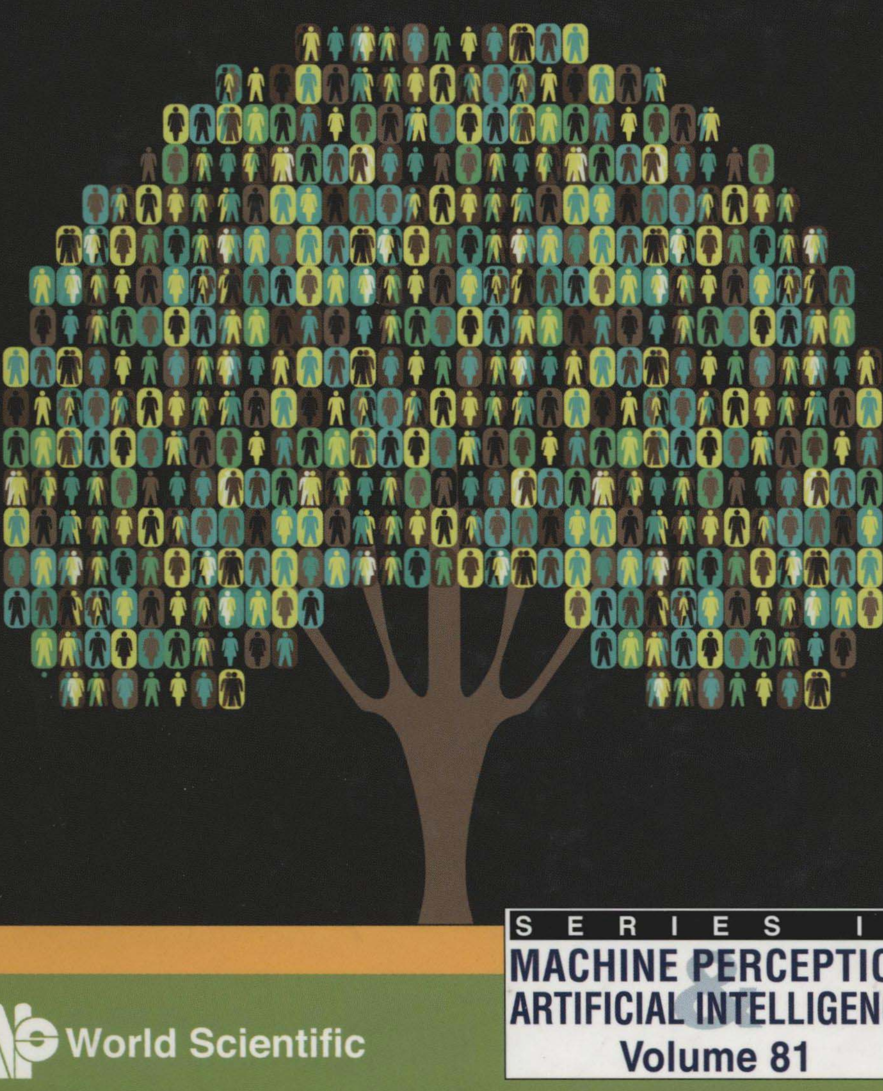
DATA MINING WITH DECISION TREES

Theory and Applications

2nd Edition

Lior Rokach • Oded Maimon



SERIES IN
MACHINE PERCEPTION & ARTIFICIAL INTELLIGENCE
Volume 81

World Scientific

Series in Machine Perception and Artificial Intelligence – Vol. 81

DATA MINING WITH DECISION TREES

Theory and Applications

2nd Edition

Lior Rokach
Ben-Gurion University of the Negev, Israel

Oded Maimon
Tel-Aviv University, Israel

World Scientific

NEW JERSEY • LONDON • SINGAPORE • BEIJING • SHANGHAI • HONG KONG • TAIPEI • CHENNAI

Published by

World Scientific Publishing Co. Pte. Ltd.
5 Toh Tuck Link, Singapore 596224
USA office: 27 Warren Street, Suite 401-402, Hackensack, NJ 07601
UK office: 57 Shelton Street, Covent Garden, London WC2H 9HE

Library of Congress Cataloging-in-Publication Data
Rokach, Lior.
Data mining with decision trees : theory and applications / by Lior Rokach (Ben-Gurion University of the Negev, Israel), Oded Maimon (Tel-Aviv University, Israel). -- 2nd edition.
pages cm
Includes bibliographical references and index.
ISBN 978-9814590075 (hardback : alk. paper) -- ISBN 978-9814590082 (ebook)
1. Data mining. 2. Decision trees. 3. Machine learning. 4. Decision support systems. I. Maimon, Oded. II. Title.
QA76.9.D343R654 2014
006.3'12--dc23
2014029799

British Library Cataloguing-in-Publication Data
A catalogue record for this book is available from the British Library.

