

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

© Springer Science+Business Media, LLC 2005, 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)

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
Beer-Sheva, Israel

Oded Maimon
Lior Rokach

April 2010

Contents

1 Introduction to Knowledge Discovery and Data Mining	
<i>Oded Maimon, Lior Rokach</i>	1

Part I Preprocessing Methods

2 Data Cleansing: A Prelude to Knowledge Discovery	
<i>Jonathan I. Maletic, Andrian Marcus</i>	19
3 Handling Missing Attribute Values	
<i>Jerzy W. Grzymala-Busse, Witold J. Grzymala-Busse</i>	33
4 Geometric Methods for Feature Extraction and Dimensional Reduction - A Guided Tour	
<i>Christopher J.C. Burges</i>	53
5 Dimension Reduction and Feature Selection	
<i>Barak Chizi, Oded Maimon</i>	83
6 Discretization Methods	
<i>Ying Yang, Geoffrey I. Webb, Xindong Wu</i>	101
7 Outlier Detection	
<i>Irad Ben-Gal</i>	117

Part II Supervised Methods

8 Supervised Learning	
<i>Lior Rokach, Oded Maimon</i>	133
9 Classification Trees	
<i>Lior Rokach, Oded Maimon</i>	149

10 Bayesian Networks

Paola Sebastiani, Maria M. Abad, Marco F. Ramoni 175

11 Data Mining within a Regression Framework

Richard A. Berk 209

12 Support Vector Machines

Armin Shmilovici 231

13 Rule Induction

Jerzy W. Grzymala-Busse 249

Part III Unsupervised Methods

14 A survey of Clustering Algorithms

Lior Rokach 269

15 Association Rules

Frank Höppner 299

16 Frequent Set Mining

Bart Goethals 321

17 Constraint-based Data Mining

Jean-Francois Boulicaut, Baptiste Jeudy 339

18 Link Analysis

Steve Donoho 355

Part IV Soft Computing Methods

19 A Review of Evolutionary Algorithms for Data Mining

Alex A. Freitas 371

20 A Review of Reinforcement Learning Methods

Oded Maimon, Shahar Cohen 401

21 Neural Networks For Data Mining

G. Peter Zhang 419

22 Granular Computing and Rough Sets - An Incremental Development

Tsau Young ('T. Y.') Lin, Churn-Jung Liau 445

23 Pattern Clustering Using a Swarm Intelligence Approach

Swagatam Das, Ajith Abraham 469

24 Using Fuzzy Logic in Data Mining	
<i>Lior Rokach</i>	505

Part V Supporting Methods

25 Statistical Methods for Data Mining	
<i>Yoav Benjamini, Moshe Leshno</i>	523
26 Logics for Data Mining	
<i>Petr Hájek</i>	541
27 Wavelet Methods in Data Mining	
<i>Tao Li, Sheng Ma, Mitsunori Ogihara</i>	553
28 Fractal Mining - Self Similarity-based Clustering and its Applications	
<i>Daniel Barbara, Ping Chen</i>	573
29 Visual Analysis of Sequences Using Fractal Geometry	
<i>Noa Ruschin Rimini, Oded Maimon</i>	591
30 Interestingness Measures - On Determining What Is Interesting	
<i>Sigal Sahar</i>	603
31 Quality Assessment Approaches in Data Mining	
<i>Maria Halkidi, Michalis Vazirgiannis</i>	613
32 Data Mining Model Comparison	
<i>Paolo Giudici</i>	641
33 Data Mining Query Languages	
<i>Jean-Francois Boulicaut, Cyrille Masson</i>	655

Part VI Advanced Methods

34 Mining Multi-label Data	
<i>Grigorios Tsoumakas, Ioannis Katakis, Ioannis Vlahavas</i>	667
35 Privacy in Data Mining	
<i>Vicenç Torra</i>	687
36 Meta-Learning - Concepts and Techniques	
<i>Ricardo Vilalta, Christophe Giraud-Carrier, Pavel Brazdil</i>	717
37 Bias vs Variance Decomposition For Regression and Classification	
<i>Pierre Geurts</i>	733

