North Carolina at Chapel Hill, USA

Geoffrey I. Webb
Faculty of Information Technology,
Monash University, Australia

Gary M. Weiss
Department of Computer and Information Science,
Fordham University, USA

Ian H. Witten
Department of Computer Science,
University of Waikato, New Zealand

Jacob Zahavi
The Wharton School,
University of Pennsylvania, USA

Arkady Zaslavsky
Centre for Distributed Systems and Software Engineering
Monash University

Peter G. Zhang
Department of Managerial Sciences,
Georgia State University, USA

Pusheng Zhang
Department of Computer Science and Engineering,
University of Minnesota, USA

Qingyu Zhang
Arkansas State University, Department of
Computer and Info. Tech.,
Jonesboro, AR 72467-0130,USA

Ruofei Zhang
Yahoo!, Inc. Sunnyvale, CA 94089

Zhongfei (Mark) Zhang
SUNY Binghamton, NY 13902-6000

Blaž Zupan
Faculty of Computer and Information Science,
University of Ljubljana, Slovenia