In-house Editor: Amanda Yun

Typeset by Stallion Press
Email: enquiries@stallionpress.com

Printed in Singapore by Mainland Press Pte Ltd.

DATA MINING WITH DECISION TREES

Theory and Applications

2nd Edition

SERIES IN MACHINE PERCEPTION AND ARTIFICIAL INTELLIGENCE*

Editors: **H. Bunke** (Univ. Bern, Switzerland)
P. S. P. Wang (Northeastern Univ., USA)

Vol. 65: Fighting Terror in Cyberspace
(*Eds. M. Last and A. Kandel*)

Vol. 66: Formal Models, Languages and Applications
(Eds. *K. G. Subramanian, K. Rangarajan and M. Mukund*)

Vol. 67: Image Pattern Recognition: Synthesis and Analysis in Biometrics
(Eds. *S. N. Yanushkevich, P. S. P. Wang, M. L. Gavrilova and S. N. Srihari*)

Vol. 68: Bridging the Gap Between Graph Edit Distance and Kernel Machines
(*M. Neuhaus and H. Bunke*)

Vol. 69: Data Mining with Decision Trees: Theory and Applications
(*L. Rokach and O. Maimon*)

Vol. 70: Personalization Techniques and Recommender Systems
(Eds. *G. Uchyigit and M. Ma*)

Vol. 71: Recognition of Whiteboard Notes: Online, Offline and Combination
(Eds. *H. Bunke and M. Liwicki*)

Vol. 72: Kernels for Structured Data
(*T Gärtner*)

Vol. 73: Progress in Computer Vision and Image Analysis
(Eds. *H. Bunke, J. J. Villanueva, G. Sánchez and X. Otazu*)

Vol. 74: Wavelet Theory Approach to Pattern Recognition (2nd Edition)
(*Y. Y. Tang*)

Vol. 75: Pattern Classification Using Ensemble Methods
(*L. Rokach*)

Vol. 76: Automated Database Applications Testing: Specification Representation for Automated Reasoning
(*R. F. Mikhail, D. Berndt and A. Kandel*)

Vol. 77: Graph Classification and Clustering Based on Vector Space Embedding
(*K. Riesen and H. Bunke*)

Vol. 78: Integration of Swarm Intelligence and Artificial Neural Network
(Eds. *S. Dehuri, S. Ghosh and S.-B. Cho*)

Vol. 79 Document Analysis and Recognition with Wavelet and Fractal Theories
(*Y. Y. Tang*)

Vol. 80 Multimodal Interactive Handwritten Text Transcription
(*V. Romero, A. H. Toselli and E. Vidal*)

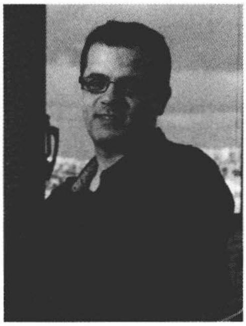
Vol. 81 Data Mining with Decision Trees: Theory and Applications
Second Edition
(*L. Rokach and O. Maimon*)

*The complete list of the published volumes in the series can be found at
http://www.worldscientific.com/series/smpai

Dedicated to our families
in appreciation for their patience and support
during the preparation of this book.

L.R.
O.M.

About the Authors



Lior Rokach is an Associate Professor of Information Systems and Software Engineering at Ben-Gurion University of the Negev. Dr. Rokach is a recognized expert in intelligent information systems and has held several leading positions in this field. His main areas of interest are Machine Learning, Information Security, Recommender Systems and Information Retrieval. Dr. Rokach is the author of over 100 peer reviewed papers in leading journals conference proceedings, patents, and book chapters. In addition, he has also authored six books in the field of data mining.

Professor **Oded Maimon** from Tel Aviv University, previously at MIT, is also the Oracle chair professor. His research interests are in data mining and knowledge discovery and robotics. He has published over 300 papers and ten books. Currently he is exploring new concepts of core data mining methods, as well as investigating artificial and biological data.

Preface for the Second Edition

The first edition of the book, which was published six years ago, was extremely well received by the data mining research and development communities. The positive reception, along with the fast pace of research in the data mining, motivated us to update our book. We received many requests to include the new advances in the field as well as the new applications and software tools that have become available in the second edition of the book. This second edition aims to refresh the previously presented material in the fundamental areas, and to present new findings in the field; nearly quarter of this edition is comprised of new materials.

We have added four new chapters and updated some of the existing ones. Because many readers are already familiar with the layout of the first edition, we have tried to change it as little as possible. Below is the summary of the main alterations:

- The first edition has mainly focused on using decision trees for classification tasks (i.e. classification trees). In this edition we describe how decision trees can be used for other data mining tasks, such as regression, clustering and survival analysis.
- The new addition includes a walk-through-guide for using decision trees software. Specifically, we focus on open-source solutions that are freely available.
- We added a chapter on cost-sensitive active and proactive learning of decision trees since the cost aspect is very important in many domain applications such as medicine and marketing.
- Chapter 16 is dedicated entirely to the field of recommender systems which is a popular research area. Recommender Systems help customers

to choose an item from a potentially overwhelming number of alternative items.