38 Mining with Rare Cases

Gary M. Weiss 747

39 Data Stream Mining

Mohamed Medhat Gaber, Arkady Zaslavsky, Shonali Krishnaswamy 759

40 Mining Concept-Drifting Data Streams

Haixun Wang, Philip S. Yu, Jiawei Han 789

41 Mining High-Dimensional Data

Wei Wang, Jiong Yang 803

42 Text Mining and Information Extraction

Moty Ben-Dov, Ronen Feldman 809

43 Spatial Data Mining

Shashi Shekhar, Pusheng Zhang, Yan Huang 837

44 Spatio-temporal clustering

Slava Kisilevich, Florian Mansmann, Mirco Nanni, Salvatore Rinzivillo 855

45 Data Mining for Imbalanced Datasets: An Overview

Nitesh V. Chawla 875

46 Relational Data Mining

Sašo Džeroski 887

47 Web Mining

Johannes Fürnkranz 913

48 A Review of Web Document Clustering Approaches

Nora Oikonomakou, Michalis Vazirgiannis 931

49 Causal Discovery

Hong Yao, Cory J. Butz, Howard J. Hamilton 949

50 Ensemble Methods in Supervised Learning

Lior Rokach 959

51 Data Mining using Decomposition Methods

Lior Rokach, Oded Maimon 981

52 Information Fusion - Methods and Aggregation Operators

Vicenç Torra 999

53 Parallel And Grid-Based Data Mining – Algorithms, Models and Systems for High-Performance KDD

Antonio Congiusta, Domenico Talia, Paolo Trunfio 1009

54 Collaborative Data Mining*Steve Moyle* 1029**55 Organizational Data Mining***Hamid R. Nemati, Christopher D. Barko* 1041**56 Mining Time Series Data***Chotirat Ann Ratanamahatana, Jessica Lin, Dimitrios Gunopulos,
Eamonn Keogh, Michail Vlachos, Gautam Das* 1049

Part VII Applications

57 Multimedia Data Mining*Zhongfei (Mark) Zhang, Ruofei Zhang* 1081**58 Data Mining in Medicine***Nada Lavrač, Blaž Zupan* 1111**59 Learning Information Patterns in Biological Databases - Stochastic Data Mining***Gautam B. Singh* 1137**60 Data Mining for Financial Applications***Boris Kovalerchuk, Evgenii Vityaev* 1153**61 Data Mining for Intrusion Detection***Anoop Singhal, Sushil Jajodia* 1171**62 Data Mining for CRM***Kurt Thearling* 1181**63 Data Mining for Target Marketing***Nissan Levin, Jacob Zahavi* 1189**64 NHECD - Nano Health and Environmental Commented Database***Oded Maimon, Abel Browarnik* 1221

Part VIII Software

65 Commercial Data Mining Software*Qingyu Zhang, Richard S. Segall* 1245**66 Weka-A Machine Learning Workbench for Data Mining***Eibe Frank, Mark Hall, Geoffrey Holmes, Richard Kirkby,
Bernhard Pfahringer, Ian H. Witten, Len Trigg* 1269**Index** 1279

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 Mathématiques et de
Génie Industriel,
École Polytechnique de Montréal,
Canada

Daniel Barbara

Department of Information and Soft-
ware 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 Mathematics 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 Cham-
paign, USA

Geoffrey Holmes

Department of Computer Science,
University of Waikato, New Zealand

Frank Höppner

Department of Information Systems,
University of Applied Sciences Braun-
schweig/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 Univer-
sity 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 Indus-
trial 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 Liao

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 Eco-
nomics
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 Tsoumakos

Dept. of Informatics, Aristotle Univer-
sity 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, Aus-
tralia

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