We apologize for the errors that have been found in the first edition and we are grateful to the many readers who have found those. We have done our best to avoid errors in this new edition. Many graduate students have read parts of the manuscript and offered helpful suggestions and we thank them for that.

Many thanks are owed to Elizaveta Futerman. She has been the most helpful assistant in proofreading the new chapters and improving the manuscript. The authors would like to thank Amanda Yun and staff members of World Scientific Publishing for their kind cooperation in writing this book. Moreover, we are thankful to Prof. H. Bunke and Prof. P.S.P. Wang for including our book in their fascinating series on machine perception and artificial intelligence.

Finally, we would like to thank our families for their love and support.

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

April 2014

Preface for the First Edition

Data mining is the science, art and technology of exploring large and complex bodies of data in order to discover useful patterns. Theoreticians and practitioners are continually seeking improved techniques to make the process more efficient, cost-effective and accurate. One of the most promising and popular approaches is the use of decision trees. Decision trees are simple yet successful techniques for predicting and explaining the relationship between some measurements about an item and its target value. In addition to their use in data mining, decision trees, which originally derived from logic, management and statistics, are today highly effective tools in other areas such as text mining, information extraction, machine learning, and pattern recognition.

Decision trees offer many benefits:

- Versatility for a wide variety of data mining tasks, such as classification, regression, clustering and feature selection
- Self-explanatory and easy to follow (when compacted)
- Flexibility in handling a variety of input data: nominal, numeric and textual
- Adaptability in processing datasets that may have errors or missing values
- High predictive performance for a relatively small computational effort
- Available in many data mining packages over a variety of platforms
- Useful for large datasets (in an ensemble framework)

This is the first comprehensive book about decision trees. Devoted entirely to the field, it covers almost all aspects of this very important technique.

The book has three main parts:

- Part I presents the data mining and decision tree foundations (including basic rationale, theoretical formulation, and detailed evaluation).
- Part II introduces the basic and advanced algorithms for automatically growing decision trees (including splitting and pruning, decision forests, and incremental learning).
- Part III presents important extensions for improving decision tree performance and for accommodating it to certain circumstances. This part also discusses advanced topics such as feature selection, fuzzy decision trees and hybrid framework.

We have tried to make as complete a presentation of decision trees in data mining as possible. However, new applications are always being introduced. For example, we are now researching the important issue of data mining privacy, where we use a hybrid method of genetic process with decision trees to generate the optimal privacy-protecting method. Using the fundamental techniques presented in this book, we are also extensively involved in researching language-independent text mining (including ontology generation and automatic taxonomy).

Although we discuss in this book the broad range of decision trees and their importance, we are certainly aware of related methods, some with overlapping capabilities. For this reason, we recently published a complementary book "Soft Computing for Knowledge Discovery and Data Mining", which addresses other approaches and methods in data mining, such as artificial neural networks, fuzzy logic, evolutionary algorithms, agent technology, swarm intelligence and diffusion methods.

An important principle that guided us while writing this book was the extensive use of illustrative examples. Accordingly, in addition to decision tree theory and algorithms, we provide the reader with many applications from the real-world as well as examples that we have formulated for explaining the theory and algorithms. The applications cover a variety of fields, such as marketing, manufacturing, and bio-medicine. The data referred to in this book, as well as most of the Java implementations of the pseudo-algorithms and programs that we present and discuss, may be obtained via the Web.

We believe that this book will serve as a vital source of decision tree techniques for researchers in information systems, engineering, computer science, statistics and management. In addition, this book is highly useful to researchers in the social sciences, psychology, medicine, genetics, business intelligence, and other fields characterized by complex data-processing problems of underlying models.

Since the material in this book formed the basis of undergraduate and graduates courses at Ben-Gurion University of the Negev and Tel-Aviv University and it can also serve as a reference source for graduate/advanced undergraduate level courses in knowledge discovery, data mining and machine learning. Practitioners among the readers may be particularly interested in the descriptions of real-world data mining projects performed with decision trees methods.

We would like to acknowledge the contribution to our research and to the book to many students, but in particular to Dr. Barak Chizi, Dr. Shahar Cohen, Roni Romano and Reuven Arbel. Many thanks are owed to Arthur Kemelman. He has been a most helpful assistant in proofreading and improving the manuscript.

The authors would like to thank Mr. Ian Seldrup, Senior Editor, and staff members of World Scientific Publishing for their kind cooperation in connection with writing this book. Thanks also to Prof. H. Bunke and Prof P.S.P. Wang for including our book in their fascinating series in machine perception and artificial intelligence.

Last, but not least, we owe our special gratitude to our partners, families, and friends for their patience, time, support, and encouragement.

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

October 2007

Contents